DEEP LEARNING-BASED DIGITAL HUMAN MODELING AND APPLICATIONS

by

Ayman Ali

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2023

Approved by:

_____
Dr. Pu Wang

_____
Dr. Minwoo Lee

_____
Dr. Dong Dai

_____
Dr. Mohsen Dorodchi

_____
Dr. Chao Wang

ABSTRACT

AYMAN ALI. Deep Learning-based Digital Human Modeling and Applications.
(Under the direction of DR. PU WANG)

Recent advancements in the domain of deep learning models have engendered remarkable progress across numerous computer vision tasks. Notably, there has been a burgeoning interest in the field of recovering three-dimensional (3D) human models from monocular images in recent years. This heightened interest can be attributed to the extensive practical applications that necessitate the utilization of 3D human models, including but not limited to gaming, human-computer interaction, virtual systems, and digital twin. The focus of this dissertation is to conceptualize and develop a suite of deep learning-based models with the primary objective of enabling the expeditious and high-fidelity digitalization of human subjects. This endeavor further aims to facilitate a multitude of downstream applications that leverage digital 3D human models.

The endeavor to estimate a three-dimensional (3D) human mesh from a monocular image necessitates the application of intricate deep-learning models for enhanced feature extraction, albeit at the expense of heightened computational requirements. As an alternative approach, researchers have explored the utilization of a skeleton-based modality, which represents a lightweight abstraction of human pose, aimed at mitigating the computational intensity. However, this approach entails the omission of significant visual cues, particularly shape information, which cannot be entirely derived from the 3D skeletal representation alone. To harness the advantages of both paradigms, a hybrid methodology that integrates the benefits of 3D human mesh and skeletal information offers a promising avenue. Over the past decade, substantial strides have been made in the estimation of two-dimensional (2D) joint coordinates derived from monocular images. Simultaneously, the application of Convolutional

Neural Networks (CNNs) for the extraction of intricate visual features from images has demonstrated its prowess in feature extraction. This progress serves as a compelling impetus for our investigation into a hybrid architectural framework that combines CNNs with a lightweight graph transformer-based approach. This innovative architecture is designed to elevate the 2D joint pose to a comprehensive 3D representation and recover essential visual cues essential for the precise estimation of pose and shape parameters.

While SOTA results in 3D Human Pose Estimation (HPE) are important, they do not guarantee the accuracy and plausibility required for biomechanical analysis. Our innovative two-stage deep learning model is designed to efficiently estimate 3D human poses and associated kinematic attributes from monocular videos, with a primary focus on mobile device deployment. The paramount significance of this contribution lies in its ability to provide not only accurate 3D pose estimations but also biomechanically plausible results. This plausibility is essential for achieving accurate biomechanical analyses, thereby advancing various applications, including motion tracking, gesture recognition, and ergonomic assessments. Our work significantly contributes to enhancing our understanding of human movement and its interaction with the environment, ultimately impacting a wide range of biomechanics-related studies and applications.

In the realm of human movement analysis, one prominent downstream task is the recognition of human actions based on skeletal data, known as Skeleton-based Human Action Recognition (HAR). This domain has garnered substantial attention within the computer vision community, primarily due to its distinctive attributes, such as computational efficiency, the innate representational power of features, and robustness to variations in illumination. In this context, our research demonstrates that, by representing 3D pose sequences as RGB images, conventional Convolutional Neural Network (CNN) architectures, exemplified by ResNet-50, when complemented by as-

tute training strategies and diverse augmentation techniques, can attain State-of-the-Art (SOTA) accuracy levels, surpassing the widely adopted Graph Neural Network models.

The domain of radar-based sensing, rooted in the transmission and reception of radio waves, offers a non-intrusive and versatile means to monitor human movements, gestures, and vital signs. However, despite its vast potential, the lack of comprehensive radar datasets has hindered the broader implementation of deep learning in radar-based human sensing. In response, the application of synthetic data in deep learning training emerges as a crucial advantage. Synthetic datasets provide an expansive and practically limitless resource, enabling models to adapt and generalize proficiently by exposing them to diverse scenarios, transcending the limitations of real-world data. As part of this research's trajectory, a novel computational framework known as "virtual radar" is introduced, leveraging 3D pose-driven physics-informed principles. This paradigm allows for the generation of high-fidelity synthetic radar data by merging 3D human models and the principles of Physical Optics (PO) approximation for radar cross-section modeling. The introduction of virtual radar marks a groundbreaking path towards establishing foundational models focused on the nuanced understanding of human behavior through privacy-preserving radar-based methodologies.

DEDICATION

In loving memory of my father, Ahmed Ali. I wish you were here to witness this milestone in my academic journey. Your unwavering support, boundless love, and belief in my potential have been the guiding force behind every step I have taken. Your wisdom and encouragement continue to resonate within me, shaping both my academic endeavors and my character.

To my family and friends, who has stood by me with unending love and support, I dedicate this work. Your strength and encouragement have been a constant source of inspiration.

This dissertation stands as a testament to the values instilled in me by my father the pursuit of knowledge, resilience in the face of challenges, and the unwavering commitment to excellence. Though you're no longer here, your spirit lives on in every achievement and every endeavor.

ACKNOWLEDGEMENTS

Embarking on this scholarly endeavor has been a transformative journey, illuminated by the generous support and guidance of numerous individuals whose contributions have shaped the contours of this dissertation.

Dr. Pu Wang, my esteemed advisor, stands as a beacon of wisdom and encouragement throughout this academic pursuit. Your guidance, expertise, and unwavering commitment to my growth as a scholar have been foundational to this work.

I extend my deepest gratitude to the members of my dissertation committee, Dr. Mohsen Dorodchi, Dr. Chao Wang, Dr. Minwoo Lee, and Dr. Dong Dai, for their scholarly insights, constructive feedback, and invaluable guidance. Their collective expertise and diverse perspectives have profoundly enriched this study.

My family deserves profound appreciation for their unwavering support, understanding, and sacrifices throughout this academic pursuit. Their encouragement has been a constant source of strength.

I express heartfelt gratitude to my friends and peers for their unwavering support, camaraderie, and understanding. Their encouragement has been invaluable.

Special recognition goes to DiLovena, whose encouragement, and belief in my capabilities were instrumental in navigating the complexities of this academic journey.

Finally, to all those whose contributions may not be explicitly named but have significantly impacted this endeavor, your support and encouragement have been deeply valued and appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D HPE	Two-Dimensional Human Pose Estimation
3D HPE	Thress-Dimensional Human Pose Estimation
AR	Augmented Reality
BRNN	Bidirectional Recurrent Neural Network
CNN	Convolutional Neural Network
DoF	Degree of Freedom
FK	Forward Kinematics
FMCW	Frequency-Modulated Continuous Wave
GCN	Graph Convolutional Network
GRU	Gated Recurrent Unit
HAR	Human Activity Recognition
HMR	Human Mesh Recovery
HPE	Human Pose Estimation
HSV	Hue-Saturation-Value
IK	Inverse Kinematics
JDM	Joint Distance Map
JTM	Joint Trajectory Map
LOS	Line Of Sight
LSTM	Long Short-Term Memory

MPJPE	Mean Per-Joint Position Error
MPVE	Mean Per-Vertex Error
PA-MPJPE	Procrustes Aligned Mean Per-Joint Position Error
RCS	Radar BackScatter
RNN	Recurrent Neural Network
SMPL	Skinned Multi-Person Linear
SNR	Signal-to-Noise Ratios
SOTA	State-Of-The-Art
ST-Transformer	Spatio-Temporal Transformer
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
VR	Virtual Reality

CHAPTER 1: Introduction to Human Modeling

Human behavior modeling, situated at the intersection of computer vision, machine learning, and behavioral sciences, epitomizes a multifaceted quest. Its core essence lies in unraveling the intricate fabric of human actions and behaviors. This dissertation embarks on a holistic expedition, aiming to decipher, replicate, and comprehend these intricate nuances through a meticulous amalgamation of estimation and simulation.

The significance of human behavior modeling transcends mere academic pursuits; it permeates across multifarious domains, propelling advancements that reverberate through industries. Acting as a linchpin, it lays the foundation for adaptive technologies that reshape paradigms in healthcare, robotics, entertainment, and beyond. This endeavor converges upon a pivotal innovation that underpins its pursuit: the profound harnessing of deep neural networks.

Deep neural networks, with their adaptive and hierarchical architectures, stand as the beacon in decoding elaborate patterns latent within complex datasets. Their prowess in discerning subtle intricacies within data and their capacity to hierarchically abstract representations have metamorphosed the very landscape of human behavior modeling. Within the tapestry of this dissertation, these neural networks take center stage, steering the exploration towards innovative frontiers in action recognition, estimation, and simulation.

The emergence of deep learning has revolutionized human pose estimation from single images, prompting a shift towards reconstructing 3D human meshes. This fascination arises from practical applications in human-computer interaction, gaming, and virtual systems. Notable advancements in 2D/3D pose estimation and mesh reconstruction, evidenced by Liu et al. [2] and Sun et al. [3], highlight the field's

burgeoning interest. Yet, reconstructing a comprehensive 3D mesh from a single image remains challenging due to depth ambiguity, complex poses, and limited real-world datasets. Key to 3D mesh reconstruction are precise pose (θ) and shape (β) parameters. The SMPL model [1], detailed by Romero et al. [4], encapsulates these parameters, governing body orientation and shape variations, derived from datasets like CAESAR [5].

The challenge lies in transforming human form into nuanced 3D representations beyond basic geometric shapes. This necessitates synthesizing body pose and shape, driving the exploration towards more accurate and computationally feasible solutions in human form modeling.

Biomechanics, merging mechanical principles with biological systems, elucidates the mechanics underlying human and animal locomotion. This field, vital in sports, rehabilitation, and product design, hinges on precise biomechanical data often obtained from specialized laboratories, demanding substantial financial investment [6,7].

The human body, a marvel of coordination, orchestrates everyday activities through intricate biomechanical principles governing stability and movement. Understanding these principles holds promise in enhancing performance, preventing injuries, and tailoring products for human capabilities [6, 8].

In sports, biomechanics dissects athletic movements, improves techniques, and aids injury prevention [9–11]. In rehabilitation, it helps tailor treatments by identifying movement irregularities and imbalances [6, 12, 13].

The challenge lies in acquiring comprehensive biomechanical data, often reliant on sophisticated lab equipment. Additionally, computer modeling and simulation have emerged as crucial tools in biomechanics, offering alternative data generation methods [6].

Enabling machines to detect and analyze human movements from video sequences or images holds vast potential across diverse applications surveillance, robotics, health-

care, sports analysis, and human-computer interaction. This pursuit has fueled a surge in research, amalgamating computer vision, machine learning, and deep learning techniques.

Recent years witnessed a transformative shift with deep learning paradigms eclipsing traditional machine learning in computer vision tasks. Notably, the focus shifted towards skeleton-based representation, leveraging the anatomical richness of skeletal joints for discriminative analysis [14–16]. These advancements showcase the potential at the intersection of deep learning methodologies and human action recognition.

Research in human activity recognition has made significant strides, finding application in surveillance, smart homes, video analytics, autopilot systems, and human-computer interaction [17–21]. HAR aims to understand and interpret user conduct, enabling proactive computational system assistance [22].

It encompasses vision-based and sensor-based approaches, with vision-based HAR leveraging optical sensors and computer vision techniques. However, challenges like illumination, occlusion, and privacy concerns persist [23, 24].

Radar-based HAR has emerged as a prominent sensor-driven methodology, offering robustness in diverse environments and safeguarding visual privacy. Its capacity to detect through barriers, coupled with its non-intrusive nature, enhances its applicability. However, the scarcity of radar-based datasets poses a challenge for its development and evaluation.

Data holds pivotal importance in deep learning, shaping its capabilities across domains. Datasets enrich deep learning models, and their quality and diversity directly influence model performance and generalization. Video-based datasets are abundant and diverse, fostering impressive progress in action recognition. Yet, radar-based datasets are limited, impeding radar's potential application.

1.1 Dissertation Questions

Human modeling, through the integration of estimation from visual data and simulation techniques, stands at the forefront of modern research, offering unprecedented insights into understanding and analyzing human actions and behaviors. This interdisciplinary field brings together computer vision, machine learning, and physics-based simulation to bridge the gap between observed behaviors and digitally simulated representations.

The significance of human modeling via estimation and simulation lies in its transformative potential. It allows us to delve deep into deciphering the complexities of human actions, interactions, and movements by harnessing the power of computational modeling. By merging insights obtained from visual data captured by single-camera systems with the realism of physics-based simulations, researchers can construct digital avatars and models that mirror real-world human behavior with increasing fidelity.

This approach opens new vistas for numerous fields. In domains ranging from healthcare, sports analytics, and entertainment to robotics and virtual environments, understanding human actions and behaviors is pivotal. Accurate and adaptable digital models of human actions enable advancements in rehabilitation techniques, predictive analytics in sports, immersive experiences in gaming and virtual reality, and the development of more responsive and human-centric robotic systems.

The impact of human modeling via estimation and simulation reverberates across various disciplines. By deciphering and replicating human actions digitally, this research not only furthers our understanding of human behavior but also facilitates the creation of sophisticated systems that interact with humans more intuitively and adaptively.

As we embark on this research journey, exploring the integration of estimation and simulation for modeling human behavior, our endeavor is driven by the quest to

unlock deeper insights into the intricacies of human actions. The research questions we pose aim to unravel the potential of this amalgamation, addressing challenges and paving the way toward more accurate, responsive, and context-aware digital models of human behavior.

Now, guided by the imperative to delve deeper into the realms of human behavior modeling, the following research questions encapsulate our pursuit to harness the synergy between estimation from visual data and physics-based simulation techniques:

- What strategies and techniques can mitigate biases and improve the fairness and inclusivity of human modeling algorithms, particularly in scenarios where biases in training data can impact model performance?

- What novel approaches or frameworks can be developed to simulate and predict human interactions and behaviors in dynamic and complex environments, enabling more accurate and context-aware human modeling?

- How can analytical inverse kinematics be effectively integrated into deep learning frameworks to enhance the precision and accuracy of 3D pose estimation in computer vision?

- What criteria and design principles should be prioritized to engineer an efficient, robust, and modular deep learning model for inferring 3D human pose from monocular images?

- How can the integration of OpenSim, a simulation tool, contribute to the development of an efficient and resilient deep learning model for precise 3D human pose estimation and prediction of associated kinematic attributes from monocular or single-camera video inputs?

- What are the performance, accuracy, and computational efficiency implications of executing advanced 3D pose estimation techniques on mobile platforms, and

how do these techniques fare within the defined constraints and parameters of mobile device applications?

- How can diverse training techniques be effectively integrated into a Convolutional Neural Network (CNN) architecture to achieve comparable performance results in action recognition when compared to state-of-the-art (SOTA) graph neural network approaches?

- What impact do diverse augmentation techniques have on the generalization and robustness of a modular CNN model designed for action recognition?

- How can privacy-preserving and non-intrusive radar-based systems be optimized and augmented to achieve robust and accurate action recognition in scenarios where visual cues are absent or limited?

- How does the utilization of virtual radar synthesis technology, combining video recordings of human actions, enhance the generation of realistic radar spectrograms for human action recognition systems?

1.2 Dissertation Structure

The dissertation encompasses a comprehensive exploration of diverse aspects within human behavior modeling, beginning with the foundational chapters:

Chapter 1 serves as an introductory prologue, elucidating the motivations driving the research endeavor. It articulates the central research inquiries driving the exploration into estimation, simulation, and radar-based methodologies in comprehending human actions and behaviors.

Chapter 2 delves into foundational knowledge essential for understanding the intricacies of human behavior modeling. It offers an extensive review of prior research, outlining methodologies, theories, and innovations in estimation, simulation, radar-based gait recognition, and CNN techniques for action recognition. This chapter

critically assesses the current state-of-the-art, laying a robust foundation for subsequent discussions.

Chapter 3 conducts an in-depth investigation into reconstructing 3D human structures from visual and 2D pose data. Navigating through methodologies, challenges, and advancements in human mesh recovery techniques, it highlights their pivotal role in advancing human behavior modeling.

Chapter 4 focuses on employing physics-based kinematic analysis and simulation techniques. This chapter sheds light on how these methodologies contribute to crafting lifelike digital simulations of human movements, enriching our comprehension of human behavior.

Chapter 5 explores sophisticated strategies embedded within CNN architectures to achieve performance akin to dominant Graph Convolutional Networks (GCN) in action recognition. Delving into innovative CNN techniques, this chapter assesses their implications for the field.

Chapter 6 investigates radar-based approaches for gait and action recognition. Discussing methodologies, significance, and proposing the use of virtual radar for generating synthetic high-resolution spectrograms, it provides crucial insights into radar-based human behavior modeling.

The final chapter culminates the dissertation, synthesizing findings from each preceding chapter. It outlines conclusions drawn from the extensive exploration and delineates future research directions, proposing avenues for further advancement in the covered domains.

CHAPTER 2: Background and Literature Review

2.1 Human Pose estimation

Human Pose Estimation (HPE), an extensively explored domain within the corpus of computer vision, encompasses the intricate task of delineating the spatial configuration of human anatomical elements. It derives its intellectual nourishment from data acquired through an array of sensory apparatuses, with a special affinity for images and videos. This veritable Rosetta Stone of HPE yields invaluable geometric and kinematic insights into the human form, thereby underpinning a diverse spectrum of applications.

The landscape of HPE has undergone a profound metamorphosis, riding the crest of the burgeoning tide of deep learning solutions in recent years. These deep-learning paradigms have successfully eclipsed their classical computer vision counterparts. Their ascendancy has ushered in a phase of remarkable performance and substantial progress within the realm of HPE. However, the domain remains encumbered by a series of formidable challenges that necessitate deft solutions.

Occlusion, a perennial nemesis in computer vision, continues to cast its shadow on HPE endeavors. Likewise, the conundrum of insufficient training data persists, hampering the trajectory towards proficiency. Additionally, the intricacies of depth ambiguity pose an obstinate challenge in 3D HPE, an arena where accurate 3D pose annotations are not as easily gleaned as their 2D counterparts. The controlled lab environments, in which motion capture systems thrive, are less accommodating to the wild, untamed scenarios of the real world. Moreover, in the context of 3D HPE from monocular RGB images and videos, the conundrum of depth ambiguity looms large, casting shadows of uncertainty. The multi-view setting, while a potent instrument,

hinges crucially on the association of viewpoints, an intellectual puzzle requiring resolute resolution.

Noteworthy endeavors have sought solace in sensors of diverse persuasions, such as depth sensors and radio frequency devices. However, these instrumental aids are often plagued by concerns of cost-effectiveness and an exigent need for specialized hardware. The alchemy of tackling such multifaceted challenges in the ambit of HPE is poised to chart new trajectories of innovation and unravel the complexities that linger in this dynamic field.

2.1.1 2D Human Pose Estimation

In the domain of two-dimensional (2D) single-person pose estimation, the primary objective revolves around the precise localization of human body joint positions, a fundamental task typically performed when analyzing single-person images. However, when presented with images featuring multiple individuals, an initial preprocessing step is mandated, involving the careful cropping of the input image to create isolated image patches or sub-images, each containing only one individual. This cropping operation can be automated and is commonly executed at the deployment.

When broadly categorizing the methodologies employed in single-person pipelines utilizing deep learning techniques, two primary categories stand out: regression methods and heatmap-based methods. Regression methods entail the creation of comprehensive end-to-end frameworks with the aim of establishing a mapping from the input image to the precise positions of body joints or the parameters governing human body models [25]. In contrast, heatmap-based methods are designed to predict the approximate locations of various body parts and joints [26], with these predictions being guided and supervised through the utilization of heatmap representations [27]. It is worth noting that heatmap-based frameworks have gained significant traction in the domain of 2D Human Pose Estimation tasks.

In contrast to single-person Human Pose Estimation, multi-person HPE presents

a more formidable and intricate challenge. This complexity primarily arises from the need to determine the number of individuals present, discern their respective spatial configurations, and resolve the puzzle of associating individual keypoints with their corresponding persons. To address these intricacies, the landscape of multi-person HPE methods can be systematically categorized into two primary paradigms: top-down and bottom-up strategies.

Top-down methods leverage pre-existing person detectors to initially glean a set of bounding boxes (each pertaining to an individual) from the input images. Subsequently, these bounding boxes serve as the canvas for the application of single-person pose estimators, thereby engendering the derivation of multi-person poses. In stark contrast, bottom-up approaches commence their journey by first meticulously pinpointing all the anatomical joints within a single image and subsequently engaging in the task of grouping these joints into their respective human subjects.

Within the top-down pipeline, it's imperative to acknowledge that the computational demands are profoundly influenced by the number of individuals present within the input image. In this regard, the bottom-up methodologies often enjoy a swifter computational pace, given their avoidance of the need to perform separate pose detection for each individual.

Delving deeper into the top-down pipeline, we encounter two pivotal components. The first component encompasses a human body detector, responsible for procuring the bounding boxes encapsulating individual persons. The second, equally critical component entails a single-person pose estimator, the task of which is to predict the spatial coordinates of keypoints within these bounding boxes. Notably, a substantial body of research endeavors is dedicated to refining and enhancing the modules within HPE networks.

In the pursuit of answering the fundamental question regarding the efficacy of simplistic methodologies in the construction of HPE networks, Xiao et al. [28] introduced

a streamlined yet highly effective structure by augmenting the ResNet (the backbone network) with a few deconvolutional layers, resulting in the generation of high-resolution representations in the form of heatmaps. In the pursuit of improving the accuracy of keypoint localization, Wang et al. introduced a two-stage model-agnostic framework named Graph-PCNN, as described in their work [29]. This framework consists of a localization subnet responsible for obtaining initial keypoint locations and a graph pose refinement module designed to enhance the representations of keypoint localization. Furthermore, Cai et al. introduced a multi-stage network that incorporates a Residual Steps Network module, as highlighted in their study [30]. This module facilitates the acquisition of detailed local representations through efficient intra-level feature fusion strategies. Complementing this is a Pose Refine Machine module, which aims to strike a balance between local and global feature representations within the network architecture.

The bottom-up pipeline, exemplified by a corpus of seminal works [31–33], unfolds in two fundamental phases. The first phase entails body joint detection, a meticulous process involving the extraction of localized features and the prediction of candidates for body joints. The second phase, the assembly of joint candidates for individual entities, is a pivotal step wherein candidate joints are organized into coherent pose representations using part association strategies.

One of the pioneering two-stage bottom-up approaches was proposed by Pishchulin et al. [34], introducing the Fast R-CNN-based body-part driven detector known as DeepCut. This approach commences by detecting all potential body part candidates and subsequently assigns labels to each part. The candidate parts are then skillfully assembled into a final pose through the deployment of integer linear programming. It is worth noting, however, that the computational demands of DeepCut are substantial.

To enhance the computational efficiency and performance, Insafutdinov et al. [35]

introduced DeeperCut as an improvement to DeepCut. DeeperCut incorporates a more robust body part detector and refines the optimization strategy, while also introducing image-conditioned pairwise terms to facilitate the grouping of body parts, resulting in improved performance and expedited execution.

Expanding upon the OpenPose framework, Zhu et al. improved the OpenPose architecture by incorporating redundant edges, which enhanced the connections between joints in Part Affinity Fields, as discussed in their work [36]. This enhancement resulted in improved performance when compared to the baseline approach. While methods based on OpenPose have demonstrated remarkable outcomes with high-resolution images, they tend to exhibit suboptimal performance in scenarios involving low-resolution images and occlusions.

In line with the seminal works of [37] and [3], Cheng et al. [33] introduced an extension of HRNet, aptly named the Higher Resolution Network. This extension ingeniously employs deconvolution to resolve the challenges posed by scale variations in the realm of bottom-up multi-person Human Pose Estimation.

In synopsis, the realm of two-dimensional Human Pose Estimation (2D HPE) has witnessed a substantial augmentation in performance, primarily propelled by the proliferation of deep learning techniques. Recent years have witnessed the emergence of deeper and more potent neural networks, significantly enhancing the efficacy of 2D single-person HPE methods, exemplified by notable instances such as DeepPose [25] and the Stacked Hourglass Network [38]. This surge in performance is mirrored in the domain of 2D multi-person HPE, as evidenced by systems like OpenPose [39]. While these achievements are commendable, it is imperative to underscore the persisting challenges that necessitate further investigation in the realm of 2D HPE.

The first challenge pertains to the reliable identification of individuals under circumstances marked by pronounced occlusion, as encountered, for instance, in crowded scenarios. In top-down 2D HPE methods, the extant person detectors may falter in

delineating the boundaries of heavily overlapped human bodies. Simultaneously, in occluded scenes, bottom-up approaches face an exacerbated difficulty in the task of keypoint association.

The second challenge revolves around the imperative of computational efficiency. Although certain methods, such as OpenPose [39], can achieve near real-time processing on specialized hardware endowed with moderate computing capabilities (e.g., an Nvidia GTX 1080 Ti GPU boasting 22 frames per second), the implementation of these networks on resource-constrained devices remains a formidable undertaking. Real-world applications, spanning the domains of gaming, augmented reality (AR), and virtual reality (VR), necessitate the deployment of more efficient HPE techniques on commercially available devices, thereby promising enhanced interactive experiences for end-users.

A third challenge is rooted in the dearth of data pertaining to infrequent or atypical poses. While the extant datasets for 2D HPE are of substantial size, typified by datasets like COCO, and encompassing an ample array of data for commonplace pose estimations (e.g., standing, walking, running), they notably lack training data for less conventional poses, such as those associated with falling. This imbalance in data distribution has the potential to induce biases within the models, ultimately resulting in suboptimal performance when confronted with these infrequent poses. Addressing this challenge necessitates the development of effective data generation or augmentation techniques geared toward the creation of supplementary pose data. This, in turn, serves to cultivate more resilient models capable of delivering robust performance across a broader spectrum of poses.

2.1.2 3D Human Pose Estimation

Three-Dimensional Human Pose Estimation (3D HPE), which is directed at the precise prediction of body joint locations within a three-dimensional spatial context, has garnered substantial attention in recent years. The heightened interest in 3D

HPE is underpinned by its capacity to furnish comprehensive structural information pertaining to the human body, thereby facilitating its applicability in diverse domains, including the realms of 3D film and animation production, virtual reality environments, and sports analysis.

While noteworthy strides have been made in the domain of two-dimensional Human Pose Estimation, the landscape of 3D HPE continues to present formidable challenges. A prevailing modality within the purview of 3D HPE involves the estimation of human body joint positions from monocular images or videos, a task fraught with intricacies. This challenge emanates from the inherent ill-posed nature of the problem, owing to the projection of three-dimensional spatial information onto a two-dimensional plane, resulting in the loss of a dimension. However, when multiple viewing angles or supplementary sensors, such as Light Detection and Ranging (LiDAR) devices, are enlisted, 3D HPE can be effectively rendered as a well-posed problem, harnessing the synergy of information fusion techniques.

Another discernible constraint is the insatiable data requirements of deep learning models, rendering them highly susceptible to the nuances of the data collection environment. In sharp contrast to 2D HPE datasets, where the acquisition of precise 2D pose annotations is a relatively straightforward endeavor, procuring accurate 3D pose annotations is a labor-intensive undertaking, marked by the impracticality of manual labeling. Additionally, the available datasets for 3D HPE are typically cultivated from controlled indoor settings, featuring a limited spectrum of routine human actions. Recent investigations have illuminated the inadequacies in the generalizability of models trained on datasets imbued with inherent biases when subjected to cross-dataset inferences.

Single-view single-person 3D Human Pose Estimation represents a domain of study that can be bifurcated into two primary categories, namely skeleton-only and Human Mesh Recovery (HMR), contingent on whether the objective is the reconstruction of

the three-dimensional human skeleton or the retrieval of the three-dimensional human mesh through the utilization of a human body model.

Skeleton-Only Methods: Within the purview of skeleton-only methods, the ultimate output revolves around the estimation of the three-dimensional positions of human joints. These methodologies refrain from the employment of human body models for the reconstruction of three-dimensional human mesh representations. This category can be further subcategorized into two subtypes: direct estimation approaches and 2D to 3D lifting approaches.

Direct estimation methods are devised to directly infer three-dimensional human pose from two-dimensional images, bypassing the intermediate step of estimating a two-dimensional pose representation. Sun et al. introduced a structure-aware regression approach, as outlined in their work [40]. In contrast to using a joint-based representation, this approach adopted a bone-based representation to enhance stability. It featured a compositional loss that harnessed the 3D bone structure to encode long-range interactions between the bones. On a similar note, Pavlakos et al. introduced a volumetric representation, as presented in their study [41]. This representation aimed to transform the highly nonlinear problem of 3D coordinate regression into a more manageable form within a discretized space. The convolutional network predicted voxel likelihoods for each joint in the volume, utilizing ordinal depth relations for human joints to reduce the reliance on precise 3D ground truth poses.

The development of 2D to 3D lifting approaches has been inspired by recent achievements in the field of 2D Human Pose Estimation and has gained prominence as a solution for 3D Human Pose Estimation. In this approach, readily available 2D HPE models are initially employed to estimate a 2D pose. Subsequently, a 2D to 3D lifting mechanism is applied in the second stage to derive the 3D pose. Pioneering contributions in this category include the work of Martinez et al. [42], who proposed the use of a fully connected residual network for regressing 3D joint locations based on

2D joint locations. Despite achieving state-of-the-art performance at the time, this method exhibited susceptibility to reconstruction ambiguities resulting from heavy reliance on the 2D pose detector.

Tekin et al. [43] and Zhou et al. [44] opted for 2D heatmaps as intermediate representations for 3D pose estimation. Wang et al. [45] introduced a pairwise ranking CNN to predict the depth ranking of pairwise human joints. Subsequently, a coarse-to-fine pose estimator was employed to regress the 3D pose based on the 2D joints and the depth ranking matrix.

In essence, the delineated methodologies represent distinctive paradigms within the domain of single-view single-person 3D HPE, with each approach bearing its unique attributes and intricacies.

In recent years, 3D Human Pose Estimation has undergone substantial advancements. A considerable proportion of 3D HPE methodologies have embraced the 2D to 3D lifting strategy, leading to notable enhancements in 3D HPE, largely attributable to the progress achieved in the domain of 2D Human Pose Estimation. Prominent 2D HPE methods, such as OpenPose [39], and HRNet [3], have found extensive utilization as 2D pose detectors in the realm of 3D HPE. Moreover, some of these methodologies extend their purview beyond 3D pose estimation, encompassing the recovery of the three-dimensional human mesh from images or videos, as evidenced in works like [46, 47].

Much like 2D Human Pose Estimation, robustness in the presence of occlusion and computational efficiency represents two pivotal challenges in the domain of 3D HPE. In scenarios characterized by high population density, the performance of contemporary 3D HPE methods experiences a significant degradation due to occlusion and potentially suboptimal each person's content.

2.2 Human Body Reconstruction

Human Mesh Recovery methods encompass the integration of parametric body models, such as SMPL (Skinned Multi-Person Linear) [1], to facilitate the reconstruction of human mesh. The SMPL model [1] plays a pivotal role in the domain of 3D Human Pose Estimation, characterized by its ability to represent natural pose-dependent deformations that encapsulate the dynamics of soft tissues. To capture how individuals deform concerning various poses, a repository of 1786 high-resolution 3D scans, featuring subjects in different poses, is leveraged alongside a template mesh within the SMPL model to optimize parameters.

In the pursuit of high-quality human mesh recovery, volumetric models come to the fore, offering additional insights into the human body's shape. Among the most renowned volumetric models is the SMPL model [1], extensively adopted in the realm of 3D HPE. This extensive usage of SMPL primarily owes to its compatibility with pre-existing rendering engines.Remarkably, Kolotouros et al. [48] departed from the conventional approach of predicting SMPL parameters and, instead, opted to regress the coordinates of the SMPL mesh vertices. They achieved this using a Graph-CNN architecture. On the other hand, Kocabas et al. [49] leveraged the extensive motion capture dataset, AMASS [50], to conduct adversarial training for their SMPL-based method known as VIBE (Video Inference for Body Pose and Shape Estimation). In the case of VIBE, AMASS played a pivotal role in distinguishing between genuine human motions and those generated by the pose regression module. Since real-world scenarios frequently feature low-resolution visual content as opposed to high-resolution imagery, existing well-trained models may falter when resolution degrades. Choi et al. [51] introduced a temporally consistent mesh recovery system, named TCMR, which serves to smoothen the output of 3D human motion through the use of a bi-directional gated recurrent unit. Furthermore, Zheng et al. [52] engineered a lightweight transformer-based approach capable of reconstructing human

mesh from 2D human pose. Notably, this approach boasts significant reductions in computation and memory costs while upholding competitive performance relative to Pose2Mesh [51].

Recent efforts have witnessed the integration of transformers in Human Mesh Recovery [53]. Lin et al. introduced METRO [53] and MeshGraphormer, amalgamating CNNs with transformer networks to regress SMPL mesh vertices from a single image. However, these methodologies primarily sought heightened accuracy at the expense of computational and memory resources. In contrast, FeatER [54] and POTTER [55] endeavored to streamline the computational and memory expenses by introducing lightweight transformer architectures. Both of these approaches manage to surpass METRO while consuming less than 10% of the total parameters and 15% of the Multiply-Accumulate operations.

Moreover, there exist several augmented SMPL-based models, developed to address the limitations of the SMPL model. These limitations encompass high computational complexity and the absence of hands and facial landmarks in the original SMPL model. In response, methods such as SMPLify [56] have arisen as optimization strategies, wherein the SMPL model is fitted to the detected 2D joints while minimizing the re-projection error. Additionally, Pavlakos et al. [57] introduced SMPL-X, an innovative model capable of predicting fully articulated hands and facial landmarks. Extending the approach of SMPLify, they introduced SMPLify-X, an improved iteration informed by the AMASS dataset [50]. Hassan et al. [58] further extended SMPLify-X, introducing PROX as a method that enforces Proximal Relationships with Object eXclusion, thereby introducing 3D environmental constraints into the framework. Kolotouros et al. [59] embraced a novel approach that integrated regression-based and optimization-based SMPL parameter estimation methods to create SPIN (SMPL oPtimization IN the loop). SPIN harnesses the estimated 2D pose as the initialization in an iterative optimization routine aimed at producing a more

accurate 3D pose and shape.

In addition to SMPL-based models, alternative approaches have demonstrated their usefulness in the pursuit of recovering 3D human pose or mesh. Chen et al. [60] introduced the Cylinder Man Model, designed for generating occlusion labels for 3D data as part of data augmentation. In this framework, a pose regularization term was incorporated to penalize inaccurately estimated occlusion labels.

Xiang et al. [61] employed the Adam model [62] for reconstructing 3D motions. They also introduced the concept of 3D Part Orientation Fields (POFs) as a representation for encoding the 3D orientation of human body parts within a 2D space.

Wang et al. [63] presented the Bone-level Skinned Model of the human mesh, which decouples bone modeling from identity-specific variations by setting bone lengths and joint angles. Furthermore, Fisch and Clark [64] introduced an orientation keypoints model proficient in computing full 3-axis joint rotations, encompassing yaw, pitch, and roll, within the context of 6D Human Pose Estimation.

2.2.1 SMPL

The SMPL (Skinned Multi-Person Linear) [1] model plays a pivotal role in the domain of human mesh recovery, demonstrating significant importance in this context. Its prominence arises from its ability to accurately represent the human body in its most fundamental form, providing researchers and practitioners with a robust and versatile tool for human body modeling. As a vertex-based linear model, SMPL excels at capturing the intricate details of minimally-clothed humans in natural poses, making it an invaluable asset in various fields such as computer graphics, computer vision, and biomechanics. Its widespread adoption within the research community underlines its relevance as the de facto standard for human body modeling. The model's compatibility with existing rendering engines ensures seamless integration into diverse applications, further solidifying its importance in the quest for accurate and realistic human mesh recovery. By precisely defining parameters for pose de-

formation and shape variation, SMPL empowers researchers to generate high-fidelity human body instances and extends its capabilities to encompass soft-tissue dynamics. Consequently, SMPL stands as an indispensable tool in the pursuit of faithful human mesh recovery, enhancing our understanding of human anatomy and facilitating advancements in fields such as virtual reality, medical imaging, and motion analysis.

The SMPL model factors deformations into shape and pose deformations, as illustrated in 2.1. It relies on two fundamental sets of parameters to control these deformations: pose parameters denoted as $\theta = [w_0^T, \ldots, w_K^T]^T$ and shape variation parameters represented by β. The pose parameters are defined by a standard skeletal rig with $K = 23$ joints, where $w_k \in \mathbb{R}^3$ represents the relative rotation of part k concerning its parent in the kinematic tree, and w_0 denotes the root orientation. The shape parameters $\beta \in \mathbb{R}^m$ are coefficients derived from the top m principal components within a low-dimensional shape space obtained through principal component analysis.

SMPL can be mathematically represented as a function $M(\cdot)$ that maps the pose parameters θ and shape parameters β to a triangulated mesh with $N = 6890$ vertices. It is formulated as an additive model within the vertex space. In essence, a posed human body instance can be generated through the following process:

$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \theta; W), \tag{2.1}$$

Where:

$$T(\beta, \theta) = \bar{T} + Bs(\beta) + Bp(\theta). \tag{2.2}$$

In this context, a rest pose $T(\beta, \theta)$ is first generated by learning corrective blend shapes, i.e., pose-dependent deformations $Bp(\theta) : \mathbb{R}^{|\theta|} \rightarrow \mathbb{R}^{3N}$ and shape-dependent deformations $Bs(\beta) : \mathbb{R}^{|\beta|} \rightarrow \mathbb{R}^{3N}$, in order to deal with standard LBS artifacts [59].

A blend shape is a vector of displacements in the mean template shape $\bar{T}$. Secondly, a linear blend skinning function W with a set of blend weights $W \in \mathbb{R}^{N \times K}$ and the pose parameters θ allow the posing of the T-shape mesh $T(\beta, \theta)$ based on its skeleton joints locations $J(\beta) : \mathbb{R}^{|\beta|} \rightarrow \mathbb{R}^{3K}$.

Furthermore, the SMPL model can be extended to encompass soft-tissue dynamics, as highlighted in the work by Pons-Moll et al. [65]. In the resulting DMPL model, dynamic deformations are parameterized using coefficients denoted as δ.

The SMPL model family has experienced significant expansion in recent times, introducing new models such as the FLAME face model [66], the MANO hands model [4], and the SMIL infant body model [67]. These models adhere to a linear blend skinning framework, which includes both shape and pose blend-shapes. Despite the success of the SMPL framework in various applications, it exhibits certain limitations. Firstly, its global blend shapes tend to capture unintended long-range correlations, resulting in non-local deformation artifacts. Secondly, SMPL does not adequately account for the interdependencies between body shape and pose-dependent shape deformation. Furthermore, SMPL relies on a linear Principal Component Analysis subspace to represent soft-tissue deformations, which presents challenges when replicating highly nonlinear deformations. These limitations have motivated numerous researchers to explore avenues for improving its descriptive capabilities, as exemplified in the work by Santesteban et al. [68].

One notable advancement in this context is the STAR model [38], which serves as a drop-in replacement for SMPL. The STAR model factorizes pose-dependent deformations into a set of sparse and spatially localized pose-corrective blend-shape functions, mitigating some of the shortcomings of SMPL. Additionally, the SoftSMPL model [68] has been introduced, defining a highly efficient nonlinear subspace for encoding tissue deformations, offering a more effective alternative to the linear descriptors found in SMPL.

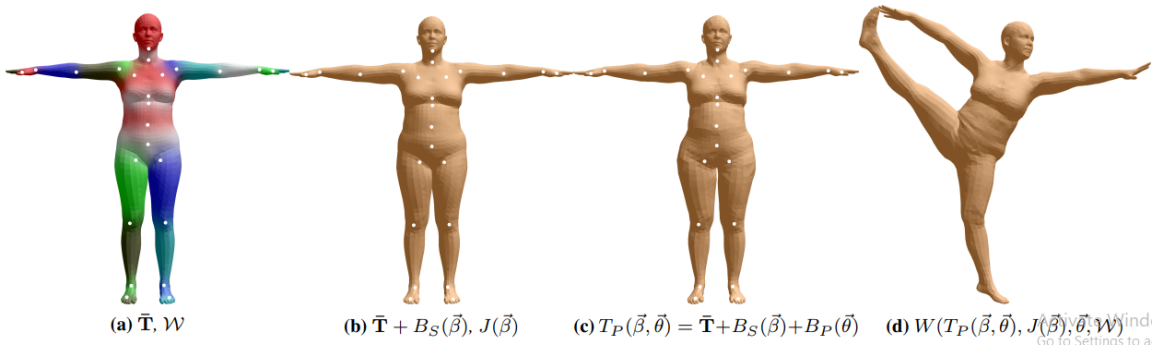

Figure 2.1: (a)A template mesh displaying blend weights represented by colors, and joints indicated in white, (b) Utilizing solely identity-driven blend shape influence, where vertex and joint positions change linearly with the shape vector, (c) Incorporating pose blend shapes in anticipation of the split pose, noticeable expansion of the hip region, (d) Adjusted vertex positions using dual quaternion skinning for the split pose. [1]

2.2.2 Datasets

Human3.6M [69] stands as a cornerstone among video-based, large-scale indoor datasets. This meticulously captured dataset employs an accurate marker-based motion-capturing system, ensuring the fidelity of motion representation. The dataset encompasses performances by 11 professional actors engaging in 17 predefined actions. As per the precedent set by [51] and [59], we judiciously selected five subjects for the training subset, reserving samples from two subjects for the evaluation subset. This selection strategy aligns with the best practices established in the field.

3DPW [70] represents a distinct dimension in the realm of human pose datasets. This outdoor dataset comprises 60 video sequences, collectively spanning 51,000 frames. What sets 3DPW apart is its capture in an uncontrolled environment, mirroring real-world scenarios. The dataset provides meticulous ground-truth 3D pose and mesh annotations, rendering it a valuable resource. In consonance with the evaluation protocol followed in seminal works like [51] and [59], we opted to focus exclusively on the test dataset. This approach ensures a robust evaluation of model performance under diverse and unstructured conditions.

MuCo-3DHP [71] is a sprawling dataset meticulously curated for advancing 3D human pose estimation and mesh reconstruction in unconstrained, real-world settings.

The dataset's sheer scale is noteworthy, with over 13 million frames of video data. This extensive corpus is the result of the contributions of five actors participating in seven distinct actions, all within diverse outdoor environments. Notably, the dataset is fortified with ground-truth 3D pose and mesh annotations, rendering it indispensable for algorithm evaluation. Our approach aligns seamlessly with the methodology set forth by [51] and [59], encompassing the usage of MuCo-3DHP for mixed training, ensuring that our models are well-prepared to tackle the complexities of the real world.

MSCOCO keypoints [72], a subset of the broader MSCOCO dataset, emerges as a linchpin in the domain of human pose estimation from images. It boasts a comprehensive repository of over 200,000 images, each meticulously annotated with ground-truth keypoint labels, spanning 17 distinct body joints. This dataset has garnered widespread acclaim and serves as the touchstone for evaluating the performance of various algorithms in human pose estimation. It serves as a robust benchmark for gauging the effectiveness of different approaches in the field. Following the lead of [51] and [59], our research has leveraged the MSCOCO keypoints dataset for mixed training, harnessing its wealth of data to enhance the versatility and real-world adaptability of our models.

2.2.2.1 Evaluation Metrics

The evaluation metrics employed in human mesh reconstruction are indispensable tools that allow researchers and developers to assess the performance and quality of their reconstruction algorithms. These metrics not only serve as benchmarks for comparing different methods but also guide the refinement and optimization of these algorithms over time. The choice and adaptation of appropriate evaluation metrics are paramount for advancing the state-of-the-art in this field.

MPJPE (mean per-joint position error) serves as a foundational metric in the evaluation of algorithms for 3D human pose estimation. This metric gauges the average

Euclidean distance between the predicted 3D joint positions and their corresponding ground-truth coordinates within a dataset. The computation of MPJPE commences with an alignment step where the predicted and ground-truth joint coordinates undergo a rigid transformation aimed at minimizing the overall spatial dissonance between the two sets of points. This alignment facilitates a meaningful point-to-point comparison, and the resulting aligned positions are employed to determine the average Euclidean distance between corresponding joint pairs. Typically expressed in millimeters, a lower MPJPE score signifies a more precise and accurate estimation of human pose.

PA-MPJPE (Procrustes aligned mean per-joint position error) represents a variant of the MPJPE metric and enjoys common usage in the assessment of 3D human pose estimation algorithms. PA-MPJPE shares the foundational principle of measuring the average Euclidean distance between predicted 3D joint positions and their ground-truth counterparts. However, what sets PA-MPJPE apart is its inclusion of an additional alignment step. Leveraging the Procrustes analysis method, this metric aligns the predicted and ground-truth joint positions through a non-rigid transformation aimed at minimizing the overall spatial divergence between the two sets of points. This enhanced alignment process yields a more robust point-to-point correspondence and, in turn, supports the computation of the average Euclidean distance between corresponding joint pairs. PA-MPJPE scores are conventionally reported in millimeters, with a lower value denoting a higher degree of accuracy in pose estimation.

MPVE (mean per-vertex error) takes center stage as a critical metric for evaluating the performance of algorithms dedicated to 3D human mesh reconstruction. It quantifies the average Euclidean distance between predicted 3D mesh vertices and the veracious ground-truth vertices within a dataset. To compute the MPVE, a pivotal alignment step aligns the predicted and ground-truth mesh vertices using a rigid transformation that minimizes the overall spatial disparity between the two sets of

points. This alignment sets the stage for calculating the average Euclidean distance between corresponding pairs of vertices. Typically denominated in millimeters, a lower MPVE score attests to the heightened accuracy and fidelity of the mesh reconstruction process.

2.3 Inverse Kinematic Solvers

In the domains of robotics and computer animation, achieving precise control over the movements of robotic manipulators or animated characters poses a fundamental and challenging endeavor. Central to the attainment of this control is the task of solving the inverse kinematics problem. Inverse kinematics, within the context of human pose and character animation, pertains to the intricate calculation of joint angles or configurations necessary to accurately position a specific body part or end-effector at a desired location and orientation. To address this pivotal challenge, various methodologies and solvers have been devised. In this section, we embark on an exhaustive exploration of diverse inverse kinematics solvers, each distinguished by its distinctive attributes and tailored applications within the realm of human pose and character animation. We will delve into the analytical inverse kinematics solver, the numerical inverse kinematic solver, the hybrid solver, and the integration of physical plausibility constraints. A comprehensive understanding of these diverse solver types, along with an awareness of their respective strengths and limitations, is imperative for the judicious selection of an appropriate solution method within the context of human pose and character animation.

2.3.1 Analytical Solvers

Analytical methods as a fundamental category of Inverse Kinematics (IK) solvers. Analytical methods aim to establish relationships between all potential solutions and specific characteristics of the mechanical system, such as the lengths of its components, the initial configuration, and rotational constraints. However, in practice, they

often rely on certain assumptions, thereby yielding only a single solution. A prototypical example employed for illustration is the planar two-link manipulator. In the case of a manipulator comprising a limited number of joints, analytical solutions can be computed explicitly by exhaustively exploring feasible link placements.

Consider a planar two-link manipulator with segment lengths l_1 and l_2, and a target position denoted as (x, y). For simplification, it is assumed that the target position resides within the operational envelope of the manipulator. The analysis reveals two viable configurations where the end effector reaches the target position (x, y). For illustrative purposes, only the rotational angles of one of these solutions are computed. The relative joint rotation angles are denoted as θ_1 and θ_2. Given the known lengths of the two links l_1 and l_2, in addition to the desired end effector position, the joint angles are determined by the following expressions:

$$\theta_1 = \cos^{-1}\left(\frac{l_1 + x^2 + y^2 - l_2^2}{2l_1\sqrt{x^2 + y^2}}\right) \tag{1}$$

$$\theta_2 = \cos^{-1}\left(\frac{l_1^2 + l_2^2 - (x^2 + y^2)}{2l_1 l_2}\right) \tag{2}$$

It is evident that as the complexity of the chain increases, the determination of relative joint angles becomes more intricate. Furthermore, the selection of an appropriate solution among multiple possible ones becomes a non-trivial challenge. A key criterion for solution selection is the stability of the derived θ values to ensure smooth and predictable motion. Transitioning from a planar to a three-dimensional space leads to an augmented number of feasible solutions, subsequently intensifying the computational complexity.

Analytical IK solutions have been a subject of extensive research in the field of robotics, addressing a wide range of manipulators and multi-body mechanisms. Numerous scholarly works have delved into the domain of analytical IK solutions, tack-

ling complex systems, including general 6R manipulators [73] and multi-body mechanisms [74]. An indispensable tool in this realm is IKFast [75], developed by Dias et al., which employs analytic approaches to solve IK equations, playing a pivotal role in real-world robotics applications. Furthermore, analytical solutions have found applications in the animation of anthropomorphic limbs. For instance, Korein's method [76] has been instrumental in manipulating human arms and legs, while Tolani and his colleagues [77] have employed the swivel representation to formulate analytic solutions for human limbs. Kallmann [78] expanded upon Tolani's work, introducing quaternion algebra to extend the analytical framework for automatically determining swivel angles. More recent developments in this field include Molla and Boulic's work [79], which introduced a middle-axis-rotation parameterization for human limbs. This innovation addresses issues of ill-conditioned cases that can arise due to the use of the swivel representation. Instead of projecting onto a fixed vector, which may lead to substantial deviations in the swivel angle, the authors define a reference coordinate system through a decomposition of rotation, effectively circumventing singularities.

In the realm of robotic motion planning, planners, often referred to as IK solvers, are mandated to efficiently handle an extensive volume of configurations, numbering in the thousands per second. Nevertheless, the intrinsic non-linear characteristics of kinematic equations, coupled with their limited scalability, render them less suited for dealing with redundant systems. These solvers frequently encounter challenges such as getting trapped in local minima, and in their rudimentary forms, they cannot effectively accommodate prioritized constraints. Analytical methods primarily cater to mechanisms with a low Degree of Freedom (DoF), failing to scale adequately to fulfill the demands posed by contemporary computer-based IK problems. For instance, even under optimal conditions, it is generally infeasible to derive a closed-form equation for the full human body, which comprises an intricate web of approximately 70 DoF. Consequently, scholars are compelled to explore alternative approaches, of-

ten relying on iterative methods to approximate viable solutions for the problem. These iterative strategies are discussed in the subsequent section, primarily founded on numerical approximations of the underlying non-linear problem.

2.3.2 Numerical Solvers

Numerical methods encompass techniques that necessitate a series of iterative steps to attain a solution meeting specific criteria of satisfaction. These iterative approaches conceptualize the problem through the formulation of a cost function, which serves as the target for minimization. Although numerical methods offer a flexible approach to tackling complex inverse kinematics problems, they are not without their limitations and challenges. One notable challenge is the potential for getting stuck in local minima during the iterative process, thus failing to reach the global solution. Additionally, numerical methods may exhibit slow convergence rates, which can be problematic when dealing with real-time applications. Moreover, the efficiency and accuracy of numerical solvers can be significantly influenced by the choice of initial guesses for the iterative process, and determining suitable initial values can be a non-trivial task. Lastly, there is the risk of encountering numerical instability in situations involving near-singular or singular configurations, further emphasizing the need for careful consideration and optimization of numerical techniques in the context of Inverse Kinematics.

2.3.2.1 Jacobian inverse Solvers

The Jacobian matrix, denoted as J, encompasses partial derivatives of the entire chain system concerning the angle parameters, θ. Jacobian solutions provide a linear approximation to the IK problem. Jacobian methods, as iterative approaches, address the IK problem by sequentially adjusting the configuration of a complete chain. This iterative process endeavors to progressively align the end effector's position and orientation with a specified target position and orientation. The differentiation of

Equation (3) yields the forward dynamics equation, expressed as:

$$\dot{s} = J(\theta)\dot{\theta}.$$

The Jacobian matrix, $J(\theta)$, is a function of the θ values and can be represented as:

$$J(\theta)_{ij} = \frac{\partial s_i}{\partial \theta_j}, \quad \text{for } i = 1, ..., k \text{ and } j = 1, ..., n,$$

where k denotes the number of end effectors and n signifies the number of joints. Therefore, J becomes a $k \times n$ matrix with vector entries, typically converted in practice to a $3k \times n$ matrix of scalar values. Following Buss's notation, the matrix entries are determined by the unit vectors, v_j, which indicate the rotation axis of the jth joint:

$$\frac{\partial s_i}{\partial \theta_j} = v_j \times (s_i - p_j),$$

where p_j represents the joint's position.

To minimize the discrepancies between actual end effector positions, $s_i(\theta)$, and target positions, t_i, we aim to find the values of θ that minimize the errors, e_i, expressed as:

$$e_i = t_i - s_i(\theta).$$

For this purpose, a small change in joint angles, denoted as $\delta\theta$, results in an approximate change in end effector positions, approximated as:

$$\delta s \approx J\delta\theta.$$

However, determining $\delta\theta$ relies on the estimation that $\delta\theta \approx J^{-1}e$. It's important to note that J may not always be square or invertible and can present singularity issues. The Jacobian method characteristically treats each joint's influence indepen-

dently of other joints, constituting a first-order approximation. As one joint undergoes a change, all segments related to it are regarded as a single rigid body. Over time, various methods have been proposed to address these challenges, particularly to circumvent singularity issues, enhance convergence, and bolster solution stability. These methods employ different approximations to resolve the IK problem. Some techniques focus on localized modifications of the inverse differential kinematic mapping near singularities, involving specifically defined mappings that relate task space to joint space, while others concentrate on smoothing overall motion [80]. Below, we provide a summary of these techniques.

The Jacobian transpose method ensures the invertibility of the Jacobian matrix by utilizing the matrix's transpose, rather than its inverse. While the Jacobian transpose method does not encounter ill-defined situations near geometric singularities, it often necessitates a substantial number of iterations to achieve convergence. Notably, it is feasible to assess the presence of singularity issues by verifying whether the Jacobian matrix contains a row of zeros.

Anomalies associated with the Jacobian transpose method have also been documented in prior research, as observed in [81]. The observed irregularities manifest as jerky movements, leading to suboptimal poses, particularly when pronounced disparities exist between the end effector positions and the specified targets. Furthermore, the transpose-based approximation does not account for the relative influence of joint variables and does not readily support strict prioritization of constrained dimensions. It's essential to consider a small value for the parameter α due to the non-linear nature of the direct kinematics model, as larger values may induce oscillations and discontinuities.

The Jacobian pseudo-inverse , also recognized as the Moore-Penrose inverse of the Jacobian, is defined by setting $\delta\theta = J^{\dagger}e$, where $J^{\dagger}$ represents an $n \times m$ matrix known as the pseudo-inverse of J. One notable advantage of the pseudo-inverse is

its applicability to all matrices J, including those that are non-square or lack full rank. The pseudo-inverse exhibits the property that the matrix $(I - J^{\dagger}J)$ performs a projection onto the null space of J. Several researchers have employed the null space method as a strategy to mitigate issues associated with singular configurations. For example, Liegeois utilized this method in [82], and Maciejewski and Klein explored its application in [83]. Additionally, Girard and Maciejewski, in their work [84], harnessed the Jacobian pseudo-inverse to address the locomotion of legged figures. Nevertheless, it is important to acknowledge that the Jacobian pseudo-inverse is not without limitations. In scenarios where a chain's configuration nears singularity, the pseudo-inverse method can lead to significant adjustments in joint angles, while the movement toward the target remains minimal. This discrepancy can result in oscillations and discontinuities in the motion.

Damped least squares is An alternative approach to the Jacobian, also recognized as the Levenberg-Marquardt algorithm, which found its early application in Inverse Kinematics through the works of Wampler in [85] and Nakamura and Hanafusa in [86]. DLS introduces stabilization to the $\delta\theta$ values, effectively mitigating many of the challenges associated with the pseudo-inverse method, particularly those related to singularities. The DLS solution can be formulated as follows:

$$\delta\theta = J^T(JJ^T + \lambda^2 I)^{-1}e, \quad \text{where } \lambda \in \mathbb{R} \text{ represents a non-zero damping constant.}$$

The choice of the damping constant λ is critical to maintain numerical stability in Equation. As noted by Buss and Kim in [87], DLS exhibits superior performance compared to both the pseudo-inverse and transpose methods. However, the effectiveness of DLS is contingent upon the selection of the damping constant λ. A larger λ value renders the solutions for $\delta\theta$ more stable near singularities but at the expense of decreased convergence rate, diminished accuracy in tracking targets, and the emergence

of oscillations and shaking.

The presence of singularities poses significant challenges to the Jacobian inversion process. To address this issue, the Singular Value Decomposition has emerged as an alternative variant of the Jacobian method, harnessing the capabilities of the pseudo-inverse matrix, as proposed in [88]. SVD is particularly advantageous for its ability to provide orthonormal bases for the fundamental subspaces of a matrix. Colome and Torras introduced a singular value filtering (SVF) approach in [89], which offers an alternative method for filtering the Jacobian matrix to ensure it maintains full rank consistently. The resulting alternative pseudo-inverse possesses singular values with lower bounds and tends to approach $J^{\dagger}$ when the singular values move away from zero.

The pseudo-inverse Damped Least Squares (DLS) method represents an extension of the DLS method, integrating the concept of Singular Value Decomposition (SVD) under the DLS framework, as introduced by Maciejewski in [90]. This enhanced approach, known as the pseudo-inverse DLS method, exhibits performance characteristics akin to the standard pseudo-inverse method when operating away from singularities. However, it excels in scenarios proximate to singularities by facilitating smoother and more stable performance.

The pseudo-inverse Damped Least Squares (DLS) method represents an extension of the DLS method, integrating the concept of Singular Value Decomposition (SVD) under the DLS framework, as introduced by Maciejewski in [90]. This enhanced approach, known as the pseudo-inverse DLS method, exhibits performance characteristics akin to the standard pseudo-inverse method when operating away from singularities. However, it excels in scenarios proximate to singularities by facilitating smoother and more stable performance.

2.3.2.2 Newton-based Sovlers

Newton methods in the context of robotics and computer graphics rely on a second-order Taylor expansion of the objective function, formulated as $f(x+\sigma) \approx f(x) + [\nabla f(x)]^T\sigma + \frac{1}{2}\sigma^T Hf(x)\sigma$, where $f(x+\sigma)$ represents the desired joint parameters, $f(x)$ signifies the current joint parameters, σ denotes the necessary modification for $f(x)$ to satisfy $f(x+\sigma)$, and $Hf(x)$ represents the Hessian matrix. This approach, as described by Nocedal and Wright in their work [91], distinguishes itself from Jacobian methods by providing a second-order approximation of the function $f(x+\sigma)$ through the quasi-Newton method, effectively circumventing singularity problems inherent in Jacobian matrices. The search direction p_N^k is determined by solving the equation $p_N^k = \nabla^2 f_k^{-1} \nabla f_k$, where $\nabla^2 f_k^{-1}$ signifies the inverted Hessian matrix and ∇f_k represents the gradient of the objective function, formulated as $\nabla f_k = J(e-g)$, with e denoting the end effector position and g representing the goal position.

However, it is essential to note that the computation of the Hessian matrix is a highly intricate process, resulting in significant computational overhead for each iteration. As a result, various alternatives have been proposed, which obviate the direct calculation of the Hessian matrix and instead rely on approximations based on function gradient values. Prominent among these methods are Broyden's method, Powell's method, Siciliano's method, and the Broyden, Fletcher, Goldfarb, and Shanno (BFGS) method [92–94].

Newton methods, identified as minimization problems, offer smooth and continuous motion without abrupt discontinuities. Moreover, they readily accommodate joint constraints. A notable method for incorporating constraints is the gradient projection approach introduced by Zhao and Badler [95], where a constrained nonlinear optimization process is employed to search for a feasible solution. Furthermore, Rose et al. [96] extended Zhao and Badler's constrained nonlinear optimization framework to handle variational constraints spanning multiple motion frames. While

Newton methods do not suffer from singularity issues associated with finding Jacobian inverses, they are characterized by complexity, implementation challenges, and a notable computational burden per iteration.

2.3.2.3 Heuristic inverse kinematics Solvers

The heuristic subset of algorithms represents a category of approaches that addresses the IK problem with a simplified methodology, avoiding the complexities of intricate mathematical equations and computations. These algorithms typically consist of elementary operations conducted iteratively, gradually converging towards a viable IK solution. Heuristic IK algorithms are characterized by their low computational overhead, allowing them to rapidly converge to the final pose. They are particularly well-suited for straightforward problems, especially when dealing with non-anthropometric skeletons, such as those resembling spiders or insects. However, a significant limitation of these heuristics lies in their propensity to produce motions and gestures that may appear unnatural or biomechanically impractical. This arises from their lack of consideration for spatio-temporal adjustments between closely positioned joints, as these algorithms tend to address each joint's constraints in isolation, without adhering to global constraints.

The Cyclic Coordinate Descent (CCD) method is an approach aimed at minimizing positional and orientational discrepancies by iteratively adjusting individual joint variables. The fundamental concept underlying CCD is to align the position of each joint with both the end effector and the target in a sequential manner. The process initiates at the end effector and progresses towards the base of the manipulator, iteratively transforming each joint angle to bring the terminal bone of the chain closer to the target position. CCD stands out for its simplicity in implementation, relying primarily on elementary mathematical operations like dot and cross products. As a result, it incurs a minimal computational burden per iteration. Moreover, CCD offers a numerically stable solution and exhibits linear-time complexity relative to the

number of degrees of freedom.

Merrick and Dwyer, as detailed in their work [97], introduced an extension of CCD capable of handling tree-articulated structures and multiple end effectors. This extended method can be sequentially applied to multiple articulated chains, segmenting the articulated structure into smaller serial chains and treating each independently.

However, CCD is not devoid of challenges. A primary issue arises from its unequal distribution of motion, as it tends to overly emphasize movements in joints nearer to the end effector, potentially resulting in the generation of unnatural postures. CCD may lead to significant joint angle rotations, which can manifest as motion with erratic discontinuities and oscillations. In certain scenarios, especially when the target is positioned close to the base, it may cause a chain to form a loop, rolling and unrolling itself before finally reaching the target. Similarly, for specific target positions, especially those requiring high precision, the algorithm may require a substantial number of iterations, resulting in a slow zigzag motion of the end effector. Even when constraints have been incorporated, the method may generate unrealistic poses, particularly in highly articulated characters, as implementing global manipulation restrictions proves to be a challenging task.

2.3.3 Hybrid Solvers

Hybrid methodologies represent an alternative category of techniques devised for addressing the IK problem. These approaches aim to mitigate the intricacy of the optimization problem by dissecting it into both analytical and numerical constituents [77]. For instance, Lee and Shin [98] devised a hybrid IK approach that amalgamates a numerical optimization methodology, predicated on penalty configuration, with an analytical strategy. The analytical component is intentionally designed to alleviate the computational load typically associated with numerical optimization. Shin et al. [99] introduced a solver tailored for full-body human puppetry, partitioning the process into three distinct sub-problems: the computation of the root position,

the determination of body posture, and the establishment of limb posture. Each of these sub-problems is independently addressed, leveraging IK techniques optimized for high performance. With a reliable root position estimate, they employ numerical optimization to fine-tune body posture, subsequently utilizing computationally efficient closed-form solutions to resolve limb postures.

Kulpa et al. [100, 101] expanded upon the approaches pioneered by Shin et al. and Tolani et al., enhancing the capability to achieve more lifelike character poses. This expansion involved extending the model to encompass control over the center of mass. They repurpose pre-recorded animations to conform to specific constraints and apply these algorithms distinctly to different body regions, encompassing the head, arms, legs, and trunk. In a similar vein, Bouenard et al. and Vahrenkamp et al. [102, 103] partitioned the problem into distinct motor tasks, each addressed independently through robotics-inspired proportional-derivative controllers.

2.4 Biomechanical Plausibility Constraints

In the context of a redundant system, exemplified by an articulated figure requiring IK solutions, it becomes imperative to incorporate joint constraints to selectively identify solutions that align with user or model-defined limitations. Diverse sets of joint and model constraints have been proposed, falling into two primary categories.

The first category of constraints pertains to high-level control attributes, encompassing considerations like positional, orientational, gaze, and balance constraints. In contrast, the second category of constraints is contingent upon goal definition, either specified by the user or dynamically determined by the environment, often arising in interaction tasks (e.g., floor contact, reaching, grasping) or collision avoidance scenarios. In the case of kinematic chains featuring multiple end effectors, an additional parameter necessitates evaluation. The presence of multiple concurrent tasks can lead to conflicts among their objectives, rendering simultaneous achievement unattainable. Consequently, an effective IK solver must weigh the significance of each task, and this

can be governed via either a prioritization or weighted approach.

The priority-based approach involves the validation of the most crucial task as a foremost priority, subsequently attempting to fulfill others only if feasible, as exemplified in works like [104, 105]. Conversely, the weighted approach seeks to derive a compromise solution by applying a weighted summation of all constraints, as outlined in [106].

In the realm of biomechanics, a joint is characterized by its positional and orientational attributes. In its most general configuration, it exhibits three Degrees of Freedom (DoF). The fundamental essence of a joint lies in its capacity to facilitate relative motion between the two limbs it connects. Within the spectrum of prevalent anthropometric joints, several distinct types merit attention.

The ball-and-socket joint, notable for its versatility, permits rotary motion in all directions within well-defined constraints. An exemplary instance of a ball-and-socket joint is found in the hip joint, enabling a broad spectrum of motion, encompassing flexion, extension, abduction, adduction, and rotation, thereby facilitating multidirectional leg movements.

Conversely, the hinge joint restricts motion to a single plane about a solitary axis, exemplified by the elbow joint. This type primarily enables flexion and extension, facilitating the bending and straightening of the forearm.

The pivot joint excels in enabling exclusive rotational movement. An illustrative example is the joint between the first and second cervical vertebrae (the atlas and axis), allowing the head to swivel from side to side.

Condyloid joints, characterized by their capacity for biaxial movements, encompass forward-backward and side-to-side motions. The wrist joint serves as a pertinent instance, affording flexion, extension, abduction, and adduction movements of the hand.

Saddle joints, akin to condyloid joints, introduce specific angle constraints that

delineate the admissible range of motion. A quintessential illustration is the carpometacarpal joint of the thumb, enabling a diverse range of movements, notably including opposition the ability to touch the thumb to the other fingers. These assorted joint types collectively assume a pivotal role in the intricate biomechanics of the human body, contributing to its diverse array of movements and functional capabilities.

In the realm of human-like models, particularly those representing legged body structures, stringent constraints are imposed on joint mobility to ensure that movements remain within plausible boundaries and avert the occurrence of unrealistic motions. Several biomechanically and anatomically accurate models have been introduced, aiming to formalize the range of motion achievable by articulated figures [107]. These models are characterized by a hierarchical structure and are distinguished by the quantity of parameters employed to delineate the motion space. Due to the inherent complexity of these models, many proposed joint configurations are subjected to simplification or approximation, often involving multiple joints. For example, the shoulder model is a complex construct composed of three different joints [108, 109]. The spine model, representing a complex arrangement of 24 vertebrae, is often simplified as a straightforward chain of joints [76, 76]. The hand model, characterized by an extensive array of joints, is the most versatile part of the body [110, 111]. Additionally, the strength model considers the forces applied from the skeletal muscles to the bones [112]. These various models collectively contribute to a deeper understanding of the intricacies of human-like biomechanical structures, offering valuable insights into the potential ranges of motion and the physiological constraints governing them.

Incorporating biomechanical constraints serves to restrict the admissible range of motion in humanoid models, thereby mitigating the occurrence of unnatural postures. However, these constraints alone do not entirely eliminate the potential for self-collisions within the model. Substantial research efforts have been directed to-

wards the detection of collisions, encompassing rigid objects [113, 114] deformable objects [115, 116], and the intricate domain of self-collisions within deformable structures, as explored by Volino and Magnenat-Thalmann [117].

Nawratil et al. [118] defined the generalized pentration depth with respect to a distance metric, enabling efficient computation. Nevertheless, when dealing with self-collisions within humanoid models, it becomes imperative to scrutinize limb constraints. This can be achieved by integrating the self-collision detection step into the IK algorithm, as demonstrated by Unzueta et al. [81]. Their approach involved the estimation of PD to preempt the penetration of elbows into the torso, thus ensuring the preservation of anatomically plausible poses.

Even when joint limits are meticulously accounted for, the attainment of unnatural postures remains a possibility. Joint limits, while crucial in defining the spatial boundaries of motion, primarily serve to reach specified target positions and adhere to rotational and translational constraints. Therefore, it becomes imperative to integrate model constraints, particularly in the context of humanoid models, which necessitate the consideration of priority factors and physiological limitations. IK, although adept at addressing kinematic constraints, falls short in directly accommodating dynamic considerations such as momentum conservation during ballistic motions. Consequently, extensive endeavors have been undertaken to design physically-grounded interactive tools that ensure the generation of plausible and biomechanically sound motions [119–121].

Physics-based character animation endeavors to guide IK solvers in producing more natural motions that remain within a physiologically feasible range. For instance, Lee and Goswami [122] harnessed momentum and inertia to enhance animation balance, while Shapiro and Lee [123] leveraged dynamic physical properties like the center of mass and angular momentum to rectify unrealistic motions. Additionally, Sok et al. [124] introduced momentum and force constraints. Inverse Kinodynam-

ics (IKD) [125] represents another kinematic workflow that encapsulates transient dynamics and permits precise space-time constraints. More recently, Rabbani and Kry [126] introduced a methodology that upholds physical principles during dynamic activities by exerting control over both the center of mass and the magnitude of the character's inertia tensor. Notably, muscle-based control methods also play a pivotal role in ensuring the plausibility and naturalness of movements [127]. Nevertheless, an alternative approach to address motion naturalness and guarantee the generation of realistic movements involves the incorporation of data-driven techniques, albeit at the expense of an intensive and time-consuming training phase.

2.5 Skeleton-based Human Action Recognition

2.5.1 Convolution Neural Network - CNN Approaches

The domain of computer vision, particularly in the context of human action recognition, relies intrinsically on the adept representation of spatial information and the intricate interplay of temporal dynamics. In this pursuit, CNNs have emerged as a preeminent framework, adept at capturing the discriminative features that underlie the complex sequences inherent to human actions. Nevertheless, akin to the challenges faced by Recurrent Neural Networks (RNNs), traditional CNN architectures confront the issue of vanishing gradients, which poses a barrier to their capacity for effective spatiotemporal modeling.

In the realm of traditional CNNs, the mitigation of the vanishing gradient predicament calls for a multifaceted array of techniques and architectural enhancements. Commencing with weight initialization strategies, such as He or Xavier initialization, is imperative to forestall premature gradient attenuation throughout the training process. The judicious selection of activation functions, encompassing ReLU, Leaky ReLU, or Parametric ReLU (PReLU), serves to circumvent the conundrum of gradient saturation. The introduction of batch normalization emerges as a pivotal facet, conferring training stability by standardizing activations and mitigating internal co-

variate shift. Architectural innovations, notably the incorporation of skip connections or residual connections as witnessed in ResNet, promote the unhindered flow of gradients, particularly in the context of deep networks. In the pursuit of capturing far-reaching temporal dependencies and abating the vanishing gradient challenge, the integration of recurrent layers like Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) proves efficacious.

The application of gradient clipping operates as a safeguard against the emergence of gradients that are excessively magnitudinous or minute, thereby upholding the stability of the training regime. A paramount facet in addressing the vanishing gradient issue resides in the deployment of attention mechanisms, which empower the network to selectively concentrate on salient features. The embrace of advanced CNN architectures, encompassing the likes of 3D CNNs or spatiotemporal CNNs meticulously tailored for the nuances of video data, innately caters to temporal dependencies. Curriculum learning, characterized by a gradual increase in task complexity throughout training, facilitates the acquisition of elementary patterns prior to embarking on the assimilation of more intricate ones, thus acting as a potent remedy for vanishing gradients.

Finally, the invocation of advanced optimization algorithms, exemplified by Adam or RMSprop, which exhibit an innate capacity to adaptively fine-tune learning rates, lends itself to the amelioration of gradient-related tribulations. In practical scenarios, the amalgamation of these methodological facets is commonplace, constituting a robust approach to effectively redress the challenge of vanishing gradients. This holistic strategy stands as a catalyst for the augmentation of the performance of traditional CNNs when confronted with spatiotemporal tasks.

Wang *et al.* [15] introduced a novel representation termed the Joint Trajectory Map (JTM), designed to encapsulate spatiotemporal intricacies within action sequences through a 2D image framework. This approach primarily focuses on the incorporation

of essential human motion attributes, specifically motion magnitude and velocity. In a departure from conventional practice, this work opts for the utilization of the Hue-Saturation-Value (HSV) color space instead of traditional RGB channels for image encoding. Within the HSV color space, heightened motion magnitude and velocity correspond to increased saturation values, thus offering an intuitive representation that directly links visual cues to dynamic motion features. The HSV color space proves instrumental in emphasizing motion-related aspects, providing a richer and more interpretable framework for spatiotemporal information representation.

Furthermore, the Joint Trajectory Map integrates this color-based encoding with comprehensive spatial mapping to represent the evolving spatial layout of key human joints and their associated motion attributes over time. This integrated approach furnishes a holistic representation, wherein spatial and spatiotemporal details are effectively fused, enabling a more comprehensive understanding of action dynamics.

Du *et al.* [14] introduced an innovative approach in which they transmute the spatiotemporal characteristics of an action sequence into an image matrix. This transformation hinges on the vertical encoding of skeletal joint coordinates into the RGB channels of a representation matrix, while the sequence's temporal evolution is embedded horizontally. This spatial-temporal encoding encapsulates the dynamic nature of the action sequence, enabling the subsequent application of CNNs for classification tasks. Notably, this encoding process includes quantization, normalization, and the conversion of the representation into an image format compatible with CNN-based networks.

Expanding on the notion of skeletal representations, Li *et al.* [128] introduced a two-stream CNN-based network that harnesses the skeleton map of skeletal motion for action recognition. This approach leverages the skeletal structure and the inherent motion patterns encoded within the skeleton map to extract discriminative features for enhanced action classification. The utilization of a two-stream architecture offers

a multi-faceted perspective, combining the advantages of both spatial and temporal skeletal information to achieve robust action recognition.

The extraction of discriminative features from the spatial relationships between skeletal joints in a human action sequence constitutes a pivotal component in action recognition. A notable approach, as demonstrated by [16], involves mapping the skeleton sequence into an image domain by quantifying the pair-wise distances between joints, a method referred to as the Joint Distance Map (JDM). For each frame in the action sequence, this approach generates a total of $m * (m-1)/2$ distances, where m represents the number of skeletal joints. These distances encapsulate critical information about the spatial layout and articulations of the human skeleton, serving as valuable descriptors for action recognition.

To address the inherent variability in sequence duration across different actions, the technique employs bilinear interpolation. This adaptive mechanism ensures that the derived JDM remains consistent across actions with varying temporal profiles, thereby enhancing the robustness and generalizability of the feature representation.

The utilization of pair-wise joint distances, as manifested through the JDM, is emblematic of the endeavors within the field of action recognition to harness detailed skeletal information to achieve finer-grained and discriminative representations for improved action classification.

Li *et al.* [129] undertook the task of addressing the intricate issue of joint co-occurrence, both within individual frames and across sequential frames, through the introduction of an end-to-end hierarchical framework meticulously crafted to facilitate enhanced feature acquisition. This framework embodies a progressive aggregation process where point-level features are systematically amalgamated into a comprehensive representation of global co-occurrence features.

The hierarchical design of this framework enables a multi-tiered approach to feature learning. At its core, it allows for the exploration of not only the individual behaviors

of skeletal joints but also their intricate interplay within the spatiotemporal context of an action sequence. By iteratively aggregating point-level information into global co-occurrence features, the framework engenders a rich, discriminative representation that encapsulates both local and global dependencies among joints.

Li *et al.* innovative hierarchical framework underscores the importance of comprehensive feature learning, shedding light on the multifaceted nature of joint co-occurrence in the context of action recognition. This approach contributes to the development of more robust and discriminating representations, ultimately enhancing the accuracy and reliability of action classification.

Ke *et al.* [130] proposed a translation, scale, and rotation invariant body part vector-based features. Essentially, they utilized the geometrical features of the skeleton to generate two sets of features 1) cosine distance and 2) normalized magnitude, which generate scaling-invariant and rotation-invariant representation. Initially, Human skeleton joints are grouped into five parts trunk, right arm, left arm, right leg, and left leg. Both cosine distance and the normalized magnitude features are computed within the body and the skeleton. Therefore, ten feature vectors are generated and fed to CNN for further feature extraction and classification.

Ke *et al.* introduced a novel approach aimed at deriving translation, scale, and rotation invariant features based on the geometric attributes of human skeletal structures. This pioneering method leverages the inherent geometrical properties of the skeletal framework to generate two distinct sets of features: 1) cosine distance features and 2) normalized magnitude features. These feature sets, in turn, provide scaling-invariant and rotation-invariant representations for improved action recognition.

The initial step in this approach involves the grouping of human skeletal joints into five distinct body parts, namely the trunk, right arm, left arm, right leg, and left leg. Within each of these body parts, both cosine distance and normalized magnitude features are systematically computed, encapsulating relevant spatial relationships.

Consequently, this method yields a total of ten feature vectors, considering both the body-specific and overall skeletal context. These feature vectors are subsequently channeled into a CNN for further extraction and classification.

Ke *et al.* innovative approach offers a comprehensive means of encoding distinctive skeletal attributes while simultaneously addressing translation, scale, and rotation invariance. By exploiting both body-specific and global skeletal information, this method contributes to the enhancement of feature extraction and, consequently, the precision of action classification.

2.5.2 Recurrent Neural Network - RNN Approaches

Human action recognition inherently hinges upon the nuanced representation of spatiotemporal dynamics. In pursuit of this objective, conventional RNN architectures have been advocated as a means to capture the discriminative features intrinsic to such sequences. However, the utility of traditional RNNs has been curtailed by the persistent challenge of the vanishing gradient problem, which impedes the effective modeling of long-range temporal dependencies.

To surmount this challenge and enhance the capacity of RNNs for spatiotemporal representation, researchers have embraced a repertoire of memory gating techniques. Prominent among these are the LSTM networks and the GRU networks. These memory gating mechanisms endow RNNs with the ability to selectively retain and update information over extended temporal spans, thus facilitating the modeling of complex and extended temporal dependencies.

The employment of memory gating techniques within RNN frameworks underscores the commitment of the research community to optimizing spatiotemporal representations for superior human action recognition. These strategies not only overcome the vanishing gradient challenge but also offer a promising avenue for capturing intricate spatiotemporal patterns.

Du *et al.* [131] introduced a sophisticated approach to address the intricate tem-

poral modeling requirements in the domain of human action recognition. Central to their method is the deployment of an end-to-end hierarchical bidirectional recurrent neural network (BRNN), meticulously designed to effectively model extensive long-term temporal sequences, crucial for understanding complex human actions. In pursuit of improved feature representation, the human body is meticulously partitioned into five distinct anatomical regions: the trunk, left arm, right arm, left leg, and right leg. This partitioning serves as the foundation for the generation of five distinct BRNN subnets, each meticulously tailored to process the nuanced temporal dynamics associated with its corresponding body part.

The approach further leverages these distinct BRNN subnets at the initial layer, enabling the comprehensive fusion of their output representations. This fusion process is meticulously reiterated in subsequent layers, thereby harnessing the collective insights derived from the diverse body parts and their associated temporal dynamics. Ultimately, the amalgamated spatiotemporal features, collectively derived from this intricate hierarchy of subnets, are seamlessly channeled into the final BRNN layer. This final BRNN layer plays a pivotal role in the precise classification of human actions, offering a robust and sophisticated solution for capturing and understanding long-range temporal dependencies associated with intricate actions.

In a related vein, Du *et al.* [132] undertook an endeavor to bolster the system's resilience and adaptability within the domain of human action recognition. To enhance the system's robustness, a pivotal enhancement was introduced during the preprocessing phase, entailing the application of random rotation transformations. This augmentation serves to diversify the dataset and fortify the model's capacity to discern actions under varying spatial orientations, contributing to a more comprehensive understanding of human actions in diverse real-world scenarios.

Moreover, to effectively address the inherent challenge of accommodating variations in human body sizes, a scale transformation technique was judiciously applied. By

integrating scale-invariant representations into the framework, the system becomes adept at recognizing actions across individuals with distinct anatomical proportions, a crucial facet of achieving robustness and generalization in human action recognition.

Salient motion patterns play a pivotal role in enhancing the efficacy of various tasks related to human action analysis. However, conventional LSTM networks encounter limitations in effectively capturing intricate spatiotemporal dynamics inherent to these patterns. In response to this challenge, Veeriah *et al.* [133] introduced an innovative approach involving differential gating mechanisms tailored for LSTM-based RNNs. The crux of this method revolves around the quantification of salient motion between consecutive frames through the concept of the derivative of state (DoS).

The introduction of differential gating for LSTM-RNN networks represents a noteworthy advancement in the field of spatiotemporal representation. The quantification of salient motion patterns via derivative of state (DoS) not only facilitates a more refined and granular understanding of motion dynamics but also significantly improves the representation of spatiotemporal information. This novel approach, as proposed by Veeriah *et al.*, constitutes a fundamental stride in the ongoing quest to extract and model intricate motion patterns, ultimately contributing to the enhancement of human action-related tasks.

The realm of skeleton-based action recognition presents a constellation of intricate challenges demanding astute solutions. Among these, the nuanced treatment of intra-frame joint spatial information constitutes a paramount concern. Typically, certain skeletal joints wield a more pronounced influence in characterizing an action, imparting disparate degrees of importance. For instance, activities like "waving" accord greater significance to the joints within the arm region in contrast to other peripheral joints within the frame.

Simultaneously, the temporal information across frames in a sequence does not uniformly hold equivalent value. Consequently, the judicious utilization of this in-

formation is indispensable. To deftly address both challenges, Song *et al.* [134] have advanced a pioneering approach, unveiling an end-to-end spatiotemporal attention model underpinned by RNN and LSTM architectures. This innovative model operates by discerning and adaptively learning the intricate interplay of intra-frame and inter-frame dependencies, ensuring the effective capture of action-specific intricacies.

The introduction of this spatiotemporal attention model stands as a hallmark in the domain of skeleton-based action recognition, facilitating the discerning selection of pivotal skeletal joints and the judicious assimilation of relevant temporal dynamics. As the field continues to grapple with its myriad challenges, this comprehensive approach has demonstrated its potential to significantly advance the understanding and recognition of complex human actions.

Human skeleton data contains unique discriminative features such as acceleration, angles, spatial coordinates, orientation, and velocity. Zhang *et al.* [135] investigated eight geometric features and evaluated them on a 3-layer LSTM network. They concluded that computing the distance between joint and selected lines outperforms the rest of the features.

Human skeletal data is replete with distinctive and discriminative attributes, encompassing diverse facets like acceleration, spatial coordinates, angular orientations, velocities, and more. This wealth of information presents a multifaceted landscape for analysis and interpretation. In a comprehensive exploration, Zhang *et al.* [135] delved into the realm of skeletal data, examining a spectrum of eight geometric features to discern their efficacy within the context of action recognition. These features were subjected to a rigorous evaluation, conducted through the lens of a 3-layer LSTM network.

The empirical findings of their investigation bore notable implications. The comprehensive evaluation revealed that among the array of geometric features scrutinized, the computation of distances between skeletal joints and selected lines emerged as the

most potent discriminator, demonstrating superlative performance when compared to other geometric attributes. This insight offers a valuable contribution to the understanding of skeletal data representation for human action recognition, unveiling a robust method for extracting salient information from the intricate skeletal structures. The work by Zhang *et al.* underscores the multifaceted nature of geometric features within skeletal data and their pivotal role in advancing the frontiers of action recognition.

2.5.3 Graph Convolution Network - GCN Approaches

The realm of Graph Convolutional Networks (GCNs) has emerged as a pivotal domain in the quest for nuanced spatiotemporal representation, particularly in the context of human action recognition. In the pursuit of this endeavor, traditional models often encounter challenges related to the effective capture of intricate spatiotemporal dynamics intrinsic to the sequences. In parallel with the analogous concerns in RNNs, the domain of GCNs confronts its unique challenge, primarily that of information diffusion and propagation across graph structures. Traditional GCNs are encumbered by the need for effective information flow, as restricted diffusion can lead to the loss of discriminative features and impede the modeling of long-range spatiotemporal dependencies.

To address these challenges and enrich the capabilities of GCNs for spatiotemporal representation, researchers have ventured into the realm of innovative techniques. Prominently, Graph Attention Networks (GATs) and Graph ConvLSTMs have risen to prominence. These methods introduce novel mechanisms for the selective aggregation and propagation of information across graph nodes, enhancing the network's capacity to capture complex and extended spatiotemporal dependencies within the data.

The integration of these advanced techniques into the GCN framework underscores the research community's commitment to optimizing spatiotemporal representations

in the context of human action recognition. These strategies not only surmount the obstacles associated with information diffusion but also open up promising avenues for modeling intricate spatiotemporal patterns, thereby advancing the state of the art in GCN-based human action recognition.

Yan *et al.* [136] introduced a pioneering advancement in the realm of spatiotemporal modeling with their innovative algorithm known as ST-GCN, denoting Spatiotemporal Graph Convolutional Networks. This seminal contribution marked a significant departure from traditional approaches by facilitating the establishment of edges not only between joints within individual frames (intra-frame) but also across frames (inter-frame). In customary graph-based techniques, the convolutional operation primarily targets neighboring joints within a receptive field, enabling effective feature extraction over adjacent spatial points. However, when applied to noneuclidean data representations, such as the spatiotemporal characteristics of human motion, this process necessitates a judicious partitioning strategy to derive a coherent and informative label map.

The ST-GCN algorithm, as proposed by Yan *et al.*, transcends the constraints of traditional convolutional operations by forging connections within and between frames, thereby enhancing the network's capacity to capture intricate spatiotemporal patterns intrinsic to human actions. This transformative approach heralds a new era in action recognition, offering an innovative perspective for graph-based modeling over dynamic, noneuclidean data.

While ST-GCN [136] has shown promise in the realm of action recognition, it is not without its limitations. ST-GCN constructs its graph representation based on the physical structure of the human body, a strategy that may not optimally capture the nuances of various human actions. For example, actions like clapping heavily rely on the relationship between the hands, a key aspect for action classification. However, ST-GCN struggles to represent such relationships effectively, as it does not establish

connections between the two hands and, in the kinematic tree, they are distantly positioned. Moreover, complex geometric features, including bone orientation and length, pose challenges for graph-based algorithms.

To address these limitations, Shi *et al.* [137] introduced an innovative approach known as the Multi-Stream Attention-Enhanced Adaptive GCN (MS-AAGCN). This end-to-end, data-driven model overcomes the shortcomings of ST-GCN by adaptively generating and learning the graph topology from input data. This adaptive graph construction allows MS-AAGCN to capture intricate dependencies and relationships, including those between nonadjacent body parts, thereby enhancing its capability to characterize diverse human actions.

In real-world scenarios, occlusion presents a formidable challenge that is often an unavoidable aspect of data acquisition. To address this challenge, Song *et al.* [138] introduced a GCN-based model designed to operate effectively on incomplete spatiotemporal skeleton data. Notably, this model incorporates a novel approach by introducing occlusion-like effects through the masking of intra-frame joints and inter-frame frames. This mimicking of occlusion scenarios enables the model to learn and adapt to the presence of missing data, thereby enhancing its robustness.

In a similar vein, Yoon *et al.* [139] introduced an innovative strategy to further fortify the model's robustness in the face of occlusion and noisy data. Their approach involves intentionally introducing noise to the skeleton data during training. This noise-injection technique assists in training the model to better handle data imperfections, ultimately improving its capacity to recognize actions in the presence of real-world occlusions and data variability.

2.5.4 Transformer-based Approaches

The domain of transformers has burgeoned as a transformative paradigm in the quest for nuanced spatiotemporal representation, especially within the ambit of human action recognition. Analogous to challenges faced by conventional models, trans-

formers confront their own set of hurdles in capturing intricate spatiotemporal dynamics intrinsic to action sequences. The key challenge is effectively modeling the dependencies between distant temporal segments while maintaining computational efficiency and avoiding vanishing gradient problems.

To circumvent these challenges and augment the capabilities of transformers for spatiotemporal representation, researchers have embarked on innovative pathways. Notably, the introduction of spatiotemporal self-attention mechanisms and sparse transformers has garnered attention. These mechanisms offer novel approaches for efficiently modeling long-range dependencies across temporal segments, enhancing the network's ability to capture complex and extended spatiotemporal patterns.

The integration of these advanced techniques into the transformer architecture underscores the research community's commitment to optimizing spatiotemporal representations in the realm of human action recognition. These strategies not only mitigate the challenges related to distant dependency modeling but also open promising avenues for the comprehensive understanding of intricate spatiotemporal dynamics, propelling the state of the art in transformer-based human action recognition.

Neimark *et al.* [140] introduced a unified model, denoted as the Video Transformer Network (VTN), aimed at addressing the challenges of weakly supervised action recognition and segmentation. The VTN is distinguished by its capacity to acquire both spatial and temporal features from video data by harnessing the power of self-attention mechanisms. Notably, the model is trained with a weakly supervised learning objective, relying solely on video-level labels, thus alleviating the need for extensive frame-level annotations. This innovative approach has yielded notable achievements, establishing new benchmarks in the realm of action recognition and segmentation.

However, it's imperative to acknowledge a limitation in the VTN's architecture. The model does not explicitly account for the interrelationships between different

perspectives or views, which can pose challenges in multi-view action recognition scenarios. In instances where the actor becomes occluded in certain viewpoints, the VTN's inability to model these inter-view dynamics may hinder its performance. Addressing this limitation to enhance the model's capacity for robust multi-view action recognition remains an area of potential research and development.

Ahn *et al.* [141] proposed spatio-temporal Transformer for human action recognition. The ST-Transformer is able to learn both spatial and temporal features from videos using a spatio-temporal attention mechanism. The ST-Transformer is able to model long-range dependencies between different frames in a video. The proposed method achieves state-of-the-art results on a variety of action recognition benchmarks.

Ahn *et al.* [141] introduced a novel and advanced approach to human action recognition with their Spatio-Temporal Transformer (ST-Transformer). This groundbreaking model exhibits a remarkable capacity for simultaneously capturing spatial and temporal features from video sequences, primarily enabled by a sophisticated spatio-temporal attention mechanism. The ST-Transformer demonstrates its excellence in modeling intricate long-range dependencies that exist among various frames within video data.

The implementation of the ST-Transformer has led to remarkable breakthroughs in the domain of action recognition, consequently establishing new benchmarks for the evaluation of action recognition methods. Its ability to holistically consider spatial and temporal aspects, together with its unique capability to handle long-range dependencies, reflects the model's prowess in addressing the complex challenges associated with human action recognition.

Although transformer architecture has demonstrated substantial advancements across diverse tasks within the realm of computer vision. However, Transformer-based methods exhibit inherent limitations in effectively learning multi-scale features from skeleton data. These multi-scale spatial-temporal features inherently encapsulate both

global and local information, which holds paramount importance for the accurate recognition of actions based on skeletal representations. In response to this challenge, Ahn *et al.* [142] delve into the multifaceted realm of multi-scale feature extraction in both spatial and temporal dimensions, ultimately proposing an innovative cross-attention mechanism designed for the seamless fusion of features across varying scales.

The proposed framework, aptly named the Multi-scale Feature Extraction and Fusion Transformer (MFEF-Former), comprises two distinct but interrelated components. First, the MFEF-SFormer module is meticulously engineered for spatial modeling, intricately capturing correlations among joints and body parts through self-attention mechanisms. Subsequently, this spatial module facilitates multi-scale spatial feature fusion via cross-attention mechanisms, fostering comprehensive understanding of the relationships between skeletal joints and anatomical body segments. Second, the MFEF-TFormer module assumes responsibility for temporal modeling, employing self-attention to capture multi-scale temporal features. These temporal features are adeptly fused through cross-attention mechanisms, facilitating an enriched understanding of temporal dynamics across various scales.

The introduction of the MFEF-Former framework represents a significant stride towards addressing the vital challenge of multi-scale feature learning in the context of skeleton-based action recognition. The framework's dual components synergistically empower it to extract, model, and fuse spatial-temporal features across diverse scales, ultimately enhancing the ability to recognize complex human actions from skeletal data.

2.5.5 Dataset

The meticulous design of a data capturing environment in the realm of action recognition hinges on the specific task at hand. The unique nuances of each task, such as gesture recognition versus human interaction analysis, necessitate tailored setups. Consider the context of a gesture recognition task, where the actions under

scrutiny are prone to manifest predominantly at the sequence's inception. In this light, an implicit assumption arises namely, the discernment of actions without the prerequisite of action detection may be deemed superfluous, thus steering the design towards a more streamlined approach.

One prevalent convention within the realm of action recognition datasets is the adoption of a fixed global sequence length, simplifying data handling and mitigating the need for supplemental preprocessing steps aimed at achieving uniform sequence length. This standardization not only streamlines dataset management but also bolsters comparability between different sequences and datasets, paving the way for seamless model evaluation.

Yet another facet of the capturing environment pertains to the presence or absence of subjects unrelated to the primary action of interest. In meticulously curated datasets, a controlled environment typically prevails, ensuring that sequences exclusively encompass the pertinent action, thus obviating the confounding influence of extraneous subjects. Such controlled settings foster data purity and minimize the potential for misleading or erroneous model inferences.

Therefore, the holistic process of designing a capturing collection environment is profoundly intertwined with the specific requirements and objectives of the action recognition task, reflecting the delicate balance between the task's inherent characteristics and the intricacies of data preparation and standardization. This interplay underpins the quality and integrity of datasets, which, in turn, serves as the bedrock for robust and reliable action recognition systems.

HOLLYWOOD Dataset [143], introduced by the INRIA Institute in France in 2008, encompasses a diverse collection of video clips, each meticulously labeled to correspond to one of eight distinct human activities. These activities encompass a spectrum of interactions and behaviors, ranging from the mundane act of getting in or out of a vehicle to more intricate actions such as answering a phone call, handshaking,

hugging, and engaging in intimate activities like kissing. Furthermore, the dataset encapsulates more quotidian activities like sitting, sitting up, and standing up. It is of particular note that the dataset was meticulously compiled from 32 distinct cinematic sources. Out of this comprehensive collection, 20 of the movies contributed to the formation of a dedicated test set, while the remaining 12 movies were harnessed to assemble the training sets. This meticulous division between training and testing sources is integral to the dataset's utility as a benchmark for action recognition and classification research.

In 2009, a notable extension of the Hollywood dataset [143], originating from the INRIA Institute, was introduced, aimed at augmenting the repertoire of action classes available for research and benchmarking in the field of action recognition. This augmented dataset [144] incorporates 12 distinct action categories, which align with the taxonomy established by the original Hollywood dataset, while also introducing four additional classes: driving a vehicle, getting in a car, partaking in an eating activity, and engaging in physical altercations or fighting. The dataset boasts a considerable compilation of 3,669 video clips sourced from a comprehensive set of 69 movies, culminating in a corpus of approximately 20 hours of footage. The richness and diversity of actions and their contextual occurrences make this dataset a valuable asset for advancing the state-of-the-art in action recognition research.

The UCF50 dataset [145], originating from the computer vision research institute at the University of Central Florida, emerged in 2012 as a noteworthy contribution to the field of action recognition datasets. The dataset's central theme revolves around an expansive collection of 50 distinct action classes, all meticulously curated from authentic YouTube videos, reflecting the true dynamism and diversity of actions encountered in the real world. Building upon the foundation of the earlier 11-category YouTube activity dataset (UCF11), this dataset markedly broadens the spectrum of action-oriented video content, encompassing a more extensive repertoire of activi-

ties and scenarios. The UCF50 dataset, with its rich and authentic source material, serves as a valuable resource for advancing the research and development of action recognition models.

In 2012, the UCF Center for Research in Computer Vision (UCF CRICV) introduced the UCF101 [146] dataset, expanding upon the UCF50 dataset [145], which originally offered 50 action classes. This extended dataset is a compilation of 13,320 videos representing 101 distinct real-world action classes, all sourced from YouTube. The UCF101 dataset stands out due to its comprehensive coverage of various actions, encompassing a wide spectrum of motion patterns and diverse perspectives, including differences in point of view and lighting conditions. This breadth of data makes it a valuable resource for action recognition research, facilitating the exploration of action-related challenges across a multitude of scenarios.

The KTH dataset [147], developed by Sweden's prestigious Royal Institute of Technology in 2004, stands as a pivotal contribution in the realm of action recognition datasets. Comprising a substantial collection of 2391 action instances captured within four distinct contextual situations, this dataset offers a rich and diverse corpus for action analysis. It encompasses a total of 25 distinct sets, each meticulously choreographed to encompass six different categories of human activities, namely running, jogging, walking, hand clapping, boxing, and waving. These activities are thoughtfully performed up to five times, featuring the dynamic engagement of 25 individual participants, collectively resulting in a multifaceted portrayal of human motion. The constituent video segments possess an average duration of 4 seconds, providing a comprehensive yet manageable timeframe for action analysis. Additionally, these video segments are consistently shot against a static backdrop, employing a single camera to maintain a consistent and controlled recording environment. The KTH dataset, with its nuanced composition and methodical design, remains an invaluable resource for advancing research in the field of action recognition.

In 2017, the DeepMind research team introduced the Kinetics dataset [148], encompassing a vast array of human activity categories. The initial release, known as Kinetics 400, featured a remarkable 400 distinct human action classes, with each class accompanied by a substantial collection of over 400 YouTube video snippets, showcasing a diverse spectrum of activities. Building upon this foundation, an enhanced iteration known as the Kinetics 600 dataset was introduced, aiming to encompass approximately 600 human action classes. For each of these action classes, a minimum of 600 video clips were meticulously curated, significantly expanding the dataset's scope. In total, the Kinetics collection comprises a staggering 500,000 short videos, each spanning approximately ten seconds in duration and meticulously labeled with a specific category. Furthermore, this dataset conveniently provides URLs for all contained video clips, ensuring easy access for researchers and facilitating the exploration of a wide range of human activities.

In 2015, a comprehensive dataset was introduced [149], featuring an extensive collection of 849 hours of films meticulously designed to showcase and categorize over 200 distinct human activities. Within this dataset, each activity class is richly represented by 137 unfiltered videos, providing a wealth of data for research and analysis. The dataset caters to the complexities of categorizing human activities through three distinct algorithms: unmodified video classification, activity classification without filtering, and activity detection. It encompasses a diverse array of intricate human activities, offering a wide spectrum of scenarios and movements, making it a valuable resource for addressing the challenges of human activity recognition and understanding.

The NTU RGB+D dataset, introduced by Shahroudy *et al.* [150] in 2016, stands as one of the most substantial multi-view action datasets available in contemporary research. This dataset encompasses a vast collection of 56,000 samples performed by 40 distinct individuals, each engaging in 60 different action classes. These actions

are thoughtfully organized into three principal categories: 40 daily activity classes, nine health-related classes, and 11 interaction classes. The dataset is recorded using Microsoft Kinect technology, offering four modalities - RGB, depth, skeleton, and infra-radiation. The data is captured from three fixed camera setups, each with a capturing angle ranging from -45 to 45 degrees. Furthermore, the distance and height of the cameras are varied to augment the view variations, enhancing the dataset's richness. Liu *et al.* [151] expanded upon this dataset, known as NTU-60, using the same capturing system and modalities. This extended dataset features 106 subjects participating in 120 action classes and encompasses an impressive total of 114,500 video samples, making it an invaluable resource for a wide range of action recognition studies.

2.5.6 Evaluation Metrics and Performance

Human activity recognition leverages several performance indicators adopted from other classification domains. We present the routinely employed metrics, encompassing precision, recall, F-score, accuracy, and the confusion matrix. Within the realm of action recognition metrics, "true positive," "false positive," "true negative," and "false negative" bear the following interpretations:

- True Positive: When both the predicted and actual activity categories align.
- False Positive: Activities that do not correspond to the target category but are erroneously projected into it.
- True Negative: Instances where both the actual and predicted activities are unrelated to the sought class.
- False Negative: Activities that should be classified under a specific category but are incorrectly predicted to be outside of that category.

The subsequent list enumerates the most prevalent performance metrics in use:

Precision: Precision gauges the ratio of true positives to the total number of predicted positive instances. It measures the model's accuracy when predicting positive cases.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

Recall: Recall, often referred to as sensitivity or true positive rate, calculates the ratio of true positives to the total number of actual positive instances. It quantifies the model's ability to identify all positive cases.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

F-score: The F-score is the harmonic mean of precision and recall, providing a balanced assessment of a model's performance on positive instances.

$$\text{F-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$$

Accuracy: Accuracy computes the ratio of correctly predicted instances (both true positives and true negatives) to the total instances. It assesses the overall correctness of the model's predictions.

$$\text{Accuracy} = \frac{\text{True Positives+True Negatives}}{\text{Total Instances}}$$

Confusion Matrix: A confusion matrix is a tabular representation of actual and predicted classifications, depicting true positives, true negatives, false positives, and false negatives. It offers a comprehensive view of the model's performance across different classes.

These metrics are fundamental tools for evaluating the effectiveness of human activity recognition systems. They help quantify the model's ability to correctly classify activities and are crucial for assessing its performance in various real-world applications.

2.5.6.1 Evaluation Protocols

Standardizing metrics is vital to ensure a fair comparison between various contributions. Depending on the task, some metrics are more intuitive than others. Furthermore, an evaluation protocol might be more suitable to show the tribulation

and complexity of the scenario. Cross-views is one of the evaluation protocols proposed by [150, 152]. Generally, this evaluation assumes samples from the same view cannot be used for training and testing. For instance, [150] proposed to use cameras 1 and 3 for training, and camera 2 is for testing only.

Cross-subject is another evaluation protocol proposed by the creator of the dataset NTU-D 60 [150, 152]. In this protocol, the subjects picked for the training cannot be selected for the testing. For example, among the 40 subjects who acted [150], 20 subjects are selected for the training, whereas the others are picked for the testing.

Lastly, Liu *et al.* [152] proposed the cross-setup evaluation protocol, which set various vertical heights, distances, and backgrounds during the capturing to include natural variation. Thus, the horizontal three-camera is fixed in terms of the capturing angle.

2.6 Radar-based Human Gait and Action Recognition

Historically, radar-based Human Activity Recognition (HAR) systems have frequently embraced classical machine learning methodologies [153]. These traditional ML algorithms are firmly grounded in theoretical principles, rendering them explicable and amenable to theoretical optimization. In contrast to their deep learning counterparts, these classical models often exhibit lower complexity, resulting in reduced computational demands. Notable among the conventional ML algorithms employed for radar-based HAR are the Support Vector Machine (SVM) [154], Dynamic Time Warping (DTW) [155], and the Random Forest Classifier [156].

For instance, in the work of Kim et al. [154], SVM was harnessed to discern human activities based on micro-Doppler signatures. The process involved the manual extraction of features from time-Doppler spectrograms. Employing a decision-tree structure, the system, comprising six SVMs, achieved success in classifying a diverse set of 12 human activities. In another study [156], a real-time gesture recognition system was constructed using a 60 GHz mm-wave radar. Here, a random forest clas-

sifier played a pivotal role in facilitating real-time gesture recognition. Additionally, in [155], an enhanced DTW algorithm was introduced for the recognition of hand gestures utilizing terahertz radar. Experimental results underscored the algorithm's proficiency in fully exploring the inherent characteristics of range profiles and Doppler signatures.

Despite being widely employed in radar-based activity recognition, traditional Machine Learning solutions exhibit notable limitations that impede the advancement of robustness and generalization. Firstly, these solutions necessitate the heuristic and manual extraction of features, heavily reliant on human expertise and domain-specific knowledge. Moreover, these hand-crafted features typically encompass low-level statistical information, including metrics such as mean, variance, frequency, and amplitude [157], rendering them highly task-specific. Consequently, models trained with such shallow hand-crafted features tend to exhibit diminished performance when applied to new datasets, particularly in comparison to their performance on the original dataset. Secondly, traditional ML methods predominantly operate on small-scale, static datasets, a scenario misaligned with the dynamic and evolving nature of real-world activity data streams. These conventional ML approaches often struggle to effectively train models in such dynamic contexts.

In contrast, the paradigm of Deep Learning, a rapidly evolving technology, has emerged as a promising avenue for overcoming these constraints. As a distinct branch of ML, DL has garnered attention for its potential to surmount these limitations. DL approaches inherently possess the capacity to automatically extract high-level, hierarchical features, eliminating the need for artificial feature engineering based on specialized domain knowledge. Furthermore, the advent of Graphics Processing Units (GPUs) has empowered DL with the ability to conduct high-speed computations on extensive datasets. DL algorithms effectively leverage parallel computing capabilities, facilitating swift data processing. DL's proficiency in feature learning from massive

datasets has not only catalyzed advancements in domains like visual object recognition, speech processing, and natural language processing [158], but it has also ushered in a new era of intelligence and versatility in Human Activity Recognition.

2.6.1 Convolutional Neural Network (CNN)

CNN draws inspiration from the structural composition of the visual cortex, characterized by a network of simple and complex cells. Shao *et al.* [159] harnessed human gait micro-Doppler features for personnel recognition utilizing a deep convolutional neural network (DCNN), achieving an impressive accuracy rate of 96.9%. Similarly, Pinyoanuntapong *et al.* [160] introduced GaitSADA, an innovative self-aligned domain adaptation approach, to enhance the generalization performance of mmWave radar-based gait recognition. This method significantly enhances identification accuracy by addressing spatial and temporal domain shifts in gait biometric data.

Lang *et al.* [161] presented a CNN-based approach for classifying human activities from microDoppler spectrograms. Their model achieved an average accuracy of 98.34% across seven activities while exploring CNN parameters and assessing noise robustness. Le *et al.* [162] proposed a learning-based technique for classifying human motions using deep learning and Bayesian optimization with radar signals. Tromme *et al.* [163] effectively employed a deep convolutional neural network (DCNN) to distinguish between the absence and presence of human gait from micro-Doppler spectrograms. Their model outperformed traditional classifiers and was validated using X-band CW radar measurements.

Kim *et al.* [164] introduced the utilization of deep convolutional neural networks (DCNNs) for human detection and activity classification through Doppler radar, achieving high accuracy rates of 97.6% for human detection and 90.9% for human activity classification. Kim *et al.* [165] investigated the optimal architecture of deep convolutional neural networks (DCNNs) for classifying human hand gestures using Doppler radar. Their model achieved an accuracy of approximately 87% through

the systematic exploration of varying hyperparameters and leveraging unique micro-Doppler signatures in spectrograms. To address challenges posed by poor lighting conditions, Zhang *et al.* [166] introduced a Doppler radar-based hand gesture recognition system employing convolutional neural networks. This approach overcomes limitations of traditional camera-based systems in low-light conditions, achieving an impressive accuracy of 98% using a cost-effective Doppler radar sensor operating at 5.8GHz. The study also investigates related factors such as recognition distance and gesture scale.

2.6.2 Micro-Doppler Signatures

The representation of micro-Doppler is of paramount significance in the context of radar-based Human Activity Recognition (HAR) [167]. This representation encapsulates crucial time-varying Doppler information, with the primary Doppler shift predominantly attributed to the torso's movement, while micro-Doppler is generated by the rotation or vibration of specific body parts, including the legs, feet, and hands. Notably, the range and velocities associated with distinct body parts often exhibit variations contingent on the particular actions executed by the target individual. Consequently, activities are discernible through the corresponding time-Doppler maps, each unique to the action being performed. These time-Doppler maps are readily attainable through the transformation of raw radar echoes utilizing Short-Time Fourier Transform (STFT) and other integrated time-frequency analysis techniques [168]. Basic human activities can be effectively identified using a simplified Continuous Wave (CW) radar system with a single transmitter and receiver, capitalizing on the time-Doppler maps. Furthermore, the intuitive and interpretable nature of time-Doppler maps has rendered them the most prevalent choice for radar-based HAR in extant literature [159, 163, 167].

In the work of Trommel et al. [163], a sophisticated 14-layer deep Convolutional Neural Network (DCNN) was employed to classify human gaits based on time-Doppler

maps. The experimental results demonstrated the ability of the DCNN architecture to effectively extract micro-Doppler features from human gaits, even when operating at lower frequencies or in conditions with low Signal-to-Noise Ratios (SNR). This performance surpassed that of SVM and artificial neural networks. Similarly, M.S. Seyfioglu et al. [169] leveraged a Convolutional Autoencoder (CAE) architecture to discriminate among 12 indoor human activities, encompassing both aided and unaided motions, often characterized by highly similar micro-Doppler signatures. The CAE model, comprised of three convolutional layers and three deconvolutional layers, was adept at capturing the subtleties within micro-Doppler signatures and achieved an impressive recognition performance of 94.2%. This approach holds promise for radar-based health monitoring systems, particularly in assisted living scenarios.

Furthermore, in a study detailed in [170], a DCNN-based system was proposed for recognizing hand gestures using time-Doppler maps. The model incorporated three convolutional layers and a fully connected layer, and its proficiency in recognizing hand gestures in uncontrolled environments was thoroughly investigated. Results underscored the variability of micro-Doppler signatures with respect to the aspect angle and distance from the radar, impacting the model's recognition performance under varying scenarios. Additionally, in reference [162], a DCNN architecture, comprising cascaded convolutional network layers for classifying human activities with time-Doppler maps, was introduced. The model's network was optimized using Bayesian optimization with a Gaussian prior process. Experimental outcomes demonstrated that this methodology outperformed three existing feature-based methods, affirming its effectiveness in radar-based HAR applications.

2.6.3 Time-range map

This particular data representation encompasses a series of temporal pulses, thereby manifesting as a temporal profile. It encapsulates the dynamic range information between the radar sensor and the subject in focus. In the context of human motion,

distinct anatomical components exhibit varying relative distances from the radar, as elucidated in Figure 9a. Consequently, although these temporal-range maps do not account for Doppler information, they retain the capacity to leverage the time varying range data pertaining to human subjects, a valuable attribute in the realm of human activity recognition [171].

In a study outlined in reference [172], temporal-range maps were employed to detect instances of falling within assisted living environments. By providing detailed range information, this approach effectively mitigated the occurrence of false alarms triggered by activities resembling falls, such as sitting. Additionally, in the work conducted by Y. Shao et al. [173], a three-layer DCNN was deployed to classify six distinct human motions, including walking, running, and boxing. This study revealed the superior robustness of temporal-range maps when compared to their temporal-Doppler counterparts, particularly in scenarios characterized by low radial velocities. Furthermore, the recognition accuracy remained stable even as the incident angle increased, an outcome attributed to the relatively invariant nature of range information in the presence of varying SNR.

2.6.4 Range-Doppler map

This representation captures the information pertaining to a mobile target at a specific temporal instant, demonstrating a remarkable capability to segregate and precisely locate distinct components of moving human body parts. Moreover, range-Doppler maps exhibit the capacity to concurrently track multiple targets, holding substantial promise for multi-faceted human activity recognition endeavors.

In an illustrative application, P. Molchanov et al. [174] harnessed a short-range monopulse Frequency-Modulated Continuous Wave (FMCW) radar system comprising one transmitter and three receivers to monitor dynamic hand gestures. The range-Doppler maps obtained from the three antennas enabled the estimation of a 4D vector representing spatial coordinates and radial velocity of the hand. Similarly,

in another study [175], a 4D vector, derived from three range-Doppler maps, was combined with a depth image mask. Subsequently, a resultant velocity layer was channeled into a three-dimensional Convolutional Neural Network (3D CNN) to discern dynamic hand gestures exhibited by car drivers. The 3D CNN adeptly extracted spatial-temporal features, an essential attribute for recognizing brief, dynamic hand gestures.

In a separate investigation detailed in [176], a strategy was proposed in which two Sparse Autoencoders (AEs) were incrementally stacked to iteratively learn sparse representations from range-Doppler maps. Classification was executed using a Softmax layer. Furthermore, reference [177] introduced a stacked AE to extract features from range-Doppler maps, followed by logistic regression to distinguish fall events from non-fall occurrences. References [176] offer pertinent examples of the application of Deep Learning techniques to range-Doppler maps in the context of Human Activity Recognition.

CHAPTER 3: A Modular Two-Stream Deep learning and Analytical Inverse Kinematics Approach From Monocular Image

3.1 Introduction

Deep learning models have ushered in a transformative era in the realm of human pose estimation from monocular images. This pursuit encompasses the endeavor to discern the coordinates of human joints within a single 2D image, a field commonly referred to as 2D pose estimation. However, in the past decade, the burgeoning interest within the research community has gravitated towards the intricate challenge of reconstructing the 3D human mesh. This newfound fascination can be attributed to the manifold practical applications of mesh representation, ranging from human-computer interaction to gaming and virtual systems. In recent years, significant advancements have been observed in the fields of 2D and 3D pose estimation, alongside the reconstruction of human meshes, as attested by the works of Liu et al. [2] and Sun et al. [3]. Nevertheless, the task of reconstructing a comprehensive mesh from a monocular image remains a formidable challenge, plagued by factors such as depth ambiguity, intricate backgrounds, complex human poses, and a paucity of annotated in-the-wild datasets.

At its essence, the reconstruction of a 3D human mesh relies on the accurate estimation of pose and shape parameters. The SMPL (Skinned Multi-Person Linear Model) [1] has garnered considerable interest among researchers as a prominent statistical human body model, with valuable insights drawn from the work of Romero et al. [4]. The pose parameters, denoted as θ, play a pivotal role in delineating the global orientation of the body's root and the relative rotations of the articulated joints, representing these rotations in an axis-angle format. In contrast, shape pa-

rameters β capture the gender-specific variations in body shape, acquired through Principal Component Analysis and informed by the CAESAR dataset [5].

The literature within the field identifies two fundamental paradigms for human mesh reconstruction: the optimization-based and the regression-based approaches. The optimization-based paradigm revolves around the minimization of the discrepancy between the 2D pose, obtained through re-projection from the human mesh, and the estimated 2D pose. In contrast, the regression-based paradigm aims to directly predict the pose and shape parameters, eliminating the requirement for intermediary optimization processes.

Significantly, the characteristics of the input provided to deep learning models play a pivotal role in shaping the system's workflow. Recent advancements in the realm of human mesh reconstruction have demonstrated a preference for an end-to-end methodology where the initial scene image serves as the input, and the ultimate objective is the prediction of the mesh. Nevertheless, this approach, owing to the inherent intricacies of deep models, frequently leads to computationally intensive procedures that may not conform to the efficiency demands of practical real-world applications, such as gaming and real-time mesh recovery.

In comparison to image-based inputs, skeleton data is sparse and poses a distinct computational challenge. Consequently, efforts have been made to harness the utility of skeleton data to alleviate these computational burdens, as expounded by Choi et al. [51]. Nonetheless, relying solely on a skeleton-based representation as input proves insufficient to tackle the computational intensity associated with mesh reconstruction. To address this, Zheng et al. [52] have proposed the utilization of a lightweight graph-based transformer architecture with a distinct emphasis on computational efficiency. This approach represents a promising stride toward making mesh reconstruction more accessible and computationally tractable within real-world, resource-constrained applications.

under the scrutinizing gaze of computational vision, the human form has been subjected to abstraction, akin to the creation of a stick figure representation [178], a canvas where cardinal anatomical landmarks in the body, hands, and visage are meticulously inscribed and artfully connected through the sinews of visual threads. However, the intrigue lies in the realization that our engagement with the world transcends the stark simplicity of key points, delving into the realm of tangible surface contacts and nuanced facial expressions. This necessitates the synthesis of both corporeal pose and shape, unveiling the profound conundrum that envelopes the modeling of human form.

In the nascent stages of exploration [179], the endeavor took shape through an exploration of diverse geometric primitives, their collective resonance echoing in the quest to approximate the essence of body shapes. A pivotal moment arrived as inspiration emanated from the watershed [180] in facial modeling, offering a blueprint for the derivation of body shape constraints from the troves of 3D scanned

3.1.1 Challenges

Reconstructing mesh parameters is challenging task due to many factors:

- Dataset attributes: the indoor/outdoor setting or number of data points do not have a significant impact on the performance of the model [181]. Instead, human pose and shape, camera characteristics, and backbone features are more important for achieving satisfactory results. Additionally, having a diverse range of these attributes can lead to better performance. Further, occlusion and SMPL fittings can help improve the accuracy of recovery.

- Mix dataset training: [181] The choice of datasets is crucial for the generalizability and accuracy of the model. It is essential to use the same combination of datasets to fairly evaluate and compare the effects of other factors, such as training algorithms. For instance, comparing two network architectures on

different dataset combinations is unfair. This is often overlooked in previous studies. To obtain a robust baseline model, it is recommended to use more challenging datasets and increase their contribution during training.

- Noisy Annotations: A substantial presence of data samples containing noise can have an adverse impact on a model's performance, especially when both the SMPL annotation and keypoints exhibit such disturbances. Nevertheless, marginally noisy SMPL data may still offer advantages in the context of training.

- Occlusion: occlusion or self-occlusion represents a formidable obstacle to the accurate reconstruction of three-dimensional human models from two-dimensional images or depth data. Occlusion refers to the scenario where a portion of the human body is concealed or obscured by objects, clothing, or other body parts, leading to an incomplete visual input. Self-occlusion, on the other hand, pertains to the scenario in which body parts, often due to complex poses or anatomical structures, obstruct the visibility of other body regions. Overcoming these challenges necessitates the development of robust computer vision algorithms and machine learning techniques capable of inferring concealed or occluded regions while preserving the spatial coherence and anatomical accuracy of the reconstructed human mesh. Addressing these issues is critical for applications such as virtual reality, biomechanical analysis, and human-computer interaction, where precise and comprehensive human mesh recovery is indispensable.

3.1.2 Our Contribution

In this undertaking, our goal is to develop a modular, end-to-end capable vision transformer with a two-streams graph-based architecture and an analytical inverse kinematic framework. This system is designed to excel in the precise estimation of human shape and pose, ultimately delivering results that are on par with the latest SOTA techniques..

- To establish a fundamental framework for the task of 3D pose estimation within the realm of deep learning, it is imperative to integrate analytical inverse kinematics, a critical component contributing to the attainment of elevated precision and accuracy in the computational modeling process.

- To engineer a deep learning model for the inference of 3D human pose from monocular images, it is imperative to adhere to a set of stringent criteria encompassing efficiency, robustness, lightweight architecture, modular design, and the facilitation of end-to-end learning.

These efforts signify our commitment to enhancing the accuracy and efficiency of human parameter estimation in computer vision applications.

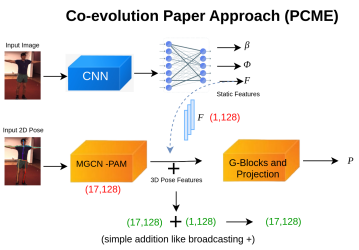

Figure 3.1: Simple addition fusion strategy

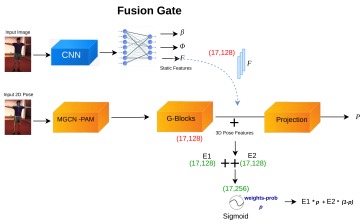

Figure 3.2: Concatenation and sigmoid activation fusion strategy

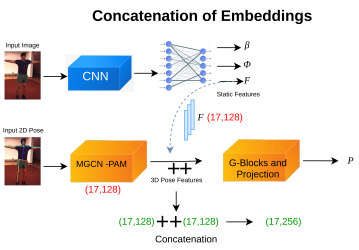

Figure 3.3: Concatenation Fusion

3.2 Camera Model

The realm of 3D human pose estimation from 2D image data is currently witnessing a profound metamorphosis, driven by the expanding frontiers of computer vision applications across an array of domains. These applications span healthcare, sports analysis, security, and entertainment, among others. The pivotal role of accurate and dependable 3D human pose estimation in these diverse domains cannot be overstated, as it facilitates the comprehension of human actions, interactions, and behaviors from visual data.

Amid the multifaceted determinants that influence the quality of pose estimation, the camera model emerges as a pivotal element with a profound and far-reaching im-

pact on the entire process. Functioning as the fundamental interface between the 3D physical world and the 2D image plane, the camera model delineates the regulations governing the projection of real-world spatial coordinates of anatomical landmarks or body joints onto the 2D image plane. Essentially, it provides the geometrical framework and context essential for the interpretation of the visual scene. Consequently, the significance of the camera model in 3D human pose estimation cannot be overemphasized.

The variations in camera parameters, including but not limited to focal length, sensor dimensions, distortion coefficients, and the intrinsic matrix, introduce distinctive characteristics and aberrations to the image data. These variations often stem from differences between individual cameras, leading to intricate issues of calibration and alignment. Neglecting to accommodate these camera-specific variations can result in substantial errors in pose estimation, directly impacting the applicability and trustworthiness of the entire system.

Camera models, encompassing diverse projection types such as perspective, orthographic, and weak-perspective, exert direct influence on the spatial and depth attributes of the estimated 3D pose. The choice of camera model, far from being an inconsequential decision, profoundly affects both the precision and scale of the reconstructed human pose. Furthermore, the selection of a camera model carries implications for the computational complexity of the pose estimation algorithm, influencing its efficiency and real-time feasibility.

In a landscape where the demand for 3D human pose estimation is witnessing remarkable growth across a spectrum of industries, the meticulous modeling of the camera becomes an imperative. In this dynamic milieu, it is essential to navigate through the intricacies of camera calibration, lens distortion correction, and the judicious selection of appropriate camera models to ensure the most accurate 3D pose estimates. The influence of the camera model transcends the realm of technicalities;

it extends its reach to practical applications encompassing human motion analysis, medical diagnostics, and gesture recognition.

This encompassing introduction underscores the pivotal role played by the camera model in the domain of 3D human pose estimation. It accentuates its profound impact on precision, robustness, computational efficiency, and the broader implications for harnessing the potential of computer vision in decoding the nuances of human behavior and interactions. It underscores the criticality of delving into the intricacies of camera modeling for the advancement and maturation of this dynamic field.

3.2.1 Coordinate Systems in Computer Vision

The precise handling of coordinate systems in computer vision is paramount to understanding the physical world, enabling accurate measurements, and facilitating the development of applications across various domains. This introductory discussion explores the importance of three fundamental coordinate systems - World Coordinates, Camera Coordinates, and Image Coordinates, as well as the challenges and significance of seamless transformations between these frames.

3.2.1.1 World Coordinates (XYZ)

In the realm of computer vision and 3D modeling, the world coordinate system plays a pivotal role as a fundamental reference frame for specifying the spatial relationships and positions of objects within a given scene. This global coordinate system establishes a fixed point of reference, independent of any particular viewpoint or image frame. Denoted as "W," the world coordinate system provides an absolute spatial reference that ensures context and consistency across various applications. It underpins the foundations of 3D object tracking, computer graphics, robotics, human pose estimation, and many other domains. The world coordinate system is characterized by several key attributes:

- Absolute Spatial Reference: One of its primary functions is to provide an abso-

lute spatial reference within a scene. This reference frame enables the precise definition of object positions and orientations without being influenced by the viewpoint of a camera or the constraints of an image frame.

- Origin and Axes: Typically, the world coordinate system is defined with a specific origin, often represented as (0, 0, 0), and a set of orthogonal axes. These axes can include the x-axis, y-axis, and z-axis, creating a right-handed or left-handed coordinate system. Standard conventions are employed to maintain consistency and compatibility across different applications.

- Consistency and Standardization: The world coordinate system ensures a consistent frame of reference across various devices, sensors, and applications. By providing this standardized reference frame, it becomes possible to seamlessly compare and integrate data from different sources.

- Transformation Hub: While objects and entities are frequently defined within the world coordinate system, there may be a need to transform them into other coordinate systems, such as camera or image coordinates, for various forms of analysis and rendering. The ability to accurately transform data between these coordinate systems is critical for mapping information captured in one frame of reference into the world coordinate system and vice versa.

- Scene Modeling: In the context of modeling complex 3D scenes, the world coordinate system acts as the foundational canvas on which objects, camera viewpoints, light sources, and other elements are positioned and defined. This simplifies the process of spatial layout and the representation of the entire scene.

However, the significance of the world coordinate system is not without its challenges. Its importance becomes evident when dealing with data from multiple sensors, devices, or sources, each having its own local coordinate system. The accurate align-

ment and registration of data from these diverse sources into the world coordinate system can be a complex and computationally demanding task.

3.2.1.2 Camera Coordinates (XYZ)

The camera coordinate system serves as a critical intermediary between the world and image coordinate systems, facilitating the mapping of 3D objects and scenes into 2D images captured by a camera. In computer vision and 3D modeling, understanding and effectively utilizing the camera coordinate system is essential for various applications, including visual perception, autonomous navigation, robotics, and augmented reality. The following are Key Characteristics of the Camera Coordinate System

- Local Reference Frame: The camera coordinate system provides a local reference frame specific to each camera's viewpoint. It defines the camera's position, orientation, and optical properties within the broader world coordinate system.

- Camera Origin: At the core of the camera coordinate system is the camera's optical center or focal point, often denoted as "C." This point represents the origin of the camera's local coordinate system and is crucial for determining the 3D positions of objects with respect to the camera.

- Coordinate Axes: The camera coordinate system typically employs a right-handed coordinate system with three principal axes: the optical axis (z-axis), which represents the camera's line of sight; the horizontal or x-axis; and the vertical or y-axis. These axes define the orientation of the camera.

- Field of View (FoV): The camera's field of view is determined within its coordinate system, specifying the extent of the scene that the camera can observe. It plays a crucial role in understanding what portion of the world is captured in the camera's image.

- Intrinsic Parameters: Within the camera coordinate system, intrinsic parame-

ters, such as the focal length, principal point, and lens distortion coefficients, are defined. These parameters influence how the 3D world is projected onto the 2D image.

Navigating the camera coordinate system introduces several intricate challenges. One of the foremost challenges is camera calibration [182], which involves determining the intrinsic and extrinsic parameters of the camera. Accurate calibration is imperative for mapping 3D world coordinates to 2D image coordinates precisely. Lens distortion, which can distort the perception of objects in images, must also be corrected. Failure to address these calibration and distortion challenges can lead to significant errors in tasks like object localization and 3D reconstruction.

Data transfer between systems [183] presents another hurdle. As cameras operate in their individual coordinate systems, there is a need for data synchronization and transformation between these systems. Coordinating multiple cameras for multi-view applications, such as 3D reconstruction and stereo vision, requires the seamless transfer of information. Achieving this synchronization, especially in real-time applications, is a complex endeavor.

Maintaining accuracy and precision [184] is an ongoing challenge. Factors such as lens imperfections, environmental conditions, and camera positioning errors can affect the accuracy of the camera's perception of the world. Mitigating these issues and ensuring consistent accuracy across varying scenarios is a constant concern.

Real-Time processing [185] is vital for applications like robotics and augmented reality, where timely decision-making and interaction are required. Managing the camera coordinate system in real-time adds to the complexity. Achieving low-latency processing while maintaining high precision is a substantial challenge.

Adapting to diverse scenarios [186] poses additional challenges. Camera setups may vary widely, and adapting the coordinate system to diverse scenarios, from controlled indoor environments to uncontrolled outdoor settings, necessitates robust solutions.

In conclusion, addressing these challenges is pivotal for the reliability of computer vision systems that rely on the camera coordinate system. Overcoming these intricacies requires a combination of accurate calibration, advanced algorithms, real-time processing capabilities, and adaptability to varying scenarios. Meeting these challenges head-on ensures that the camera coordinate system can be harnessed effectively for a wide range of real-world applications, from robotics and augmented reality to 3D reconstruction and beyond.

3.2.1.3 Image Coordinates (xy)

The Image Coordinate System, a fundamental framework within the realm of computer vision, plays a pivotal role in translating the visual world captured by cameras into digital information. It is a 2D plane used to represent the spatial positions of objects, features, and entities perceived by imaging devices. This coordinate system is essential for a wide array of computer vision applications, such as object recognition, tracking, and scene analysis, and serves as a bridge between the real-world environment and the digital domain. The following are Key Characteristics of the Camera Coordinate System

The Image Coordinate System (ICS) is an essential component of computer vision, providing a foundational framework for representing two-dimensional spatial information within the image plane. In the context of ICS, spatial positions of objects and features are described by two coordinates, conventionally denoted as (x, y). Here, pixels serve as the fundamental discrete units, with each pixel corresponding to a specific location in the image. The origin of the ICS is usually established at the top-left corner of the image, marked as (0, 0), where the x-coordinate increases from left to right, and the y-coordinate increases from top to bottom, mirroring the conventions employed by many display systems. This fundamental framework has been widely adopted in computer vision applications and serves as the basis for processing image data [187].

The spatial resolution of the ICS is determined by the total number of pixels in the image, effectively defining the precision with which spatial locations can be represented. The distance between any two points within the ICS is calculated using the Euclidean distance formula. This system also allows for geometric transformations, including scaling, rotation, and translation, which are instrumental in addressing lens distortions, aligning images, and making other essential geometric corrections [188].

One of the primary advantages of the ICS is its capacity to map real-world objects and features to specific pixel locations in the image. This mapping forms a crucial bridge between the two-dimensional image plane and the three-dimensional world coordinates, enabling the translation of real-world scenes into digital representations. This capability has profound implications for a wide range of computer vision applications, as it facilitates the extraction of depth and spatial information from two-dimensional images, contributing to critical tasks such as object detection, image segmentation, and feature extraction [189].

However, the ICS is not without its challenges. Issues such as transformation ambiguity [], stemming from the difficulty in accurately recovering 3D information from 2D images, and image distortions caused by lens effects or other environmental factors are central concerns. Overcoming these challenges is crucial for achieving precise 3D reconstructions and measurements.

In conclusion, the Image Coordinate System serves as the cornerstone for numerous computer vision tasks and image processing applications. Understanding its characteristics and conventions is essential for effectively working with digital images and leveraging the power of computer vision in diverse fields, from robotics and augmented reality to medical imaging and autonomous vehicles.

3.2.2 Coordinate System Transformations

In the field of computer vision, the ability to navigate between different coordinate systems is paramount for tasks such as 3D human pose estimation, object recognition,

and image processing. Understanding how to convert between world, camera, and image coordinate systems is essential for accurate spatial analysis. This section will delve into the mathematics and methodologies of such conversions.

As the demand for robust and accurate 3D analysis continues to expand across various industries, mastering the intricacies of coordinate systems, their transformations, and the associated challenges becomes pivotal. The ability to seamlessly transfer data between different frames of reference empowers applications ranging from autonomous navigation and robotics to virtual reality and augmented reality. In this context, the world coordinate system remains a cornerstone of modern computer vision and 3D modeling, providing a reliable and consistent global reference frame for spatial understanding.

The transformation from the world coordinate system (WCS) to the camera coordinate system (CCS) is a crucial step [190]. It involves accounting for the position and orientation of the camera in the world and is typically achieved through mathematical operations such as rotations and translations. Proper calibration and camera matrix computations, as discussed by Zhang *et al.* [191], are key aspects to ensure that real-world objects are accurately projected into the camera's image plane.

Once data resides in the camera coordinate system, mapping it to image coordinates is essential. Mapping from camera to image coordinates is a process that considers the intrinsic parameters of the camera, including focal length and optical center, as explained by Faugeras *et al.* [192]. Geometric transformations, such as perspective projection, allow the mapping of 3D points in the camera frame to their 2D positions in the image plane.

Addressing lens distortions is another crucial aspect of the transformation from camera to image coordinates, a topic well-covered by Heikkila *et al.* [193]. Radial and tangential distortions can significantly affect the accuracy of pixel positions. Correction for these distortions is vital, ensuring that straight lines in the real world

remain straight in the image, and spatial measurements are precise.

Each of these coordinate system conversions is critical for various computer vision applications, from augmented reality and autonomous navigation to medical imaging and 3D reconstruction. They form the backbone of spatial analysis in computer vision and highlight the importance of understanding these transformations for accurate, real-world applications. These conversions enable the seamless integration of computer vision into diverse domains.

World to Camera Coordinate system transformation : The transformation from world coordinates to camera coordinates is a fundamental aspect of computer vision and 3D geometry, playing a pivotal role in various applications ranging from robotics and augmented reality to computer graphics and 3D reconstruction. This section delves into the importance of this transformation, highlighting its essential role and real-world applications.

For 3D reconstruction from 2D images, converting world coordinates to camera coordinates is a fundamental step. It allows the projection of 3D points onto a 2D image plane, facilitating tasks like structure-from-motion and multi-view stereo. These techniques are utilized in fields like photogrammetry, medical imaging, and cultural heritage preservation.

The transformation is typically achieved through a combination of rotation and translation, where the world coordinates are converted to camera coordinates. This transformation can be mathematically represented as follows:

$$X_c = R \cdot X_w + T$$

which also represented as:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix}$$

Where X_c represents a 3D point or vector in the camera coordinate system, X_w represents the same 3D point in the world coordinate system, R is the camera rotation matrix, describing the orientation of the camera relative to the world coordinate system, and T is the translation vector, representing the displacement of the camera origin with respect to the world origin.

In this context, R and T are often referred to as the extrinsic parameters of the camera. The extrinsic parameters are crucial for understanding the spatial relationship between the camera and the world. They enable accurate 3D-to-2D projections, which are essential for tasks like human pose estimation, object localization, and scene reconstruction.

In practical applications, obtaining precise values for the extrinsic parameters often involves camera calibration procedures. These procedures ensure that the transformation from world to camera coordinates is accurate, allowing for reliable analysis of the 3D-to-2D mapping.

In conclusion, This transformation from world coordinates to camera coordinates significantly influences the performance of computer vision systems, impacting the accuracy and reliability of tasks that rely on spatial information. Accurate extrinsic parameters are critical for tasks such as human action recognition, object tracking, and 3D scene reconstruction.

Camera to Image Coordinate system transformation : The transformation from the camera coordinate system to the image coordinate system holds significant importance in the realm of computer vision. This essential process facilitates the projec-

tion of 3D information onto a 2D image plane, a fundamental step for a multitude of applications in computer vision and beyond. Understanding the implications of this transformation and the parameters that underpin it is pivotal for the accuracy and reliability of various computer vision tasks.

The camera coordinate system, representing the 3D world, is defined by the camera's position and orientation. In contrast, the image coordinate system serves as the bridge that translates real-world spatial coordinates into pixel positions in the 2D image. This translation is made possible through several key parameters.

First and foremost, the focal length (f) of the camera lens plays a central role in this transformation. It essentially determines the camera's field of view and directly affects how 3D objects are projected onto the 2D image plane. A longer focal length corresponds to a narrower field of view, influencing the scale and representation of objects in the image.

The principal point (u_o, v_o), another critical parameter, represents the optical center of the camera. This point marks the intersection of the camera's optical axis with the image plane. The principal point serves as the origin of the image coordinates and is essential for addressing lens distortions and realigning the image.

Lastly, the intrinsic matrix (K) combines the focal length, principal point, and other intrinsic parameters. This matrix holds the key to accurately mapping 3D points in the world into the 2D image plane. Its precise calibration is paramount in ensuring the fidelity of the transformation.

The significance of this transformation ripples through diverse computer vision applications. Object detection relies on it for accurate localization and identification by projecting 3D object locations into the 2D image. Pose estimation, which determines object positions and orientations, depends on precise projections to establish relationships between 2D image data and real-world coordinates. Furthermore, scene reconstruction for purposes such as augmented reality and environment mapping hinges on

this transformation to construct precise 3D models from 2D image data.

Mathematically, Given a 3D point in camera coordinates, denoted as $X_c = [X, Y, Z]$, the corresponding 2D point in the image coordinates, denoted as $x = [u, v]$, can be obtained through the intrinsic matrix K. The intrinsic matrix is defined as:

$$K = \begin{bmatrix} f & 0 & u_o \\ 0 & f & v_o \\ 0 & 0 & 1 \end{bmatrix}$$

Where: - f is the focal length, (u_o, v_o) are the coordinates of the principal point, and The last row is typically $[0, 0, 1]$ for normalization.

The transformation is represented mathematically as:

$$x = K \cdot X_c$$

The accuracy of this transformation greatly influences the precision of object localization and measurement in the image, making it a critical component in applications such as robotics, augmented reality, and 3D reconstruction. It is essential to ensure that the intrinsic parameters in the intrinsic matrix K are accurately calibrated to obtain reliable and accurate results.

Challenges and potential sources of error in this transformation encompass lens distortions, variations in focal length, and calibration inaccuracies [182]. Ensuring accurate calibration and a deep comprehension of intrinsic parameters are critical steps in mitigating these challenges [184]. Real-world scenarios further complicate matters by introducing perspective distortions and occlusions, underscoring the need for an exacting transformation process [183] .

In conclusion, the transformation from the camera coordinate system to the image coordinate system forms the bedrock of computer vision. It functions as the conduit between the 3D world and 2D image data, enabling a wide spectrum of applications.

Profoundly grasping the parameters at play and their role is essential for achieving the precision and dependability required for various computer vision tasks.

3.2.3 Camera projections

Camera models play a pivotal role in mapping the 3D physical world to 2D image space, influencing the accuracy and efficacy of the reconstruction process. In this section, we delve into the importance of various camera models in human mesh reconstruction, the assumptions underlying their use, and the implications of choosing one model over another.

Camera models serve as the bridge between the 3D world, inhabited by humans and their movements, and the 2D images captured by cameras. They provide the geometrical framework required for the mapping of human body landmarks to image coordinates. Accurate human mesh reconstruction depends on the precise modeling of this relationship. Various camera models, including perspective, orthographic, weak-perspective, and others, offer distinct advantages and are chosen based on the specific requirements of the application.

The choice of camera model comes with implicit assumptions about the nature of the scene and the human subjects. For example, the perspective camera model assumes a pinhole camera capturing the scene, which may not hold true in all situations. The consequences of these assumptions can be profound. Choosing a perspective camera model for a scene that deviates from its assumptions, such as objects at varying depths or wide-angle views, can lead to significant distortions in the reconstruction. Therefore, careful consideration of the scene and the model's suitability is imperative.

The selection of a camera model should align with the specific requirements of the human mesh reconstruction task. Perspective projection is well-suited for scenes with objects at different depths, where the perspective effect is significant. Orthographic projection, on the other hand, is valuable for technical applications and scenarios where maintaining object sizes is crucial. Weak-perspective projection finds utility

when objects are relatively far from the camera. Choosing the appropriate camera model enhances the accuracy and reliability of the reconstruction.

Existing methods for 3D HPS estimation often resort to making a series of simplifying assumptions, including the application of a weak-perspective projection, a constant and large focal length, and the assumption of zero camera rotation. Regrettably, these assumptions tend to inadequately represent the complexities inherent in the image formation process, resulting in errors in the reconstructed 3D shape and pose.

The inherently ill-posed nature of recovering high-dimensional 3D structures from 2D observations compounds this challenge. To mitigate this, existing data for comprehensive 3D supervision often relies on controlled laboratory settings, synthetic image data, or, more recently, data acquired in the wild for reference [46, 67]. Nevertheless, despite the substantial advancements, it is noteworthy that many state-of-the-art (SOTA) methods [194–197], persistently rely on a range of simplifications regarding the image formation process.

Primarily, these methods universally adopt the weak-perspective or orthographic projection assumption, leading to a simplified camera model characterized by merely three parameters, designed to capture the camera's translation concerning the subject. In addition, some of these methods [49, 57, 59] enforce the focal length to be a predefined large constant for every input image. Furthermore, they collectively assume zero camera rotation, blurring the distinction between body and camera rotations, thereby introducing formidable challenges in accurately estimating the body's orientation in 3D.

While these assumptions may hold true for images in which subjects are approximately perpendicular to the principal axis and located at a considerable distance from the camera, they prove inadequate for most real-world images depicting individuals. In such images, perspective effects, such as foreshortening, are conspicuously

manifest. The disregard for perspective projection inevitably results in pronounced errors in the estimation of pose, shape, and global orientation.

3.2.3.1 Perspective Camera Model

The Perspective Projection Camera Model stands as a fundamental pillar within the domains of computer vision and computer graphic [190]s. In an age defined by the synthesis of reality and virtuality, this model plays a critical role in enabling the transformation of three-dimensional worlds into two-dimensional representations. The ubiquity of this model, from photography to augmented reality, attests to its indomitable significance in modern technology.

In the intricate realms of computer vision, computer graphics, and imaging sciences, the Perspective Camera Projection Model stands as a fundamental concept. It serves as the cornerstone of understanding and translating the complexities of the three-dimensional world into the two-dimensional confines of an image. With its rigorous mathematical foundations, versatile applications, and inherent trade-offs, this model plays a pivotal role in modern imaging technologies.

At the core of the Perspective Camera Projection Model lies a set of precise mathematical equations that dictate the transformation from a point in three-dimensional space to its corresponding position on a two-dimensional image plane. These equations, underpinned by principles of geometry and optics, ensure that the model captures the visual accuracy of a scene, mirroring the way humans perceive the world. The mathematical formulation can be succinctly expressed as follows:

3.2.3.2 Orthographic Camera Model

The Orthographic Camera Model is a fundamental abstraction in computer vision and computer graphics that characterizes a camera's projection of three-dimensional (3D) points onto a two-dimensional (2D) image plane [190]. Unlike perspective projection models, which incorporate depth-related distortion, the Orthographic Camera

Model assumes a parallel projection, making it a valuable tool in various applications, such as computer-aided design, remote sensing, and orthographic rendering.

Mathematically, the Orthographic Camera Model can be represented as follows:

Consider a 3D point in the world coordinate system, denoted as $X = [X, Y, Z]^\top$, where X, Y, and Z represent the coordinates along the x, y, and z axes, respectively. The projection of this 3D point onto the 2D image plane yields its image coordinates $x = [x, y]^\top$. The camera intrinsic parameters are encapsulated in the Orthographic Projection Matrix, $\mathbf{M}$, defined as:

$$\mathbf{M} = \begin{bmatrix} s & 0 & 0 & 0 \\ 0 & s & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Here, s is a scaling factor that determines the scale of the projection. The Orthographic Camera Model effectively eliminates the depth component (Z) from the projection, resulting in a parallel projection of 3D points onto the image plane.

The mathematical representation of the Orthographic Camera Model is characterized by the following equation:

$$\begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \mathbf{M} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Where $\mathbf{M}$ represents the Orthographic Projection Matrix. The resulting 2D coordinates $[x, y]$ are a function of the 3D world coordinates $[X, Y, Z]$.

The Orthographic Camera Model is valuable in scenarios where it is essential to preserve parallel lines and relative object sizes in the image. However, it is important to note that this model discards depth information, making it unsuitable for

applications that require accurate depth perception. Nonetheless, its simplicity and predictability make it a versatile tool in many domains, particularly when depth distortions are to be avoided.

3.2.3.3 Weak-Perspective Camera Model

The Weak-Perspective Camera Model is a widely employed abstraction in computer vision and photogrammetry that approximates the projection of three-dimensional (3D) points onto a two-dimensional (2D) image plane. This model is particularly valuable for scenarios where the camera-object distance is considerably greater than the focal length, thereby resulting in a weak-perspective projection [190].

Mathematically, the Weak-Perspective Camera Model can be expressed as follows: Consider a 3D point in the world coordinate system represented as $X = [X, Y, Z]^\top$, where X, Y, and Z denote the coordinates along the x, y, and z axes, respectively. This 3D point is projected onto the image plane, yielding its 2D image coordinates $x = [x, y]^\top$. The camera intrinsic parameters, denoted by $\mathbf{K}$, include the focal length f, principal point coordinates c_x and c_y, and skew coefficient α_c. The Weak-Perspective Camera Model can then be expressed as:

$$s \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

Here, s represents a scaling factor that accounts for the depth of the point, and $\mathbf{K}$ is the camera calibration matrix:

$$\mathbf{K} = \begin{bmatrix} f & \alpha_c & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

The camera calibration matrix encapsulates the intrinsic parameters of the camera,

such as the focal length, principal point, and skew coefficient, which are critical for accurate projection.

In practice, the Weak-Perspective Camera Model is advantageous in scenarios where the camera-object distances are large compared to the focal length. This model simplifies the perspective projection by neglecting the nonlinear distortion inherent in strong-perspective projection models. While the model offers an approximation of the true projection, it is suitable for applications like 2D object detection, image stitching, and structure-from-motion tasks. Nevertheless, it is crucial to recognize its limitations, particularly when dealing with scenes characterized by significant depth variations or when high-precision 3D reconstruction is required, as it does not account for depth-related distortions.

In summary, the choice of a camera model in human mesh reconstruction is a critical decision that significantly influences the accuracy and reliability of the reconstruction process. The implicit assumptions of each camera model should be carefully considered, and the model selected based on the specific requirements of the scene and the application. Understanding the implications of each camera model and addressing the associated challenges are essential for advancing the field of human mesh reconstruction.

3.3 Forward Kinematics

In the realm of human body modeling, forward kinematics is the essential process of ascertaining the three-dimensional spatial coordinates and orientations of distinct body components, including limbs and joints. This determination is based on the input of joint angles and bone dimensions within a structured skeletal hierarchy. The significance of this methodology extends to various fields, including animation, biomechanics, and motion capture. It empowers the creation of lifelike and dynamic portrayals of human motion in computer graphics and biomechanical simulations, facilitating the realistic and accurate representation of human movement within these

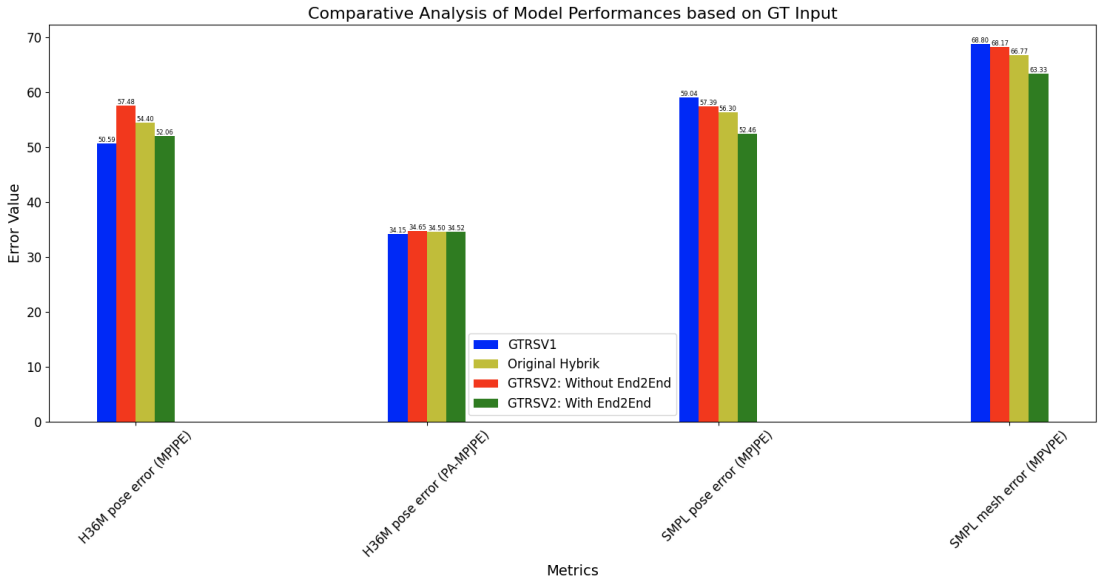

Figure 3.4: comparative Analysis of Model Performances

domains.

The hierarchical structure plays a pivotal role in the context of forward kinematics by providing a systematic and intuitive representation of the intricate connections within the human body. This hierarchical arrangement, often resembling a kinematic tree, is of paramount importance in understanding and modeling the human skeletal system. It encapsulates the dependencies and articulations between body parts, enabling a clear delineation of the parent-child relationships among joints and bones. This organization is fundamental for solving the forward kinematics problem, as it simplifies the process of determining the 3D positions and orientations of various body components. By breaking down the complex human body into a structured hierarchy, forward kinematics can be efficiently applied to calculate the transformations of individual elements, making it possible to represent natural movements and poses accurately. Moreover, the hierarchical structure allows for modularity and ease of manipulation, contributing to its significance in animation, biomechanics, and medical simulations. Overall, the hierarchical structure serves as a cornerstone in the successful application of forward kinematics, facilitating the realistic and dynamic

modeling of human motion in a wide range of fields.

A "kinematic chain" serves as a subsidiary component within the overarching kinematic tree structure, embodying a contiguous sequence of interconnected joints and bones. Typically initiated from a specific joint, a kinematic chain progresses along an unbroken series of parent-child relationships. As an illustrative example, a kinematic chain could embody the representation of an arm, originating at the shoulder joint and proceeding seamlessly through the elbow and wrist joints until reaching the fingertips. These kinematic chains are of fundamental importance when tackling the task of solving forward kinematics, as they play a pivotal role in isolating and facilitating the calculation of transformations specific to distinct body segments or anatomical parts.

In practice, forward kinematics in human body modeling involves traversing the kinematic tree or specific kinematic chains to calculate the transformation matrices that describe the 3D positions and orientations of various body parts. This process allows for the realistic and dynamic representation of human movements and poses in computer graphics, biomechanical simulations, and other applications. The hierarchical structure and the use of kinematic trees and chains are foundational elements in achieving accurate and natural human body modeling and animation.

Mathematically, the forward kinematics transformation from the root to a specific bone can be represented as:

$$T_{\text{root to bone}} = T_0^1 \cdot T_1^2 \cdot T_2^3 \cdot \ldots \cdot T_{n-1}^n$$

Where: - $T_{\text{root to bone}}$ is the transformation matrix from the root bone to the specific bone you're interested in. - T_i^{i+1} is the transformation matrix from $joint_i$ to $joint_{i+1}$ in the skeletal hierarchy

The general mathematical representation of a 4x4 homogeneous transformation matrix is given as:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{d} \\ \mathbf{0} & 1 \end{bmatrix}$$

Where: - $\mathbf{R}$ is a 3x3 rotation matrix representing the orientation component of the transformation. - $\mathbf{d}$ is a 3x1 translation vector representing the position component of the transformation. - $\mathbf{0}$ is a 1x3 zero vector to maintain the matrix dimensions. - The bottom-right element is always 1 for normalization.

To demonestrate the concept of the forward kinematic, Assuming we want to calculate the forward kinematics of a chain from the shoulder to the hand using a 3D Cartesian coordinate system. We'll assume a simplified representation with two joints: the shoulder joint (S) and the elbow joint (E). We're interested in finding the transformation from the shoulder (S) to the hand (H). Given the joint parameters θ_S and θ_E that represents Shoulder joint rotation angle and Elbow joint rotation angle respectively. Similarly, The length of the upper arm and the length of the forearm is defined by $L1$ and $L2$ respectively. The transformation Matrices could be represented by:

$$T_{SE} = \begin{bmatrix} \cos(\theta_S) & -\sin(\theta_S) & 0 & 0 \\ \sin(\theta_S) & \cos(\theta_S) & 0 & 0 \\ 0 & 0 & 1 & L1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_{EH} = \begin{bmatrix} \cos(\theta_E) & -\sin(\theta_E) & 0 & L2 \\ \sin(\theta_E) & \cos(\theta_E) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

To find the transformation from the shoulder to the hand (T_{SH}), we multiply the

transformation matrices for each joint in the chain as follow:

$$T_{SH} = T_{SE} \cdot T_{EH}$$

The resulting matrix T_{SH} contains the transformation that describes the position and orientation of the hand relative to the shoulder.

3.4 Recovering Human Mesh Pose and Shape

Human Mesh Recovery (HMR) from monocular images, devoid of additional devices like depth sensors, presents a formidable challenge due to various complexities, including depth ambiguity, intricate background elements, and the diverse range of human poses. Typically, a deep learning network is diligently trained with the objective of regressing human pose parameters concerning a vertex-based parametric human model, exemplified by SMPL (Skinned Multi-Person Linear) [1].

An approach introduced by *Kanazawa et al.* proposes an innovative strategy that centers on the minimization of 2D reprojection loss with respect to the anatomical joints. Crucially, this approach circumvents the necessity for paired 2D-to-3D supervision to effectively estimate the intricate parameters governing pose and shape directly from an input image. On a parallel front, the work of *Kolotouros et al.* [49] introduces a method that meticulously aligns the reconstructed mesh with the original image at the pixel level. This alignment is achieved through an iterative optimization process that refines the estimated parameters, resulting in an accurate mesh representation. Notably, *Kocabas et al.* augment their approach by integrating adversarial training techniques. These techniques contribute to enhancing the quality of the estimated human mesh, particularly leveraging a comprehensive motion capture dataset known as AMASS [50].

The rapid evolution of 2D pose estimation techniques has catalyzed a line of inquiry among researchers. This exploration centers on the feasibility of harnessing

off-the-shelf 2D pose detectors, such as HRNet [3], to estimate critical human mesh parameters. Specifically, HRNet facilitates the estimation of a 2D skeletal structure derived from the input image. This 2D skeletal representation serves as the foundational input to the network responsible for deducing and elaborating upon the intricate human mesh parameters, further advancing the state of the art in human mesh recovery from monocular imagery.

3.4.1 Preliminaries

The primary aim of this meticulously crafted pipeline was to conceptualize and develop a modular, multi-stages, end-to-end enabled graph-based vision transformer depicted in Fig 3.7. The pipeline was engineered to tackle the formidable challenge of accurately estimating the intricate shape and pose of the human figure. It was designed with the express intent of delivering results that stand shoulder to shoulder with the state-of-the-art (SOTA) methodologies in the field, highlighting a commitment to achieving, and in some cases even surpassing, the benchmarks set by the best-performing existing techniques.

The initial stage of this pipeline harnesses the power of 2D pose estimation, adeptly mapping the observed human pose from a 2D visual domain. This two-dimensional pose is then thoughtfully conveyed to a Graph Convolutional Network (GCN). The GCN plays a pivotal role in facilitating the transition from the two-dimensional to the three-dimensional domain, effectuating a process commonly referred to as "2D-to-3D lifting." The outcome of this stage culminates in the generation of a three-dimensional pose representation, which serves as a foundation for subsequent stages in the pipeline.

In the second module of this meticulously designed pipeline, an innovative and sophisticated Transformer Mesh Regression module comes to the fore. This module is instrumental in predicting the critical parameters governing the pose and shape of the human mesh. The Shape and Pose parameters, often represented through the Skinned Multi-Person Linear (SMPL) model, are skillfully deduced with an emphasis

on precision and accuracy.

Ultimately, the outcomes from both stages of this cutting-edge pipeline converge to pave the way for the meticulous reconstruction of the human mesh. This integrated approach not only showcases the synergy between deep learning techniques and geometric understanding but also presents a notable leap forward in the domain of human mesh recovery from 2D pose estimations. The overarching objective is to create high-fidelity human mesh representations, significantly advancing our capabilities in understanding and modeling human movements and interactions from visual data.

In the pursuit of refining and enhancing our pipeline for human pose and shape estimation, a crucial insight emerged from our meticulous analysis of the initial version. While the first iteration exhibited commendable results, adeptly rivaling the current state-of-the-art (SOTA) techniques, we began to recognize the intrinsic limitations of relying solely on skeleton data for the acquisition of accurate shape and pose representations.

This pivotal realization led to the inception of a second iteration of our pipeline, one that was meticulously engineered to address these limitations and extend the frontiers of our capabilities. The crux of this advancement lay in the incorporation of image-based data to augment our understanding of the human form. In this evolved version, our pipeline skillfully leveraged the information embedded in images to facilitate the learning of both shape and twist angles. The articulated body aspect of our model continued to be adeptly gleaned from the 2D pose estimations, which had proven reliable in the previous version.

However, the most notable deviation from our initial framework lies in the approach we employed for predicting the human mesh from the pose and shape parameters. In this second version, we opted for an analytical inverse kinematic approach, eschewing the transformer-based mesh regression module that was a hallmark of our first

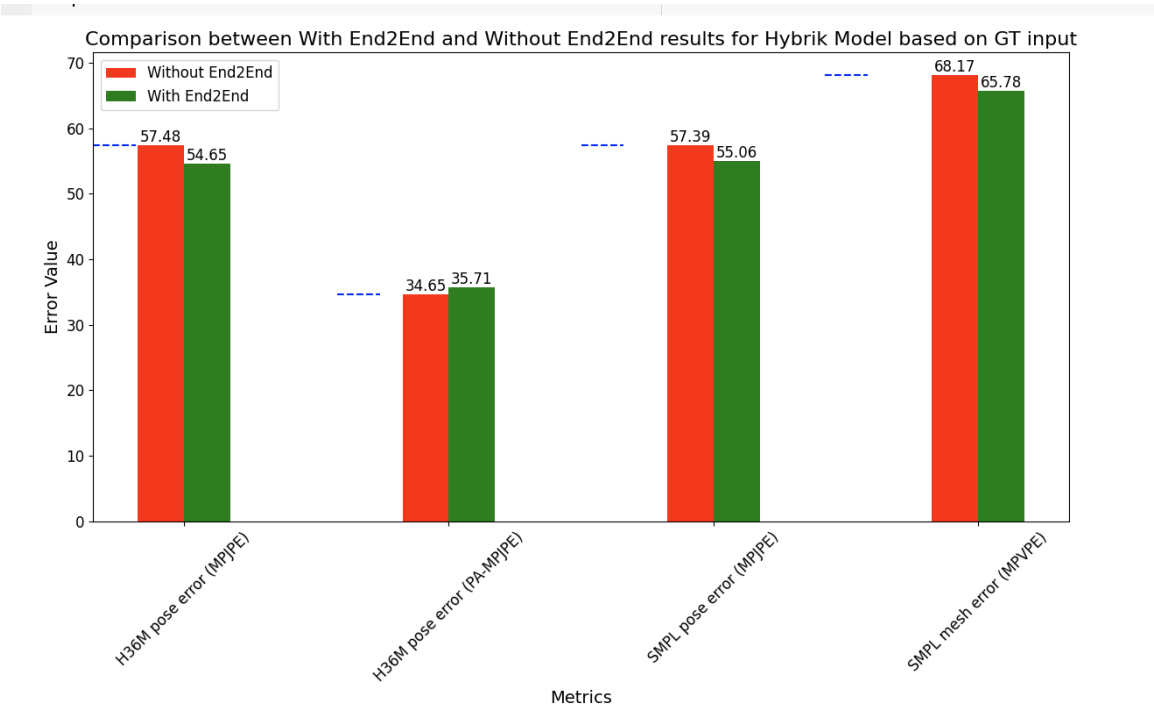

Figure 3.5: Comparative analysis of our model performance based on ground-truth data

iteration.

This evolutionary step was born out of the conviction that by tapping into the rich visual cues present in images, we could unlock a more comprehensive and nuanced understanding of the complex interplay between shape and pose. It is our belief that this synergy between skeletal information and image data, coupled with the adoption of analytical inverse kinematics, will propel our pipeline to new heights. Through this refined approach, we strive to not only match but exceed the capabilities of the current SOTA methods, pushing the boundaries of what can be achieved in the realm of human pose and shape estimation.

3.4.2 System Description

The objective of this deep learning pipeline is to perform human mesh reconstruction for the SMPL (Skinned Multi-Person Linear) model. This entails recovering essential parameters, including 3D pose, twist angles, and shape from input data, facilitating the generation of a 3D human mesh model. The pipeline involves two

versions, each contributing to different aspects of the reconstruction process.

In our second version of the pipeline depicted in Fig 3.6, we adopt a distinct approach to the reconstruction process. This iteration consists of two primary branches, each specializing in the acquisition of specific parameters for human mesh reconstruction.

The first branch leverages 2D pose estimation, serving as the foundation for the subsequent steps. Following the estimation of the 2D pose, a 2D pose lifter, a Graph Convolutional Network (GCN), and a transfer-based 2D-to-3D lifter are employed in succession. These components work sequentially, with the lifter and GCN processing the pose information and transforming it into a 3D pose representation. The 3D pose estimation is pivotal to the overall reconstruction process, contributing to the eventual generation of the 3D human mesh.

In the second branch, the pipeline takes a novel approach by utilizing raw image data as input. Here, a ResNet50 or HRNet48 backbone is employed, harnessing its capabilities to extract detailed information from the input images. Through a transformer-based mesh regression module, the model predicts SMPL pose and shape parameters, which are crucial in reconstructing the human mesh.

However, following a comprehensive analysis of our initial version, it became evident that relying solely on skeleton data may not suffice for capturing the intricate nuances of human shape and pose. This realization prompted the development of a second version of the pipeline, which combines both 2D pose lifter and image-based approaches to learn the shape and twist angles more comprehensively. The articulated body representation continues to be derived from 2D pose estimations.

The most notable alteration in this second version lies in the method of predicting the human mesh from the pose and shape parameters. In this iteration, we opt for an analytical inverse kinematic approach, eschewing the transformer-based mesh regression module featured in the initial version. This strategic decision is driven

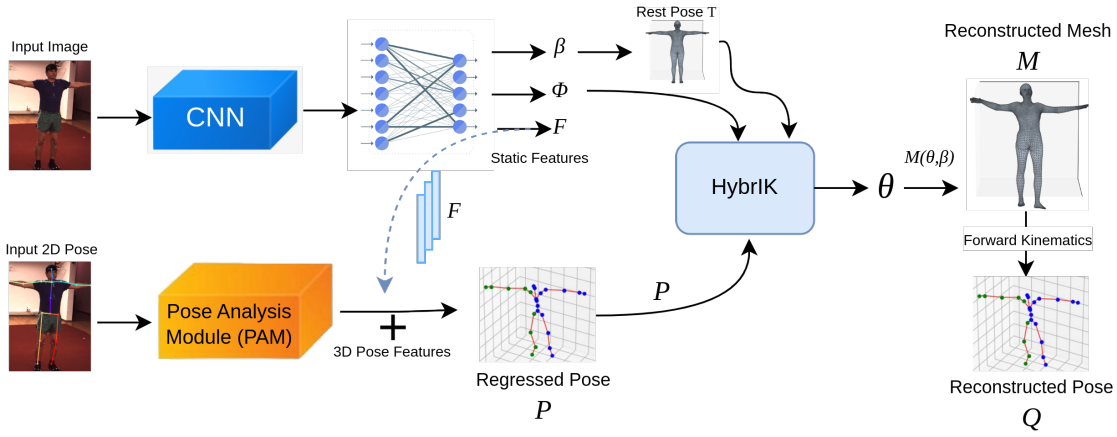

Figure 3.6: Given an image, static image features F are extracted by a pre-trained CNN, and 2D poses are detected by an off-the-shelf 2D pose detector. These static features, enriched with depth details, are then integrated with the Pose Analysis Module to refine 3D pose estimation P. In parallel, shape parameters β and twist angles ϕ are discerned from the visual image through fully connected layers. This compiled information is channeled into the HybrIK setup to resolve relative rotations, translating into specific pose parameters θ. Finally, with the pose and shape parameters, we can obtain the reconstructed body mesh M, and the reconstructed pose Q via a further Forward Kinematics process.

by the belief that a deeper fusion of skeletal and image data, in conjunction with analytical inverse kinematics, can further propel the capabilities of our pipeline.

Incorporating the analytical inverse kinematic solution, powered by Hybrik, marks a significant step forward, allowing us to unravel the complex relationship between pose and shape in a more comprehensive and nuanced manner. Through this refined approach, we aim to not only match but exceed the capabilities of the current state-of-the-art methods, pushing the boundaries of what can be achieved in the realm of human pose and shape estimation.

3.4.3 Implementation Details

3.4.3.1 Upper-branch 3D-lifter Module

Within this module, we harnessed the potential of a graph transformer, as expounded in [52], for the explicit purpose of capturing the intricate interplay between global and localized features inherent in the human pose. The operational sequence commences with the implementation of a projection layer that diligently encodes the

pose data into a latent space representation. Subsequently, the resultant features undergo propagation through multiple parallel graph transformer blocks, thoughtfully configured with compact embedding dimensions. This strategic dimensionality choice is pivotal, serving to optimize the overall efficiency of the module. Ultimately, the concurrent fusion of these parallel streams engenders a rich feature set that the module adeptly leverages in the pursuit of predicting the 3D pose estimation.

In the context of the 2D-to-3D lifting endeavor, the starting point is the presence of a 2D skeleton denoted as $S \in \mathbb{R}^{j\times 2}$, where 'j' embodies the quantity of joints, and the '2' characterizes the joint coordinates residing within the X, Y plane. The process unfolds under the aegis of the Pose Analysis Module (PAM), deftly delivering a collection of invaluable outcomes that include the introduction of feature embedding, expressed as $F \in \mathbb{R}^{J\times D}$, where 'J' faithfully represents the count of joints, while 'D' delineates the dimensions of the latent feature space, as well as the revelation of 3D joint positions denoted as $P \in \mathbb{R}^{J\times 3}$, with 'J' standing as the number of joints, and '3' portraying the coordinates that uniquely position these joints within a 3D spatial domain governed by the axes $[X, Y, Z]$.

The realm of Graph Convolution Networks (GCN) has garnered notable attention in recent years, largely attributable to its inherent aptitude for intuitive data modeling, particularly in contexts involving articulated models. In the domain of 3D human pose estimation, a plethora of endeavors have leveraged GCN-based approaches, as underscored by contributions such as [198] and [199]. Notably, a graph, fundamentally, represents a structured assemblage of vertices and edges. In the context of human skeleton modeling, this entails the mapping of joints to vertices and bones to edges. This structural depiction finds its mathematical instantiation as $\mathcal{G} = v, \epsilon$. Building upon the insights delineated in [52], the application of GCN in this setting revolves around the modeling of 2D pose features. The output produced by the GCN layer adheres to the mathematical expression:

$$X' = \sigma(A \times W) \tag{3.1}$$

Herein, $\sigma(.)$ signifies the Gaussian Error Linear Unit (GELU) activation function, A is representative of the adjacency matrix, and W denotes the weight matrix subject to learnable adjustments.

The transformer architecture, as presented in [200], has led to a substantial reconfiguration of various deep learning models in the field of computer vision. Its inherent capacity to encapsulate holistic contextual understanding in the feature space has resulted in significant advancements.

Self-attention can be described as a function for computing the attention matrix given a query matrix $Q \in \mathbb{R}^{N\times d}$, key matrix $K \in \mathbb{R}^{N\times d}$, and value matrix $V \in \mathbb{R}^{N\times d}$. Here, N denotes the number of vectors in the sequence, and d represents the dimension. To mitigate the unwarranted amplification of the softmax function, a scaling factor $\frac{1}{\sqrt{d}}$ is employed. The scaled-dot product attention can be expressed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Moreover, the concept of multi-headed self-attention extends this paradigm by employing multiple parallel scaled-dot attention mechanisms. This approach projects the queries, keys, and values h times with different linear projections to dimensions d_k, d_k, and d_V. The concatenation of the h head attention outputs can be expressed as:

$$\text{MSA}(Q, K, V) = \text{Concat}(H_1, H_2, ..., H_h)W_{out}$$

$$\text{where } H_i = \text{Attention}(Q_i, K_i, V_i), i \in [1, 2, ..., h]$$

These intricate operations are instrumental in enhancing the model's ability to capture complex dependencies in the data, which is particularly advantageous for a

Table 3.1: An Ablation Study with GTRV2 on Human 3.6M dataset

3D Joints P GT	Shape β GT	Twist Angles ϕ GT	H36M MPJPE	H36M PA-MPJPE	SMPL MPJPE	SMPL MPVPE
✓			22.94	13.59	21.18	27.98
	✓		46.43	32.66	46.99	57.57
		✓	50.74	32.81	52.25	59.63

wide array of computer vision tasks.

3.4.4 Experimental Settings and Results

The preliminary results, as shown in Table 1, demonstrate significant improvements in the accuracy and reliability of human pose estimation techniques for the Human 3.6M dataset. Notably, the GTRSV2 model, equipped with a resnet101 backbone and a dual-branch system, exhibits substantial advancements over its predecessor, GTRS. When challenging HRNET 2D joint detection is used as input, GTRSV2 consistently outperforms the GTRS across multiple metrics, such as MPJPE, PA-MPJPE, SMPL MPJPE, and SMPL MPVPE, highlighting its refined accuracy in estimating 3D human poses.

A key factor contributing to GTRSV2's success is the decoupling of learning tasks within the model, enhancing its capability for precise pose alignment, an essential element in 3D human pose reconstruction. The incorporation of advanced transformer mechanisms allows GTRSV2 to handle spatial dependencies more effectively, leading to lower error rates, especially in noisy 2D detection. These findings not only represent incremental improvements but a significant leap in the model's interpretability.

The ablation study delineates the importance of individual components within the pose detection mechanism, namely 3D Joints, Shape, and Twist Angles, as showcased in Table 2. In comparing the results when these components are utilized as ground truth (GT) versus when they are predicted by the GTRSV2 model, a number of key insights are unveiled. Most notably, employing GT for 3D joints substantially improves the outcomes, thereby underscoring their pivotal role in attaining precise

pose detection. While the incorporation of GT for shape β does result in enhanced scores, its impact is not as pronounced as that of the 3D joints. Moreover, the application of GT for Twist angles ϕ exhibits a slight reduction in error metrics, indicating that the GTRSV2 model provides a reasonable prediction for these angles. This study serves to illuminate potential avenues for refinement within the GTRSV2 approach, particularly with regard to optimizing the prediction mechanisms for 3D joints and shape, in order to further elevate the model's performance.

We adhered to the prescribed methodologies outlined for the evaluation of our pipeline's performance on the Human3.6 dataset. Our rigorous assessment demonstrates the superior efficacy of our pipeline when compared to the contemporary state-of-the-art benchmarks, as visually represented in 3.4.

The outcomes presented in 3.6 substantiate our preliminary analysis, which posited that enhancing the performance of the pipeline could be achieved by furnishing the network with a 2D pose estimation as a prior during the learning process.

Additionally, our findings indicate that the network exhibits elevated accuracy when employing more advanced backbone architectures for the image-based branch, such as HRNet48 or ResNet-101.

Table 3.2: Benchmark of state-of-the-art models on Human 3.6M dataset

Method	Human 3.6M Dataset			
	H36M MPJPE	H36M PA-MPJPE	SMPL MPJPE	SMPL MPVPE
GTRS [52]	50.59	34.15	59.04	68.80
GTRSV2 (ours)	**51.71**	**34.13**	**52.56**	**61.84**

3.4.4.1 Fusion Strategies

GTRSV2 emerges as a notable player in the realm of 3D pose estimation. This approach capitalizes on convolutional network backbones like resnet and hrnet to predict specific parameters such as beta and twist angles. Central to the methodology is the integration of static features via the Pose Analysis Module (PAM).

H igh-dimensional Fusion: In the first fusion method, the normalized 2D pose, denoted as P_{2D}, is mapped to a high-dimensional space represented by $X \in \mathbb{R}^{J \times C_1}$. This is achieved using graph convolutions and attention mechanisms. Parallelly, static image features F are transformed into F', which is then concatenated with X to derive a dimensionality of $J \times C_2$. This high-dimensional representation, however, has a caveat. Although it captures rich feature interplay, it poses challenges like overfitting, especially when projecting the 3D Pose P on unseen data. When empirically evaluated on the Human3.6 dataset with the HRNet48 backbone, it resulted in a 69.11 MPJPE and 39.84 PA-MPJPE.

F usion with Depth Information: The second fusion strategy centers on the incorporation of depth information from the static feature into the 3D pose estimation process. In this approach, after transforming the static image features F into F', a fusion gate is employed. This gate, characterized by a sigmoid activation, effectively assigns weights to the information derived from both embeddings. The fusion process accentuates the discernment of which details from static features and 2D pose estimations hold greater significance. When subjected to testing on the Human3.6 dataset, particularly with a less complex ResNet34 backbone, the results exhibited promise. The model achieved a 52.06 MPJPE and 34.5 PA-MPJPE, surpassing the performance of the former fusion method. An experiment involving the more robust ResNet101 also demonstrated notable improvements, suggesting a 2.35% enhancement in MPJPE.

Within the realm of the GTRSV2 approach, the high-dimensional fusion stands as a notable strength. This technique engenders a rich and layered interplay of features. By elevating the feature representation into higher dimensions, the approach affords the model the luxury to tap into and possibly capture more intricate and nuanced relationships between features. Such depth not only amplifies the granularity of feature interactions but also provides a robust platform to discern patterns pivotal for precise

3D pose estimation. In essence, this depth of feature interplay can potentially resolves complexities and relationships in the data that might otherwise remain obscured in a lower-dimensional space.

The second approach introduces an elegant solution to feature integration through its fusion gate mechanism. Unlike conventional methods, this mechanism doesn't just merge features; it discerningly weighs and prioritizes them. Such a strategy ensures that every piece of information is not just assimilated but is done so based on its relevance and importance. This selective weighing and integration of features make the model more adaptive and responsive, ensuring that it is finely tuned to capture the most pertinent features that significantly influence the outcome. In a way, it introduces a layer of refinement, refining the model's predictive capabilities by focusing on what truly matters.

While the high-dimensional fusion in the first approach offers a plethora of benefits, it also comes with its set of challenges. The very nature of highdimensional embeddings, despite being comprehensive, introduces the risk of overfitting. By representing features in such a vast space, the model can become overly adaptive to the training data, capturing even its noise. This, in turn, can compromise the model's ability to generalize well on unseen or new data, making it susceptible to poor performance in real-world scenarios or diverse datasets.

The second approach, though innovative in its feature integration strategy, is not without its pitfalls. Its heavy reliance on the fusion mechanism can be a problematic. On the one hand, it ensures a refined and weighted integration of features, but on the other, it might inadvertently bias the model towards the 2D pose projections, especially during the training phase. Such a bias can make the model overly reliant on certain feature sets, thereby affecting its ability to generalize well during testing. In essence, while the fusion introduces a layer of sophistication, it also runs the risk of making the model focus on one projection, potentially compromising its broader

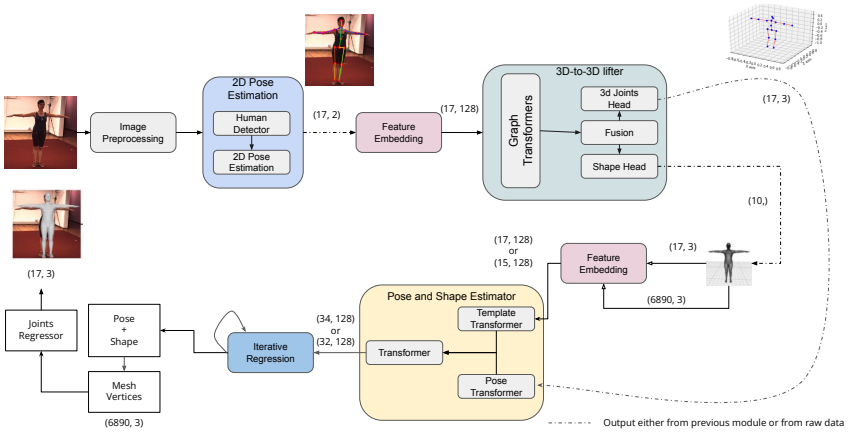

Figure 3.7: GTRSv1- Modular Multi-stage Lightweight Graph Transformer Network for Human Pose and Shape Estimation from 2D Human Pose

applicability.

3.5 Conclusion

In conclusion, our research endeavors have led to the development of a novel modular, two-streamed, end-to-end enabled pipeline for human parameter estimation. We introduced a pioneering approach to learn human pose through analytical inverse kinematics, leveraging the capabilities of a graph-based vision transformer. This holistic system is designed to estimate both human shape and pose, aiming to achieve results that are on par with state-of-the-art (SOTA) methods in the field of computer vision.

Our contributions signify a significant step towards addressing the challenges of accurate and efficient human parameter estimation. By combining analytical techniques and advanced deep learning models, we've bridged gaps in the existing methodologies, providing a robust solution for a wide range of computer vision applications. This work not only contributes to the advancement of the field but also opens up avenues for further exploration and innovation in the domain of human parameter estimation.

As our research continues, we remain dedicated to refining and expanding our approach, with the ultimate goal of pushing the boundaries of what is achievable in the realm of human shape and pose estimation. We believe that this work lays a solid

Table 3.3: Our extensive experiments showed our pipeline score comparable results to various contributions

	Methods		Humans.6M	
			MPJPE	PA-MPJPE
Image Based	HMR	CVPR 2018	88.0	56.8
	GraphCMR	CVPR 2019	-	50.1
	SPIN	ICCV 2019	-	41.1
	METRO	CVPR 2021	54.0	36.7
	MeshGraphormer	ICCV 2021	**51.2**	**34.5**
	PyMAF	ICCV 2021	57.7	40.5
Video Based	VIBE	CVPR 2020	65.6	41.4
	TCMR	CVPR 2021	62.3	41.1
Pose Based	Pose2Mesh	ECCV 2020	64.9	47.0
	Baseline	-	68.3	50.0
	GTRS	-	64.3	45.4
	Our	-	65.9	47.13

foundation for future developments and holds the potential to significantly impact computer vision research and its practical applications.

3.6 Future Work

An intriguing future direction in human pose estimation is the integration of physics-based models to attain plausible and context-aware pose estimations. Physics-based models consider the biomechanical constraints and interactions between body parts, the environment, and external forces. By incorporating these models, we can enhance the realism and naturalness of estimated poses. This approach becomes particularly valuable in applications like computer animation, medical simulations, and virtual reality, where achieving lifelike and contextually plausible human poses is critical. By fusing physics-based models with data-driven deep learning techniques, we can enable human avatars to exhibit more natural movements and adapt to dynamic environments realistically. Moreover, these models can contribute to pose refinement and consistency, especially in cases involving occlusions, complex poses, or challenging environments. This fusion of physics and deep learning paves the way for a new era of human pose estimation that delivers not only accuracy but also plausibility

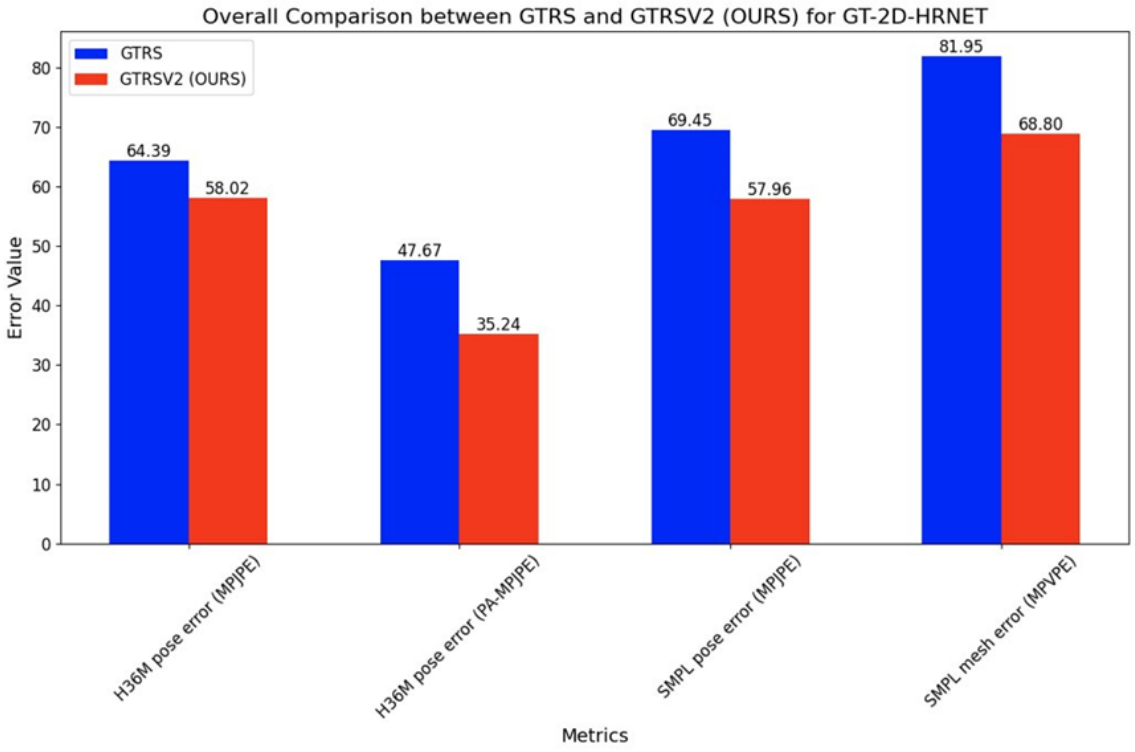

Figure 3.8: Comparative results between GTRSv1 and GTRSv2

and adaptability.

A fruitful avenue for future research is the exploration of advanced inverse kinematics solvers to refine human pose estimation. Analytical inverse kinematics, though widely adopted, may not always capture the intricate nuances of complex human motions, leading to inaccuracies in pose estimation. Advanced IK solvers, such as numerical, optimization-based, or learning-based methods, offer the advantage of handling more complex scenarios. Use cases where analytical IK might struggle include highly constrained environments, such as rehabilitation robotics, where individuals have limited joint mobility, and physically demanding sports analysis, where extreme poses and dynamic motions challenge conventional IK solutions. Additionally, scenarios involving interactions with objects or external forces, like human-robot collaboration and exoskeleton control, benefit from advanced solvers capable of adapting to real-time changes in pose and environment. By delving into advanced IK techniques, we

can achieve more robust and accurate human pose estimations across a spectrum of challenging applications.

Combining physics-based models with advanced Inverse Kinematics (IK) solvers holds significant potential for enhancing the generalizability and robustness of human pose estimation pipelines. By integrating biomechanical principles and dynamics into the pipeline, physics-based models provide a solid foundation for understanding human movement, allowing for more effective generalization across diverse datasets. This approach is particularly advantageous in scenarios with dataset-specific variations or challenges, ensuring the system's adaptability and robustness. The synergy between physics-based modeling and advanced IK solvers not only promotes realism and consistency in estimated poses but also enables the system to adapt to complex environments and interactions. This combination bridges the gap between data-driven methods and domain-specific knowledge, making it well-suited for applications in gaming, virtual reality, robotics, sports analysis, and human-computer interaction, where accurate and adaptable pose estimation is essential.

CHAPTER 4: Real-Time Kinematic Sequence Analysis via Monocular 3D Human Pose and Kinematics Estimation

4.1 Introduction

Biomechanics, an interdisciplinary field that melds mechanical principles with the investigation of biological systems, plays a pivotal role in advancing our comprehension of the mechanics underlying human and animal locomotion [6]. It delves into the intricate interplay of the body's intricate anatomical structure and the musculoskeletal system, elucidating the mechanisms governing motion, stability, and overall physical performance [201]. The analytical scrutiny of biomechanics proves indispensable in diverse domains such as sports science, rehabilitation medicine, ergonomics, and product design [202]. However, it is paramount to acknowledge that acquiring the essential biomechanical data, integral to the efficacy of these applications, predominantly transpires within specialized laboratories equipped with sophisticated apparatus designed for data capture, and therein lies the substantial financial investment [7].

The human body, characterized as an intricate marvel of engineering and complexity [6], undergoes meticulous coordination of various anatomical components comprising bones, muscles, tendons, and ligaments [6]. This orchestration culminates in the myriad of activities we undertake daily, spanning from running and jumping to lifting objects and even mere ambulation [6]. Biomechanical principles meticulously expound on how the human body generates forces and preserves stability, and these principles find comprehensive elucidation through biomechanical analyses [6]. The utmost significance of this analytical investigation rests in its potential to optimize human physical performance, prevent injuries, and tailor products to better suit the

inherent capabilities and constraints of the human physique [8].

Within the sphere of sports, biomechanics stands as an indispensable discipline [9]. Athletes and their coaches harness biomechanical analysis as a tool for dissecting athletic movements, critically appraising strengths and weaknesses, and refining techniques to maximize athletic performance [10]. Given the recurrent incidence of injuries in sports, biomechanical scrutiny extends its utility in discerning injury risk factors and contributing to injury prevention [11]. Furthermore, biomechanics significantly informs the development of sports equipment and footwear, engineering them for enhanced efficiency and safety.

Rehabilitation medicine relies extensively on the tenets of biomechanical analysis [6]. When individuals sustain injuries or musculoskeletal disorders, an understanding of the underlying biomechanical intricacies proves indispensable for devising rehabilitation protocols [12]. Biomechanical assessments empower therapists to discern deviations in movement patterns, irregularities in gait, or imbalances in muscle function, facilitating the tailoring of individualized treatment regimens to expedite recovery [13].

The linchpin in the aforementioned applications of biomechanics resides in the availability of comprehensive biomechanical data [6]. The efficacy of analytical scrutiny hinges on the precise collection, processing, and interpretation of data encompassing parameters such as forces, motion, and physiological attributes including joint angles and muscle activity. Technological advancements have significantly augmented data collection capabilities through sophisticated equipment encompassing motion capture systems, force plates, and electromyography. Nonetheless, these sophisticated technologies are predominantly housed within specialized laboratories, and their procurement and maintenance invariably entail substantial financial investments.

Additionally, the realm of biomechanical analysis has witnessed the pivotal integration of computer modeling and simulation as indispensable tools for the investigation

of biomechanical data. Researchers and practitioners employ simulation tools as a viable alternative route to generate biomechanical data, particularly in scenarios where the collection of experimental data is either logistically challenging or financially burdensome. The quality and quantity of available biomechanical data directly influence the accuracy and reliability of biomechanical analyses, underscoring the paramount importance of meticulous data acquisition and processing.

Typically, biomechanics constitutes an indispensable discipline for the enhancement of human physical performance, injury prevention, and the optimization of product and environmental design. The acquisition of accurate and comprehensive biomechanical data is not a superfluous luxury but, rather, an imperative requisite for the successful conduct of analysis within these domains. Ongoing technological advancements and research methodologies, though accompanied by the requirement for substantial financial investment in specialized laboratory equipment, are poised to perpetuate the pivotal role of biomechanical analysis in augmenting our comprehension of human locomotion and well-being.

In order to surmount the formidable obstacles associated with the establishment of a specialized biomechanical laboratory for the acquisition and analysis of kinematic data, we have engineered a comprehensive framework harnessing the capabilities of a widely recognized simulation tool in the biomechanics domain, namely 'OpenSim [121]' This endeavor is propelled by the overarching aim of formulating an efficient and robust deep learning methodology capable of providing cost-effective, real-time, and high-precision estimations of kinematic data. These estimations are derived from monocular video inputs and are tailored for deployment on readily available mobile devices, including smartphones and tablets, equipped with a single camera. By capitalizing on the capabilities of the OpenSim simulator, we empower consumers with the ability to perform diverse biomechanical analyses at their fingertips

4.1.1 Challenges

Conducting a biomechanical kinematics analysis presents a multitude of challenges. In this contribution, we endeavor to address several of these challenges:

- The execution of 3D kinematic analysis is traditionally conducted within a dedicated physical biomechanics laboratory, wherein a costly marker-based motion capture system is deployed. This system utilizes a network of synchronized cameras to infer the 3D human poses, encompassing the spatial positioning of joints. This inference is achieved through the tracking of markers affixed to the subject. It is important to note that this conventional approach demands a substantial financial outlay and necessitates the establishment of specialized laboratory configurations.

- One of the formidable challenges lies in the attainment of precise and high-fidelity 3D pose estimation, upon which the kinematic analysis fundamentally relies. Precise 3D pose estimation is essential, as errors can propagate through the entire analysis.

- Occlusion occurs when certain body parts or markers are temporarily hidden from the view of the motion capture system's cameras. This can happen when one body part obstructs the view of another during complex movements. Occlusion introduces missing data, which can lead to inaccuracies in the 3D pose estimation. Dealing with occlusion often involves sophisticated algorithms for filling in the gaps in the data or predicting the obscured positions of body parts. In addition, Self-occlusion specifically refers to situations where a body part or a marker becomes hidden or obscured by other parts of the body itself. For example, when a person raises their arm to touch their head, the markers on the arm might become obscured by the head. Self-occlusion is particularly challenging because it involves complex interactions within the body. Mitigat-

ing self-occlusion may require careful marker placement, advanced modeling techniques, and specialized camera configurations.

- Smartphones have constrained computational resources compared to desktop computers. Running complex biomechanical algorithms, particularly those involving deep learning and 3D pose estimation, can strain the device's processing capabilities.

4.1.2 Our Contribution

In this research endeavor, our primary objectives encompass:

- To establish a comprehensive framework leveraging the OpenSim simulation tool, aimed at the development of a highly efficient and resilient deep learning model. This model is designed to achieve the precise estimation of 3D human poses while concurrently predicting associated kinematic attributes. The intent is to facilitate this process through monocular or single-camera video inputs, fostering deployment on ubiquitous mobile devices, including smartphones and tablets.

- To undertake a rigorous evaluation of the execution of advanced 3D pose estimation techniques. This evaluation encompasses the intricacies of running these techniques in the context of mobile device applications. The emphasis is on assessing their performance, accuracy, and computational efficiency within the defined constraints and parameters of mobile platforms.

4.2 Rotation Angles Representations

Understanding various rotation angle representations is of paramount importance when tackling Inverse Kinematic (IK) problems. IK deals with the complex task of determining joint configurations that achieve a desired end-effector pose. Different applications and contexts may require different representations of rotations, such as

Euler angles or axis-angles. Proficiency in these representations is crucial for several reasons. Firstly, it ensures compatibility with diverse software, hardware, and robotics systems that may employ different rotation conventions. Secondly, it facilitates interoperability when exchanging data or animations between systems, preventing misinterpretation of rotations. Thirdly, a sound understanding of various representations helps mitigate the risk of encountering gimbal lock, a limitation of Euler angles that can hinder accurate orientation representation. Furthermore, distinct rotation representations possess unique mathematical properties, with quaternions, for example, excelling in smooth interpolation. In the context of IK solving, different rotation representations can be employed to tailor solutions to specific constraints or requirements. In this section, we delve into an examination of the diverse rotation representations commonly utilized in the literature.

4.2.1 Rotation Matrix

Rotation matrices are essential tools in linear algebra and geometry for describing and understanding 3D rotations. At their core, rotation matrices provide a mathematical representation of how points or vectors in 3D space transform when an object is rotated. The intuition behind rotation matrices lies in their ability to preserve lengths, angles, and shapes, making them ideal for modeling rigid-body transformations.

The mathematical representation of a 3D rotation matrix involves a 3x3 matrix, typically denoted as R. To rotate a 3D vector $\mathbf{v}$ using the rotation matrix R, you simply perform the matrix-vector multiplication:

$$\mathbf{v}' = R\mathbf{v}$$

where $\mathbf{v}'$ represents the rotated vector, $\mathbf{v}$ is the original vector, and R is the rotation matrix.

The rotation matrix R is typically orthogonal, meaning its columns (or rows) are

orthonormal, ensuring that lengths and angles are preserved during the transformation. The orthonormality of rotation matrices guarantees that they are rigid-body transformations, which maintain the relative distances between points and preserve the dot products and cross products of vectors.

An example of a rotation matrix for a rotation about the Z-axis by an angle θ (in the counterclockwise direction when viewed from above) is:

$$R_z(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Here, $\cos(\theta)$ and $\sin(\theta)$ are the cosine and sine of the rotation angle θ, respectively.

A rotation matrix is a square matrix that, when multiplied with a 3D vector, produces a new vector representing the rotated position of the original vector. This matrix encodes information about the orientation of an object in 3D space, allowing us to apply rotations in a consistent and precise manner. Matrix representation serves as an unequivocally explicit means of representing orientation, affording several notable advantages.

First and foremost, matrix representation enables the immediate rotation of vectors. A pivotal characteristic of the matrix form lies in its capability to effectuate vector rotations between object and canonical spaces. This property stands unparalleled among other orientation representations, mandating the conversion of orientation into matrix form for vector rotations.

Moreover, graphics Application Programming Interfaces (APIs) extensively employ matrix representations to articulate orientation. Given the intrinsic need to interface with graphics hardware through APIs, the necessity emerges to express transformations as matrices during communication. This aligns with the standard practice in graphics programming. Although the internal storage of transformations is subject

to programmer discretion, adherence to matrix representation is inevitable at some point within the graphics processing pipeline.

Additionally, matrices offer an advantageous mechanism for the concatenation of multiple angular displacements. By employing matrices, the nesting of coordinate space relationships can be efficiently collapsed. For instance, when the orientation of object A concerning object B and that of object B concerning object C are known, matrices facilitate the determination of the orientation of object A concerning object C.

Furthermore, matrix inversion becomes a feasible operation when angular displacements are represented in matrix form. The computation of the inverse angular displacement is made possible, with the orthogonal nature of rotation matrices rendering the process trivial. Mathematically, this operation can be expressed as:

$$R^{-1} = R^T$$

where R^{-1} represents the inverse rotation matrix, and R^T denotes the transpose of the original rotation matrix.

The explicit nature of matrix representation confers several benefits, as previously expounded. However, matrices employ nine numerical components to encapsulate an orientation, while it is theoretically feasible to parameterize orientation with a mere three numbers. The surplus numerical data within a matrix can engender specific challenges.

Matrices necessitate greater memory allocation, a consideration of paramount significance, particularly in scenarios involving the storage of numerous orientations. For instance, in the context of animating a complex model, such as a human segmented into 15 distinct body parts, animation control hinges upon the precise definition of the orientation of each component relative to its parent part. In an illustrative scenario

where one orientation is stored for each part per frame at a modest animation rate of 15 Hz, 225 orientations per second must be accommodated. Employing matrices with 32-bit floating-point precision, each frame would entail 8,100 bytes of memory. Conversely, utilizing Euler angles, the identical data can be represented with a mere 2,700 bytes. For a brief 30-second animation sequence, the utilization of matrices would translate to an excessive 162K of additional memory usage compared to Euler angles.

Moreover, matrices pose a challenge in terms of human intuitiveness. Their numerical nature, coupled with the wide range of values encompassing the full real number space, renders matrices less intuitive for direct human interaction. Human cognition naturally gravitates towards orientation conceptualization in the realm of angles, while matrices express orientation in terms of vectors. Although proficiency can be developed to decipher orientation from a given matrix, this process remains notably more intricate in comparison to Euler angles. Constructing the matrix for a nontrivial orientation manually is a time-consuming endeavor. In summary, matrices do not align with the innate human thought process regarding orientation.

Furthermore, matrices are susceptible to being ill-formed. A matrix inherently accommodates six degrees of redundancy, necessitating satisfaction of six constraints for the matrix to be deemed "valid" for representing an orientation. These constraints involve the requirement for matrix rows to constitute unit vectors and to be mutually orthogonal. The occurrence of ill-formed matrices can precipitate numerical anomalies, distorted graphical output, and other unexpected behavioral consequences.

In conclusion, the utilization of matrices, while advantageous in certain aspects, presents challenges related to memory usage, human interpretability, and the potential for ill-formed matrices. These challenges should be considered in the context of selecting an appropriate orientation representation for specific applications.

4.2.2 Euler Angles

Euler angles provide a way to break down a 3D rotation into a series of simpler rotations. The idea is to describe the rotation of an object in terms of rotations about its principal axes. In 3D space, we have three principal axes: X, Y, and Z. Euler angles allow us to describe how much the object rotates about each of these axes, one after the other. These individual rotations can be intuitively understood as follows:

1. Roll (ϕ or Phi): This is the first rotation and typically represents a rotation about the object's X-axis. It's like tilting your head from side to side, causing the object to roll around its forward direction.

2. Pitch (θ or Theta): The second rotation is often about the Y-axis. It's similar to nodding your head up and down, causing the object to pitch around its side-to-side direction.

3. Yaw (ψ or Psi): The final rotation typically occurs around the Z-axis. It's akin to turning your head left or right, causing the object to yaw around its up-and-down direction.

These three angles together describe the complete orientation of the object in 3D space. However, it's crucial to note that the order in which these rotations occur matters and can lead to different results. This sequence of rotations is referred to as the "Euler angle sequence" and can vary based on conventions and applications.

The mathematical representation of Euler angles involves a sequence of three rotations. Let's denote the three Euler angles as ϕ (roll), θ (pitch), and ψ (yaw). The transformation matrix for a sequence of Euler angles can be represented as a product of three rotation matrices:

$$R = R_z(\psi) \cdot R_y(\theta) \cdot R_x(\phi)$$

Here, $R_x(\phi)$, $R_y(\theta)$, and $R_z(\psi)$ are the rotation matrices for the individual Euler angle rotations about the X, Y, and Z axes, respectively.

For instance, $R_x(\phi)$ is given by:

$$R_x(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\phi) & -\sin(\phi) \\ 0 & \sin(\phi) & \cos(\phi) \end{bmatrix}$$

similarly, $R_y(\theta)$ is given by:

$$R_y(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix}$$

Additionally, $R_y(\psi)$ is given by:

$$R_z(\psi) = \begin{bmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Where cos and sin are the cosine and sine trigonometric functions, respectively.

Euler angles serve as a parameterization method for orientation, characterized by their utilization of merely three numerical values, representing angles. These distinctive traits of Euler angles offer several advantages compared to alternative approaches for expressing orientation.

Euler angles are notably user-friendly, providing an intuitively comprehensible means of orientation representation, which is considerably more accessible to humans in contrast to matrices or quaternions. This inherent ease of use stems from the fact that Euler angles directly embody angles, a concept closely aligned with human perception of orientation. By judiciously selecting conventions tailored to specific

scenarios, practical angular measures can be unambiguously expressed. For example, the heading-pitch-bank system directly conveys the angle of declination. This user-friendliness is a substantial advantage, particularly when numerical orientation representation or manual entry is necessitated, rendering Euler angles the preferred choice.

A pivotal advantage of Euler angles is their minimalistic representation. A mere three numerical values suffice to describe a 3D orientation. This succinct parameterization is the most compact approach achievable for encoding orientation, as no alternative system can effectively encapsulate 3D orientation using fewer than three numbers. In situations where memory resources are constrained, Euler angles emerge as the most economically efficient choice for orientation representation.

Furthermore, the utilization of Euler angles is advantageous in terms of space conservation. These numerical values can be efficiently compressed into a reduced number of bits using a straightforward fixed-precision scheme. Owing to their angular nature, the data loss stemming from quantization is evenly distributed. In contrast, matrices and quaternions entail the utilization of diminutive numerical values, as they encapsulate sines and cosines of angles. This leads to non-proportional differences between values, rendering matrices and quaternions less amenable to integration into a fixed-point system.

In summary, when the imperative is to store copious amounts of 3D rotational data within constrained memory, a frequent requirement in animation data handling, Euler angles (or the forthcoming exponential map format discussed in Section 8.4) emerge as the most judicious selections.

It is noteworthy that any set of three numerical values is deemed valid in Euler angles. Randomly chosen triplet values inherently constitute a legitimate set of Euler angles, interpretable as an expression of orientation. Consequently, there exists no concept of invalid Euler angles, although the correctness of the specific values may

be subject to scrutiny. In contrast, such a leniency is not applicable to matrices and quaternions, where the concept of validity is more rigidly defined.

Although Euler angles is widely adopted, the aliasing problem in Euler angles arises due to the non-uniqueness of the representation of a given orientation. For a single orientation, there exist multiple Euler angle triples that can be employed to describe it, leading to ambiguity in interpretation. This phenomenon, known as aliasing, can pose challenges in various applications. Basic questions, such as whether two different Euler angle triples represent the same angular displacement, become intricate to answer due to aliasing. Additionally, a more troublesome form of aliasing emerges from the interdependence of the three Euler angles. For instance, a particular orientation, like pitching down by 135Â°, can be equivalently represented as heading 180Â°, pitching down by 45Â°, and then banking 180Â°. To mitigate this aliasing issue, practitioners often establish canonical sets of Euler angles by constraining the ranges of the individual angles. These canonical sets ensure a unique representation for any given orientation, simplifying various computations and tests. Nonetheless, aliasing can still manifest itself, particularly in situations leading to the notorious Gimbal lock, where an angle choice for the second rotation can cause rotations around the same axis. Addressing this aliasing is essential for achieving unambiguous orientation representations using Euler angles.

Another disadvantage of Euler angles is Gimbal lock which is a specific and noteworthy issue associated with Euler angles. It occurs when certain choices of Euler angles lead to a loss of one degree of freedom, resulting in a situation where two rotational axes become aligned, effectively reducing the system's ability to represent orientations uniquely. A classic example of Gimbal lock is when an orientation sequence begins with a rightward rotation about one axis (e.g., heading), followed by a pitch-down rotation, where the pitch angle approaches 90 degrees. In this scenario, further changes in orientation may inadvertently lead to unintended rotations about

the same axis. This phenomenon can result in significant challenges in applications that require continuous and seamless 3D rotations. To address Gimbal lock, conventions are often established within the canonical set of Euler angles, ensuring that certain angle combinations, such as a pitch of 90 degrees, are accompanied by specific behaviors, such as constraining the bank angle to zero. Effectively managing Gimbal lock is crucial for preserving the integrity of orientation representations and avoiding unexpected transformations in systems employing Euler angles.

4.2.3 Axis-angles

Axis-angle rotation is a mathematical representation of 3D rotations. It is defined by two primary components: the axis of rotation and the associated angle. The axis represents the direction around which the rotation occurs, typically expressed as a unit vector (a vector with a magnitude of 1). The angle specifies the magnitude of the rotation and is commonly measured in radians. This representation provides an intuitive and straightforward way to conceptualize orientation changes in three-dimensional space. It's a natural approach to thinking about how objects or entities are oriented.

Axis-angle rotation finds applications in various fields, making it a fundamental concept in 3D modeling and analysis. For instance, In the realm of computer graphics, axis-angle rotation is a fundamental tool for animating 3D objects and characters. It allows for smooth and realistic transitions between different orientations, ensuring that animations are visually appealing and lifelike. Moreover, In physics simulations, physics modeling, and engineering applications, axis-angle rotations are used to describe and manipulate the orientation of objects and systems. This is particularly important in fields where the accurate representation of 3D rotations is vital.

One of the significant strengths of axis-angle rotation is its intuitive interpretation. It directly conveys the axis of rotation and the angle of rotation, which aligns with how humans naturally think about orientation changes. When we think about turn-

ing an object or reorienting ourselves, we often visualize a rotation around an axis. This direct correspondence makes axis-angle rotation easy to grasp and work with. Furthermore, it offers a compact representation, using only four values, which makes it memory-efficient. This is particularly valuable in applications where memory usage is a concern. Additionally, axis-angle rotations allow for straightforward linear interpolation between orientations. This means that you can smoothly transition from one orientation to another, which is crucial in animations and simulations.

Axis-angle rotation is interoperable with various other rotation representations, including rotation matrices and quaternions. It can be converted to and from these representations, allowing for flexibility and compatibility with different rotation systems and frameworks. This interoperability is particularly useful in applications where different tools and libraries use different rotation representations, as it enables seamless communication and data exchange.

One of the limitations of axis-angle rotation is the non-uniqueness of the representation. Multiple unit vectors can represent the same rotation, which can lead to non-uniqueness. In some cases, this non-uniqueness may cause ambiguities or complications. Additionally, axis-angle rotations can experience issues similar to gimbal lock when the rotation axis aligns with one of the coordinate axes. Gimbal lock occurs when one degree of freedom is lost in the rotation, and it can limit the ability to represent certain orientations without introducing singularities.

Mathematically, Axis-angle rotation is a representation of 3D rotations that combines two crucial elements: the rotation axis and the rotation angle. These elements are represented as follows:

1. Rotation Axis (Unit Vector): The rotation axis is specified as a unit vector, typically denoted as $\mathbf{u} = (u_x, u_y, u_z)$. A unit vector is a vector with a magnitude of 1, and it defines the direction around which the rotation occurs. The components (u_x, u_y, u_z) indicate the direction of the axis in 3D space.

2. Rotation Angle (θ): The rotation angle θ represents the magnitude of the rotation and is usually measured in radians. It quantifies the extent of the rotation around the defined axis. The rotation angle can be thought of as the 'strength' of the rotation.

Defining the Axis from the Unit Vector: The unit vector $\mathbf{u}$ represents the rotation axis. To define this axis, we calculate its components using the following equation:

$$\mathbf{u} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

Where:

- $\mathbf{u}$ is the unit vector representing the rotation axis.
- $\mathbf{w}$ is a 3D vector that can represent the axis of rotation. Let's consider an example where $\mathbf{w} = (1, 1, 0)$, which represents an axis that points diagonally in the xy-plane.
- $\|\mathbf{w}\|$ is the magnitude of vector $\mathbf{w}$, calculated as $\sqrt{1^2 + 1^2 + 0^2} = \sqrt{2}$, which is approximately 1.414.

By dividing the vector $\mathbf{w}$ by its magnitude, we obtain the unit vector $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2}, 0)$, which represents the rotation axis in a unit form.

Determining the Rotation Angle (θ): The rotation angle θ is determined based on the magnitude of the vector $\mathbf{w}$, which is not a unit vector:

$$\theta = \|\mathbf{w}\|$$

In the example we considered, $\theta = \sqrt{2}$, which is approximately 1.414 radians. This angle quantifies the extent of the rotation.

4.2.4 Exponential Map

Exponential map rotation is a mathematical representation used to describe 3D rotations by mapping them onto a tangent space. In exponential map rotation, a 3D rotation is represented as a 3D vector, often denoted as $\mathbf{w}$, which lies in a tangent space of the rotation manifold. This vector encodes both the axis and the angle of rotation. The magnitude of the vector $\mathbf{w}$ represents the rotation angle, while the direction of $\mathbf{w}$ specifies the rotation axis. This representation provides a compact and efficient way to describe rotations, allowing for smooth interpolation between orientations.

Exponential map rotation finds applications in various domains. In computer graphics, it's used for animating 3D objects and character movements, providing smooth transitions between keyframes. In robotics, it plays a critical role in robot kinematics, control, and path planning. In computer vision, exponential map rotation is utilized for 3D object tracking, structure-from-motion, and camera pose estimation.

One of the strengths of exponential map rotation is its ability to represent rotations compactly. Unlike other representations that require multiple parameters (e.g., Euler angles), exponential map rotation condenses the information into a single 3D vector, reducing memory and computational requirements. Moreover, it naturally handles small-angle rotations, making it suitable for applications where precision is crucial.

Exponential map rotation can be less intuitive for human understanding compared to representations like Euler angles. While it excels in compactness, it may not be as straightforward for visualization or manual editing of orientations. Additionally, singularities can be encountered at large rotation angles, which can lead to numerical instability.

Exponential map rotation is particularly valuable in scenarios where efficiency and compactness are paramount. Real-time biomechanics analysis frameworks, which require rapid computations for human movement assessment, benefit from this repre-

sentation. It is also instrumental in skeleton-based human action recognition, where the focus is on extracting meaningful features efficiently. Radar-based sensing for human monitoring leverages exponential map rotation to track and analyze human movements in real time.

One notable drawback of exponential map rotation is its susceptibility to singularities at 180-degree rotations, which can lead to mathematical instabilities. Handling these singularities requires careful consideration and additional computational effort.

Exponential map rotation can be converted to other rotation representations, such as rotation matrices or quaternions, for compatibility with different frameworks and systems. This interoperability allows seamless integration of exponential map rotation into a wide array of applications.

Mathematically, In exponential map rotation, a 3D rotation is represented as a 3D vector, often denoted as $\mathbf{w}$. The magnitude of this vector, $\|\mathbf{w}\|$, represents the rotation angle, denoted as θ. The direction of $\mathbf{w}$ specifies the rotation axis. The mathematical relationship between $\mathbf{w}$, θ, and the unit rotation axis $\mathbf{u}$ can be expressed as follows:

$$\mathbf{w} = \theta \cdot \mathbf{u}$$

Where:

- $\mathbf{w}$ is the 3D vector encoding the axis and angle of rotation.
- θ is the rotation angle in radians.
- $\mathbf{u}$ is the unit vector representing the rotation axis.
- The multiplication of θ by $\mathbf{u}$ scales the unit vector to represent the rotation.

As a demonstration, Let's consider an example where we want to represent a 90-degree counterclockwise rotation around the z-axis using exponential map rotation. In this case, $\theta = \frac{\pi}{2}$ radians (since 90 degrees is $\frac{\pi}{2}$ radians), and the unit vector $\mathbf{u}$

corresponds to the z-axis, which is $(0, 0, 1)$. Using the formula, we can calculate $\mathbf{w}$ as follows:

$$\mathbf{w} = \frac{\pi}{2} \cdot (0, 0, 1) = (0, 0, \frac{\pi}{2})$$

So, the exponential map representation for a 90-degree counterclockwise rotation around the z-axis is $\mathbf{w} = (0, 0, \frac{\pi}{2})$.

In conclusion, exponential map rotation provides an efficient and compact way to represent 3D rotations by combining the rotation axis and angle into a single 3D vector. Its mathematical representation simplifies interpolation and transformation, making it a valuable tool in a wide range of applications.

Exponential map and axis-angle representations are often used interchangeably, despite being conceptually distinct, due to their close relationship in representing 3D rotations. Both representations encode essential information about rotations, such as the rotation axis and angle, and this shared information allows for a seamless transition between them. In practice, users often leverage this interchangeability to facilitate various tasks. For instance, they may employ exponential map rotation for its efficiency and compactness and then convert it to axis-angle representation when human interpret

4.2.5 Quaternions

Quaternions are a hypercomplex number system consisting of four components: a scalar part (real) and a vector part (imaginary) in 3D space. Quaternion rotation primarily concerns unit quaternions, which have a magnitude of one. These unit quaternions can be represented as $q = w + xi + yj + zk$, where w is the scalar part and (x, y, z) is the vector part. Quaternion multiplication efficiently combines rotations and provides a concise means of representing 3D orientations. Quaternion rotation is based on Hamilton's original quaternion algebra, as developed in the 19th

century.

Quaternions offer several advantages that make them ideal for applications such as digital twins and biomechanics. They avoid the problem of gimbal lock, a limitation encountered in Euler angles. Quaternion interpolation ensures smooth and stable transitions between rotations, which is crucial for realistic simulations and accurate biomechanical analysis. Additionally, quaternion multiplication is computationally efficient, facilitating real-time and high-performance computations in digital twins and biomechanical models.

While quaternions offer numerous strengths, they can be less intuitive for human interpretation compared to representations like Euler angles. Users often require visualization tools or conversion functions to understand quaternion rotations. Additionally, quaternion normalization and handling the sign ambiguity (i.e., two quaternions representing the same rotation) can be complex and may require additional computational steps.

Quaternions are hypercomplex numbers represented as $q = w + xi + yj + zk$, where w is the scalar part and (x, y, z) is the vector part. For quaternion rotation, unit quaternions are utilized, meaning they have a magnitude of one: $\|q\| = 1$. Quaternion multiplication combines rotations effectively, offering a concise way to represent 3D orientations.

The mathematical relationships within quaternions are as follows:

First, The quaternion Norm which represents magnitude or length of a quaternion q whcih could be calculated as:

$$\|q\| = \sqrt{w^2 + x^2 + y^2 + z^2}$$

For unit quaternions, $\|q\| = 1$.

Second, The conjugate of a quaternion $q = w + xi + yj + zk$ is given by:

$$q^* = w - xi - yj - zk$$

The conjugate is important for quaternion inversion.

To ensure a quaternion is a unit quaternion (having $\|q\| = 1$), normalization is performed by dividing each quaternion component by its magnitude:

$$q_{\text{normalized}} = \frac{q}{\|q\|}$$

For the sake of demonstration, Let's consider an example where we want to represent a 90-degree counterclockwise rotation about the z-axis using quaternion rotation. In this case, the unit quaternion can be expressed as:

$$q = \cos\left(\frac{\theta}{2}\right) + \sin\left(\frac{\theta}{2}\right) k$$

Here, $\theta = \frac{\pi}{2}$ radians (90 degrees in radians), and the quaternion axis corresponds to the z-axis.

Plugging in the values:

$$q = \cos\left(\frac{\pi}{4}\right) + \sin\left(\frac{\pi}{4}\right) k$$

$$q = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}k$$

So, the quaternion representation for a 90-degree counterclockwise rotation about the z-axis is $q = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}k$.

4.2.6 Rotation Angles Transformation

The conversion between various rotation representations is a fundamental concept in the field of 3D graphics, computer vision, robotics, biomechanics, and many other areas where the accurate depiction of orientation and motion is essential. Under-

standing and seamlessly transitioning between different rotation representations is of paramount importance for several reasons.

First and foremost, different applications and domains often require different rotation representations. For example, representing the orientation of an object in 3D space for computer graphics might best be achieved using quaternions, while biomechanical analysis may rely on Euler angles. As such, the ability to convert between these representations facilitates the integration of techniques and data across disciplines, fostering collaboration and enabling multi-disciplinary research.

Furthermore, the choice of rotation representation can significantly impact computational efficiency. Some representations are more suitable for real-time applications, such as gaming and virtual reality, while others may be better suited for high-precision scientific simulations. Conversions between representations allow for the optimization of performance and the selection of the most appropriate representation for a given task.

In addition, error accumulation is a common concern when working with rotations. Converting between representations can help mitigate error propagation, ensuring that the final result remains as accurate as possible. This is particularly crucial in fields where precise motion tracking and orientation control are imperative, such as robotics and motion capture.

The ability to convert between rotation representations also enhances interoperability in the context of data exchange and communication between software, hardware, and research teams. Standardized conversions allow researchers, developers, and engineers to share data seamlessly, regardless of the chosen rotation representation in their respective applications.

In essence, the ability to convert between diverse rotation representations stands as an indispensable asset within the arsenal of professionals and researchers immersed in the realm of 3D rotations and orientations. It not only facilitates seamless collabo-

ration but also enhances computational efficiency, mitigates error accumulation, and fosters interoperability. This multifaceted utility extends its impact across a spectrum of fields, spanning from the realms of computer graphics and computer vision to the domains of biomechanics and robotics. The subsequent segments of this section offer a comprehensive exploration of various rotation angles transformations.

4.2.6.1 Euler angles to a matrix

Euler angles can be converted to a 3x3 rotation matrix using the following formulas, depending on the order of rotations (often denoted as roll-pitch-yaw or yaw-pitch-roll):

For ZYX rotation order (yaw-pitch-roll):

$$\mathbf{R} = \mathbf{R}_z(\psi) \cdot \mathbf{R}_y(\theta) \cdot \mathbf{R}_x(\phi)$$

Where: - $\mathbf{R}_z(\psi)$ represents the rotation matrix for a yaw (rotation about the z-axis) by an angle ψ. - $\mathbf{R}_y(\theta)$ represents the rotation matrix for a pitch (rotation about the y-axis) by an angle θ. - $\mathbf{R}_x(\phi)$ represents the rotation matrix for a roll (rotation about the x-axis) by an angle ϕ.

For other rotation orders (such as XYZ), you would change the order of multiplication accordingly.

Let's consider a specific example. Suppose we have a set of Euler angles representing a 3D rotation as follows:

Roll (ϕ): 45° Pitch (θ): 30° Yaw (ψ): 60°

We want to convert these Euler angles to a 3x3 rotation matrix $\mathbf{R}$. Using the ZYX rotation order, we can apply the formula:

$$\mathbf{R} = \mathbf{R}_z(60^\circ) \cdot \mathbf{R}_y(30^\circ) \cdot \mathbf{R}_x(45^\circ)$$

Now, let's calculate each individual rotation matrix:

$$\mathbf{R}_z(60^\circ) = \begin{bmatrix} \cos(60^\circ) & -\sin(60^\circ) & 0 \\ \sin(60^\circ) & \cos(60^\circ) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{R}_y(30^\circ) = \begin{bmatrix} \cos(30^\circ) & 0 & \sin(30^\circ) \\ 0 & 1 & 0 \\ -\sin(30^\circ) & 0 & \cos(30^\circ) \end{bmatrix}$$

$$\mathbf{R}_x(45^\circ) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(45^\circ) & -\sin(45^\circ) \\ 0 & \sin(45^\circ) & \cos(45^\circ) \end{bmatrix}$$

Now, multiply these matrices in the correct order:

$$\mathbf{R} = \mathbf{R}_z(60^\circ) \cdot \mathbf{R}_y(30^\circ) \cdot \mathbf{R}_x(45^\circ)$$

This multiplication will yield the 3x3 rotation matrix $\mathbf{R}$ representing the given Euler angles.

4.2.6.2 Matrix to Euler angles

To convert a 3x3 rotation matrix $\mathbf{R}$ to Euler angles (roll-pitch-yaw order), you can use the following formulas:

$$\phi = \arctan 2(R_{32}, R_{33})$$

$$\theta = -\arcsin(R_{31})$$

$$\psi = \arctan 2(R_{21}, R_{11})$$

Where: - ϕ represents the roll angle. - θ represents the pitch angle. - ψ represents the yaw angle. - R_{ij} denotes the elements of the rotation matrix $\mathbf{R}$.

Example: Suppose you have a rotation matrix $\mathbf{R}$:

$$\mathbf{R} = \begin{bmatrix} 0.866 & -0.5 & 0 \\ 0.5 & 0.866 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

You want to find the corresponding Euler angles. Using the formulas:

$$\phi = \arctan 2(0, 0) = 0 \text{ radians}$$
$$\theta = -\arcsin(0.5) = -\frac{\pi}{6} \text{ radians}$$
$$\psi = \arctan 2(0.5, 0.866) = \frac{\pi}{3} \text{ radians}$$

Converting the angles from radians to degrees:

$$\phi = 0^\circ$$
$$\theta = -30^\circ$$
$$\psi = 60^\circ$$

So, the Euler angles for the given rotation matrix are roll (ϕ) = 0°, pitch (θ) = -30°, and yaw (ψ) = 60°.

4.2.6.3 Quaternion to a Matrix

To convert from a quaternion to a 3x3 rotation matrix mathematically, you can use the following formulas:

$$R_{11} = q_1^2 - q_2^2 - q_3^2 + q_4^2$$

$$R_{12} = 2(q_1 q_2 + q_3 q_4)$$

$$R_{13} = 2(q_1 q_3 - q_2 q_4)$$

$$R_{21} = 2(q_1 q_2 - q_3 q_4)$$

$$R_{22} = -q_1^2 + q_2^2 - q_3^2 + q_4^2$$

$$R_{23} = 2(q_2 q_3 + q_1 q_4)$$

$$R_{31} = 2(q_1 q_3 + q_2 q_4)$$

$$R_{32} = 2(q_2 q_3 - q_1 q_4)$$

$$R_{33} = -q_1^2 - q_2^2 + q_3^2 + q_4^2$$

Where: - R_{ij} represents the elements of the 3x3 rotation matrix. - q_1, q_2, q_3, q_4 are the quaternion components.

Here's an example of converting a quaternion to a rotation matrix:

Suppose you have a quaternion:

$$q = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i + 0j + 0k$$

You want to find the corresponding 3x3 rotation matrix. Using the formulas above:

$$R_{11} = \left(\frac{1}{\sqrt{2}}\right)^2 - \left(\frac{1}{\sqrt{2}}\right)^2 - 0^2 + 0^2 = 0$$
$$R_{12} = 2\left(\frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + 0 \cdot 0\right) = 1$$
$$R_{13} = 2\left(\frac{1}{\sqrt{2}} \cdot 0 - \frac{1}{\sqrt{2}} \cdot 0\right) = 0$$
$$R_{21} = 2\left(\frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} - 0 \cdot 0\right) = 1$$
$$R_{22} = -\left(\frac{1}{\sqrt{2}}\right)^2 + \left(\frac{1}{\sqrt{2}}\right)^2 - 0^2 + 0^2 = 0$$
$$R_{23} = 2\left(0 \cdot 0 + \frac{1}{\sqrt{2}} \cdot 0\right) = 0$$
$$R_{31} = 2\left(\frac{1}{\sqrt{2}} \cdot 0 + 0 \cdot 0\right) = 0$$
$$R_{32} = 2\left(0 \cdot 0 - \frac{1}{\sqrt{2}} \cdot 0\right) = 0$$
$$R_{33} = -\left(\frac{1}{\sqrt{2}}\right)^2 - \left(\frac{1}{\sqrt{2}}\right)^2 + 0^2 + 0^2 = -1$$

So, the 3x3 rotation matrix corresponding to the given quaternion is:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

4.2.6.4 Matrix to a Quaternion

To convert from a 3x3 rotation matrix to a quaternion mathematically, you can use the following formulas:

$$q_1 = \frac{1}{2}\sqrt{1 + R_{11} - R_{22} - R_{33}}$$
$$q_2 = \frac{1}{2}\sqrt{1 - R_{11} + R_{22} - R_{33}}$$
$$q_3 = \frac{1}{2}\sqrt{1 - R_{11} - R_{22} + R_{33}}$$
$$q_4 = \frac{1}{2}\sqrt{1 + R_{11} + R_{22} + R_{33}}$$

Where: - q_1, q_2, q_3, q_4 are the quaternion components. - R_{ij} represents the elements of the 3x3 rotation matrix.

Here's an example of converting a 3x3 rotation matrix to a quaternion:

Suppose you have a rotation matrix:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

You want to find the corresponding quaternion components. Using the formulas above:

$$q_1 = \frac{1}{2}\sqrt{1+0-0-1} = 0$$
$$q_2 = \frac{1}{2}\sqrt{1-0+0-1} = 0$$
$$q_3 = \frac{1}{2}\sqrt{1-0-0+1} = \frac{1}{2}$$
$$q_4 = \frac{1}{2}\sqrt{1+0+0+1} = \frac{1}{2}$$

So, the quaternion corresponding to the given 3x3 rotation matrix is:

$$q = 0 + 0i + \frac{1}{2}j + \frac{1}{2}k$$

4.2.6.5 Euler Angles to a Quaternion

To convert from Euler angles to a quaternion mathematically, you can use the following formulas:

For the ZYX representation (yaw-pitch-roll), you can calculate the quaternion components as follows:

$$q_w = \cos(\frac{\theta_x}{2}) \cdot \cos(\frac{\theta_y}{2}) \cdot \cos(\frac{\theta_z}{2}) + \sin(\frac{\theta_x}{2}) \cdot \sin(\frac{\theta_y}{2}) \cdot \sin(\frac{\theta_z}{2})$$

$$q_x = \sin(\frac{\theta_x}{2}) \cdot \cos(\frac{\theta_y}{2}) \cdot \cos(\frac{\theta_z}{2}) - \cos(\frac{\theta_x}{2}) \cdot \sin(\frac{\theta_y}{2}) \cdot \sin(\frac{\theta_z}{2})$$

$$q_y = \cos(\frac{\theta_x}{2}) \cdot \sin(\frac{\theta_y}{2}) \cdot \cos(\frac{\theta_z}{2}) + \sin(\frac{\theta_x}{2}) \cdot \cos(\frac{\theta_y}{2}) \cdot \sin(\frac{\theta_z}{2})$$

$$q_z = \cos(\frac{\theta_x}{2}) \cdot \cos(\frac{\theta_y}{2}) \cdot \sin(\frac{\theta_z}{2}) - \sin(\frac{\theta_x}{2}) \cdot \sin(\frac{\theta_y}{2}) \cdot \cos(\frac{\theta_z}{2})$$

Where: - q_w represents the scalar (real) part of the quaternion. - q_x, q_y, q_z represent the vector (imaginary) parts of the quaternion. - θ_x represents the roll angle. - θ_y represents the pitch angle. - θ_z represents the yaw angle.

For example, consider Euler angles representing a 3D rotation:

$$\theta_x = 45^\circ$$

$$\theta_y = 30^\circ$$

$$\theta_z = 60^\circ$$

To convert these Euler angles to a quaternion in the ZYX representation, you can use the provided formulas. Substitute the angles θ_x, θ_y, and θ_z into the equations to obtain the corresponding quaternion components q_w, q_x, q_y, q_z.

In this example, the calculated quaternion will represent the orientation in the ZYX representation.

4.2.6.6 Quaternion to Euler Angles

To convert from a quaternion to Euler angles mathematically, you can use the following formulas, depending on the desired Euler angle representation (e.g., ZYX):

For the ZYX representation (yaw-pitch-roll), you can calculate the Euler angles as follows:

$$\theta_x = \text{atan2}\left(2(q_w q_x + q_y q_z), 1 - 2(q_x^2 + q_y^2)\right)$$

$$\theta_y = \text{asin}\left(2(q_w q_y - q_x q_z)\right)$$

$$\theta_z = \text{atan2}\left(2(q_w q_z + q_x q_y), 1 - 2(q_y^2 + q_z^2)\right)$$

Where: - θ_x represents the roll angle. - θ_y represents the pitch angle. - θ_z represents the yaw angle. - q_w, q_x, q_y, q_z are the quaternion components.

For example, consider a quaternion with the following components:

$$q_w = 0.866$$

$$q_x = 0.25$$

$$q_y = 0.353$$

$$q_z = 0.288$$

To convert this quaternion to Euler angles in the ZYX representation, you can use the above formulas:

$$\theta_x = \text{atan2}\left(2(0.866 \cdot 0.25 + 0.353 \cdot 0.288), 1 - 2(0.25^2 + 0.353^2)\right)$$

$$\theta_y = \text{asin}\left(2(0.866 \cdot 0.353 - 0.25 \cdot 0.288)\right)$$

$$\theta_z = \text{atan2}\left(2(0.866 \cdot 0.288 + 0.25 \cdot 0.353), 1 - 2(0.353^2 + 0.288^2)\right)$$

After solving these equations, you will obtain the Euler angles θ_x, θ_y, and θ_z representing the given quaternion rotation.

In this example, the calculated Euler angles will represent the orientation in the ZYX representation.

4.2.6.7 Axis-angles to Matrix

To convert from axis-angles to a 3x3 rotation matrix mathematically, you can use the following formula:

$$R = I + \sin(\theta)K + (1 - \cos(\theta))K^2$$

Where: - R is the 3x3 rotation matrix. - I is the identity matrix. - θ is the rotation angle. - K is the skew-symmetric matrix derived from the unit axis vector $\mathbf{u} = [x, y, z]$:

$$K = \begin{bmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{bmatrix}$$

For example, consider an axis-angle representation where the rotation axis is $\mathbf{u} = [1, 0, 0]$ (i.e., rotation about the x-axis) and the rotation angle is $\theta = \frac{\pi}{4}$ radians. To convert this to a rotation matrix:

$$\mathbf{u} = [1, 0, 0]$$

$$\theta = \frac{\pi}{4}$$

Calculate K using the axis:

$$K = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

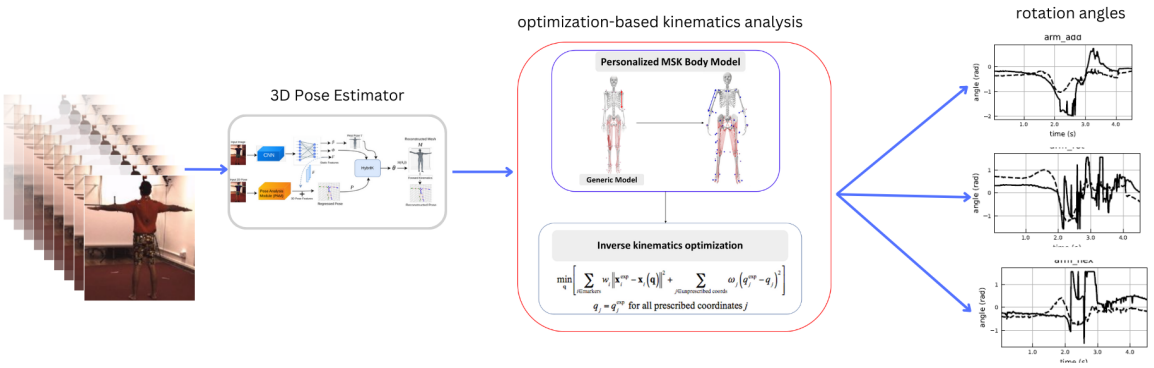

Figure 4.1: System Overview

Apply the Rodrigues' rotation formula to obtain the rotation matrix:

$$R = I + \sin\left(\frac{\pi}{4}\right) K + \left(1 - \cos\left(\frac{\pi}{4}\right)\right) K^2$$

After solving this equation, you will have the 3x3 rotation matrix representing the given axis-angle rotation.

In this example, the rotation matrix R will represent a 45-degree counterclockwise rotation about the x-axis.

4.3 Methodology

4.3.1 preliminary

When dealing with a set of marker trajectories that correspond to a specific motion of interest, software applications are tasked with the intricate process of reconstructing a digital representation of the experimental subject. This digital counterpart, often referred to as a "scaled" model, must closely match the subject's actual segment dimensions. This initial step is termed "scaling." For the attainment of precise kinematic results, adjustments are then necessary for marker positions on the scaled digital twin. These adjustments account for variations stemming from 3D pose estimation and the inherent differences in human subjects' dimensions; this is commonly known as "marker registration." Subsequently, determining the positions and orienta-

tions of body segments over time is a crucial undertaking. This is typically achieved through an optimization process known as "inverse kinematics." Notably, inverse kinematics algorithms tend to yield more accurate outcomes when constrained by an underlying skeletal model. This relationship between scaling, marker registration, and inverse kinematics necessitates an iterative process, where experts iteratively refine each step, making incremental adjustments to each value until the desired level of accuracy is achieved. For instance, altering the length of the upper arm segment in a subject's digital twin necessitates corresponding adjustments in marker registrations for forearm and hand markers, as a lengthened upper arm would affect their positions. Furthermore, this alteration in segment length would inherently impact the resultant motion derived through inverse kinematics. This process typically demands substantial subjective input from an expert.

The automation of the scaling and registration procedures has been explored in previous research, notably by Reinbolt *et al.* [203] and *Charlton et al.* [204]. These authors employed gradient-free optimization techniques to autonomously estimate body segment scales and marker registrations while repetitively solving gradient-based inverse kinematics problems in an inner-loop to assess optimization progress. It is worth noting that these methods require significant computational resources since each iterative decision made by the outer optimizer regarding body segment scaling and marker offsets necessitates solving a computationally intensive inner optimization problem, namely inverse kinematics, to evaluate the quality of the decision.

Recently, markerless motion capture systems based on video recordings have gained prominence, primarily owing to their cost-effectiveness as they circumvent the need for expensive motion capture equipment [205]. Notably, these methodologies dispense with the tracking of optical markers, instead concentrating on the amalgamation of markerless motion capture techniques, such as pose detection, with appropriately scaled musculoskeletal models. This synthesis is aimed at encompassing the intricacies

of physiological joint constraints [206]. It is essential to acknowledge that these novel approaches still necessitate the resolution of an inverse kinematics problem. However, this process is facilitated through the utilization of keypoints extracted from pose detection algorithms rather than traditional optical markers.

Furthermore, it is worth emphasizing that the precision of scaled models holds paramount significance, as it enables more profound and intricate biomechanical analyses in conjunction with markerless motion capture techniques. These analyses extend to the estimation of kinetic parameters, encompassing joint moments and muscle forces [206]. The marriage of accurate scaled models with markerless motion capture not only enhances the cost-efficiency of motion analysis but also broadens the spectrum of biomechanical insights that can be gleaned from such data.

4.3.2 System Overview

As depicted in 4.1, our kinematic analysis system is composed of two principal components. In the first component, 2D images are processed using the newly developed monocular 3D pose estimator, GTRSV2 as elaborated in 3, to estimate 3D human poses. The second component employs physics-based optimization techniques, specifically inverse kinematics, to refine this data and calculate joint rotation angles.

4.3.3 Optimization-Based Inverse Kinematic Solver

The 3D pose estimation from GTRS v2 can be analogized to experimental markers placed on the body. The Inverse Kinematic (IK) module is tasked with computing the joint angles based on the positions of these experimental markers. The IK module consists of two core components: musculoskeletal model scaling and inverse kinematic optimization solver. In the first step, a generic musculoskeletal model must be scaled to adjust its physical dimensions to align with the unique anthropometry of the specific human subject being analyzed. This scaling process is a crucial phase in the context of inverse kinematics, as the precision achieved during this stage has a

significant impact on the accuracy of subsequent solutions. Once the musculoskeletal model has been appropriately scaled, the IK tool then proceeds to process each time step (or frame) of motion. During this phase, it calculates the joint rotation angle values that position the model markers in a pose that best corresponds to the experimental marker values for that specific time step. Mathematically, this "best match" is formulated as a weighted least squares problem as shown in 4.3.3, the objective of which is to minimize thecoordinate discrepancies between the experimental and model markers. To resolve this optimization problem, a commercial non-linear optimization solver IPOPT is employed. The outputs of the solver are the rotation angles, such as the arm rotation angles of a baseball pitcher, as illustrated in 4.2.

$$\min_{\mathbf{q}} \left[\sum_{i \in \text{ markers}} w_i \left\| \mathbf{x}_i^{\text{exp}} - \mathbf{x}_i(\mathbf{q}) \right\|^2 + \sum_{j \in \text{ unprescribed coords}} \omega_j \left(q_j^{\text{exp}} - q_j \right)^2 \right]$$

$$q_j = q_j^{\text{exp}} \text{ for all prescribed coordinates } j$$

4.3.3.1 Musculoskeletal Model

Individuals exhibit substantial anatomical variation, encompassing attributes such as stature, limb proportions, body mass, and muscular constitution, all of which exert a profound influence on biomechanical dynamics. Failure to calibrate a model to an individual's unique anatomical characteristics compromises its fidelity in representing their locomotion and forecasting the repercussions of imposed mechanical loads on their physique.

To tailor the comprehensive musculoskeletal framework delineated by Rajagopal [207] to our particular research requirements, notable adaptations were instituted. The originally prescribed marker set was supplanted by the SMPL [1] marker configuration, ensuring markers were judiciously situated on the body in close proximity. At this developmental stage, muscular elements were deliberately excluded from considerations. Further amendments included the liberalization of restrictions on pelvis translation and subtalar angle rotation, rendering these degrees of freedom unre-

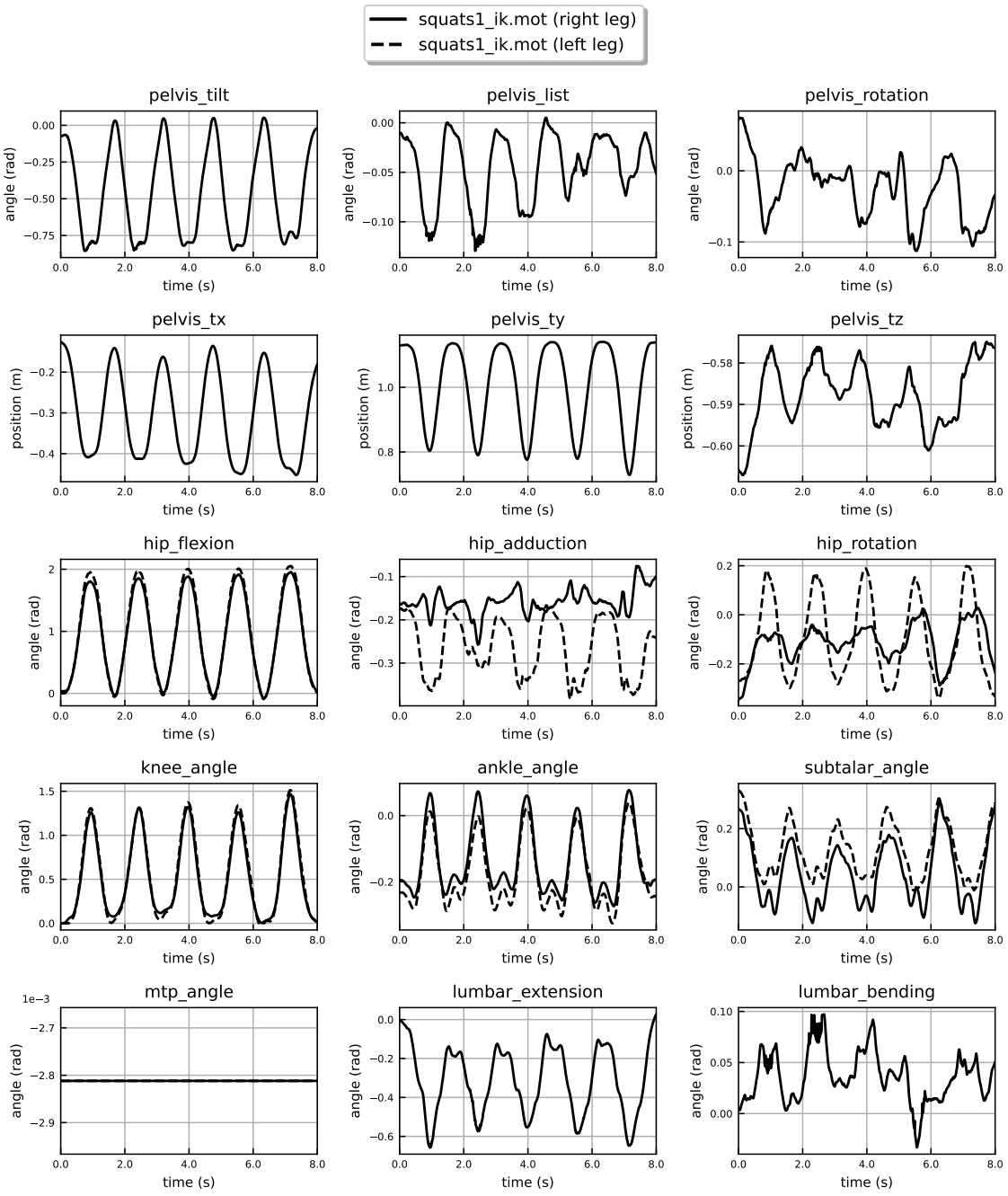

Figure 4.2: Kinematic rotation angle for various human joints based on Motion Capture System

stricted.

The inception of our musculoskeletal model, fashioned after the Skinned Multi-Person Linear (SMPL) model, is driven by the pervasive adoption of SMPL [1] as a benchmark in crafting human mesh representations. By adhering to the foundational principles of SMPL, our model ensures seamless integration and familiarity within the research community. Our primary motivation lies in creating a musculoskeletal framework that not only aligns with the prevalent use of SMPL but also lays the groundwork for future optimization endeavors. Given SMPL's proven success in generating versatile human mesh representations, our model serves as an extension, aiming to maintain compatibility while concurrently positioning itself for subsequent refinement and enhancement. This strategic alignment with SMPL underscores our commitment to contributing to the ongoing evolution of human motion analysis and simulation techniques.

4.3.3.2 Scaling Musculoskeletal Model

Scaling operations in biomechanical modeling are paramount for ensuring the accuracy of kinematic analyses. Properly scaled models are essential to precisely represent an individual's unique anatomical proportions and characteristics. This is particularly crucial in understanding and interpreting joint kinematics, as inaccuracies in model scaling can lead to misrepresentations of joint angles and movement patterns. By addressing the scaling intricacies through optimization, the proposed workflow aims to enhance the fidelity of biomechanical analyses, thereby contributing to more accurate and reliable assessments of human movement.

To scale the model, we leveraged AddBiomechanic platform [208] whereThe initial phase of the workflow involves addressing model scaling and inverse kinematics challenges through a series of linear and bilevel optimization problems. The objective is to iteratively determine optimal solutions for crucial parameters, including body segment scale factors, marker registrations, and joint kinematics. The intricacy of

these interdependent problems necessitates a sequential resolution approach, where the outcome of each optimization problem serves as the initial guess for the subsequent one. Furthermore, if the user provides ground reaction force data, the AddBiomechanics process undertakes an additional step. This step entails a linear optimization to estimate the center of mass trajectory and overall subject mass. Subsequently, a non-convex optimization step is executed to minimize residual forces and refine the original model scaling and joint kinematics solution.

4.3.4 Mobile Application

The proposed workflow for mobile 3D pose estimation adopts a two-tiered methodology encompassing a sequential 2D pose estimation phase and a transformer-based model designed for 3D pose estimation. In the initial component, we employ real-time 2D pose estimation algorithms, specifically YOLOv8 [209], to capture the pose of a subject in a two-dimensional representation, derived from a single data frame. This initial step furnishes us with a foundational portrayal of the subject's posture in a 2D domain. The subsequent component, which incorporates a transformer-based 2D-to-3D transformation using our GTRS 3 model, takes these 2D pose estimations and elevates them into the three-dimensional spatial domain. This transformative process assumes paramount significance in achieving 3D pose estimations at a remarkable rate of 10 frames per second, as illustrated in [figure].

4.4 Conclusion

In conclusion, our two-stage approach has proven to be a robust and versatile methodology for generating diverse kinematics and dynamic indicators, laying a solid foundation for subsequent in-depth analyses. Through a meticulous integration of linear and bilevel optimization techniques, we successfully addressed the intricacies of model scaling, inverse kinematics, and dynamic estimation. The sequential resolution of these interdependent problems has allowed us to obtain nuanced insights

Figure 4.3: 2D Pose estimation

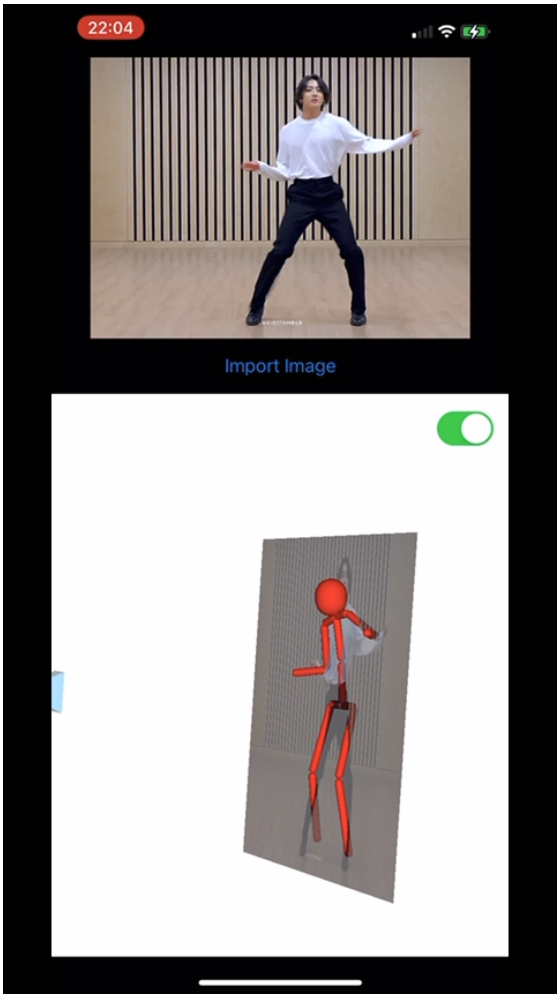

Figure 4.4: 3D pose estimation

Figure 4.5: IOS Human Pose Estimation

into body segment scale factors, marker registrations, joint kinematics, and dynamic characteristics.

Furthermore, our evaluation of transformer-based 3D pose estimation on mobile devices signifies a pivotal step towards the practical implementation of our methodologies. The deployment of pose estimation on phones opens new avenues for real-time biomechanical assessments in varied settings, transcending traditional laboratory setups. The efficiency and accuracy demonstrated in our evaluation highlight the potential of leveraging modern computational techniques for on-the-go biomechanical analysis.

In essence, our work not only contributes to the refinement of biomechanical modeling but also paves the way for practical applications in mobile-based biomechanical assessments. The generated kinematics and dynamic indicators serve as valuable resources for future research endeavors, offering a comprehensive understanding of human movement that extends beyond traditional constraints. This marks a significant stride towards enhancing the accessibility and applicability of biomechanical analyses in diverse contexts.

4.5 future Work

In the current state, our iOS application demonstrates the capability to generate 3D pose estimations at a commendable speed of 10 frames per second. However, recognizing the continuous pursuit of performance enhancement, our future endeavors will delve into exploring advanced model compression techniques. Specifically, we plan to investigate methodologies such as model pruning and knowledge distillation to optimize the inference speed further.

Beyond the scope of local device optimizations, a pivotal expansion lies in the establishment of a cloud server infrastructure. This server will serve as a computational hub, receiving 3D pose estimations from mobile devices and executing the IPOPT solver for efficient inverse kinematics optimizations. This architectural shift

not only offloads computational burdens from individual devices but also harnesses the scalability and processing capabilities of cloud-based solutions.

Simultaneously, recognizing the significance of real-time optimization on mobile platforms, we aim to implement an inverse kinematics optimizer directly on iOS devices. This will be facilitated by leveraging SciPy, a versatile Python library renowned for providing powerful non-linear optimization solvers. By integrating SciPy into our iOS application, we anticipate achieving on-device optimization capabilities, thereby enhancing the efficiency and autonomy of biomechanical analyses. These collective efforts underline our commitment to advancing both the speed and accessibility of 3D pose estimation and biomechanical optimization processes.

Enhancing the overall pipeline by integrating physics-based constraints into the pose estimation process represents a compelling and forward-thinking direction. By imbuing the pose estimation with biomechanical principles and adherence to the laws of physics, this approach holds the promise of producing more realistic and plausible human poses. The incorporation of physics laws not only contributes to increased biomechanical accuracy but also ensures that the estimated poses align cohesively with natural human movement patterns. This strategy has the potential to elevate the precision and reliability of pose estimations, reducing ambiguity and fostering consistency with real-world behavior. While introducing physics-based constraints adds a layer of complexity, the potential benefits in terms of improved accuracy, generalization across diverse scenarios, and enhanced interpretability make it a noteworthy avenue for advancing the efficacy and realism of biomechanical analyses and applications.

CHAPTER 5: Skeleton-based Human Action Recognition via Convolutional Neural Network

5.1 Introduction

The endeavor to imbue machines with human-like visual perception capabilities has engendered a profound scientific pursuit, yielding a multifaceted array of technological innovations, algorithms, and methodologies. This concerted effort has led to remarkable achievements across various paradigms within the realm of computer vision, exemplified by milestones in image classification [210] and object detection [211]. One particularly intricate facet of computer vision is the domain of human action recognition, characterized by a taxonomy predicated on the temporal dynamics and intricacy inherent to the actions observed.

Empowering machines with human-like vision capabilities has been a driving force in the field of computer vision and artificial intelligence. It encompasses the automated detection and analysis of human movements, gestures, and actions, often from video sequences or image data. Understanding and accurately categorizing human actions hold immense potential across a broad spectrum of applications, including surveillance, robotics, healthcare, sports analysis, and human-computer interaction. The quest to develop robust human action recognition systems has spurred a wealth of research activities, drawing from the intersection of computer vision, machine learning, and deep learning techniques. In this introductory exploration, we embark on a comprehensive journey to unravel the intricacies of human action recognition, delving into its significance, challenges, historical evolution, and the state of the art in this dynamic field.

The initial taxonomy delineates the category of "human gestures," characterized by

actions of relatively diminished complexity and brevity. Such actions often find their manifestations in commonplace demonstrations, such as hand waving or head nodding, symbolizing the incipient layer of human behavioral dynamics. A subsequent layer, referred to as "human actions," encapsulates actions of protracted duration and enhanced complexity, typically involving the orchestrated engagement of multiple anatomical segments. Moreover, the concatenation of distinct actions within a temporal sequence is defined as "human activity," representing a higher-level manifestation of human action recognition. Beyond these classifications, "human interaction" pertains to the engagement of individuals with their environmental milieu, encapsulating the complex and multifaceted interactions between humans and their surroundings.

Human action recognition serves as a critical building block in the ongoing evolution of intelligent systems. By enabling machines to comprehend and interpret human actions, we pave the way for a plethora of transformative applications across diverse domains. One of the most prevalent applications is in the realm of surveillance and security, where automated systems can detect and classify suspicious or anomalous activities in real-time video streams, thereby enhancing public safety [212]. In the domain of healthcare, human action recognition plays a pivotal role in monitoring patient activities, ensuring adherence to prescribed exercises, and tracking the progression of rehabilitation [213]. Furthermore, it has far-reaching implications in the entertainment industry, where motion capture technology based on human action recognition is widely employed in animation and virtual reality applications [214]. Lastly, The field's significance also extends to sports analysis, facilitating the detailed examination of athletes' performances, and the development of immersive gaming experiences that respond to human movements [215].

Historically, the purview of human action recognition predominantly encompassed the realm of machine learning algorithms. Conventionally, the task of human motion

analysis necessitated the intricate process of capturing and articulating spatiotemporal information, a process often conducted through manual feature extraction. These extracted features were subsequently integrated into classical machine learning frameworks, endowing algorithms with the capacity to undertake various recognition tasks. The manifold features enlisted in these efforts encompassed attributes such as joint coordinates [216], the center of gravity [217], the angles defining inter-joint relationships [218], motion velocity [219], co-occurrence features [129], and a myriad of other attributes ardently explored and assimilated by the research community. The pivotal role of algorithm selection in the machine learning paradigm is underscored by the deployment of methodologies such as Support Vector Machine (SVM) [220], Linear Discriminant Analysis (LDA) [221], Naive Bayes Nearest Neighbor [222], Logistic Regression [223], and K-Nearest Neighbors (KNN) [224] in the comprehensive landscape of human behavior analysis. Nevertheless, it is noteworthy that machine learning approaches are often hindered by the challenge of generalization and necessitate extensive feature engineering endeavors.

In recent years, the ascendancy of deep learning-based paradigms has transmuted the landscape of human action recognition, demonstrably surpassing traditional machine learning methods across an array of computer vision tasks, including image classification [210], object detection [211], action recognition [225], and action detection [226]. Notably, the discerning focus has gravitated toward the realm of skeleton-based representation, which has garnered significant attention due to the intrinsic richness and discriminative attributes that can be gleaned from the anatomical articulation of skeletal joints. This novel approach encompasses various methodologies, such as the "Skeleton map" [14], "Joint Trajectory Map" [15], "Joint Distance Map" [16], and an array of additional features extracted exclusively from skeletal joint data. These advancements serve as exemplars of the transformative potential intrinsic to the confluence of deep learning methodologies and human action recognition.

5.1.1 Challenges

5.1.1.1 Intra-class and inter-class variations

In the sphere of action recognition, the formidable challenges posed by intra-class and inter-class variations demand rigorous consideration. Intra-class variations encompass the diversity inherent in the execution of the same action, whether enacted by disparate individuals or under varying contextual conditions. Conversely, inter-class variations encompass the subtle distinctions between different action categories, often characterized by shared visual attributes that engender perceptual overlap. These intricacies introduce a conundrum, as they complicate the fine-grained discrimination within identical action categories and the differentiation of visually akin actions across classes.

To navigate this multifaceted challenge, a multidimensional strategy is requisite. Primarily, the cultivation of robust feature representations stands as an imperative directive. Advanced deep learning paradigms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have underscored their prowess in extracting salient, invariant features from intricate spatiotemporal data. The integration of attention mechanisms and graph neural networks further enhances the model's acumen for capturing subtle action nuances, ameliorating the discrimination of intra-class variations and augmenting the distinctiveness of inter-class actions.

In tandem, the requisition of expansive, meticulously curated datasets assumes a pivotal role. These datasets should encompass a diverse gamut of actions enacted by a profusion of individuals across heterogeneous settings, affording comprehensive coverage of the rich tapestry of intra-class and inter-class variations. This diverse corpus of data avails the action recognition models the opportunity to acquire heightened generalization prowess, as they assimilate a breadth of visual subtleties.

The imperative role of data augmentation techniques endures. By synthetically generating a panoply of action variations, including alterations in temporal dynam-

ics, scale, speed, or viewpoint, these augmentations instill the model with resilience against the vicissitudes of intra-class and inter-class variations. These synthetic diversifications encapsulate the authentic variability intrinsic to actions, thereby priming the model to accommodate an expansive array of real-world scenarios.

Furthermore, the ascendancy of fine-grained action recognition models is discernible. These models delve into the minute intricacies of actions, endowing them with the capacity to differentiate actions within the same category predicated on nuanced disparities, thus adroitly confronting intra-class variations. The incorporation of temporal modeling techniques, exemplified by long short-term memory (LSTM) networks and temporal convolutions, potentiates a heightened comprehension of the intricate temporal dynamics inherent in actions.

In summation, the mitigation of the challenges inherent in intra-class and inter-class variations within the purview of action recognition necessitates a holistic and multifaceted approach. This multifarious stratagem encompasses the refinement of feature representations, the embrace of expansive and variegated datasets, the judicious implementation of data augmentation techniques, and the adoption of fine-grained recognition models. This collective endeavor empowers action recognition systems to unravel the complex fabric of action categories, fortify their accuracy, and enhance their resilience in the face of inherent intra-class and inter-class variations.

5.1.1.2 Viewpoint variations

The challenge of viewpoint variations in the domain of action recognition represents a formidable hurdle, characterized by the inherent inconsistencies in action recognition when confronted with differing angles, distances, or orientations of observation. This phenomenon is predicated upon the intricate three-dimensional nature of human actions, coupled with the inherent diversification in the execution of actions by individuals, thereby compounded by the diverse array of potential viewpoints. To address this intricacy, a spectrum of innovative solutions has been conceived.

Foremost among these is the employment of multi-view action recognition systems. These systems harness the capabilities of multiple cameras or viewpoints to capture actions from a diverse array of angles, imparting a more comprehensive comprehension of the action. Through the amalgamation of information derived from multiple viewpoints, these systems manifest an elevated capacity for bolstering the accuracy of recognition, especially in situations where a solitary viewpoint may inadequately encapsulate the entirety of the action's dynamics.

Data augmentation techniques proffer another invaluable stratagem. By artificially generating a spectrum of action variations from distinct viewpoints, these techniques enrich the training dataset, thereby endowing the model with the aptitude to assimilate resilient features that remain invariant amidst changes in viewpoint. Data augmentation methodologies encompass spatial transformations, encompassing operations like rotation and scaling, alongside manipulations simulating changes in observational perspective, thus simulating the appearance of actions from a diverse ensemble of vantage points.

The realm of deep learning has emerged as a pivotal instrument in ameliorating the challenges posed by viewpoint variations. Various deep learning models such as CNN have evinced their potential to acquire hierarchical and spatiotemporal features exhibiting diminished susceptibility to alterations in viewpoint. Moreover, the integration of attention mechanisms and transformer architectures serves to amplify the discernment of intricate and nuanced actions, independent of the perspective of observation.

In summation, the resolution of the viewpoint variations challenge within the ambit of action recognition necessitates a confluence of methodologies, including the embrace of multi-view systems, the application of data augmentation techniques, and the adoption of sophisticated deep learning models. Collectively, these stratagems contribute to the cultivation of action recognition systems characterized by augmented resilience

and precision, accommodating the diverse panorama of observational perspectives encountered in practical contexts.

5.1.1.3 Different coordinate system representation

The challenge of disparate coordinate system representations in the domain of action recognition stands as a prominent concern, emblematic of the intricate nature inherent in the interpretation of human actions across a spectrum of perspectives and environmental conditions. This challenge is rooted in the inherent variability characterizing how actions are represented and encoded within distinct coordinate systems, spanning the gamut from Cartesian, polar, or alternative mathematical frameworks. Such variability may emanate from disparities in skeletal joint representations, the employment of various motion capture technologies, or divergent sensor configurations. The consequence of this divergence is the potential for actions that are intrinsically identical to be construed differently within disparate coordinate systems, thereby engendering the possibility of complications in the consistent and accurate recognition of actions.

To surmount this challenge, a suite of methodologies has surfaced. Foremost among these is the imperative development of adaptable feature representations. Notably, deep learning paradigms, most prominently convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have exhibited the potential to glean features that remain impervious to transformations across coordinate systems. These models possess the innate capacity to autonomously adapt to the nuances of varying coordinate representations by discerning spatiotemporal patterns and distinctive features ingrained within the data.

The standardization of data and preprocessing techniques emerges as a pivotal directive. The transformation of assorted coordinate system representations into a unified and standardized framework serves to augment the harmonization of action data. This standardization process may entail the alignment of joint positions, the

normalization of data, or the application of geometric transformations to establish a uniform reference frame as depicted in 5.2 (D, E, F)

Furthermore, domain adaptation techniques have arisen as a potent mechanism for the acclimatization of models to an assortment of coordinate systems. These methodologies empower models to transfer acquired knowledge from one coordinate system to another, thereby enhancing their capacity for generalization and proficient recognition.

To conclude, the challenge of divergent coordinate system representation in action recognition necessitates the cultivation of adaptable feature representations, the pursuit of standardization, and the deployment of domain adaptation techniques. These multifaceted strategies collectively contribute to the fruition of action recognition systems characterized by the robust capacity to navigate varying coordinate system representations and confer accurate recognition across a spectrum of perspectives and environmental conditions.

5.1.1.4 Occlusions

The challenge of occlusions in action recognition is a pervasive concern, manifesting as the obstruction or partial concealment of crucial visual information during the execution of actions. Occlusions can occur when objects or body parts involved in an action are hidden from view, leading to the loss of vital cues required for accurate recognition. This challenge is exacerbated in real-world scenarios, where occlusions are commonplace due to objects in the environment, interactions with other individuals, or self-occlusions, where body parts obstruct each other. To address this challenge, researchers have developed innovative solutions.

One key approach is the incorporation of spatial and temporal context. By considering the broader context of actions, models can infer missing information during occlusions. Temporal modeling techniques, such as long short-term memory (LSTM) networks, enable the understanding of action dynamics, aiding in the prediction of

occluded body parts or objects based on their previous and subsequent states.

Multi-modal data fusion is another promising strategy. By combining information from multiple sensors or modalities, such as RGB and depth cameras or wearable sensors, models can mitigate the impact of occlusions. These modalities offer complementary information that can fill in the gaps created by occlusions, enhancing action recognition accuracy.

Additionally, generative models like variational autoencoders (VAEs) have been employed to synthesize plausible data during occlusions. By learning the underlying distribution of unoccluded actions, VAEs can generate reconstructions of actions even when some parts are occluded.

In summary, addressing the challenge of occlusions in action recognition demands a multi-faceted approach. This includes the incorporation of context, multi-modal data fusion, and generative models. These solutions collectively empower action recognition systems to contend with occlusions and maintain accuracy in the face of obscured visual information, fostering robust performance in real-world scenarios.

5.1.1.5 Sptio-Temporal variations

The challenge of spatio-temporal variations in action recognition is a fundamental concern, encapsulating the complexity associated with the dynamic interplay of spatial and temporal aspects within human actions. Spatio-temporal variations manifest as differences in the spatial configuration and timing of actions, even when they belong to the same category. These variations can be influenced by diverse factors such as the execution speed, scale, viewpoint, and individual style, introducing formidable challenges for accurate recognition. To address this challenge, researchers have devised several innovative solutions.

One critical approach involves the use of 3D convolutional neural networks (CNNs) and spatiotemporal models. These models are specifically designed to capture the intricate interdependencies between spatial and temporal elements within actions.

They excel at learning spatiotemporal features and patterns, enabling more robust recognition even in the presence of variations.

Another strategy is the integration of optical flow information. Optical flow provides insights into motion patterns and the displacement of pixels over time. By fusing RGB and optical flow data, models can better understand the motion dynamics, compensating for spatio-temporal variations.

Attention mechanisms have emerged as a valuable tool in addressing spatio-temporal variations. These mechanisms allow models to focus on relevant spatiotemporal regions, enhancing the ability to capture discriminative features and patterns.

Furthermore, graph convolutional networks (GCNs) have gained prominence in handling spatio-temporal data. GCNs enable the modeling of spatiotemporal relationships within actions, offering a more comprehensive understanding of their dynamics.

In sum, addressing the challenge of spatio-temporal variations in action recognition necessitates the application of 3D CNNs, the integration of optical flow information, attention mechanisms, and graph convolutional networks. These solutions collectively empower action recognition systems to contend with the intricacies of spatio-temporal variations, fostering more accurate and robust recognition in the face of dynamic action attributes.

5.1.2 Our Contribution

Our objective in this work is to develop a modular, end-to-end enabled Convolutional Neural Network (CNN) for the action recognition task to achieve comparable results to SOTA.

- Although most of the SOTA contributions in action recognition task is based on graph neural network, We show that embracing various training techniques can lead to comparable results using CNN.

- We demonstrate that applying diverse augmentation techniques leads to better generalization and robustness of the model.

- We show that using margin-based cosine loss function over the conventional cross-entropy loss leads to performance gain.

5.2 Problem Formulation

Given a skeleton sequence S, where $s \in (1, ..., S)$, and the j-th skeleton joint on the t-th frame is denoted as $s_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}]$, with $j \in (1, ..., J)$ and $t \in (1, ..., T$, where J represents the number of joints in a frame, and T represents the total number of frames in sequence S.

Drawing inspiration from the approach by Du *et al.* [14], we embark on the transformation of raw skeleton data into a skeleton map image. The primary objective is to encapsulate the spatiotemporal information effectively while preserving the structural integrity, as illustrated in Figure 5.3. This transformation entails a meticulous process. Commencing with an RGB image characterized by dimensions $[H, W, C]$, where H corresponds to the height, W signifies the width, and C encapsulates the conventional RGB image channels, we intricately map the action sequence S. This sequence is inherently represented in the dimension $T \times N \times 3$, where T signifies the number of frames and seamlessly aligns with the image's height. Simultaneously, N corresponds to the number of joints, aligning harmoniously with the image's width. The last dimension manifests as the 3D joint coordinate of the frame, which equivalently maps to the image's three channels.

In addressing a key aspect of the transformation, it is essential to consider potential variations in value ranges between raw skeleton data and image data. To ensure that the values of the mapped data are well-structured within the range of 0 to 255, a pivotal step is introduced, known as pixel normalization. This normalization process adheres to the following mathematical expression:

$$P_{t,j} = \text{floor}\left(255 \times \frac{s_{t,j} - C_{\min}}{C_{\max} - C_{\min}}\right)$$

This equation systematically normalizes the pixel values ($P_{t,j}$) within the designated range while considering the minimum ($C_{\min}$) and maximum ($C_{\max}$) values observed in the transformation, further ensuring that the data remains within the prescribed limits.

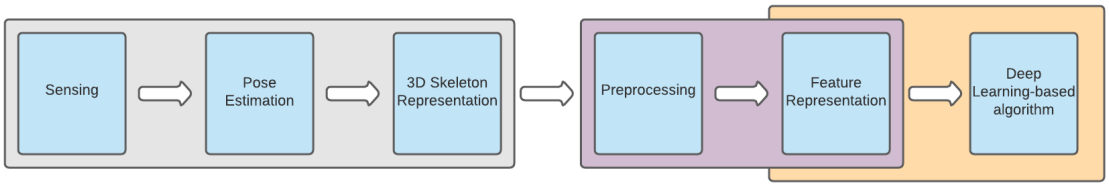

Figure 5.1: Abstracted Action Recognition Pipeline

5.2.1 Data Preprocessing

Skeleton data is conventionally represented within the camera coordinate system, leading to disparate representations when sequences are captured from different viewing angles, as illustrated in Fig. 5.2 (A, B, C). To mitigate the impact of these viewpoint variations, a common practice is to transform the skeleton data into a unified coordinate system, as evidenced in Fig. 5.2 (D, E, F) [227–230]. This transformation process typically involves a sequence of geometric translation, rotations, and normalization operations.

In the literature, diverse transformation strategies have been proposed. One approach, the frame-based technique, applies transformation operations individually to each frame within the sequence. However, this method may result in the loss of relative motions. For instance, implementing this process during a walking action would resemble walking on a treadmill. In contrast, the sequence-based strategy leverages the initial frame within the sequence as a reference frame. Subsequent transformations are performed relative to this reference frame, preserving more realistic human

skeleton motions.

These transformation strategies play a crucial role in addressing the challenge of view variation in skeleton-based action recognition.

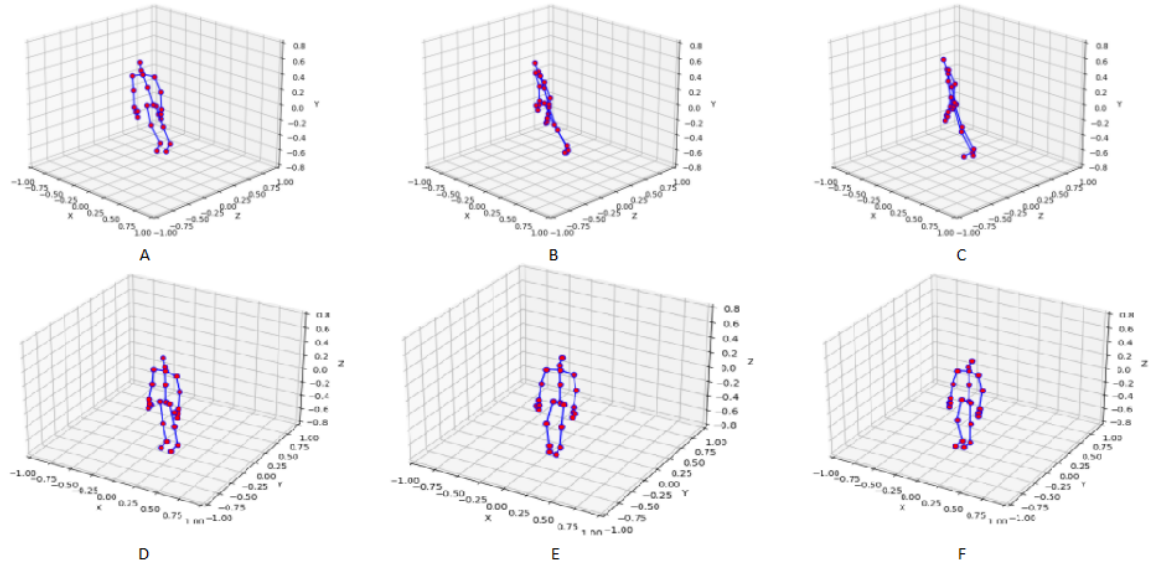

Figure 5.2: Action representation from NTU-D 60 dataset A) -45°skeleton visualization, B) 0 °skeleton visualization, C) 45°skeleton visualization. (D, E, F) are the transformed skeleton for the same skeletons in (A, B, C)

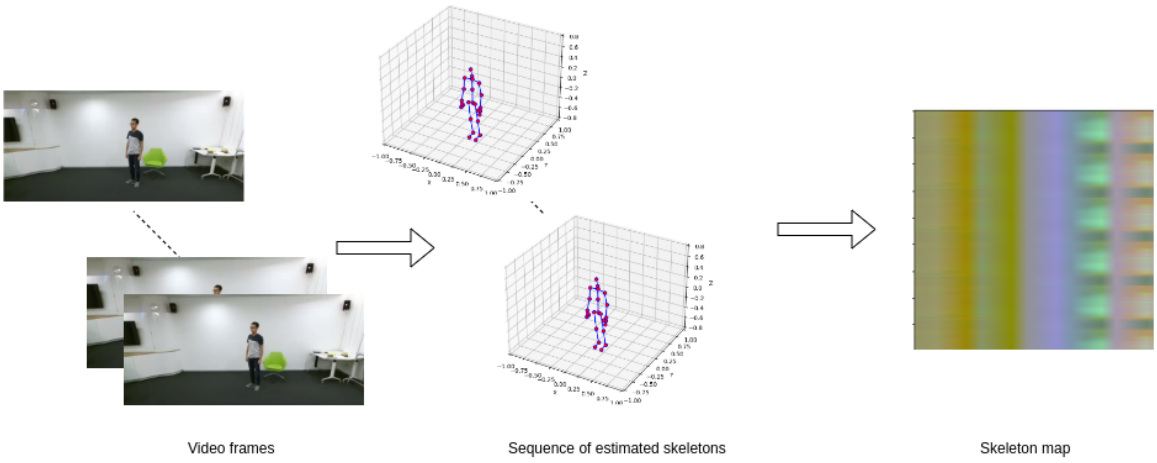

Figure 5.3: The pipeline of generating the skeleton map image

5.2.2 Data Augmentation

Data augmentation is a pivotal technique extensively employed to enhance the performance of machine learning algorithms, notably through an augmentation of the diversity inherent in training data. The core principle underlying data augmentation revolves around the generation of synthetic data samples from an existing training dataset. This is accomplished through the application of a diverse array of transformations, including but not limited to translation, rotation, scaling, and various deformations. Subsequently, these transformed data samples are seamlessly integrated into the training set, leading to an expansion of the dataset's volume and diversification. The underlying objective of data augmentation is to bolster the generalization capacity of machine learning algorithms and foster their resilience when exposed to variations within the input data. The practice of data augmentation finds widespread utilization across a gamut of deep-learning applications, spanning image classification, object detection, and natural language processing.

In the realm of skeleton-based data augmentation, this strategy becomes particularly pertinent when the input data is rooted in the realm of 3D pose information. Such a scenario is frequently encountered in tasks exemplified by 3D human pose estimation. The essence of skeleton-based data augmentation lies in the application of an assortment of transformations to the 3D joint positions characterizing a skeleton. This process generates an array of synthetic pose samples that contribute to diversifying the training dataset. Distinguishing itself from its counterpart, image-based data augmentation, which pertains to scenarios where input data is constituted by images, such as image classification and object detection.

This research endeavor involved an extensive and systematic exploration of a wide spectrum of image-based and skeleton-based augmentation techniques. The primary impetus for exploring this dual approach is rooted in the practice of encoding skeleton data into an image-based representation. This transformative step served as

the inspiration for leveraging augmentation methodologies from both these domains. Drawing inspiration from the concepts underpinning RandAugmentation [231], which entail the injection of the training pipeline with randomly predefined augmentations, this methodology incorporates a set of augmentation parameters that determine the count and magnitude of augmentations to be applied. The magnitude of augmentation plays a critical role in determining the extent of data variation introduced during training. While a moderate augmentation magnitude can provide models with a more robust understanding of the data, excessively high magnitudes may lead to overfitting or introduce unrealistic variations that hinder model generalization. Therefore, carefully selecting the magnitude of augmentation is crucial. By balancing the introduction of diversity with the risk of overfitting, practitioners can harness the full potential of RandAugmentation to enhance model adaptability and effectiveness in various machine learning tasks.

The Flipping sequence augmentation, as introduced by Rao *et al.* [232], constitutes a valuable technique for generating novel synthetic pose sequences with enhanced diversity. This augmentation methodology operates by performing a horizontal reflection of the input pose sequences. In the context of skeleton data, this involves mirroring the 3D joint positions across a horizontal plane. In the case of image data, it extends to dual reflections across both horizontal and vertical planes. The outcome of this operation is the creation of a new pose sequence wherein the same actions are faithfully replicated but performed in the opposite direction.

Mathematically, the flipping sequence augmentation can be succinctly formulated as follows. Given an initial pose sequence denoted as S, where S_t represents the pose data at time step t, the horizontal flipping operation can be expressed as:

$$S_t^{'} = F_h(S_t)$$

Where $S_t^{'}$ signifies the horizontally flipped pose data, and $F_h(\cdot)$ represents the

horizontal flipping function.

In the case of image data, the dual reflection operation across both horizontal and vertical planes can be mathematically represented as:

$$S_t^{'} = F_h(F_v(S_t))$$

Here, $F_v(\cdot)$ and $F_h(\cdot)$ represent vertical and horizontal flipping functions, respectively. This operation effectively mirrors the image data both horizontally and vertically, generating the horizontally and vertically flipped image, denoted as $S_t^{'}$.

Geometric Rotation augmentation, as introduced by Cubuk *et al.* [231], stands as a valuable technique for generating novel synthetic samples with augmented diversity in datasets. This augmentation approach is characterized by the application of random rotations to the input data, thereby enriching the dataset with variations resulting from different orientations. The application of these rotations can involve the rotation of images or 3D skeleton data, utilizing specified rotation angles and centers of rotation.

Mathematically, the Geometric Rotation augmentation can be defined as follows. Given an initial data sample represented as D, where D_i indicates the data at sample i, the application of a random rotation can be expressed as:

$$D_i^{'} = R(D_i, \theta_i, C_i)$$

Here, $D_i^{'}$ signifies the data sample after the rotation, $R(\cdot)$ represents the rotation function, θ_i denotes the rotation angle applied to the i-th sample, and C_i represents the center of rotation for that specific sample.

The Geometric Rotation augmentation technique offers significant utility in enhancing the diversity and robustness of datasets in the context of machine learning tasks. It enables models to be more resilient to variations in orientation, contributing

to improved generalization and performance.

Cutout augmentation, originally introduced by DeVries *et al.* [233], is a regularization technique widely employed in the realm of deep learning. This technique holds the capability to enhance the performance and robustness of deep neural networks, particularly in the domain of image processing tasks. The fundamental principle underlying Cutout augmentation involves the random masking or zeroing-out of a portion of an image during the training process.

Mathematically, Cutout augmentation can be expressed as follows. Given an input image represented as I, the application of Cutout involves selectively masking out a region, resulting in a modified image $I^{'}$. This operation can be represented as:

$$I^{'}(x, y, c) = \begin{cases} I(x, y, c) & \text{if } (x, y) \text{ is outside the cutout region} \\ 0 & \text{if } (x, y) \text{ is within the cutout region} \end{cases}$$

Here, $I^{'}(x, y, c)$ denotes the pixel value at position (x, y) in channel c of the modified image, and $I(x, y, c)$ represents the pixel value in the original image.

The utilization of Cutout augmentation during training encourages the model to learn to disregard the masked-out regions and focus on the unmasked portions. This, in turn, enhances the model's ability to generalize and perform effectively on new and unseen data. Cutout is particularly valuable in image classification tasks, where it aids the model in recognizing objects despite variations in their appearance.

The simplicity and effectiveness of Cutout augmentation render it a widely adopted technique in the realm of deep learning. Furthermore, its straightforward implementation makes it a convenient choice in various deep learning frameworks.

Zoom augmentation, as proposed by Cubuk *et al.* [231], constitutes a fundamental transformation in the realm of data augmentation techniques. This augmentation method serves the purpose of rescaling input images or 3D objects using a specified zoom factor. The zoom factor, denoted as Z, precisely determines the magnitude of

zooming to be applied to the input data.

Mathematically, Zoom augmentation can be expressed as follows. Given an input image or 3D object, each point in the data (representing coordinates in 2D or 3D space) undergoes a scaling operation, as shown below:

$$X^{'}(x, y, z) = X(x/Z, y/Z, z/Z)$$

Here, $X^{'}(x, y, z)$ represents the position of a point in the scaled or zoomed data, and $X(x/Z, y/Z, z/Z)$ signifies the original point's position divided by the zoom factor Z.

Notably, Zoom augmentation is a versatile technique suitable for both image and skeleton data. It enables the augmentation of data by modifying the scale, which can be instrumental in enhancing the model's ability to generalize and perform well on diverse data.

Similarly, Shear augmentation, another transformation detailed by Cubuk *et al.* [231], is accomplished through linear transformations that map the x-axis or y-axis coordinates of the input data to new positions. Shear transformations can be expressed as shear matrices that determine the degree of shearing in the data.

Mathematically, Shear augmentation can be represented as follows, where the shear matrix S is applied to the data:

$$X^{'} = S \cdot X$$

Here, $X^{'}$ represents the transformed data, S is the shear matrix, and X is the original data.

Lastly, Translate augmentation, detailed by Cubuk *et al.* [231], involves shifting the x-axis, y-axis, or both coordinates of the input data by a specified amount. The translation vectors Δx and Δy control the extent of the translation along the x-axis and y-axis, respectively.

Mathematically, Translate augmentation can be represented as follows, where Δx and Δy determine the shifts:

$$X^{'}(x, y) = X(x + \Delta x, y + \Delta y)$$

Here, $X^{'}(x, y)$ signifies the position of a point in the translated data, and $X(x + \Delta x, y + \Delta y)$ denotes the original point's position shifted by the amounts Δx and Δy.

These augmentation techniques, encompassing Zoom, Shear, and Translate, are valuable tools for modifying data to improve the performance and robustness of machine learning models.

The inclusion of various forms of noise during training serves as an effective strategy to enhance the generalizability and robustness of machine learning models, as described in the survey paper by Shorten *et al.* [234].

Salt and pepper noise, as presented by Liu *et al.* [235], is a distinctive noise augmentation method that simulates the effects of corrupted or missing pixels in the input data. Mathematically, Salt and pepper noise can be expressed as follows:

$$I'(x, y) = \begin{cases} 0 & \text{with probability } p/2 \\ 255 & \text{with probability } p/2 \\ I(x, y) & \text{with probability } 1 - p \end{cases}$$

In the equation, $I'(x, y)$ denotes the pixel at coordinates (x, y) in the noisy image, $I(x, y)$ is the original pixel intensity, and p represents the probability of replacing a pixel with either the minimum (0) or maximum (255) intensity value.

Localvars noise augmentation, as detailed by Shorten *et al.* [234], involves generating random noise samples from a specified distribution. Mathematically, this can be described as:

$$I'(x, y) = I(x, y) + N(x, y)$$

Here, $I'(x, y)$ is the noisy pixel at coordinates (x, y), $I(x, y)$ is the original pixel value, and $N(x, y)$ represents the generated noise at the same coordinates.

Speckle augmentation, as explained by Shorten *et al.* [234], introduces random noise samples following a Poisson distribution with a specified mean. The mathematical formulation is as follows:

$$I'(x, y) = I(x, y) + N(x, y)$$

In this equation, $I'(x, y)$ is the noisy pixel at coordinates (x, y), $I(x, y)$ is the original pixel value, and $N(x, y)$ is the generated noise from a Poisson distribution.

Lastly, Gaussian noise augmentation, another noise-based approach by Shorten *et al.* [234], generates random noise samples from a Gaussian distribution with a specified mean (μ) and standard deviation (σ). The mathematical expression is as follows:

$$I'(x, y) = I(x, y) + \mathcal{N}(\mu, \sigma)$$

Here, $I'(x, y)$ is the noisy pixel at coordinates (x, y), $I(x, y)$ is the original pixel value, and $\mathcal{N}(\mu, \sigma)$ represents random noise sampled from a Gaussian distribution with mean μ and standard deviation σ.

The utilization of these noise-based augmentation techniques, encompassing Salt and pepper, Localvars, Speckle, and Gaussian noise, facilitates the introduction of controlled variations in the training data, leading to models that are more resilient to noisy and real-world input data.

Bone shuffling augmentation is primarily applied to skeleton data, aiming to introduce diversity in body configurations during training. It achieves this by randomly permuting the 3D joint positions of the skeleton. This operation results in a new pose

sequence that retains the same actions but presents the subject in a different body configuration. Mathematically, bone shuffling can be expressed as follows:

Given a skeleton sequence S with T frames and J joints, and $s_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}]$ representing the 3D joint positions for the j-th joint at the t-th frame, bone shuffling involves generating a new sequence S' by permuting the joint indices. The mathematical formulation for bone shuffling can be described as:

$$S'_t = S_{\pi(t)}, \text{ where } \pi(t) \text{ is a random permutation of } \{1, 2, \ldots, T\}$$

Here, S'_t denotes the frame at time t in the shuffled sequence, and $\pi(t)$ represents the permutation function that shuffles the frames.

Another bone-related augmentation technique is bone masking, which is applied to skeleton data to simulate occlusions or partial information scenarios. Bone masking selectively masks a subset of the bones by setting the 3D joint positions to zero for those specific bones. The mathematical formulation for bone masking is as follows:

Given a skeleton sequence S with T frames and J joints, bone masking involves generating a new sequence S' by zeroing out the joint positions for a randomly selected subset of the bones at each frame. This can be mathematically represented as:

$$S'_t = \begin{cases} [s_{t,j} = [0, 0, 0]] & \text{if } j \text{ belongs to the masked subset} \\ s_{t,j} & \text{otherwise} \end{cases}$$

This operation selectively zeroes out the joint positions for the chosen bones, creating a sequence that retains partial information. Moreover, similar masking can be applied at the frame level, rather than the bone level, depending on the specific augmentation requirements.

The use of bone shuffling and bone masking techniques in skeleton data augmentation contributes to the robustness of models in action recognition tasks by enhancing

their adaptability to varying body configurations and occlusion scenarios.

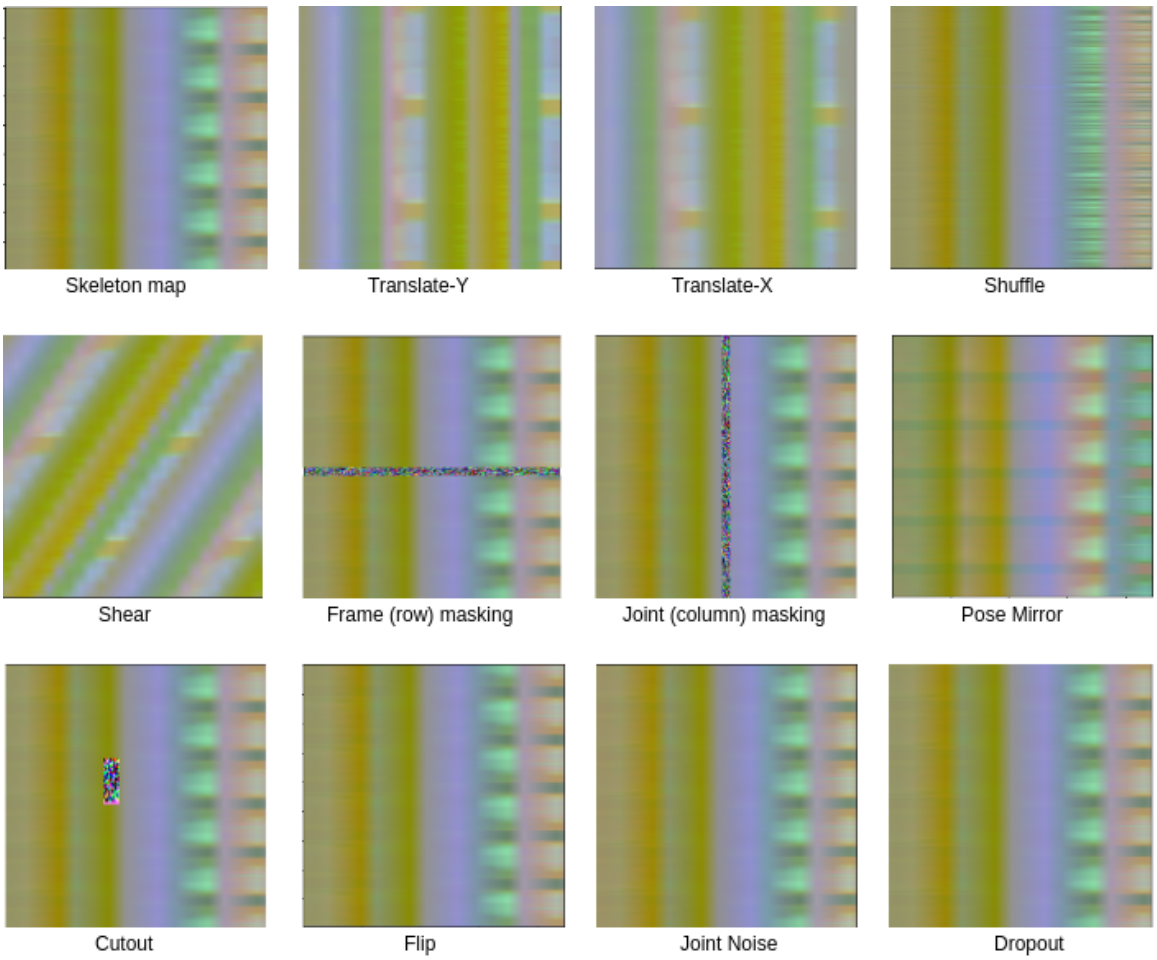

Figure 5.4: Various augmentation implementation

5.2.3 Loss function

In the realm of deep learning, the efficacy of a model's predictions is gauged by its ability to minimize a designated loss function. This function serves as the compass for model training, guiding the search for the optimal set of parameters that yield minimal loss and, consequently, enhance the model's predictive capabilities on unseen data. When dealing with classification tasks, particularly in action recognition, the widely adopted loss function is cross-entropy. It measures the dissimilarity between predicted class probabilities and the true class labels, effectively quantifying the information divergence between them.

Mathematically, cross-entropy loss for binary classification is defined as:

$$\text{Cross-Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right)$$

Where: - N is the number of samples in the dataset. - y_i is the true binary label of the i-th sample (0 or 1). - p_i is the predicted probability that the i-th sample belongs to class 1.

While effective in many scenarios, cross-entropy exhibits certain limitations. Its sensitivity to the relative magnitude of predicted and true values complicates its utility in specific cases.

A noteworthy observation while employing cross-entropy is the convergence of distances between samples from different classes. Consequently, models trained with cross-entropy may be prone to erroneous predictions. An intuitive and alternative approach to mitigating this challenge is through the application of metric learning techniques, as expounded upon by Suarez *et al.* [236]. Unlike traditional machine learning, which endeavors to learn a function mapping inputs to outputs, metric learning focuses on acquiring a metric that can effectively measure the distances between data points. This learned distance metric, in turn, plays a pivotal role in the prediction process. By adopting such techniques, one can maximize the inter-class distance while simultaneously minimizing the intra-class distance, thereby improving classification accuracy and the overall performance of the model.

The Additive Angular Margin Loss (AAML) [237] loss function is widely used in deep learning for face recognition. It is designed to learn a discriminative feature representation for face images by maximizing the angular margin between the features of the same identity and minimizing the angular margin between the features of different identities. The AAML loss function is based on the idea of learning a weight vector and a feature vector for each identity, such that the angle between the weight and feature vectors is maximized for the same identity and minimized for different

identities. The loss function encourages the weight and feature vectors to lie on the surface of a hypersphere. The large margin between the vectors of the same and different identities helps improve the discriminative power of the learned feature representation. The AAML loss function has been shown to outperform other loss functions for face recognition on several benchmarks. Therefore, we observed a gain of 1.5% in the classifier accuracy.

Mathematical Formulation

The AAML loss function is formulated as follows:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s\cdot\cos(\theta_{y_i,i}+m)}}{e^{s\cdot\cos(\theta_{y_i,i}+m)}+\sum_{j=1,j\neq y_i}^{N}e^{s\cdot\cos(\theta_{j,i})}}\right) \tag{5.1}$$

Where: - N: Number of training samples - i: Sample index - s: Scaling parameter - m: Angular margin - y_i: Ground truth class label of the i-th sample - θ: Feature vectors - Cosine similarity is used to measure the angle between feature vectors.

In general, the Additive Angular Margin Loss (AAML) is a powerful tool in deep learning, primarily applied in face recognition tasks. AAML excels in learning a discriminative feature representation for face images by maximizing the angular margin between features of the same identity and minimizing it between different identities. Its capabilities encompass various aspects, including discriminative feature learning, robustness to noise, improved generalization, and scalability to large-scale identification tasks. AAML achieves discriminative feature learning by ensuring that features from the same class are brought close in angular similarity, promoting effective clustering. This loss function demonstrates remarkable robustness, making it highly reliable even in the presence of noisy or mislabeled data. AAML enhances generalization by creating more distinct class boundaries, helping models perform effectively on unseen data. Its applicability extends to large-scale identification tasks, proving its effectiveness across numerous identities. In essence, AAML serves as a pivotal component for

enhancing the performance of deep learning models in the context of face recognition.

In conclusion, Additive Angular Margin Loss is effectively encourages weight and feature vectors to reside on the surface of a hypersphere. By introducing a substantial margin between the vectors corresponding to the same and distinct identities, AAML significantly enhances the discriminative power of the acquired feature representation. Empirical results have highlighted the superiority of the AAML loss function over other alternatives in the context of face recognition, leading to a noteworthy classifier accuracy improvement of 1.5

5.2.4 Deep-Learnig Optimizer and Schedulers

5.2.4.1 Training Optimizers

Within the realm of deep learning, optimizers serve as the linchpin for model training, playing a pivotal role in the pursuit of model convergence and performance optimization. These algorithms embark on a quest to discover the most suitable configuration of a model's parameters, one that effectively minimizes the loss function associated with a given task. In essence, optimizers are the guiding hands that mold a model to best match the provided data, with their efficacy profoundly influencing the success of the entire training process.

The journey of an optimizer is characterized by a meticulous interplay between mathematics and data. It meticulously fine-tunes the model's parameters in an iterative fashion, making incremental adjustments at each step. This gradual refinement is propelled by the optimizer's grasp of the intricacies of the loss landscape, discerning how changes in parameters influence the loss. Through these iterations, the optimizer endeavors to reach a state where the model offers robust and accurate predictions on unseen data.

While there exists a diverse array of optimization algorithms, the choice of a specific optimizer hinges on the nuances of the given task. Stochastic gradient descent (SGD), a foundational optimization technique, optimizes parameters using gradients

computed on small, random subsets of the data. Adam [238] is an adaptive optimization algorithm that goes a step further by employing dynamic learning rates for each parameter. Meanwhile, RMSprop [239] adapts learning rates differently for each parameter and tends to work well in scenarios with sparse gradients.

Selecting the appropriate optimizer, tailored to the task at hand, is not a mere implementation detail; it's a crucial determinant of a model's success. Different optimizers excel in various contexts, and their impact on model convergence, training efficiency, and generalization performance should be carefully considered when designing and training deep learning models.

MadGrad [240] is a variant of the commonly adopted Adam [238] optimizer used to train deep learning models. MadGrad introduces multiplicative noise to the gradients during training, which enhances the generalization performance of deep learning models. This noise injection helps models escape local minima and find better solutions, especially in complex, high-dimensional datasets. One of the significant advantages of MadGrad is its superiority over the widely used Adam optimizer in various tasks, including image classification and natural language processing. This feature makes it an attractive choice for those seeking improved model performance. MadGrad's dynamic noise scaling is another advantage. It allows for noise reduction over time, resulting in more adaptive and effective training. This adaptability ensures a balanced exploration of the optimization landscape, leading to better convergence and higher accuracy. It has been shown that MadGrad outperforms the Adam optimizer on several tasks, including image classification and natural language processing. Therefore, in this contribution, we validated the optimizer's effectiveness, leading to an overall 1.1% accuracy gain.

However, MadGrad is not without its disadvantages. It introduces an additional layer of complexity due to the noise injection mechanism. While this complexity can be advantageous for experienced deep learning practitioners, it may pose a challenge

for beginners. Understanding the optimizer's behavior and parameters is essential for effective usage. Additionally, the effectiveness of MadGrad may not be universal. Its benefits may vary depending on the dataset and model architecture. In simpler or smaller-scale tasks, the noise injection can be counterproductive.

MadGrad is best adopted in scenarios where it can leverage its strengths. Image classification and natural language processing tasks have shown notable improvements with MadGrad. Its noise injection mechanism is particularly valuable when dealing with complex, noisy, or large datasets. In such cases, traditional optimizers may struggle to find the global minima, while MadGrad's dynamic noise scaling can be advantageous. Additionally, researchers and practitioners exploring optimization strategies or seeking to push the boundaries of model performance will find MadGrad a valuable tool for experimentation. Its unique approach to gradient optimization can lead to new insights and discoveries in the field of deep learning.

In conclusion, MadGrad's approach of introducing multiplicative noise to gradients offers advantages such as improved generalization and outperforming traditional optimizers like Adam in specific tasks. However, its complexity and context-dependent effectiveness should be considered when choosing to adopt it for a given task. Researchers and practitioners should carefully evaluate its suitability for specific scenarios, especially in complex datasets and experiments aimed at advancing the state of the art in deep learning.

5.2.4.2 Learning rate schedulers

The primary purpose of employing a learning rate scheduler in the realm of deep learning is to dynamically regulate the learning rate throughout the training process. This learning rate, a crucial hyperparameter, dictates the step size by which the optimizer updates the model's parameters. An excessively high learning rate can lead the optimizer to take large and erratic steps, thereby risking convergence to suboptimal solutions. Conversely, an excessively low learning rate can result in small,

painstaking steps that significantly slow down the training procedure. Learning rate schedulers are instrumental in mitigating these challenges by autonomously adapting the learning rate during training.

Various strategies can be employed to schedule the learning rate effectively. One prevalent approach involves adopting a fixed schedule that maintains a constant learning rate over the entire training process. Alternatively, more adaptive schedules can be devised based on training progress or even on the model's performance. The integration of the Cosine Annealing scheduler with the reduced plateau strategy in this contribution has not only served to stabilize the learning process but also optimize the model's convergence. This combined approach enhances the training dynamics by ensuring that the learning rate aligns with the model's learning curve and adapts to the intricacies of the training data. Consequently, it aids in expediting the convergence to an optimal solution while mitigating the risks associated with excessively high or low learning rates.

The ReducedLR scheduler, as proposed by Al et al. [241], constitutes a valuable component in the domain of deep learning, specifically designed to fine-tune the learning rate throughout the training process. This scheduler's operation is rooted in a strategic approach, which involves decrementing the learning rate by a predefined factor in instances where the training loss exhibits stagnation over a specified number of consecutive epochs. By doing so, the ReducedLR scheduler aims to mitigate the peril of the training procedure becoming ensnared in a local minimum, thereby enhancing the overall generalization capabilities of the model under consideration.

Notably, the ReducedLR scheduler is seamlessly incorporated into deep learning frameworks as a callback function. This user-friendly feature streamlines its integration into the training loop, simplifying its adoption within various deep learning tasks. Furthermore, it provides users with the flexibility to configure the precise factor by which the learning rate should be reduced, as well as the number of epochs to endure

before triggering this reduction. Such configurability empowers practitioners with a potent tool for exerting precise control over the training process, ultimately resulting in the amelioration of deep learning models' performance.

In essence, the ReducedLR scheduler emerges as an indispensable asset in the realm of deep learning, offering a sophisticated mechanism for learning rate adjustment. By diligently monitoring the training loss, it averts the potential quagmire of local minima while enhancing the model's capacity to generalize. Furthermore, the scheduler's seamless integration into deep learning workflows, coupled with its configurable parameters, renders it a versatile tool for optimizing the training dynamics and enhancing the performance of deep learning models.

The CosineAnnealing scheduler, introduced by Cazenave *et al.* [242], is an astute approach to learning rate scheduling in the realm of deep learning. This scheduler operates by orchestrating a dynamic modulation of the learning rate, following the graceful trajectory of a cosine curve. Commencing with an initially high learning rate, it then orchestrates a gradual descent, culminating in a reduced learning rate as the training progresses. This unique approach contributes to the optimization of the training process in several essential ways.

One of the paramount advantages of the CosineAnnealing scheduler lies in its ability to foster efficient convergence during training. The dynamic reduction of the learning rate prevents the occurrence of premature saturation, a scenario where the learning rate diminishes excessively, impeding the training process's efficacy. This aspect is particularly salient in the training of deep learning models, where a judicious learning rate schedule can make the crucial difference between convergence and stagnation. By offering a mechanism that carefully navigates the learning rate, the CosineAnnealing scheduler ensures that the model can make consistent and substantial progress in the optimization process.

The CosineAnnealing scheduler is a vital asset in the deep learning toolbox. It

leverages the elegance of cosine-based annealing to deliver a learning rate schedule that promotes the efficient convergence of deep learning models. This approach mitigates the challenges associated with premature saturation and maintains the training process's vigor, leading to models that are well-optimized and capable of tackling complex tasks with efficacy.

In summary, a learning rate scheduler plays a pivotal role in the training of deep learning models by dynamically adjusting the learning rate to optimize convergence and mitigate the challenges posed by both excessively high and low learning rates. The specific choice of scheduler can vary depending on the nature of the dataset and the model architecture. In the context of this contribution, the combination of the Cosine Annealing scheduler and the reduced plateau strategy has proven particularly effective in stabilizing the learning process and facilitating model convergence.

5.2.5 Regularization

Regularization stands as an indispensable facet within the realm of machine learning, addressing the critical challenge of overfitting. This predicament unfolds when a model excels on the training dataset but falters when tasked with previously unseen data. The root of this issue lies in the model's tendency to grasp patterns confined solely to the training dataset, failing to extrapolate its knowledge effectively. Regularization takes center stage in rectifying this conundrum by imparting a corrective influence upon the loss function during the model's training process.

In essence, regularization administers a judicious penalty to the model's loss function. This penalty serves as a directive, steering the model toward acquiring a more generalizable ensemble of parameters. As a consequence, the model's newfound adaptability equips it to exhibit superior performance when faced with novel, unencountered data, transcending the limitations of mere memorization. The spectrum of regularization techniques is extensive, encompassing an array of sophisticated strategies, each tailored to address specific facets of overfitting.

One notable contender in the arsenal of regularization techniques is label smoothing regularization, which introduces controlled uncertainty to the model's predictions, thereby discouraging it from making overconfident and overly precise estimations. Another influential method is dropout, which temporarily deactivates randomly selected neurons during training, enforcing robustness in the model's architecture. Meanwhile, batch normalization optimizes internal layer activations, mitigating the potential for overfitting. Lastly, early stopping intervenes when the model's performance on the validation data begins to deteriorate, ensuring that training halts at an optimal juncture.

It is worth emphasizing that these regularization techniques are not mutually exclusive; their combined deployment forms a formidable defense against overfitting, fostering machine learning models that generalize with remarkable efficacy across diverse datasets. The holistic orchestration of these strategies caters to the overarching goal of safeguarding model performance against the pitfalls of overfitting, thus engendering a more reliable and adaptable machine learning paradigm.

Label smoothing [243]represents a pivotal regularization strategy harnessed within the domain of deep learning, meticulously tailored to augment the classifier's prowess in generalization. This technique intricately curtails the classifier's proclivity toward unwavering confidence in its predictions, fostering a more adaptable and cautious demeanor when confronted with novel data. The landscape where label smoothing finds prominent application is image classification, a bastion where robust generalization carries paramount significance.

Traditionally, classification tasks are typified by one-hot encoded labels, exemplified by a vector such as [0, 0, 1, 0], which unambiguously designates an image's association with a particular class. The novelty of label smoothing manifests in the transformation of this crisp, one-hot encoded vector into a nuanced probabilistic representation. For instance, an image purportedly belonging to class 3 would have its

label vector metamorphose from [0.1, 0.1, 0.8, 0.1], signifying a substantial probability concentration on class 3 while simultaneously entertaining the prospect of affiliation with other classes.

This seemingly subtle alteration bears profound implications. It functions as a regularization tool, vigilantly shielding the classifier from the perilous shoals of overfitting, a scourge that ensnares models in an overly confident and narrow viewpoint derived from their training data. In essence, label smoothing advocates for a realm where one-hot encoded training labels refrain from endorsing zero values for indices outside their designated class, instilling a more cautious and open-minded approach in the classifier. Consequently, this technique operates as a critical instrument, fine-tuning the classifier's predictive abilities and enhancing its proficiency in navigating the intricate landscape of image classification.

Early stopping [244], a well-established regularization technique in the realm of deep learning, orchestrates a strategic intervention during the model's training process, artfully truncating the iterative march towards convergence before the model attains its zenith performance. This surgical halt is meticulously choreographed by the vigilant monitoring of the model's performance vis-Ã -vis a dedicated validation dataset. Early stopping commences its role with unwavering commitment, diligently scrutinizing the model's progress. The moment it discerns any signs of stagnation or the ominous descent into performance decrepitude, it unfailingly intervenes.

The underlying rationale behind this deliberate interruption is profound, primarily geared towards staving off the perils of overfitting, an insidious malaise that plagues models with an uncanny ability to excel on their training data yet fumble when confronted with previously unseen data. Early stopping capitalizes on the principle that by truncating the training process before the model's inexorable convergence to its absolute best fit for the training data, we, paradoxically, foster a more resilient and adaptable model.

This truncated, unperfected model, emerging from the chrysalis of early stopping, demonstrates an enhanced ability to generalize and perform capably on hitherto unencountered data, transcending the limitations that could have been imposed by unbridled convergence. The beauty of early stopping lies in its simplicity and efficacy, rendering it a readily deployable technique within the arsenal of most deep learning frameworks, poised to foster improved generalization and overall model performance.

Dropout [245], a well-established gem within the toolbox of deep learning, unfurls as a potent regularization technique designed to imbue neural networks with the prowess of enhanced generalization. Anchored in a brilliant concept, it orchestrates the deliberate, stochastic omission of a pre-specified fraction of neuron activations within the network during training. In effect, this stratagem bestows resilience and diversity to the network's feature representations, countering the perilous throes of overfitting that can shackle a model to the idiosyncrasies of its training data.

The art of dropout unfolds predominantly within the domain of fully-connected layers, where its salutary impact is most pronounced. Nevertheless, it's versatile enough to extend its beneficence to other architectural constituents of neural networks, whether convolutional layers or recurrent layers, thereby offering a panacea for overfitting across diverse model architectures. In practice, dropout is administered with a substantial dropout rate during the training phase, often hovering around 0.5, instigating an appreciable level of neuron omission. However, come the phase of inference, this rate plummets to a nominal value, approximately 0.1 or even vanishing altogether, thereby permitting the unfettered employment of all neurons.

This duality of operation, marked by restraint during training and full participation during inference, lies at the heart of dropout's efficacy. It imparts models with the ability to traverse the tightrope between learning intricate details from the training data while retaining the flexibility to gracefully accommodate novel inputs, thus fortifying their generalization performance.

Dropout is a regularization technique used in deep learning to improve the generalization performance of a neural network. It is based on the idea of randomly dropping out (i.e., set to zero) a specified proportion of the activations of the neurons in the network during training. This reduces the co-adaptation of the neurons and makes the network less sensitive to the specific weights of individual neurons. Training the network with dropout can learn more robust and diverse feature representations that are less overfitted to the training data. Dropout is commonly applied to the fully-connected layers of a neural network but can also be applied to other types of layers, such as convolutional layers and recurrent layers. It is typically used with a high dropout rate (e.g., 0.5) during training and a low dropout rate (e.g., 0.1) or no dropout during inference.

Batch normalization [246], a cornerstone of modern deep learning, stands as a transformative technique meticulously designed to foster neural networks' performance and reliability. Its core principle revolves around normalizing the inputs to each layer within the network, a seemingly innocuous act that harbors a wealth of benefits.

The orchestration of batch normalization unfurls as follows an ingenious layer is interjected strategically between the network's input and output. Its mission is to perform real-time normalization on the activations of the preceding layer. This normalization leverages the mean and variance computed from the present mini-batch of data, bestowing it with an adaptable nature. By enlisting these statistics, batch normalization mitigates the scourge of internal covariate shift a phenomenon that plagues neural networks as the distribution of inputs to each layer undergoes unsettling fluctuations during training.

The merits of this discipline are manifold. For starters, it expedites the convergence process, magnifying the speed at which the network gobbles up the wealth of knowledge embedded in the training data. What's more, it adds a vital layer of stability, staunching the tumultuous fluctuations that often afflict unnormalized

networks. Batch normalization renders the network impervious to the whims of initialization parameters, infusing it with a sense of self-sufficiency, eliminating the need for meticulous tuning.

As a confluence of these virtues, the neural network surges forward with newfound vigor, sipping from the elixir of stability, speed, and resilience provided by batch normalization. Its impact is so profound that it's seamlessly integrated into the most intricate of neural architectures, promising and delivering superior performance.

5.2.6 Experimental Settings and Results

In our rigorous experimental setup, we harnessed the formidable capabilities of deep convolutional neural networks, specifically employing the ResNet-50 architecture. This well-regarded model comes pretrained on the vast and diverse ImageNet dataset, endowing it with extensive feature representation. Our primary objective was to conduct a meticulous evaluation of the model's performance on the formidable NTU RGB+D dataset. This dataset intricately captures a rich spectrum of human actions, considering varying camera perspectives and viewpoints. Our research, in particular, directed its focus towards cross-view evaluation, a challenging paradigm that scrutinizes the model's efficacy in recognizing actions from divergent camera angles.

Inspired by the innovative work of Du et al. [14], who ingeniously represent skeleton sequences as images, we incorporated this approach into our pipeline. To ensure methodological rigor, we adhered to the 'Cross-views' protocol, one of the evaluation protocols recommended by Shahroudy *et al.* [150] and Liu *et al.* [152]. This evaluation paradigm mandates that samples from the same view cannot be used for both training and testing. For instance, Shahroudy et al. proposed a specific camera configuration for training and testing, such as using cameras 1 and 3 for training, while reserving camera 2 exclusively for testing.

Our efforts culminated in the comprehensive performance evaluation presented in

Table 1. The crux of our analysis revolved around cross-view action recognition accuracy. Through meticulous scrutiny, our method demonstrated competitive performance when juxtaposed against existing state-of-the-art approaches. This outcome underscores the profound efficacy of employing ResNet-50, pretrained on ImageNet, as the foundational backbone for our action recognition pipeline.

we unveil the impactful influence of various optimization techniques and learning rate scheduling strategies. Our pursuit led us to experiment with advanced methodologies, including the novel MadGrad optimizer, a suite of regularization techniques (comprising early stopping, label smoothing, batch normalization, and dropout), and the judicious utilization of reducedLR and Cosine Annealing schedulers. This endeavor was undertaken with the explicit goal of conducting a granular analysis of their impacts on action recognition. The results were indeed enlightening, revealing improved action recognition accuracy, augmented stability during training, and the pivotal role of these techniques in counteracting overfitting, ultimately leading to enhanced model generalization.

Table 2 exemplifies the depth and breadth of our experimental investigations. It provides a meticulous breakdown of the outcomes of our concerted efforts to enhance data augmentation. Our examination encompassed the effects of diverse augmentation techniques, scrutinizing their performance under both strong and weak augmentation magnitudes. Furthermore, we explored the differential impact of augmentations when applied to image data, skeleton data, or a combination of both modalities. These meticulous investigations unveiled the nuanced dynamics of data augmentation within the context of cross-view action recognition, ultimately providing valuable insights into the augmentation strategies instrumental in enhancing the model's adaptability to diverse viewpoints and conditions.

In summary, our extensive experiments, leveraging ResNet-50 pretrained on ImageNet, for action recognition on the NTU RGB+D dataset, with a specific emphasis

Table 5.1: Effectiveness (in accuracy (%)) of applying (regularlization, Madgrad Optimizer, multiple learning schedulers)

Architecture	Technique	NTU	
		CS	CV
RNN-based	S-trans+RNN	76.0	82.3
	S-trans+RNN (aug.)	77.0	85.0
	VA-RNN	79.4	87.6
	VA-RNN (aug.)	79.8	88.9
CNN-Based	S-trans+CNN	87.5	92.2
	S-trans+CNN (aug.)	87.9	93.5
	VA-CNN	88.2	93.8
	VA-CNN (aug.)	88.7	94.3
	Our + ArcFace	-	**95.0**

on cross-view evaluation, have yielded profound insights and promising results. The model showcased competitive performance when benchmarked against state-of-the-art methodologies. Our comprehensive analysis of optimization techniques, learning rate scheduling, and data augmentation strategies deepens our understanding of the model's prowess in the domain of human action recognition. This research contributes significantly to the burgeoning field of computer vision and action recognition, underlining the potential and value of our approach.

5.3 Conclusion

The rapid evolution of sensing devices has opened up abundant opportunities for generating datasets across diverse modalities. Among these, the skeleton-based modality has garnered significant attention due to its computational efficiency and the rich feature set it offers, encapsulating the intricate human skeletal structure. The process of estimating human poses from video data is imperative for extracting comprehensive 3D skeleton information from disparate modalities. This valuable resource forms the bedrock for a myriad of applications, with action recognition being a prime use case.

Our research endeavors have elucidated the compelling potential of convolutional neural networks (CNNs) when equipped with an array of advanced training tech-

Table 5.2: Various augmentations applied on the encoded skeleton image map

Augmentation	Image-based	Weak	Strong
Flipping	Both	93.70	93.62
Rotation	Both	94.21	94.1
Zoom	Image	93.60	93.40
Shear	Image	94.03	94.20
Translate-x	Image	93.71	94.03
Translate-y	Image	94.24	94.1
Cutout	Image	94.03	94.1
Salt and pepper noise	Image	93.52	-
Bone Shuffling	Pose	94.13	93.96
Bone Masking	Pose	93.80	93.8
Frame Masking	Both	93.8	93.75
Gaussian Noise	Image	93.5	-
Speckle Noise	Image	93.45	-
Localvars Noise	Image	93.5	-
Salt noise	Image	93.5	-
Pepper Noise	Image	93.5	-

niques. Notably, we've demonstrated their ability to yield state-of-the-art (SOTA) results in the realm of action recognition. This success is particularly noteworthy as it showcases that CNNs can rival the performance of graph neural networks, a hitherto dominant approach in this domain. In parallel, we've delved into the realm of data augmentation and its pivotal role in enhancing model generalization and robustness. Our findings affirm that judiciously applying augmentation techniques can significantly bolster the model's adaptability, enabling it to excel when confronted with previously unseen data or input variations and distortions.

Furthermore, our exploration extends into the optimization and loss function domains. The adoption of MadGrad as the optimizer, coupled with the strategic use of learning rate schedulers, has revealed noteworthy improvements in model accuracy. This optimization strategy brings forth higher precision and reliability, which are crucial for the model's performance. In tandem, we've pioneered the use of a margin-based cosine loss function, departing from the traditional cross-entropy loss. This groundbreaking approach has exhibited its prowess by enhancing the model's

predictive acumen, thereby translating into superior overall performance. Additionally, we underscore the significance of regularization techniques in guarding against overfitting, an ever-present challenge in machine learning. By mitigating overfitting, these techniques enhance the model's ability to excel on new, previously unseen data.

In sum, our research unfolds a spectrum of insights into action recognition methodologies. From CNNs to optimization techniques and data augmentation strategies, our work augments the understanding of what drives performance in this field, shedding light on how to advance the state-of-the-art.

5.4 Future Directions

For future work in the domain of human action recognition, several promising avenues emerge to enhance the robustness and efficacy of the existing framework. Firstly, the utilization of mixed dataset training stands out as a paramount strategy. By amalgamating diverse datasets, encompassing a wide spectrum of human actions and scenarios, the model's capacity to generalize across varied contexts is bolstered. This approach not only ensures a richer and more comprehensive training experience but also fortifies the model against potential biases inherent in singular datasets.

Moreover, recognizing the potential limitations of existing skeleton data accuracy, a crucial future endeavor involves integrating more precise and physically plausible 3D pose estimation techniques. Leveraging advanced methods that adhere closely to biomechanical principles and the laws of physics can significantly enhance the fidelity of the skeletal representations, consequently augmenting the model's capacity to discern subtle nuances in human actions.

Furthermore, the integration of kinematic analysis features into the action recognition task emerges as a frontier for deeper insights. By incorporating kinematic parameters such as joint angles, velocities, and accelerations, the model gains access to richer contextual information about the underlying biomechanics. This not only refines the discriminative capabilities of the deep learning model but also facilitates

a more nuanced understanding of the intricate dynamics inherent in human actions. Overall, these proposed future directions collectively contribute to advancing the accuracy, generalization, and interpretability of human action recognition systems.

CHAPTER 6: High-fidelity Radar Data Synthesis and Restoration

6.1 Introduction

Research in the field of human activity recognition (HAR) has witnessed substantial advancements in the past decade. Noteworthy applications of HAR encompass surveillance [17], smart home environments [18], video analytics [19], autopilot systems [20], and human-computer interaction [21]. The primary objective of HAR is to discern and comprehend a user's conduct, thereby enabling computational systems to proactively offer assistance to the user [22].

HAR can be broadly categorized into two principal domains: vision-based and sensor-based approaches [23]. Vision-based HAR leverages the advantages of optical sensors with high resolution and the continuous evolution of computer vision (CV) techniques, yielding remarkable outcomes in recent studies [247]. Despite the notable accomplishments of vision-based HAR, various challenges persist, including issues related to illumination, occlusion, and privacy breaches [24].

Radar-based human activity recognition (HAR), as one of the sensor-driven methodologies, has garnered significant attention in recent research endeavors [248]. Several compelling reasons underscore its increasing prominence.

First and foremost, radar technology exhibits remarkable robustness in adverse environmental conditions, making it suitable for deployment in challenging settings. Its ability to function effectively under varying light and weather conditions renders it particularly versatile.

Secondly, radar offers a unique advantage in safeguarding visual privacy. Rather than capturing explicit visual representations of individuals, it captures and modulates signals reflected from the target. These modulated signals carry a wealth of

time-varying data related to range and velocity, enabling precise analysis of human activities [249].

Additionally, radar technology possesses the remarkable capability to detect human presence even through solid barriers, such as walls. This distinctive feature expands the applicability of radar-based HAR to a broader range of scenarios and settings.

Furthermore, radar-based HAR systems eliminate the need for attaching any tags or sensors to the human body, enhancing user-friendliness and ease of adoption. Consequently, radar-based HAR has witnessed a growing trend in recent years, with increased recognition and utilization.

Data stands as the cornerstone of paramount importance within the domain of deep learning, encapsulating its pivotal role and efficacy across diverse academic and practical domains. Deep learning models, notably neural networks, heavily rely on extensive, meticulously curated, and diverse datasets enriched with precise annotations. These datasets form the bedrock upon which these models are nurtured, honed, and empowered to make intelligent, data-driven decisions. The centrality of data in deep learning remains a fundamental tenet, facilitating the development of sophisticated algorithms capable of tackling a wide spectrum of tasks, ranging from image and speech recognition to natural language understanding and autonomous decision-making.

Moreover, the quality and diversity of the data hold paramount significance, for they bear a direct impact on the model's ability to generalize and exhibit robust performance in real-world contexts. In essence, data serves as the essential catalyst propelling the potential of deep learning, propelling innovation and empowering the creation of intelligent systems with far-reaching applications. The ongoing pursuit of data acquisition, its judicious curation, and its systematic integration remain pivotal, not only for the field's academic advancement but also for its practical, real-world implementation across domains as varied as healthcare and autonomous systems, heralding a promising era of data-driven discovery and technological innovation.

The landscape of datasets for action recognition presents a striking disparity between video-based and radar-based approaches. On one hand, video-based datasets are abundant and diverse, offering a rich tapestry of human actions in various contexts, encompassing activities, viewpoints, and environmental conditions. These datasets have catalyzed significant progress in the field, enabling deep learning models to achieve impressive accuracy and robustness. On the other hand, radar-based human action recognition contends with a stark contrast, as the availability of dedicated radar datasets remains notably limited. Radar-based datasets are sparse and often constrained in their diversity, representing a fraction of the wide-ranging actions captured in video datasets. This scarcity poses a considerable challenge for researchers seeking to develop and evaluate radar-based recognition systems. Bridging this gap in dataset availability is essential to further advance radar-based action recognition and unlock its potential in a broader spectrum of applications, particularly in domains where radar's unique capabilities, such as privacy preservation and through-wall sensing, offer distinct advantages.

In this research we propose a pioneering approach to bridge the existing chasm between the abundant availability of video-based datasets and the scarcity of radar-based equivalents. By introducing a theoretically grounded framework for generating dense feature representations from skeletal data, the approach simultaneously upholds the consistency of temporal and spatial data representation while ensuring visual interpretability. It accomplishes this by emulating a virtual radar system and deriving spectrograms from skeletal data. These spectrograms maintain a steadfast representation of skeletal data, resilient to variations in frame rates and the number of skeletons present in a given scene.

One of the notable contributions of this methodology is its potential to transform any video-based dataset into a synthetic radar-based dataset. This transformation enhances the accessibility of radar-related data, thereby mitigating the limitations

imposed by the paucity of dedicated radar datasets. By doing so, this innovative approach not only expands the horizons of research in radar-based action recognition but also fosters the creation of diverse and comprehensive datasets. This transformational capability, harmonizing video-based and radar-based data, is a significant step forward in advancing the efficacy and robustness of radar-based action recognition systems. It is poised to empower researchers with a wealth of data resources, ultimately accelerating progress in this field and unlocking the full potential of radar technology for a broader spectrum of applications.

6.1.1 Our Contribution

In this research, we have made substantial contributions in several key areas that advance the field of radar-based human action recognition:

- Virtual Radar synthesis technology: We have introduced a groundbreaking virtual radar synthesis technology, a novel computational framework designed to generate high-fidelity synthetic radar data. This technology leverages video recordings of human actions and electrocardiogram (ECG) waveforms of cardiac activities to simulate realistic radar spectrograms. By bridging the gap between video-based and radar-based datasets, this innovation not only enriches the availability of radar-related data but also provides a means to explore the synergy between these two modalities, unlocking new avenues for research and development.

- NTU-60 Radar-Based Dataset: To facilitate research in radar-based human action recognition, we have created the NTU-60 radar-based dataset. This dataset comprises a diverse range of human actions and activities, captured through the virtual radar synthesis technology. By providing this dataset to the research community, we aim to catalyze further investigations and experimentation in the field, fostering a deeper understanding of radar-based action recognition.

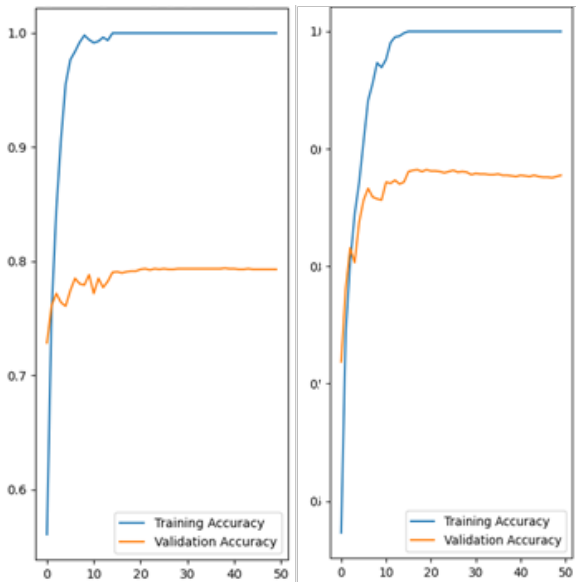

Figure 6.1: Left: Training and validation curve based on nine medical classes, Right: Training and validation curve after excluding three confusing actions

- Evaluation and Medical Action Baseline: We have meticulously evaluated the effectiveness of our proposed approach by establishing a baseline performance standard. This baseline was specifically tested on the "medical actiona" a subset of the NTU-60 radar-based dataset, which encompasses a focused set of actions related to medical and healthcare scenarios. By conducting this targeted evaluation, we have provided a robust performance reference for radar-based recognition of medical actions. This focused evaluation not only aids in benchmarking the capabilities of our technology but also serves as a valuable resource for researchers in the medical field, offering insights into the recognition of critical healthcare actions.

6.2 Radar Foundations

Radar systems emit an electromagnetic (EM) signal towards an object and capture the signal reflected back from the object. The time delay in the return signal allows the radar to determine the distance to the object, while any motion of the object causes a change in the frequency of the received signal, a phenomenon known as the Doppler

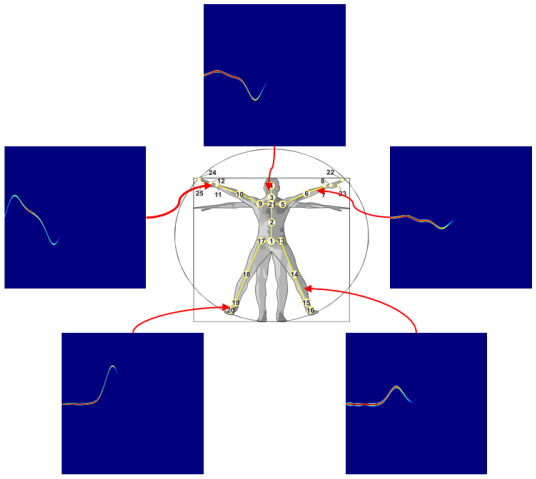

Figure 6.2: Per bone spectrogram generation

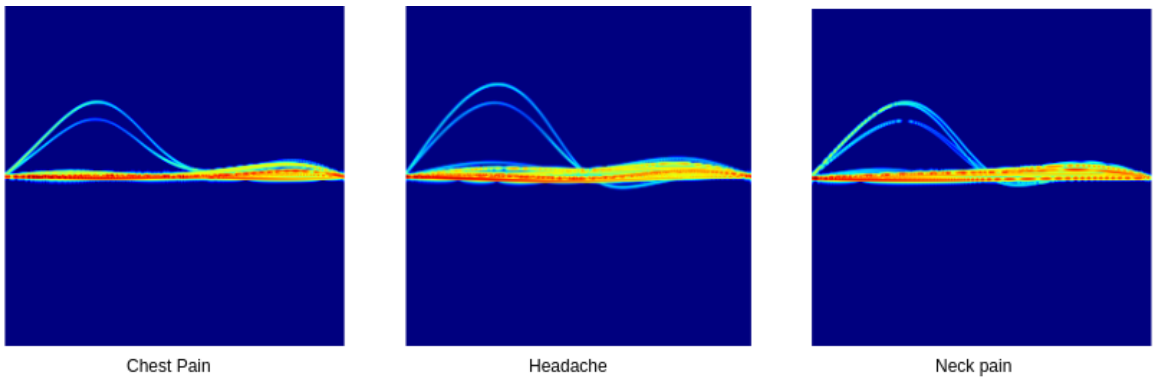

Figure 6.3: Similar spectrogram representation for closely similar actions

effect [250]. The magnitude of this Doppler frequency shift corresponds to the radial velocity of the object, which signifies the velocity component along the line of sight (LOS). In instances where the object or any of its structural components exhibit oscillatory motion in addition to its overall movement, these oscillations introduce further frequency modulation to the reflected signal. Consequently, this generates sidebands surrounding the Doppler-shifted frequency of the transmitted signal due to the object's bulk motion. This supplementary Doppler modulation is referred to as the micro-Doppler effect [251].

The determination of the Doppler frequency shift generally involves a frequency domain analysis achieved through the application of the Fourier transform on the received signal [249]. Within the resulting Fourier spectrum, the most prominent component signifies the Doppler frequency shift attributable to the radial velocity of the object's motion. Moreover, the spread or width of Doppler frequency shifts offers an approximation of the velocity dispersion attributed to the micro-Doppler effect. To ensure the precise tracking of phase information in radar-received signals, it is imperative that the radar transmitter operates in conjunction with a highly stable frequency source to sustain full phase coherency.

The micro-Doppler effect serves as a valuable tool for determining the kinematic attributes of an object. For instance, it enables the detection of vibrations stemming from a vehicle's engine by analyzing the surface vibrations of the vehicle structure. Through the examination of the micro-Doppler characteristics associated with these surface vibrations, one can ascertain the engine's speed, facilitating the identification of specific vehicle types, such as distinguishing between a tank with a gas turbine engine and a bus powered by a diesel engine. The micro-Doppler effect manifested in an object can be discerned by its distinctive signature, denoting the intricate frequency modulation generated by the object's structural components. This signature is typically represented in the combined time and Doppler frequency domain.

The Micro-Doppler Signature of Objects commonly denote the characteristic representation of an object or a process, serving to encapsulate its distinct attributes. When scrutinizing the Doppler phenomenon within an object, unique micro-Doppler characteristics serve as evidence for identifying the object's motion. The micro-Doppler signature encompasses the object's distinct movement features, presenting itself as a complex frequency modulation profile depicted in both the time and Doppler frequency domain. These discernible characteristics, embodied in the micro-Doppler signature, are the defining attributes that grant an object its individual identity.

In recent years, the utilization of micro-Doppler signatures has gained prominence in diverse applications involving the extraction of distinctive features and the detection and identification of specific targets. Within the extensive literature pertaining to the micro-Doppler effect in radar, select references have significantly advanced our understanding of its theoretical properties [249], while others have expanded the scope of investigation from monostatic to bistatic and multistatic micro-Doppler features [252]. A substantial body of research has further contributed to the processing and analysis of micro-Doppler signatures associated with both rigid bodies exhibiting oscillatory motion and nonrigid bodies manifesting articulated motion [249]. Additionally, the domain of target classification, recognition, and identification grounded in micro-Doppler signatures has emerged as a crucial research area, with numerous studies dedicated to these topics [253].

6.3 Virtual Radar

The process of converting a video into a spectrogram encompasses the orchestration of several interconnected modules as depicted is 6.4, a comprehensive exploration of which is presented in this section. This operation yields three notable advantages. Firstly, it effectively mitigates environmental influence, as it exclusively synthesizes radar returns stemming from interactions with the human body, yielding a synthetic spectrogram that is not only devoid of noise but also exhibits high-resolution at-

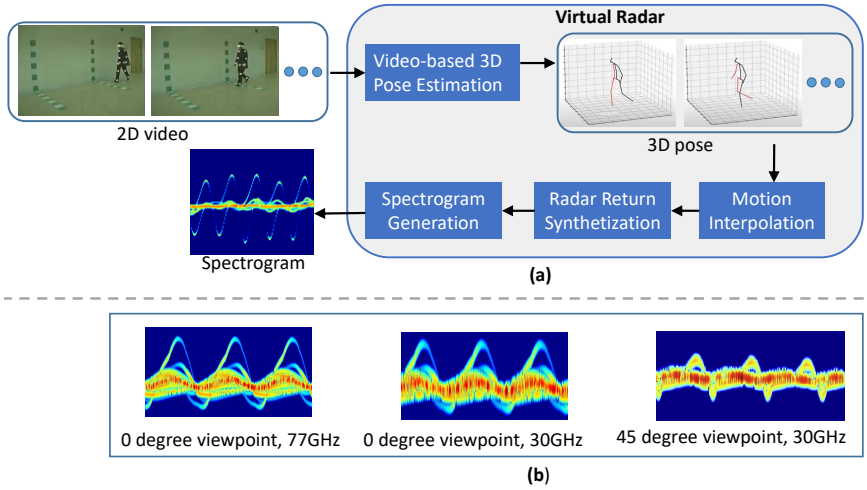

Figure 6.4: (a) Virtual radar system. (b) The resolutions of synthesized mD signature can be adjusted freely by changing the operating frequency of virtual radar (from 60GHz to 30 GHz). Multi-view mD signatures can be easily created by changing the virtual radar viewpoints (from 0° to 45°).

tributes. Secondly, the virtual radar configuration affords the flexibility to manipulate radar perspectives, thus facilitating the creation of multi-view spectrograms. Thirdly, the capacity to adjust virtual radar parameters, including carrier frequency and sampling rate, allows for the generation of spectrograms with diverse resolutions. In this research endeavor, we seek to enhance the authenticity of the synthesized radar return signals by incorporating a human kinematic model [254, 255] to facilitate motion interpolation.

The initial stage in the proposed methodology entails the creation of a 3D pose skeleton-based representation. In pursuit of this objective, two distinct approaches are available. The first approach involves the utilization of an off-the-shelf 2D pose estimation technique, which estimates the 2D human pose. Subsequently, a transformer-based lifter is employed to transform these 2D estimations into a 3D pose representation. The second alternative entails the use of Poseformer [256], a specialized model tailored for the direct regression of 3D pose information from video sequences. It is noteworthy that the latter approach tends to yield less erratic results, owing to the inherent smoothness achieved by the integrated spatiotemporal transformer.

Conversely, the former approach may introduce some degree of inconsistency in pose estimation across successive frames.

The subsequent stage involves the strategic application of cubic interpolations to facilitate motion interpolation, a crucial operation within the proposed solution. This choice is predicated on the distinctive characteristics of many activity recognition datasets, which often provide action recordings characterized by relatively short durations. As a consequence, the number of frames available for generating a high-resolution spectrogram is often insufficient.

Cubic interpolation, in this context, plays a pivotal role in addressing this limitation. It is a mathematical technique employed to estimate values between known data points within a given dataset. In the realm of video-to-radar spectrogram conversion, it is leveraged to bridge temporal gaps and to produce intermediary frames. The primary motivation behind employing cubic interpolation is to ensure a smooth motion estimation. This smoothness in motion is essential for the generation of realistic radar spectrograms that accurately capture the nuances of human activities. It helps mitigate potential jitters or abrupt transitions that might arise when simply subsampling the available frames.

In the third and critical stage, an ellipsoid-based human backscattering model is harnessed to compute the radar return signals, constituting a pivotal element in the backscattering model. Of particular significance is the application of the Physical Optics (PO) approximation for radar cross-section, which, in turn, is employed to synthesize radar return signals. This synthesis culminates in the generation of a simulated radar spectrogram, accomplished through the application of the Short-Time Fourier Transform (STFT) on the synthetically derived radar returns. The resulting spectrogram serves as a foundational component of the radar-based dataset, encapsulating the essential characteristics necessary for subsequent action recognition tasks.

The process of generating radar spectrograms within the proposed solution relies on the Short-Time Fourier Transform STFT, a well-established technique in signal processing. The STFT divides the simulated radar return signals into overlapping short segments or windows and applies a Fourier Transform to each segment. This procedure enables the capture of the signal's frequency components as they evolve over time, resulting in a time-frequency representation. The choice of window function and size allows for customization of the trade-off between time and frequency resolution in the spectrogram. Overlapping windows ensure a smooth representation. The spectrogram's horizontal axis represents time, the vertical axis depicts frequency, and the intensity at each point indicates the amplitude of the corresponding frequency component at a specific time, making it a valuable tool for creating informative radar-based datasets for action recognition.

Spectrograms serve as the conventional visualization technique for radar signals, offering a comprehensive representation. In the spectrogram representation, radar signals are mapped, associating an object's velocity with the vertical axis and temporal information with the horizontal axis. The intensity of each pixel in the spectrogram signifies the energy of the reflected signal, essentially reflecting the amount of radar signal interaction with the object. This pixel intensity assumes critical significance due to the distinct material properties of objects, which result in variations in radar signal absorption and reflection. Objects that introduce longer signal propagation times before returning to the radar exhibit a corresponding decay in energy. Spectrograms, thus configured, prove particularly advantageous for action recognition tasks. They maintain a dedicated temporal axis, enable the creation of a single-image representation in scenarios featuring multiple individuals, and readily adapt to diverse frame rates for each skeleton within a given dataset, enhancing their utility in the context of action recognition.

In the heart of our work lies the pivotal component, the virtual radar layer, respon-

sible for the transformation of skeleton data into spectrograms. The conventional approach to generating spectrograms from radar data necessitates the employment of actual radar equipment. In contrast, our approach leverages well-established mathematical models [249] for radar signal propagation, enabling the calculation of radar signals stemming from hypothetical objects in a spatial context. These models, while primarily tailored to basic geometric forms like spheres and ellipsoids, offer a means to represent skeleton data effectively. Consequently, the skeleton dataset possesses the capability to emulate a virtual radar signal and the corresponding spectrogram. To elucidate, an ellipsoid is delineated as a representation of the skeletal structure, facilitating the translation of skeleton data into radar-like signals and their subsequent spectrogram visualization. Mathematically, an Ellipsoid is defined by:

$$\left(\frac{x-x_0}{a}\right)^2+\left(\frac{y-y_0}{b}\right)^2+\left(\frac{z-z_0}{c}\right)^2=1, \tag{6.1}$$

Where (x_0, y_0, z_0) represent the central coordinates of the ellipsoid, while a, b, c correspond to the lengths of its semi-principal axes. The process of generating realistic skeleton data involves the estimation of each body segment's length, a computation based on the average body ratios relative to the overall body height. During human activity, individual body parts undergo spatial positional changes, which can be quantified through a series of joint-based calculations encompassing translation, rotation, and flexing operations [257]. The mathematical formulation for radar signals reflected by such ellipsoids, also referred to as radar backscatter (RCS), is delineated as follows:

$$RCS=\frac{\pi a^2b^2c^2}{((a\sin\theta\cos\phi)^2+(b\sin\theta\sin\phi)^2+(c\cos\theta)^2)^2}. \tag{6.2}$$

With regard to the radar reflections, the computation of the complex value representation for radar data follows a specific procedure. In this context, as depicted in

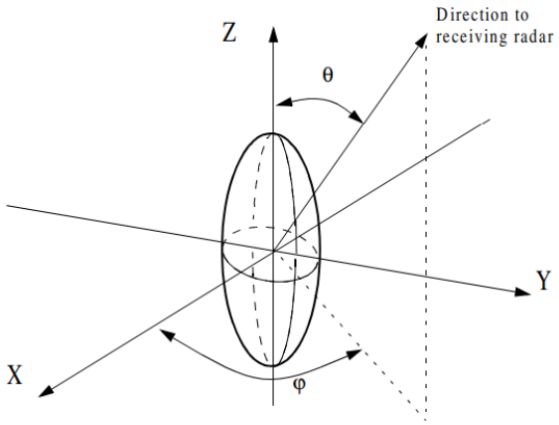

Figure 6.5: Section ellipsoid

6.5, the symbol θ denotes the angle formed between the z-axis of the ellipsoid and the direction of the radar receiver, while Φ represents the angle between the x-axis of the ellipsoid and the direction of the radar receiver [258].

$$Phase = \sqrt{RCS} \times e^{\frac{-j4\pi d}{\lambda}}. \tag{6.3}$$

Here, the variable d signifies the 2-norm distance extending from the ellipsoid's center to the radar, while λ stands for the radar wavelength. In the concluding step, a Short-Time Fourier Transform (STFT) is employed to process the radar signals, culminating in the creation of a spectrogram. Notably, the resulting spectrogram exhibits a remarkable congruence with those derived from actual radar signals reflected by an ellipsoid in a real-world scenario.

The transformation of skeleton data into a spectrogram entails an intricate process. Initially, the skeleton's structural elements are transmuted into ellipsoidal representations. Subsequently, the radar reflections originating from each individual ellipsoid are meticulously computed and subsequently aggregated, yielding a comprehensive radar signal that emulates what one might observe in a real-world radar scenario. These synthesized radar signals are further subjected to the STFT operation, ultimately producing a spectrogram.

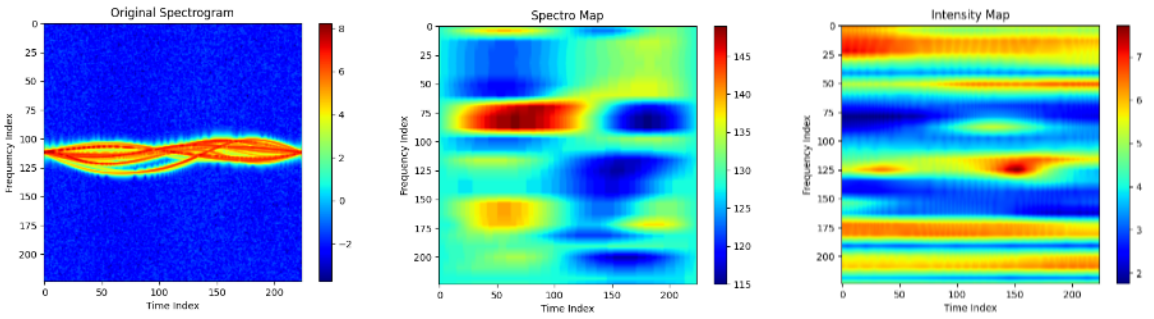

Figure 6.6: Different spectrogram representation

Conventionally, the resulting spectrogram tends to amalgamate all signals into a unified representation, often leading to the potential overlap or replacement of representations of distinct body segments when transcribed into images. In contrast, an alternative approach is adopted, wherein we disassemble the same spectrogram into discrete components, specifically a frequency map and an intensity map, as visually depicted in Figure 6.6. This novel method offers a differentiated perspective on the spectrogram's composition.

6.4 Experiment

To substantiate the validity of our methodology, we undertook the creation of the NTU-radar dataset, harnessing the skeletal representations derived from the ntu-60 dataset. As previously alluded, video-based action recognition datasets often feature relatively short action durations, thereby impeding the successful generation of high-resolution spectrograms. In response to this challenge, we adopted interpolation techniques to augment the dataset, ensuring an adequate volume of data for utilization by the virtual radar layer in the spectrogram generation process. Our evaluation approach encompasses a qualitative comparison between the patterns exhibited by the synthetic spectrogram and the original spectrogram generated through mmWave radar. Notably, our findings indicate that our approach consistently yields spectrograms of higher resolution when contrasted with those generated through the original mmWave radar technology. This outcome underscores the efficacy of our

novel approach in enhancing the quality of radar spectrogram generation.

To validate our approach, we conducted preliminary experiments on a medical dataset extracted from NTU dataset [259], which includes nine classes. Following the prescribed cross-view evaluation protocol as stipulated by NTU, our experimental setup involved the utilization of data samples from two distinct cameras for the purpose of training our model, while the third camera was reserved for validation. This rigorous evaluation protocol enables a robust assessment of the model's performance across varying camera views, ensuring its generalization and reliability under diverse viewing conditions. By employing pretrained resent18 on imagenet and we obtained an accuracy of 78% on the cross-view protocol. Further, we tried to analyze the accuracy degradation compared to the skeleton map representation, which only represents the raw joints coordination. From a radar perspective, the reflected signal from the skeleton data may have a similar velocity representation as shown in Fig 6.3.

After conducting a meticulous investigation to pinpoint the causes behind the performance degradation of the classifier, our analysis unveiled a noteworthy revelation from the skeletal perspective. Specifically, we observed that, within the class categories of 'chest pain,' 'headache pain,' and 'neck pain,' the estimated 3D poses exhibited striking similarity. Consequently, this inherent similarity in the skeletal representations translated to a marked convergence in the generated spectrograms for these specific action classes, as visually depicted in Figure 6.3. This finding underscores the significance of the skeletal perspective and its direct influence on the resulting radar spectrograms, shedding light on the nuances of the recognition process for these closely related actions. As a result of this meticulous analysis and optimization, we achieved an impressive classification accuracy of 87%, surpassing the challenges posed by these highly similar actions.

Given that this approach heavily relies on the skeleton representation of human

subjects, the quality and precision of the pose estimation are of paramount importance to ensure optimal results. It is worth noting that the accuracy of the pose estimation directly influences the classification accuracy, as inaccuracies or inaccurately estimated poses can propagate errors throughout the classification process. To address this, we have incorporated a series of skeleton-specific preprocessing techniques outlined in the broader context of action recognition. Furthermore, the NTU dataset's skeleton representation framework has been leveraged to tackle the challenge of varying action lengths. This is achieved by artificially extending the action sequences to a uniform length, ensuring comprehensive representation for all actions. Consequently, multiple cycles of the same action are incorporated into the dataset. As an essential component of our preprocessing pipeline, we perform sequence trimming to retain solely the initial cycle of each action, a strategy aimed at optimizing data consistency and facilitating subsequent analysis.

Furthermore, our approach offers the flexibility not only to produce a unified spectrogram for all skeletal figures but also the capability to generate spectrograms that are specific to individual bones, as visually illustrated in Figure 6.2. This form of decomposition proves to be instrumental when the primary aim is to recognize particular actions based on a selected subset of bones. It also becomes invaluable when the focus of the analysis is narrowed down to a specific anatomical region or body part, enabling a more granular exploration of human actions and their respective radar spectrograms. This adaptability adds a layer of precision and customization to the recognition process, accommodating various research and application scenarios.

6.5 Conclusion

In conclusion, our study has introduced a groundbreaking method for radar-based human action recognition that addresses the inherent challenges and limitations associated with the scarcity of radar-based datasets. The core of our approach revolves around the transformation of video-based data into synthetic radar spectrograms,

which extends the boundaries of radar-driven recognition systems. The method is characterized by a meticulous and multi-step process, starting with the generation of 3D pose skeleton-based representations. This step serves as a critical foundation for subsequent stages, as the quality and accuracy of the pose estimation are paramount for achieving optimal results. Several skeleton-specific preprocessing techniques, inspired by the broader field of action recognition, have been adapted to enhance the quality of the pose estimation.

One of the remarkable features of our approach is its capacity to generate high-resolution, noise-free spectrograms, providing a clearer and more detailed representation of human actions. Additionally, it offers flexibility in terms of viewpoint, enabling multi-view perspectives, which can be crucial for various applications. Moreover, we have introduced the ability to adjust the virtual radar parameters, such as carrier frequency and sampling rate, thereby allowing for the creation of spectrograms with varying resolutions. This adaptability makes our approach highly versatile and suitable for a wide range of applications.

To validate the effectiveness of our method, we have created the NTU-radar dataset and conducted comprehensive cross-view evaluations. The preliminary experiments on a medical subset of the NTU dataset demonstrated our approach's potential, achieving an accuracy of 78% following the cross-view protocol. Additionally, we performed a detailed analysis of the classifier's performance, revealing that the quality of the generated spectrogram plays a pivotal role in the recognition process. Notably, we observed that certain actions with similar skeleton-based representations resulted in similar spectrograms. This insight underscores the importance of robust pose estimation and its direct influence on the recognition outcome.

To validate the effectiveness of our method, we have undertaken a robust evaluation process, which included the creation of the NTU-radar dataset and comprehensive cross-view assessments. In our preliminary experiments, we focused on a medical sub-

set of the NTU dataset, which allowed us to showcase the potential of our approach. In this context, we achieved an impressive accuracy of 87% following the cross-view protocol, highlighting the method's prowess in action recognition. Furthermore, our rigorous performance analysis of the classifier uncovered a crucial factor: the quality of the generated spectrogram plays a pivotal role in the recognition process. Notably, we observed that certain actions with similar skeleton-based representations led to the production of spectrograms exhibiting remarkable resemblance. This revelation underscores the paramount importance of robust pose estimation, as it directly influences the recognition outcome.

Our work not only overcomes the challenges posed by the limited availability of radar-based datasets but also paves the way for novel applications across domains such as healthcare and security. We anticipate that our approach will inspire further research and development in the realm of radar-based human action recognition, offering new possibilities for real-world implementation and fostering advancements in this burgeoning field. As we continue to refine and expand our method, we envision a future where radar-based recognition systems find broader utility and make significant contributions to the fields of technology, healthcare, and beyond.

6.6 Future Work

To enhance the veracity of synthesized radar data, we propose an optimization strategy for the virtual radar that bridges the gap between simulation and real-world radar data. Presently, a noticeable disparity exists between the spectrograms generated by the virtual radar and those acquired from physical hardware. To narrow this gap, we suggest the application of a CycleGAN architecture, a deep learning model designed to perform high-to-low resolution translation. This model serves the dual purpose of capitalizing on the respective strengths of both high-resolution virtual radar-generated spectrograms and their lower-resolution physical hardware counterparts. Furthermore, it offers a streamlined approach to deploy physical radar solutions

without the necessity of collecting new data for each unique problem instance. This translation mechanism not only aids in aligning the fidelity of radar data synthesis with real-world data but also ensures flexibility in addressing various data resolution requirements.

Establishing a dedicated testbed designed to seamlessly integrate cameras and radar systems, thereby facilitating the creation of a paired dataset that combines high-resolution and low-resolution spectrograms, is a pivotal step in our methodology. This unique dataset serves the crucial purpose of enabling the training of deep learning models, specifically tailored to perform both sim-to-real and real-to-sim translations. The pairing is achieved by coupling the camera with the physical radar hardware. The camera will contribute to generating the video data used to derive skeleton information, a key component in generating synthetic spectrograms. By meticulously pairing high-resolution virtual radar-generated spectrograms with their real-world low-resolution counterparts, this testbed provides the necessary ground truth data for training and validating the performance of the proposed translation models. Consequently, the development of an effective sim-to-real and real-to-sim translator hinges on the creation of this paired dataset, which captures the essential nuances of high and low-resolution radar spectrograms, thereby bridging the gap between virtual radar simulations and actual hardware output.

CHAPTER 7: Conclusion and Future Directions

7.1 Dissertation Conclusion

The culmination of our research efforts heralds significant advancements in diverse domains within computer vision, biomechanics, human parameter estimation, and radar-based human action recognition. Across these multifaceted domains, our methodologies, innovations, and insights signify substantial progress and open doors for further exploration and application.

Our endeavors in human parameter estimation embrace a modular, two-streamed, end-to-end pipeline. This pioneering approach harnesses graph-based vision transformers and analytical inverse kinematics, promising results that rival the state-of-the-art in computer vision. By bridging existing gaps in methodologies, we present a robust solution for varied computer vision applications and pave the way for future innovations in this domain.

In the realm of biomechanics, our two-stage approach stands as a testament to the robustness and versatility of generating diverse kinematics and dynamic indicators. The meticulous integration of optimization techniques has enabled nuanced insights into body segment scale factors, marker registrations, joint kinematics, and dynamic characteristics. Moreover, our evaluation of transformer-based 3D pose estimation on mobile devices promises real-time biomechanical assessments outside traditional laboratory setups.

Moving to action recognition, our exploration delves deep into convolutional neural networks, optimization techniques, loss functions, and data augmentation strategies. The groundbreaking success of CNNs in action recognition showcases their potential to rival graph neural networks' performance. Our innovative optimization strategy

and novel loss function pave the way for superior model precision, reliability, and generalization.

Lastly, our work in radar-based human action recognition transcends the limitations imposed by the scarcity of radar-based datasets. The transformation of video-based data into synthetic radar spectrograms not only extends radar-driven recognition systems but also introduces a method for adjusting virtual radar parameters. Our method's adaptability, validated through comprehensive evaluations and the creation of the NTU-radar dataset, underscores its potential across varied applications.

Collectively, these contributions signify a transformative era in computer vision, biomechanics, and radar-based human action recognition. They not only advance the state-of-the-art but also lay the foundation for future innovations and practical applications in technology, healthcare, surveillance, and beyond.

7.2 Future Directions

Future endeavors in human pose estimation involve integrating physics-based models to achieve more context-aware and realistic estimations. By considering biomechanical constraints and interactions, these models can enhance pose realism, particularly in applications like computer animation and medical simulations. Additionally, exploring advanced Inverse Kinematics (IK) solvers beyond analytical methods could address complex scenarios better. Numerical, optimization-based, or learning-based IK methods could be pivotal in scenarios with constrained environments, sports analysis, human-robot collaboration, and interactions with external forces.

Integrating physics-based constraints into pose estimation represents a forward-thinking direction. This approach aims to imbue pose estimations with biomechanical principles, aligning poses with natural human movements for increased accuracy and consistency. While introducing physics-based constraints might add complexity, the potential benefits in terms of improved accuracy, generalization across diverse scenarios, and enhanced interpretability make it a noteworthy avenue.

For the iOS application, future efforts aim to optimize the inference speed by exploring model compression techniques such as model pruning and knowledge distillation. Establishing a cloud server infrastructure could offload computational burdens from individual devices and leverage the scalability of cloud-based solutions. Simultaneously, implementing an on-device inverse kinematics optimizer could enhance the efficiency and autonomy of biomechanical analyses directly on iOS devices.

The future of human action recognition involves leveraging mixed dataset training to enhance model generalization across varied contexts and fortify against biases present in singular datasets. Additionally, integrating more precise 3D pose estimation techniques that adhere closely to biomechanical principles could significantly improve the model's capacity to discern subtle nuances in human actions. Incorporating kinematic analysis parameters like joint angles, velocities, and accelerations may refine the model's discriminative capabilities.

Enhancements in radar data synthesis can be achieved by applying a CycleGAN architecture to bridge the gap between virtual radar-generated spectrograms and physical hardware. A dedicated testbed, coupling high-resolution virtual radar-generated spectrograms with real-world low-resolution counterparts, could facilitate sim-to-real and real-to-sim translations for improved radar data synthesis.

REFERENCES

[1] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[2] R. Liu, J. Shen, H. Wang, C. Chen, S.-c. Cheung, and V. Asari, "Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5064–5073.

[3] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[4] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.

[5] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming, "Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary," Sytronics Inc Dayton Oh, Tech. Rep., 2002.

[6] D. A. Winter, *Biomechanics and motor control of human movement*. John wiley & sons, 2009.

[7] D. D. Harrison, S. O. Harrison, A. C. Croft, D. E. Harrison, and S. J. Troyanovich, "Sitting biomechanics part i: review of the literature," *Journal of manipulative and physiological therapeutics*, vol. 22, no. 9, pp. 594–609, 1999.

[8] Y.-c. Fung, *Biomechanics: mechanical properties of living tissues*. Springer Science & Business Media, 2013.

[9] J. Hay, *The biomechanics of sports techniques*. Prentice-Hall, 1978.

[10] J. Campos, P. Poletaev, A. Cuesta, C. Pablos, and V. Carratalá, "Kinematical analysis of the snatch in elite male junior weightlifters of different weight categories," *The Journal of Strength & Conditioning Research*, vol. 20, no. 4, pp. 843–850, 2006.

[11] C. Woods, R. Hawkins, S. Maltby, M. Hulse, A. Thomas, and A. Hodson, "The football association medical research programme: an audit of injuries in professional footballâanalysis of hamstring injuries," *British journal of sports medicine*, vol. 38, no. 1, pp. 36–41, 2004.

[12] C. H. Ho, R. J. Triolo, A. L. Elias, K. L. Kilgore, A. F. DiMarco, K. Bogie, A. H. Vette, M. L. Audu, R. Kobetic, S. R. Chang *et al.*, "Functional electrical stimulation and spinal cord injury," *Physical Medicine and Rehabilitation Clinics*, vol. 25, no. 3, pp. 631–654, 2014.

[13] J. Goodfellow and J. O'Connor, "The mechanics of the knee and prosthesis design," *The Journal of Bone & Joint Surgery British Volume*, vol. 60, no. 3, pp. 358–369, 1978.

[14] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*. IEEE, 2015, pp. 579–583.

[15] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 102–106.

[16] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.

[17] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[18] Q. Wan, Y. Li, C. Li, and R. Pal, "Gesture recognition for smart home applications using portable radar sensors," in *2014 36th annual international conference of the IEEE engineering in medicine and biology society*. IEEE, 2014, pp. 6414–6417.

[19] J. J. Wang and S. Singh, "Video analysis of human dynamicsâa survey," *Real-time imaging*, vol. 9, no. 5, pp. 321–346, 2003.

[20] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 1–7.

[21] L. Chen, M. Zhou, M. Wu, J. She, Z. Liu, F. Dong, and K. Hirota, "Three-layer weighted fuzzy support vector regression for emotional intention understanding in human–robot interaction," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 2524–2538, 2018.

[22] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, pp. 1–33, 2014.

[23] X. Huang and M. Dai, "Indoor device-free activity recognition based on radio signal," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5316–5329, 2016.

[24] A. Markman, X. Shen, and B. Javidi, "Three-dimensional object visualization and detection in low light illumination using integral imaging," *Optics letters*, vol. 42, no. 16, pp. 3068–3071, 2017.

[25] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[26] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," *Advances in neural information processing systems*, vol. 27, 2014.

[27] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 648–656.

[28] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[29] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-pcnn: Two stage human pose estimation with graph pose refinement," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 492–508.

[30] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 455–472.

[31] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 205–214.

[32] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, and P. Luo, "Differentiable hierarchical graph grouping for multi-person pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 718–734.

[33] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5386–5395.

[34] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4929–4937.

[35] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in

Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. Springer, 2016, pp. 34–50.

[36] X. Zhu, Y. Jiang, and Z. Luo, "Multi-person pose estimation for posetrack with enhanced part affinity fields," in *ICCV PoseTrack Workshop*, vol. 7, 2017, p. 4321.

[37] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *Advances in neural information processing systems*, vol. 30, 2017.

[38] A. A. Osman, T. Bolkart, and M. J. Black, "Star: Sparse trained articulated human body regressor," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 598–613.

[39] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[40] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2602–2611.

[41] G. Pavlakos, X. Zhou, and K. Daniilidis, "Ordinal depth supervision for 3d human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7307–7316.

[42] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2640–2649.

[43] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, "Learning to fuse 2d and 3d image cues for monocular body pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3941–3950.

[44] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2344–2353.

[45] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, "Drpose3d: Depth ranking in 3d human pose estimation," *arXiv preprint arXiv:1805.08973*, 2018.

[46] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, "3d human mesh regression with dense correspondence," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7054–7063.

[47] K. Zhou, B. L. Bhatnagar, and G. Pons-Moll, "Unsupervised shape and pose disentanglement for 3d meshes," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 341–357.

[48] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4501–4510.

[49] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.

[50] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[51] H. Choi, G. Moon, and K. M. Lee, "Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *European Conference on Computer Vision*. Springer, 2020, pp. 769–787.

[52] C. Zheng, M. Mendieta, P. Wang, A. Lu, and C. Chen, "A lightweight graph transformer network for human mesh reconstruction from 2d human pose," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5496–5507.

[53] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1954–1963.

[54] C. Zheng, M. Mendieta, T. Yang, G.-J. Qi, and C. Chen, "Feater: An efficient network for human reconstruction via feature map-based transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 945–13 954.

[55] C. Zheng, X. Liu, G.-J. Qi, and C. Chen, "Potter: Pooling attention transformer for efficient human mesh recovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1611–1620.

[56] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6050–6059.

[57] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[58] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3d human pose ambiguities with 3d scene constraints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2282–2292.

[59] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2252–2261.

[60] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 723–732.

[61] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 965–10 974.

[62] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8320–8329.

[63] H. Wang, R. A. Güler, I. Kokkinos, G. Papandreou, and S. Zafeiriou, "Blsm: A bone-level skinned model of the human mesh," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 1–17.

[64] M. Fisch and R. Clark, "Orientation keypoints for 6d human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10 145–10 158, 2021.

[65] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, "Dyna: A model of dynamic human shape in motion," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 1–14, 2015.

[66] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[67] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger *et al.*, "Learning an infant body model from rgb-d data for accurate full body motion analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 792–800.

[68] I. Santesteban, E. Garces, M. A. Otaduy, and D. Casas, "Softsmpl: Data-driven modeling of nonlinear soft-tissue dynamics for parametric humans," in *Computer Graphics Forum*, vol. 39, no. 2. Wiley Online Library, 2020, pp. 65–75.

[69] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[70] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 601–617.

[71] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *2017 international conference on 3D vision (3DV)*. IEEE, 2017, pp. 506–516.

[72] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[73] M. Raghavan and B. Roth, "Inverse kinematics of the general 6r manipulator and related linkages," 1993.

[74] R. P. Paul and B. Shimano, "Kinematic control equations for simple manipulators," in *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*. IEEE, 1979, pp. 1398–1406.

[75] R. Diankov, "Automated construction of robotic manipulation programs," Ph.D. dissertation, Carnegie Mellon University, The Robotics Institute Pittsburgh, 2010.

[76] J. U. Korein, *A geometric investigation of reach*. MIT press, 1986.

[77] D. Tolani, A. Goswami, and N. I. Badler, "Real-time inverse kinematics techniques for anthropomorphic limbs," *Graphical models*, vol. 62, no. 5, pp. 353–388, 2000.

[78] M. Kallmann, "Analytical inverse kinematics with body posture control," *Computer animation and virtual worlds*, vol. 19, no. 2, pp. 79–91, 2008.

[79] E. Molla and R. Boulic, "Singularity free parametrization of human limbs," in *Proceedings of Motion on Games*, 2013, pp. 187–196.

[80] B. Siciliano, O. Khatib, and T. Kröger, *Springer handbook of robotics*. Springer, 2008, vol. 200.

[81] L. Unzueta, M. Peinado, R. Boulic, and Á. Suescun, "Full-body performance animation with sequential inverse kinematics," *Graphical models*, vol. 70, no. 5, pp. 87–104, 2008.

[82] A. Liegeois *et al.*, "Automatic supervisory control of the configuration and behavior of multibody mechanisms," *IEEE transactions on systems, man, and cybernetics*, vol. 7, no. 12, pp. 868–871, 1977.

[83] A. A. Maciejewski and C. A. Klein, "Obstacle avoidance for kinematically redundant manipulators in dynamically varying environments," *The international journal of robotics research*, vol. 4, no. 3, pp. 109–117, 1985.

[84] M. Girard and A. A. Maciejewski, "Computational modeling for the computer animation of legged figures," *ACM SIGGRAPH Computer Graphics*, vol. 19, no. 3, pp. 263–270, 1985.

[85] C. W. Wampler, "Manipulator inverse kinematic solutions based on vector formulations and damped least-squares methods," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, no. 1, pp. 93–101, 1986.

[86] Y. Nakamura and H. Hanafusa, "Inverse kinematic solutions with singularity robustness for robot manipulator control," 1986.

[87] S. R. Buss and J.-S. Kim, "Selectively damped least squares for inverse kinematics," *Journal of Graphics tools*, vol. 10, no. 3, pp. 37–49, 2005.

[88] S. A. Teukolsky, B. P. Flannery, W. Press, and W. Vetterling, "Numerical recipes in c," *SMR*, vol. 693, no. 1, pp. 59–70, 1992.

[89] A. Colomé and C. Torras, "Redundant inverse kinematics: Experimental comparative review and two enhancements," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5333–5340.

[90] A. A. Maciejewski, "Dealing with the ill-conditioned equations of motion for articulated figures," *IEEE Computer Graphics and Applications*, vol. 10, no. 3, pp. 63–71, 1990.

[91] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.

[92] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2000.

[93] B. Siciliano, "A closed-loop inverse kinematic scheme for on-line joint-based robot control," *Robotica*, vol. 8, no. 3, pp. 231–243, 1990.

[94] W. Kwan, "Closed-form and generalized inverse kinematic solutions for animating the human articulated structure," *Curtin University of Technology*, 1996.

[95] J. Zhao and N. I. Badler, "Inverse kinematics positioning using nonlinear programming for highly articulated figures," *ACM Transactions on Graphics (TOG)*, vol. 13, no. 4, pp. 313–336, 1994.

[96] C. Rose, B. Guenter, B. Bodenheimer, and M. F. Cohen, "Efficient generation of motion transitions using spacetime constraints," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 147–154.

[97] D. Merrick and T. Dwyer, "Skeletal animation for the exploration of graphs," in *Proceedings of the 2004 Australasian symposium on Information Visualisation-Volume 35*. Citeseer, 2004, pp. 61–70.

[98] J. Lee and S. Y. Shin, "A hierarchical approach to interactive motion editing for human-like figures," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 39–48.

[99] H. J. Shin, J. Lee, S. Y. Shin, and M. Gleicher, "Computer puppetry: An importance-based approach," *ACM Transactions on Graphics (TOG)*, vol. 20, no. 2, pp. 67–94, 2001.

[100] R. Kulpa and F. Multon, "Fast inverse kinematics and kinetics solver for human-like figures," in *5th IEEE-RAS International Conference on Humanoid Robots, 2005.* IEEE, 2005, pp. 38–43.

[101] R. Kulpa, F. Multon, and B. Arnaldi, "Morphology-independent representation of motions for interactive human-like animation," in *Eurographics*, 2005.

[102] A. Bouënard, S. Gibet, and M. M. Wanderley, "Hybrid inverse motion control for virtual characters interacting with sound synthesis: Application to percussion motion," *The Visual Computer*, vol. 28, pp. 357–370, 2012.

[103] N. Vahrenkamp, T. Asfour, and R. Dillmann, "Efficient inverse kinematics computation based on reachability analysis," *International Journal of Humanoid Robotics*, vol. 9, no. 04, p. 1250035, 2012.

[104] H. Hanafusa, T. Yoshikawa, and Y. Nakamura, "Analysis and control of articulated robot arms with redundancy," *IFAC Proceedings Volumes*, vol. 14, no. 2, pp. 1927–1932, 1981.

[105] P. Baerlocher and R. Boulic, "Task-priority formulations for the kinematic control of highly redundant articulated structures," in *Proceedings. 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems. Innovations in Theory, Practice and Applications (Cat. No. 98CH36190)*, vol. 1. IEEE, 1998, pp. 323–329.

[106] M. Retargetting, "Retargetting motion to new characters."

[107] G. Monheit and N. I. Badler, "A kinematic model of the human spine and torso," 1990.

[108] W. Maurel and D. Thalmann, "Human shoulder modeling including scapulo-thoracic constraint and joint sinus cones," *Computers & Graphics*, vol. 24, no. 2, pp. 203–218, 2000.

[109] N. Klopčar, M. Tomšič, and J. Lenarčič, "A kinematic model of the shoulder complex to evaluate the arm-reachable workspace," *Journal of biomechanics*, vol. 40, no. 1, pp. 86–91, 2007.

[110] R. M. Murray, Z. Li, and S. S. Sastry, *A mathematical introduction to robotic manipulation*. CRC press, 2017.

[111] A. Aristidou, "Hand tracking with physiological constraints," *The Visual Computer*, vol. 34, pp. 213–228, 2018.

[112] N. I. Badler, C. B. Phillips, and B. L. Webber, *Simulating humans: computer graphics animation and control*. Oxford University Press, 1993.

[113] S. Quinlan, "Efficient distance computation between non-convex objects," in *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*. IEEE, 1994, pp. 3324–3329.

[114] S. Gottschalk, M. C. Lin, and D. Manocha, "Obbtree: A hierarchical structure for rapid interference detection," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 171–180.

[115] G. v. d. Bergen, "Efficient collision detection of complex deformable models using aabb trees," *Journal of graphics tools*, vol. 2, no. 4, pp. 1–13, 1997.

[116] J. Mezger, S. Kimmerle, and O. Etzmuß, "Hierarchical techniques in collision detection for cloth animation," 2003.

[117] P. Volino and N. M. Thalmann, "Efficient self-collision detection on smoothly discretized surface animations using geometrical shape regularity," in *Computer graphics forum*, vol. 13, no. 3. Wiley Online Library, 1994, pp. 155–166.

[118] G. Nawratil, H. Pottmann, and B. Ravani, "Generalized penetration depth computation based on kinematical geometry," *Computer Aided Geometric Design*, vol. 26, no. 4, pp. 425–443, 2009.

[119] O. Khatib, L. Sentis, J. Park, and J. Warren, "Whole-body dynamic behavior and control of human-like robots," *International Journal of Humanoid Robotics*, vol. 1, no. 01, pp. 29–43, 2004.

[120] L. Liu, M. V. D. Panne, and K. Yin, "Guided learning of control graphs for physics-based characters," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, pp. 1–14, 2016.

[121] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, "Opensim: open-source software to create and analyze dynamic simulations of movement," *IEEE transactions on biomedical engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.

[122] S.-H. Lee and A. Goswami, "Reaction mass pendulum (rmp): An explicit model for centroidal angular momentum of humanoid robots," in *Proceedings 2007 IEEE international conference on robotics and automation*. IEEE, 2007, pp. 4667–4672.

[123] A. Shapiro and S.-H. Lee, "Practical character physics for animators," *IEEE computer graphics and applications*, vol. 31, no. 4, pp. 45–55, 2010.

[124] K. W. Sok, K. Yamane, J. Lee, and J. Hodgins, "Editing dynamic human motions via momentum and force," in *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer animation*. Citeseer, 2010, pp. 11–20.

[125] P. G. Kry, C. Rahgoshay, A. Rabbani, and K. Singh, "Inverse kinodynamics: Editing and constraining kinematic approximations of dynamic motion," *Computers & Graphics*, vol. 36, no. 8, pp. 904–915, 2012.

[126] A. H. Rabbani and P. G. Kry, "Physik: Physically plausible and intuitive keyframing." in *Graphics Interface*, 2016, pp. 153–161.

[127] T. Geijtenbeek, M. Van De Panne, and A. F. Van Der Stappen, "Flexible muscle-based locomotion for bipedal creatures," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 1–11, 2013.

[128] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 597–600.

[129] ——, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055*, 2018.

[130] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "Skeletonnet: Mining deep part features for 3-d action recognition," *IEEE signal processing letters*, vol. 24, no. 6, pp. 731–735, 2017.

[131] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[132] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, 2016.

[133] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4041–4049.

[134] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[135] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 148–157.

[136] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[137] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[138] Y.-F. Song, Z. Zhang, and L. Wang, "Richly activated graph convolutional network for action recognition with incomplete skeletons," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1–5.

[139] Y. Yoon, J. Yu, and M. Jeon, "Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition," *Applied Intelligence*, pp. 1–15, 2021.

[140] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3163–3172.

[141] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "Star-transformer: a spatio-temporal cross attention transformer for human action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3330–3339.

[142] Z. Lin, Y. Gao, and D. Li, "Cross-attention multi-scale spatial temporal transformer for skeleton-based action recognition," 2023.

[143] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[144] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2929–2936.

[145] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine vision and applications*, vol. 24, no. 5, pp. 971–981, 2013.

[146] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[147] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.

[148] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[149] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[150] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[151] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[152] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[153] M. Zenaldin and R. M. Narayanan, "Radar micro-doppler based human activity classification for indoor and outdoor environments," in *Radar Sensor Technology XX*, vol. 9829. SPIE, 2016, pp. 364–373.

[154] Y. Kim and H. Ling, "Human activity classification based on micro-doppler signatures using a support vector machine," *IEEE transactions on geoscience and remote sensing*, vol. 47, no. 5, pp. 1328–1337, 2009.

[155] Z. Zhou, Z. Cao, and Y. Pi, "Dynamic gesture recognition with a terahertz radar based on range profile sequences and doppler signatures," *Sensors*, vol. 18, no. 1, p. 10, 2017.

[156] K. A. Smith, C. Csech, D. Murdoch, and G. Shaker, "Gesture recognition using mm-wave sensor for human-car interface," *IEEE sensors letters*, vol. 2, no. 2, pp. 1–4, 2018.

[157] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern recognition letters*, vol. 119, pp. 3–11, 2019.

[158] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and trends® in signal processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[159] Y. Shao, Y. Dai, L. Yuan, and W. Chen, "Deep learning methods for personnel recognition based on micro-doppler features," in *Proceedings of the 9th International Conference on Signal Processing Systems*, 2017, pp. 94–98.

[160] E. Pinyoanuntapong, A. Ali, K. Jakkala, P. Wang, M. Lee, Q. Peng, C. Chen, and Z. Sun, "Gaitsada: Self-aligned domain adaptation for mmwave gait recognition," *arXiv preprint arXiv:2301.13384*, 2023.

[161] Y. Lang, C. Hou, Y. Yang, D. Huang, and Y. He, "Convolutional neural network for human micro-doppler classification," in *Proc. Eur. Microw. Conf.*, 2017, pp. 1–4.

[162] H. T. Le, S. L. Phung, A. Bouzerdoum, and F. H. C. Tivive, "Human motion classification with micro-doppler radar and bayesian-optimized convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2961–2965.

[163] R. Trommel, R. Harmanny, L. Cifola, and J. Driessen, "Multi-target human gait classification using deep convolutional neural networks on micro-doppler spectrograms," in *2016 European Radar Conference (EuRAD)*. IEEE, 2016, pp. 81–84.

[164] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE geoscience and remote sensing letters*, vol. 13, no. 1, pp. 8–12, 2015.

[165] Y. Kim and B. Toomajian, "Application of doppler radar for the recognition of hand gestures using optimized deep convolutional neural networks," in *2017 11th European Conference on Antennas and Propagation (EUCAP)*. IEEE, 2017, pp. 1258–1260.

[166] J. Zhang, J. Tao, and Z. Shi, "Doppler-radar based hand gesture recognition system using convolutional neural networks," in *Communications, Signal Processing, and Systems: Proceedings of the 2017 International Conference on Communications, Signal Processing, and Systems*. Springer, 2019, pp. 1096–1113.

[167] D. Tahmoush, "Review of micro-doppler signatures," *IET Radar, Sonar & Navigation*, vol. 9, no. 9, pp. 1140–1146, 2015.

[168] V. C. Chen and S. Qian, "Joint time-frequency transform for radar range-doppler imaging," *IEEE transactions on aerospace and electronic systems*, vol. 34, no. 2, pp. 486–499, 1998.

[169] M. S. Seyfioğlu, A. M. Özbayoğlu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1709–1723, 2018.

[170] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[171] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3278–3289, 2018.

[172] B. Erol, M. Amin, Z. Zhou, and J. Zhang, "Range information for reducing fall false alarms in assisted living," in *2016 IEEE Radar Conference (RadarConf)*. IEEE, 2016, pp. 1–6.

[173] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human motion classification based on range information with deep convolutional neural network," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 2017, pp. 1519–1523.

[174] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range fmcw monopulse radar for hand-gesture sensing," in *2015 IEEE Radar Conference (RadarCon)*. IEEE, 2015, pp. 1491–1496.

[175] ——, "Multi-sensor system for driver's hand-gesture recognition," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[176] B. Jokanovic, M. Amin, and B. Erol, "Multiple joint-variable domains recognition of human motion," in *2017 IEEE Radar Conference (RadarConf)*. IEEE, 2017, pp. 0948–0952.

[177] B. Jokanović and M. Amin, "Fall detection using deep learning in range-doppler radars," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 1, pp. 180–189, 2017.

[178] D. Drover, R. MV, C.-H. Chen, A. Agrawal, A. Tyagi, and C. Phuoc Huynh, "Can 3d pose be learned from 2d projections alone?" in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[179] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International journal of computer vision*, vol. 99, pp. 190–214, 2012.

[180] M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, and R. Cucchiara, "Compressed volumetric heatmaps for multi-person 3d pose estimation," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7204–7213.

[181] H. E. Pang, Z. Cai, L. Yang, T. Zhang, and Z. Liu, "Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[182] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal on Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[183] Y. Zhang and C. Kambhamettu, "Stereo matching with segmentation-based cooperation," in *Computer VisionâECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part II 7*. Springer, 2002, pp. 556–571.

[184] Z. Zhao, D. Ye, X. Zhang, G. Chen, and B. Zhang, "Improved direct linear transformation for parameter decoupling in camera calibration," *Algorithms*, vol. 9, no. 2, p. 31, 2016.

[185] K. Pulli, A. Baksheev, K. Kornyakov, and V. Eruhimov, "Real-time computer vision with opencv," *Communications of the ACM*, vol. 55, no. 6, pp. 61–69, 2012.

[186] F. Fraundorfer and D. Scaramuzza, "Visual odometry: Part i: The first 30 years and fundamentals," *IEEE Robotics and Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[187] K. R. Castleman, *Digital image processing*. Prentice Hall Press, 1996.

[188] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[189] D. A. Forsyth and J. Ponce, *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

[190] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[191] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[192] Q.-T. Luong and O. Faugeras, "The geometry of multiple images," *MIT Press, Boston*, vol. 2, no. 3, pp. 4–5, 2001.

[193] J. Heikkila and O. Silvén, "A four-step camera calibration procedure with implicit image correction," in *Proceedings of IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1997, pp. 1106–1112.

[194] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 561–578.

[195] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 20–40.

[196] R. A. Guler and I. Kokkinos, "Holopose: Holistic 3d human reconstruction in-the-wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 884–10 894.

[197] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, "Coherent reconstruction of multiple humans from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5579–5588.

[198] Z. Zou and W. Tang, "Modulated graph convolutional network for 3d human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 477–11 487.

[199] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3425–3435.

[200] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[201] D. GRIEVE, "Biomechanics of the musculo-skeletal system, 2nd edn. edited by benno m. nigg and walter herzog,(pp. xii+ 643; illustrated;£ 60 hardback; isbn 0 471 97818 3.) chichester: Wiley. 1999." *The Journal of Anatomy*, vol. 195, no. 2, pp. 315–317, 1999.

[202] R. M. Enoka, *Neuromechanics of human movement*. Human kinetics, 2008.

[203] J. A. Reinbolt, J. F. Schutte, B. J. Fregly, B. I. Koh, R. T. Haftka, A. D. George, and K. H. Mitchell, "Determination of patient-specific multi-joint kinematic models through two-level optimization," *Journal of biomechanics*, vol. 38, no. 3, pp. 621–626, 2005.

[204] I. W. Charlton, P. Tate, P. Smyth, and L. Roren, "Repeatability of an optimised lower body model," *Gait & posture*, vol. 20, no. 2, pp. 213–221, 2004.

[205] R. M. Kanko, E. K. Laende, E. M. Davis, W. S. Selbie, and K. J. Deluzio, "Concurrent assessment of gait kinematics using marker-based and markerless motion capture," *Journal of biomechanics*, vol. 127, p. 110665, 2021.

[206] S. D. Uhlrich, A. Falisse, Ł. Kidziński, J. Muccini, M. Ko, A. S. Chaudhari, J. L. Hicks, and S. L. Delp, "Opencap: Human movement dynamics from smartphone videos," *PLoS computational biology*, vol. 19, no. 10, p. e1011462, 2023.

[207] A. Rajagopal, C. L. Dembia, M. S. DeMers, D. D. Delp, J. L. Hicks, and S. L. Delp, "Full-body musculoskeletal model for muscle-driven simulation of human gait," *IEEE transactions on biomedical engineering*, vol. 63, no. 10, pp. 2068–2079, 2016.

[208] K. Werling, N. A. Bianco, M. Raitor, J. Stingel, J. L. Hicks, S. H. Collins, S. L. Delp, and C. K. Liu, "Addbiomechanics: Automating model scaling, inverse kinematics, and inverse dynamics from human motion data through sequential optimization," *bioRxiv*, 2023.

[209] Y. Li, Q. Fan, H. Huang, Z. Han, and Q. Gu, "A modified yolov8 detection network for uav aerial image recognition," *Drones*, vol. 7, no. 5, p. 304, 2023.

[210] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[211] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[212] S. Karthickkumar and K. Kumar, "A survey on deep learning techniques for human action recognition," in *2020 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2020, pp. 1–6.

[213] G. Ogbuabor and R. La, "Human activity recognition for healthcare using smartphones," in *Proceedings of the 2018 10th international conference on machine learning and computing*, 2018, pp. 41–46.

[214] V. Bloom, "Multiple action recognition for video games (marvig)," Ph.D. dissertation, Kingston University, 2015.

[215] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, 2022.

[216] J. Huang, X. Xiang, X. Gong, B. Zhang *et al.*, "Long-short graph memory network for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 645–652.

[217] B. N. Raj, A. Subramanian, K. Ravichandran, and D. N. Venkateswaran, "Exploring techniques to improve activity recognition using human pose skeletons," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2020, pp. 165–172.

[218] D. Q. Huynh, "Metrics for 3d rotations: Comparison and analysis," *Journal of Mathematical Imaging and Vision*, vol. 35, no. 2, pp. 155–164, 2009.

[219] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton-based action synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4394–4402.

[220] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[221] M. Liu, Q. He, and H. Liu, "Fusing shape and motion matrices for view invariant action recognition using 3d skeletons," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3670–3674.

[222] J. Weng, C. Weng, and J. Yuan, "Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2017, pp. 4171–4180.

[223] N. C. Tang, Y.-Y. Lin, J.-H. Hua, M.-F. Weng, and H.-Y. M. Liao, "Human action recognition using associated depth and skeleton information," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4608–4612.

[224] S. Ubalde, F. Gómez-Fernández, N. A. Goussies, and M. Mejail, "Skeleton-based action recognition using citation-knn on bags of time-stamped pose descriptors," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3051–3055.

[225] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," *arXiv preprint arXiv:2104.13586*, 2021.

[226] X. Wang, Z. Qing, Z. Huang, Y. Feng, S. Zhang, J. Jiang, M. Tang, C. Gao, and N. Sang, "Proposal relation network for temporal action detection," *arXiv preprint arXiv:2106.11812*, 2021.

[227] A. M. De Boissiere and R. Noumeir, "Infrared and 3d skeleton feature fusion for rgb-d action recognition," *IEEE Access*, vol. 8, pp. 168 297–168 308, 2020.

[228] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1012–1020.

[229] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5832–5841.

[230] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.

[231] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[232] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition," *Information Sciences*, vol. 569, pp. 90–109, 2021.

[233] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[234] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[235] X. Liu, H. Wang, Y. Zhang, F. Wu, and S. Hu, "Towards efficient data-centric robust machine learning with noise-based augmentation," *arXiv preprint arXiv:2203.03810*, 2022.

[236] J. L. Suárez, S. García, and F. Herrera, "A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges," *Neurocomputing*, vol. 425, pp. 300–322, 2021.

[237] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[238] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[239] M. C. Mukkamala and M. Hein, "Variants of rmsprop and adagrad with logarithmic regret bounds," in *International conference on machine learning*. PMLR, 2017, pp. 2545–2553.

[240] A. Defazio and S. Jelassi, "Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization," *Journal of Machine Learning Research*, vol. 23, pp. 1–34, 2022.

[241] A. Al-Kababji, F. Bensaali, and S. P. Dakua, "Scheduling techniques for liver segmentation: Reducelronplateau vs onecyclelr," *arXiv preprint arXiv:2202.06373*, 2022.

[242] T. Cazenave, J. Sentuc, and M. Videau, "Cosine annealing, mixnet and swish activation for computer go," in *Advances in Computer Games*. Springer, 2022, pp. 53–60.

[243] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[244] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69.

[245] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[246] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[247] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern recognition*, vol. 61, pp. 295–308, 2017.

[248] Y. Lin, J. Le Kernec, S. Yang, F. Fioranelli, O. Romain, and Z. Zhao, "Human activity classification with radar: Optimization and noise robustness with iterative convolutional neural networks followed with random forests," *IEEE Sensors Journal*, vol. 18, no. 23, pp. 9669–9681, 2018.

[249] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and electronic systems*, vol. 42, no. 1, pp. 2–21, 2006.

[250] A. Eden and A. Eden, *The search for christian doppler*. Springer, 1992.

[251] C. ChenV, "Analysisofradarmicro-dopplersignaturewith time-frequencytransform," *ProceedingsonStatisticalSignal andArrayProcessing*, p. r466, 2000.

[252] G. E. Smith, K. Woodbridge, and C. J. Baker, "Multistatic micro-doppler signature of personnel," in *2008 IEEE Radar Conference*. IEEE, 2008, pp. 1–6.

[253] Y. Yang, J. Lei, W. Zhang, and C. Lu, "Target classification and pattern recognition using micro-doppler radar signatures," in *Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06)*. IEEE, 2006, pp. 213–217.

[254] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei, "Model-based deep hand pose estimation," 2016, available: arXivpreprintarXiv:1606.0685.

[255] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," *ECCV Worktp on Geometry Meets Deep Learning*, 2016.

[256] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 656–11 665.

[257] R. Boulic, N. M. Thalmann, and D. Thalmann, "A global human walking model with real-time kinematic personification," *The visual computer*, vol. 6, no. 6, pp. 344–358, 1990.

[258] E. F. Knott, J. F. Schaeffer, and M. T. Tulley, *Radar cross section.* SciTech Publishing, 2004.

[259] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.