EFFICIENTHRNET: EFFICIENT AND SCALABLE NETWORKS FOR REAL-TIME POSE ESTIMATION AND SEGMENTATION

by

Aneri Parag Sheth

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Electrical Engineering

Charlotte

2020

Approved by:

_____
Dr. Hamed Tabkhi

_____
Dr. Andrew Willis

_____
Dr. Chen Chen

ABSTRACT

ANERI PARAG SHETH. EfficientHRNet: Efficient and Scalable Networks for real-time Pose Estimation and Segmentation. (Under the direction of DR. HAMED TABKHI)

Recent years have brought great advancement in 2D human pose estimation. However, bottom-up approaches that do not rely on external detectors to generate person crops, tend to have large model sizes and intense computational requirements, making them ill-suited for applications where large computation costs can be prohibitive. Lightweight approaches are exceedingly rare and often come at the price of massive accuracy loss.

This thesis presents EfficientHRNet, a family of lightweight 2D human pose estimators that unifies the high-resolution structure of state-of-the-art HigherHRNet, a multi-scale high resolution network with the highly efficient model scaling principles of EfficientNet to create high accuracy models with significantly reduced computation costs. In addition, it provides a formulation for jointly scaling the backbone EfficientNet below the baseline B0 and the rest of EfficientHRNet with it. Ultimately, this work is able to create a family of highly accurate and efficient 2D human pose estimators that is flexible enough to provide a lightweight solution for a variety of application and device requirements. The baseline H0 model achieves 64.8% accuracy on COCO dataset and overall, EfficientHRNet proves to be more computationally efficient than other bottom-up 2D human pose estimation approaches, while achieving highly competitive accuracy.

Moreover, inspired by creating a family of EfficientHRNet based models for pose estimation, this work also provides a similar formulation for creating models in another popular computer vision application, image segmentation. Pose estimation and image segmentation models created using these methods are further used in the edge video analytics pipeline as a front-end to evaluate the performance of an end-to-end

real time system. This thesis also carries out simulation of pose estimation and segmentation model into the real-time vision pipeline.

DEDICATION

This work is wholeheartedly dedicated to my father Dr. Parag Sheth and my mother Dr. Hetal Sheth who have been my source of encouragement and gave me strength to overcome any challenge faced, who continually provide their moral, emotional, spiritual and financial support.

To my brothers, sisters, relatives, mentors, friends and colleagues who shared their wisdom, advice and enthusiasm to finish this thesis.

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my advisor Dr. Hamed Tabkhivayghan, BS, MS, PhD, Assistant Professor, Electrical and Computer Engineering Department, the William States Lee College of Engineering, University of North Carolina at Charlotte, for giving me the opportunity to carry out research and providing invaluable guidance throughout this thesis. His vision, intelligence, and motivation have deeply inspired me. He has truly taught me some valuable aspects of carrying out research in an ethical and sincere manner. It was a great privilege and an honor to work under him and his research lab, TeCSAR. I would also like to thank his friendship, empathy and a great sense of humor.

I am extremely grateful to my parents for their love, caring and sacrifices for educating and providing me with a future. I am thankful to my relatives and my friends for constantly supporting me. My special thanks to Christopher Neff, PhD candidate, Electrical and Computer Engineering, University of North Carolina at Charlotte, for his patience, ideas and mentoring throughout this work. I would also like to thank Steven Fergerson and John Middleton, Electrical and Computer Engineering, University of North Carolina at Charlotte for their valuable inputs into this research. I would also like to acknowledge Google Meet for providing a platform for online conferencing during my defense presentation and throughout the last semester, which was unavoidable due to the coronavirus pandemic. Lastly, I would like to thank my best friend, my companion, Maharsh Patel for his emotional and moral support throughout my MS journey.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF ABBREVIATIONS xi

CHAPTER 1: INTRODUCTION 1

1.1. Motivation 4

1.2. Contributions 5

1.3. Thesis Outline 6

CHAPTER 2: RELATED WORK 7

2.1. Top down pose estimation methods 7

2.2. Bottom-up pose estimation methods 7

2.3. Multi-scale High Resolution Networks 8

2.4. Model Scaling 9

CHAPTER 3: EFFICIENTHRNET 10

3.1. Network Architecture and Formulation 10

3.2. Compound Scaling Method 15

CHAPTER 4: EXPERIMENTAL RESULTS 18

4.1. Classification for Compact EfficientNet 18

4.2. 2D Human Pose Estimation for EfficientHRNet 20

4.3. Image Segmentation for EfficientHRNet 23

4.4. Qualitative Results 25

4.4.1. Qualitative Results of EfficientHRNet on COCO 25

4.4.2. Qualitative Results of EfficientHRNet in video analytics pipeline 26

4.4.3. Challenges in the current pipeline 27

CHAPTER 5: CONCLUSIONS AND FUTURE WORK 28

5.1. Conclusion 28

5.2. Future Work 28

REFERENCES 30

LIST OF TABLES

TABLE 3.1: Efficient scaling configs for EfficientHRNet 16

TABLE 4.1: Compact EfficientNet performance on ImageNet and CIFAR-100 datasets. 19

TABLE 4.2: Comparisons with bottom-up methods on COCO2017 val dataset 20

TABLE 4.3: Comparisons with state-of-the-art bottom-up methods on COCO2017 test-dev dataset. 22

TABLE 4.4: EfficientHRNet Pose Estimation Performance Evaluation 23

TABLE 4.5: EfficientHRNet Segmentation on Cityscapes dataset. 24

LIST OF FIGURES

FIGURE 1.1: Video Analytics Application 1

FIGURE 1.2: A comparison of computational complexity and accuracy between bottom-up human pose estimation methods. Accuracy measured on COCO2017 val dataset. X-axis is logarithmic in scale. 3

FIGURE 1.3: Edge video analytics pipeline 4

FIGURE 3.1: A detailed illustration of the EfficientHRNet architecture for multi-person pose estimation. Consisting of a backbone EfficientNet, a High-Resolution Network with three stages and four branches (denoted by different colors), and a Heatmap Prediction Network. EfficinetHRNet is completely scalable, allowing network complexity to be customized for target applications. 11

FIGURE 3.2: A detailed illustration of the EfficientHRNet architecture for segmentation. The backbone EfficientNet, a High-Resolution Network with three stages and four branches (denoted by different colors) remain the same, but the Heatmap Prediction Network is now replaced by Segmentation Network showing that EfficinetHRNet is completely scalable, allowing network complexity to be customized for target applications. 14

FIGURE 4.1: Qualitative Results for EfficientHRNet models on COCO2017 test dataset. Top to bottom: Simple, Medium and Complex examples. 25

FIGURE 4.2: EfficientHRNet Pose Estimation and Segmentation in Vision pipeline. Left: Pose Estimation, Right: Segmentation. 26

LIST OF ABBREVIATIONS

AI Artificial Intelligence

AP Average Precision

CIFAR Canadian Institute for Advanced Research

CNN Convolutional Neural Network

COCO Common Objects in Context

FPS Frames Per Second

GLOPS Giga Floating Point Operations

GPU Graphics Processing Unit

GT Ground Truth

IoT Internet of Things

IOU Intersection Over Union

OKS Object Keypoint Similarity

SGD Stochastic Gradient Descent

CHAPTER 1: INTRODUCTION

Real-time edge video analytics demand accuracy, performance and power efficiency. Edge video analytics capture video streams from surveillance cameras installed at a desired location, for instance a parking lot, in case of pedestrian tracking and then the video is processed on the edge servers and GPUs for further AI processing. The AI processing includes video analytics applications like pose estimation, image segmentation, human path prediction, person re-identification and action detection. Due to bulky computations and computations happening on the edge node, it has become extremely important to have Convolutional Neural Networks (CNNs) customized for video analytics to achieve real-time performance on power-constrained devices. Figure 1.1 shows an example of pose estimation, object detection, and segmentation applications running on a video frame. It shows how these three applications can be used to get useful information about the entire image for a surveillance application. In this thesis, pose estimation algorithm design and image segmentation design will be deeply explored and qualitative and quantitative results are shown.

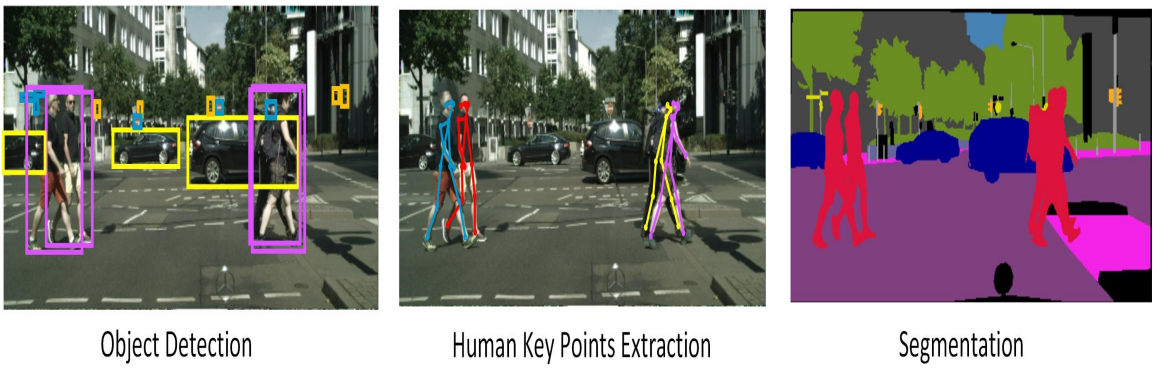

Figure 1.1: Video Analytics Application

Two-dimensional human pose estimation is a common task used in many popular smart applications and has made substantial progress in recent years. There are two primary approaches to 2D human pose estimation. The first is a top-down approach, where cropped images of humans are provided and the network uses those cropped images to produce human keypoints. Top-down approaches rely on object detectors to provide initial human crops, thus they often come with relatively higher computation cost, and are not truly end-to-end. The second is a bottom-up approach, where a network works off the original image and produces human keypoints for all people in the image. While these methods often do not quite reach the accuracy that is possible with state-of-the-art top-down approaches, they come with relatively lower model size and computational overhead.

Even so, state-of-the-art bottom-up approaches are still quite large and computationally expensive. The current state-of-the-art [1] having 63.8 million parameters and requiring 154.3 billion floating-point operations. Even though many emerging applications, such as self-driving vehicles, intelligent surveillance, and augmented reality, require lightweight multi-person human pose estimation, there has been much less attention towards developing lightweight bottom-up methods. This means that if an application's real-time requirements and resource constraints do not match up well to the existing models, then a sub-optimal solution is the only available choice. To address this gap, there is a need for a family of lightweight human pose estimation models that achieves comparable accuracy to the state-of-the-art approaches.

This study presents EfficientHRNet, a family of lightweight scalable networks for high-resolution and efficient bottom-up multi-person pose estimation. EfficientHRNet unifies the principles of state-of-the-art EfficientNet and HRNet, and presents a new formulation that enables near state-of-the-art human pose estimation while being more computationally efficient than all other bottom-up methods. Similar to HRNet, EfficientHRNet uses multiple resolutions of features to generate keypoints, but in a

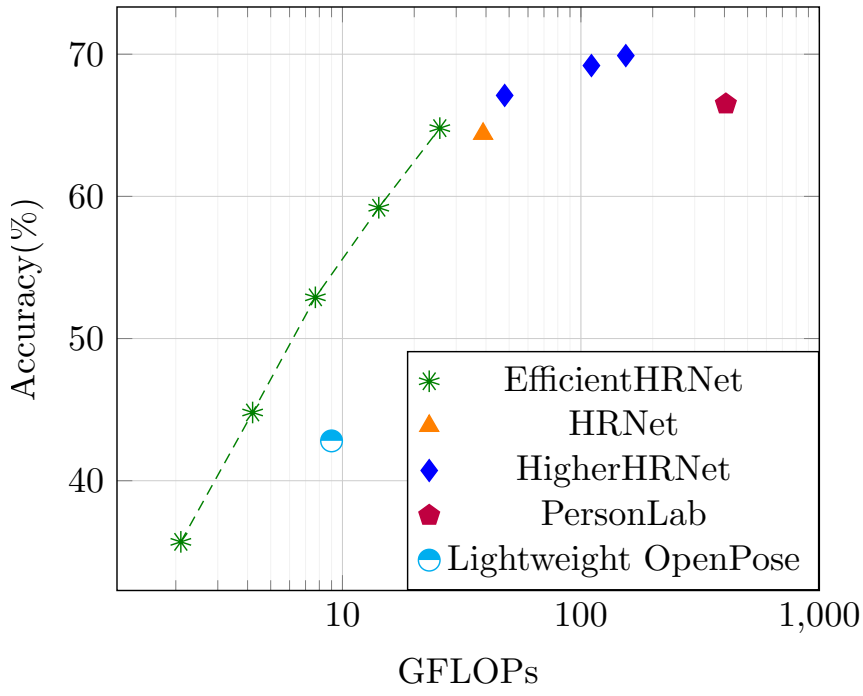

Figure 1.2: A comparison of computational complexity and accuracy between bottom-up human pose estimation methods. Accuracy measured on COCO2017 val dataset. X-axis is logarithmic in scale.

much more efficient manner. At the same time, it uses EfficientNet as a backbone and adapts its scaling methodology to be better suited for human pose estimation. To enable agile lightweight execution, EfficientHRNet further expands the EfficientNet formulation to not only scale below the baseline, but also to jointly scale down the input resolution, High-Resolution Network, and Heatmap Prediction Network. Through this, a family of networks is created that can address the entire domain of lightweight 2D human pose estimation while being flexible towards the accuracy and computation requirements of an application. The models are evaluated on the COCO dataset [2]. Figure 1.2 demonstrates how EfficientHRNet models provide equivalent or higher accuracy at lower computational costs than their direct peers. When comparing to state-of-the-art models, baseline EfficientNet competes in accuracy while requiring much less computation. Compared to HRNet [3], EfficientHRNet achieves 0.4% higher accuracy while requiring only 66% the number of operations. When

comparing to HigherHRNet [1] and PersonLab [4], EfficientHRNet sees between a 1.7% to 5.1% decrease in accuracy, while only requiring between 7% to 17% of the computation. Even when comparing to models designed specifically for lightweight execution, such as Lightweight OpenPose [5], a scaled down EfficientHRNet is able to achieve 10.1% higher accuracy while further reducing computation by 15%. In addition, EfficientHRNet architecture is applied for image segmentation which achieves 71% accuracy, comparable to the state of the art. Lastly, these two models are added into the vision pipeline in order to get accurate and efficient results.

1.1 Motivation

Most of the emerging applications like self-driving cars, intelligent surveillance and more, require lightweight front-end for selecting areas of interest i.e. segmenting people from scene, understanding the scene, detecting people, etc. Similar to this, for video surveillance system for privacy aware human pedestrian tracking and action detection, lightweight models for detecting human and scene i.e. multi-person pose estimation and image segmentation are required. Figure 1.3 demonstrates a vision

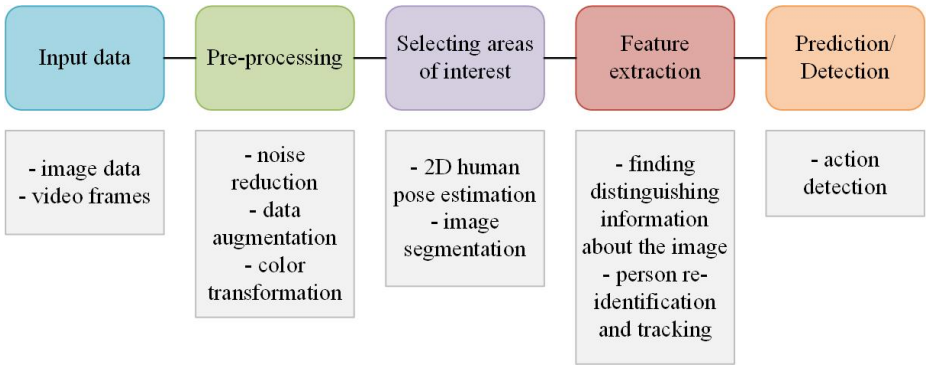

Figure 1.3: Edge video analytics pipeline

pipeline flow and shows the importance of performance for an end-to-end system. The focus of this study is in selecting areas of interest - detecting people - multi-person pose estimation and understanding the scene elements - image segmentation. There has been much less attention towards developing lightweight pose estimation

and segmentation. This means that if an applicationâs real-time requirements and resource constraints do not match up well to the existing models, then a sub-optimal solution is the only available choice. To address this gap, there is a need for a family of lightweight human pose estimation models as well as segmentation models that achieves comparable accuracy to the state-of-the-art approaches and are much more power-efficient than existing methods.

1.2 Contributions

EfficientHRNet is a family of lightweight scalable networks for high-resolution and efficient bottom-up multi-person pose estimation. EfficientHRNet unifies the principles of state-of-the-art EfficientNet and HRNet, and presents a new formulation that enables near state-of-the-art human pose estimation while being more computationally efficient than all other bottom-up methods. EfficientHRNet is also customizable for different computer vision applications like segmentation and object detection and have proven to achieve high performance (FPS) as compared to other approaches. In summary, this study has the following major contributions:

- Towards front-end algorithms:
  - Proposing a family of efficient and scalable models called EfficientHRNet for lightweight multi-person pose estimation by unifying the principles of popular EfficientNet and HRNet.
  - Providing a scaling methodology to scale down the backbone and multi-scale network to generate lightweight models that run efficiently on low power IoT devices.
  - Developing EfficientHRNet based segmentation model to support the pose estimation in identifying people in the scene for the overall edge video analytics pipeline.

- Towards front-end simulation:
  - Integrating pose estimation and segmentation models into the vision pipeline and simulating results on this project's own image frames in order to evaluate the qualitative and FPS performance.

1.3 Thesis Outline

This thesis is outlined as follows: Chapter 2 gives a detailed report on background related work of pose estimation methods, multi-resolution networks and model scaling methods. Chapter 3 introduces the network architecture of EfficientHRNet - both for pose estimation and segmentation. It explains the backbone CNN as well as the HRNet networks along with the scaling formulation. Chapter 4 is the experimental results section where an exhaustive evaluation of five different EfficientHRNet models on the challenging COCO dataset is conducted and the models compared to state-of-the-art methods. Segmentation model is also evaluated on Cityscapes dataset. Experimental results also include performance numbers and qualitative results on this project's own curated videos illustrating both where the models excel and where they fall short. Chapter 5 discusses the conclusion and future work of this study.

CHAPTER 2: RELATED WORK

This section first presents related work relevant to the field of top-down and bottom-up methods for 2D human pose estimation. Then, a survey on multi-scale high-resolution networks, particularly for computer vision applications, is presented. Lastly, popular model scaling techniques that have emerged in recent years are discussed.

2.1 Top down pose estimation methods

Top-down methods rely on first identifying all the persons in an image using an object detector, and then detecting the keypoints for a single person within a defined bounding box. These single person [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] and multi-person [17], [18], [19], [20], [21] pose estimation methods often generate person bounding boxes using object detector [22], [23], [24], [25]. For instance, RMPE [20] adds symmetric spatial transformer network on top of single person pose estimator stacked hourglass network [13] to get high-quality regions from inaccurate bounding boxes and then detects poses using parametric non-maximum suppression.

2.2 Bottom-up pose estimation methods

Bottom-up methods [26], [27], [28], [29], [4], [30], [31], [32], [33], [34] first detect identity-free keypoints in an image and then group them into persons using various keypoints grouping techniques. Methods like [31] and [32] perform grouping by integer linear program and non-maximum suppression. This allows for much faster inference times as compared to top-down methods with almost similar accuracies. Other methods further improve upon prediction time by using greedy grouping techniques, along with other optimizations, as seen in [26], [27], [28], [29], [4]. For instance, OpenPose

[26], [27] is a multi-stage network where one branch detects keypoints in the form of heatmaps, while the other branch generates Part Affinity Fields that are used to associate keypoints with each other. Grouping is done by calculating the line integral between all keypoints and grouping the pair that has the highest integral. Lightweight OpenPose [5] replaces larger backbone with MobileNets to achieve real-time performance with fewer parameters and FLOPs while compromising on accuracy. PifPaf [28] uses Part *Intensity* Fields to detect body parts and Part *Associative* Fields for associating parts with each other to form human poses. In [29], a stacked hourglass network [13] is used both for predicting heatmaps and grouping keypoints. Grouping is done by assigning each keypoint with an embedding, called a tag, and then associating those keypoints based on the L_2 distance between the tag vectors. In this paper, we mainly focus on a highly accurate, end-to-end multi-person pose estimation method as in [29].

2.3 Multi-scale High Resolution Networks

Feature pyramid networks augmented with multi-scale representations are widely adopted for complex and necessary computer vision applications like segmentation and pose estimation [35], [36], [37], [38], [39], [40]. Recovering high-resolution feature maps using techniques like upsampling, dilated convolution, and deconvolution are also widely popular for object detection [38], semantic segmentation [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51] and pose estimation [13], [52, 53], [54], [55], [39], [40], [56], [31], [32]. Moreover, there are several works that focus on generating high-resolution feature maps directly [57], [58], [59], [60], [61], [62], [3], [1]. HRNet [3], [62] proposes to maintain high-resolution feature maps throughout the entire network. HRNet consists of multiple branches with different resolutions across multiple stages. With multi-scale fusion, HRNet is able to generate high resolution feature maps and has found its application in object detection, semantic segmentation, and pose estimation [3], [61], [62] thereby achieving remarkable accuracies. Recently, HigherHRNet

for multi-person pose estimation [1] is proposed which uses HRNet as base network to generate high resolution feature maps, and further adds a deconvolution module to predict accurate, high-quality heatmaps. HigherHRNet achieves state-of-the-art accuracy on the COCO dataset [2], surpassing all existing bottom-up methods. In this study, the principles of HigherHRNet are adopted for generating high-resolution feature maps with multi-scale fusion for predicting high-quality heatmaps.

2.4 Model Scaling

Previous works on bottom-up pose estimation [26], [27], [1], [3], [29], [13] often rely on either large backbone networks, like ResNet [63] or VGGNet [64], or large input resolutions and multi-scale training for achieving state-of-the-art accuracy. Some recent works [3],[1] show that increasing the channel dimension of otherwise identical models can further improve accuracy. EfficientNet [65] and RegNet [66] show that by jointly scaling network width, depth, and input resolution, better efficiency for image classification can be achieved compared to previous state-of-the-art networks using much larger models. More recently, EfficientNet's lite models remove elements, such as squeeze and excite and swish layers, to make the network more hardware friendly. Inspired by EfficientNet, EfficientDet [67] proposes a compound scaling method for object detection along with efficient multi-scale feature fusion. We observe that there is a lack of an efficient scaling method for multi-person pose estimation, especially for embedded devices. Lightweight pose estimation models which are scalable and comparatively accurate are needed for computer vision applications which focus on real-time performance. This study proposes compound scaling which is also inspired by EfficientNet, a method that jointly scales the width, depth, and input resolution of our network, as well as the repetition within the high-resolution modules, explained in Chapter 3. In addition, this compound scaling allows the EfficientNet backbone to scale below the baseline B0, creating even lighter weight models.

CHAPTER 3: EFFICIENTHRNET

EfficientHRNet is a family of scalable and lightweight models customizable for various computer vision applications like multi-person pose estimation and segmentation. In this chapter, firstly a brief review of the proposed architecture will be provided and then the new compound scaling method in order to generate lightweight models for EfficientHRNet is described.

3.1 Network Architecture and Formulation

EfficientHRNet comprises of three sub-networks: (1) Backbone Network, (2) High-Resolution Network, and (3) Heatmap prediction network (pose estimation)/Segmentation network. The first stage of the network is the backbone, consisting of EfficientNet [65]. The backbone EfficientNet model outputs four different resolution feature maps of decreasing resolutions $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ the size of the input image. These feature maps are passed into the main body of the network, called the High-Resolution Network. The High-Resolution Network is inspired by HRNet [3], [62] and HigherHRNet [1]. Borrowing the principles of these higher resolution networks brings two major advantages:

1. By maintaining multiple high-resolution feature representations throughout the network, heatmaps with a higher degree of spatial precision are generated.

2. Repeated multi-scale fusions allow for high-resolution feature representations to inform lower-resolution feature representations, and vise versa, resulting in robust, multi-resolution feature representations that are ideal for multi-person human pose estimation as well as segmentation.

Figure 3.1 presents a detailed architecture illustration of EfficientHRNet for the

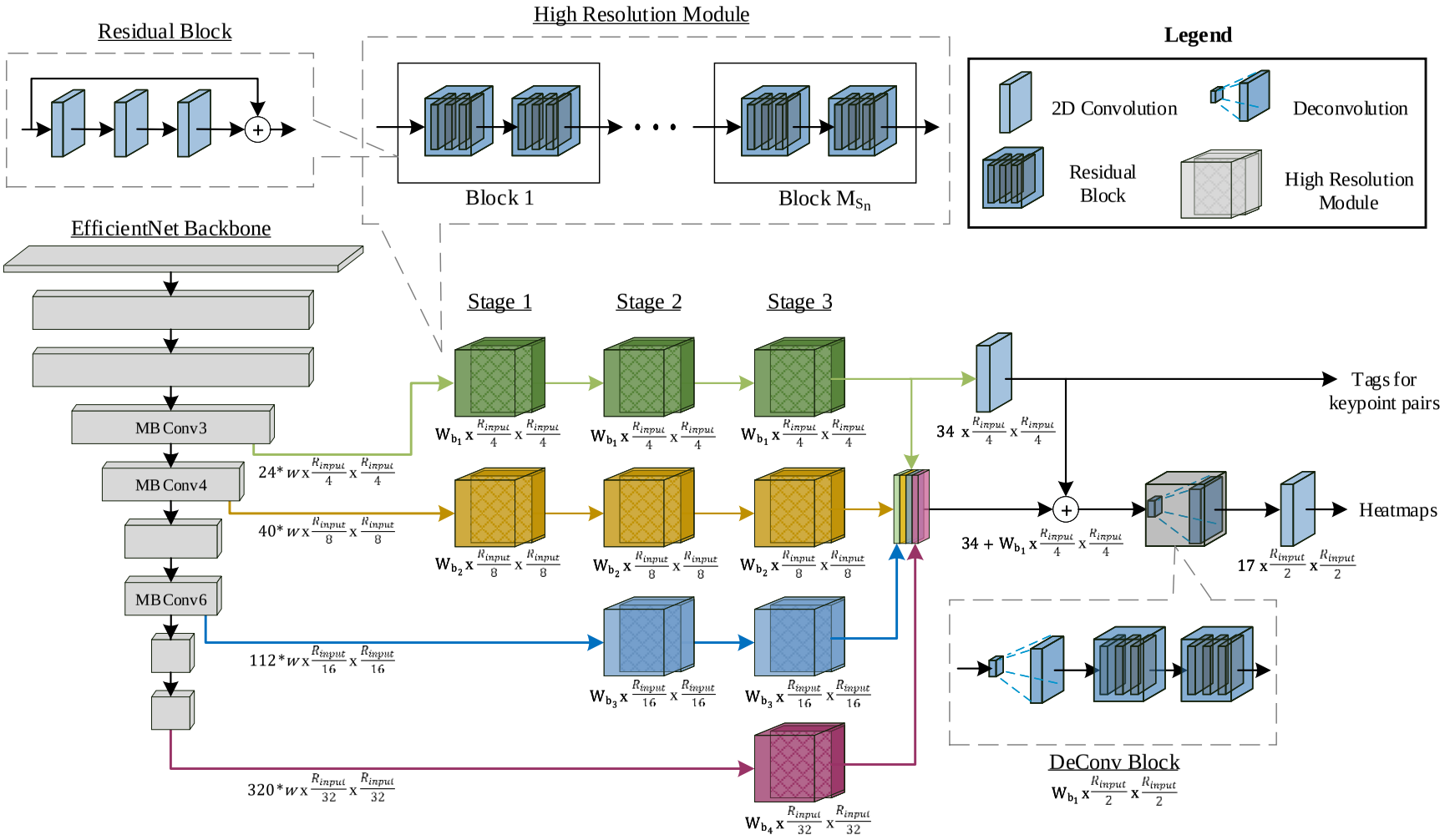

Figure 3.1: A detailed illustration of the EfficientHRNet architecture for multi-person pose estimation. Consisting of a backbone EfficientNet, a High-Resolution Network with three stages and four branches (denoted by different colors), and a Heatmap Prediction Network. EfficinetHRNet is completely scalable, allowing network complexity to be customized for target applications.

application of multi-person human pose estimation. It shows the three sub-networks - the Backbone Network, the High-Resolution Network, and the Heatmap Prediction Network - in detail. It also provides equations showing how the network scales the input resolution R_{input} and width of feature maps W_{b_n}, which will be further explained in the next section.

As seen in Figure 3.1, the High-Resolution Network has three stages s_1, s_2, and s_3, containing four parallel branches b_1, b_2, b_3, and b_4 of different resolutions. The first stage s_1 starts with two branches b_1 and b_2, with each consecutive stage adding an additional branch, until all four branches are present in s_3. These four branches each consist of *high resolution modules* with a width of W_{b_n}. Each branch b_n contains feature representations of decreasing resolutions that mirror the resolutions output by the Backbone Network, as shown in Figure 3.1 and the following equation:

$$W_{b_n} \times \frac{R_{input}}{2^n + 1} \tag{3.1}$$

For instance, stage 2 (s_2) has three branches of resolutions $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ of the original input image resolution and a width W_{b_n} as seen in Figure 3.1. Moreover, each *high resolution module* is made up of a number of blocks, M_{s_n}, each containing two *residual blocks*, each performing three convolution operations with a residual connection.

In order to predict more accurate heatmaps, a *DeConv block* is added on top of the High-Resolution Network, as proposed in [1]. Transposed convolution is used to generate high quality feature maps which are $\frac{1}{2}$ the original input resolution. The input to the *DeConv block* is the concatenation of the feature maps and predicted heatmaps from the High-Resolution Network, as shown in the equation below:

$$34 + W_{b_1} \times \frac{R_{input}}{4} \times \frac{R_{input}}{4} \tag{3.2}$$

Two *residual blocks* are added after the *DeConv block* to refine the up-sampled feature

maps, as seen in Figure 3.1.

Lastly, the Heatmap Prediction Network is used to generate human keypoint predictions. After the end of the *DeConv block*, a 1x1 convolution is used to predict heatmaps and tagmaps in a similar fashion to [29], the feature map size of each shown below:

$$\begin{aligned} T_{size} &= 34 \times \frac{R_{input}}{4} \times \frac{R_{input}}{4} \\ H_{size} &= 17 \times \frac{R_{input}}{2} \times \frac{R_{input}}{2} \end{aligned} \tag{3.3}$$

The grouping process clusters keypoints into multiple persons by grouping keypoints whose tags have minimum L_2 distance. Moreover, much like [1], the High-Resolution Network is scale-aware and uses multi-resolution supervision for heatmaps during training to allow the network to learn with more precision, even for small-scale persons. From the ground truth, heatmaps for different resolutions are generated to match the predicted keypoints of different scales. Thus, the final heatmaps loss is the sum of mean squared errors for all resolutions. However, as high resolutions tagmaps do not converge well, tagmaps are trained on a resolution $\frac{1}{4}$ of the original input resolution, as in [29].

In order to show how EfficientHRNet customizes for other applications, this study shows its application on image segmentation. As seen in Figure 3.2, the backbone network and the High-Resolution Network provide the same outputs for the head network but the head is now customized for Segmentation Network. The output from the High-Resolution Network having 4 different resolutions - $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ are now upsampled using bilinear transformation. The upsampled outputs are concatenated based on the width and $\frac{1}{2}$ of the original input resolution. Lastly, a final 3x3 convolution is applied on the concatenated output to get the segmented image with probabilities of 19 different classes of the Cityscapes dataset [68].

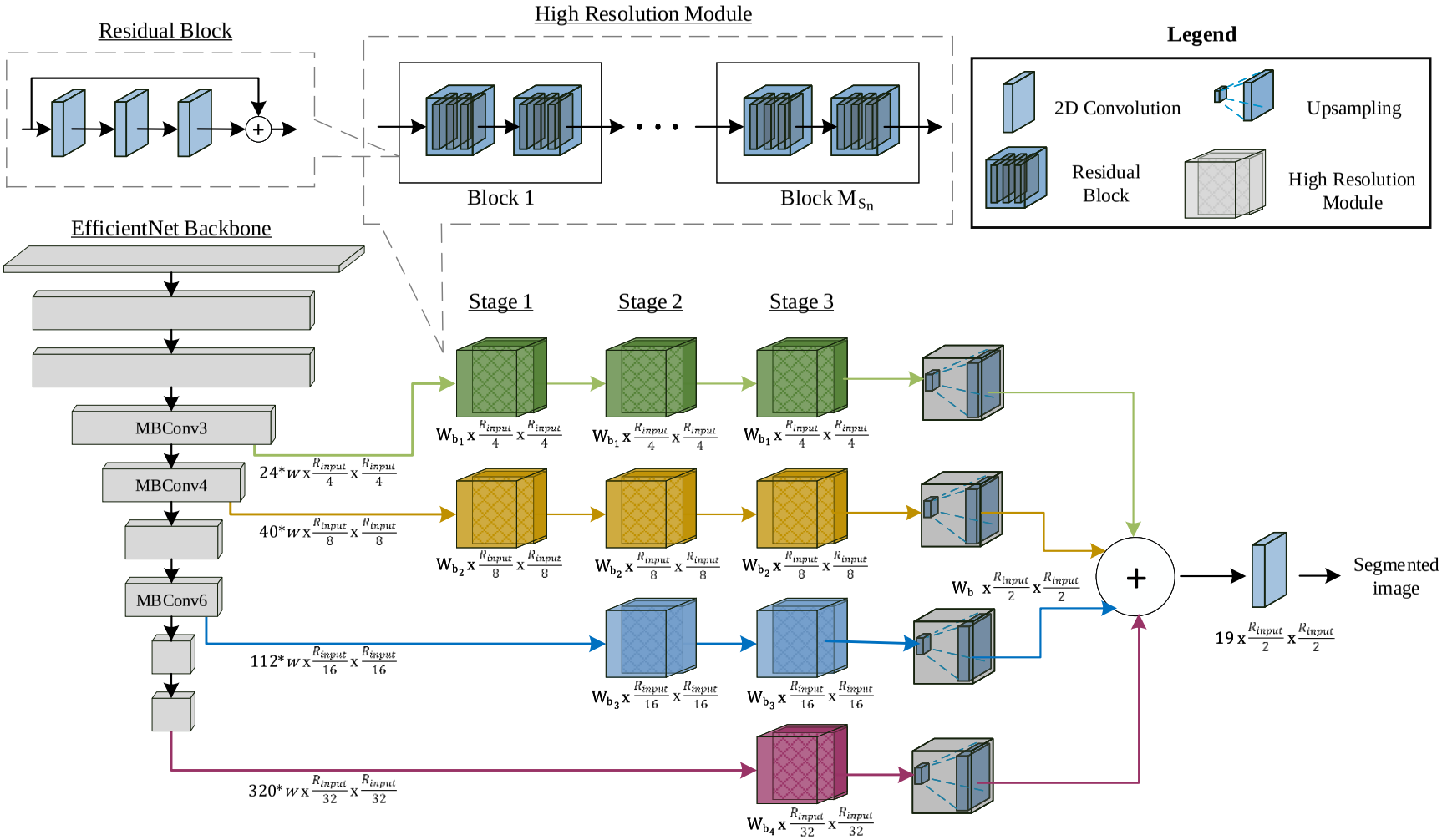

Figure 3.2: A detailed illustration of the EfficientHRNet architecture for segmentation. The backbone EfficientNet, a High-Resolution Network with three stages and four branches (denoted by different colors) remain the same, but the Heatmap Prediction Network is now replaced by Segmentation Network showing that EfficinetHRNet is completely scalable, allowing network complexity to be customized for target applications.

3.2 Compound Scaling Method

In this part of EffcientHRNet, the compound scaling method is described, which jointly scales all parts of EfficientHRNet, as seen in Figure 3.1. The aim is to develop a family of models optimized for both accuracy and efficiency, which can be scaled to meet a diverse set of memory and compute constraints.

Previous works on bottom-up pose estimation mostly scale the base network by using bigger backbone networks like ResNet [63] and VGGNet [64], using large input image sizes, or using multi-scale training to achieve high accuracies. However, these methods rely on scaling only a single dimension, which has limited effectiveness. Recent works [65], [66] show notable performance on image classification by jointly scaling the width, depth, and input image resolution. Inspired by EfficientNet, EfficientDet [67] proposes a similar compound scaling method for object detection, which jointly scales the backbone network, multi-scale feature network, and the object detector network. In this study, a heuristic-based compound scaling method for bottom-up pose estimation is proposed which is then applied to segmentation, based on [65], [67], using a scaling coefficient ϕ to jointly scale the Backbone Network, the High-Resolution Network, and the Heatmap Prediction Network/Segmentation Network. More precisely, the EfficientNet backbone is scaled down below the baseline and scale down the overall network in order to maintain near state-of-the-art accuracy while creating lightweight and flexible networks.

Backbone Network. The same width and depth scaling coefficients are maintained as in EfficientNet [65]. In order to meet the demands of running models on constrained devices, a new formulation for scaling EfficientNet below the baseline into a more compact model is provided.

Table 3.1: Efficient scaling configs for EfficientHRNet

Model	Input size (R_{input})	Backbone network	Width per branch (W_{b_1}, W_{b_2}, W_{b_3}, W_{b_4})	# blocks per stage (M_{s_2}, M_{s_3}, M_{s_4})	Tags (T_{size})	Heatmaps (H_{size})
H_0 ($\phi = 0$)	512	$B0$	32, 64, 128, 256	1, 4, 3	128	256
H_{-1} (ϕ = -1)	480	B_{-1}	26, 52, 103, 206	1, 3, 3	120	240
H_{-2} (ϕ = -2)	448	B_{-2}	21, 42, 83, 166	1, 2, 3	112	224
H_{-3} (ϕ = -3)	416	B_{-3}	17, 34, 67, 133	1, 1, 3	104	208
H_{-4} (ϕ = -4)	384	B_{-4}	14, 27, 54, 107	1, 1, 2	96	192

Starting with the baseline EfficientNet-B0 scaling coefficients:

$$\begin{aligned} depth : d &= 1.2^{\phi} \\ width : w &= 1.1^{\phi} \\ resolution : r &= 1.15^{\phi} \end{aligned} \tag{3.4}$$

ϕ, i.e. ϕ = -1, -2, -3, -4, is inverted to calculate the scaling multipliers for the compact EfficientNet models, which is symbolized as B_{-1}, B_{-2}, B_{-3} and B_{-4} respectively. As an example, in order to take the baseline resolution, 224, and scale it down for compact EfficientNet B_{-1} model, take r, from Equation 3.4, with $\phi = -1$. This would result in a resolution scaling coefficient of 1.15^{-1}, i.e. 0.87, leaving a scaled resolution size of $ceil(224 * 0.87) = 195$. This pattern repeats for B_{-2} — B_{-4}, and can be seen in Table 4.1. These compact EfficientNet models (B_{-1} to B_{-4}) are trained on ImageNet and the resulting weights are used for the Backbone Network in EfficientHRNet models.

High-Resolution Network. The High-Resolution Network has three stages and four branches with four different feature map sizes. Each branch n also has a different width W_{b_n} and our baseline H_0 model has a width of 32, 64, 128, and 256 for each branch respectively. We selectively pick a width scaling factor of 1.25 and scale down the width using the following equation:

$$W_{b_n} = (n \cdot 32) \cdot (1.25)^{\phi} \tag{3.5}$$

where n is a particular branch number and ϕ is the compound scaling coefficient.

Furthermore, within each stage, each *high resolution module* has multiple *blocks* M_{s_n} which repeat a number of times, as seen in Table 3.1. In our baseline H_0 model, *blocks* within each stage repeat 1, 4, and 3 times respectively. We found that the number of repetitions in stage 3 had the largest impact on accuracy. Therefore, the number of repetitions within a *high resolution module* M_{s_2} decreases linearly as the models are scaled down, starting with stage 2 until reaching a single repetition and then moving on to stage 3, as shown in Table 3.1.

Heatmap Prediction Network. The *DeConv block* is scaled in the same manner as the width of the High Resolution Network (Equation 3.5). The Heatmap Prediction Network outputs tags and heatmaps whose width remains fixed across all the models.

Segmentation Network. The Segmentation Network head involves upsampling and concatenating the outputs from the High-Resolution Network and the final output is the the segmented image along with 19 labeled classes which remain the same for all the scalable models.

Input Image Resolution. The EfficientNet layers downsample the original input image resolution by 32 times. Thus, the input resolution of EfficientHRNet must be dividable by 32, and is linearly scaled down as shown in Equation 3.6.

$$R_{input} = 512 + 32 \cdot \phi \tag{3.6}$$

Based on Equations 3.4, 3.5, and 3.6, a group of pose estimation models are developed from H_0 to H_{-4} called EfficientHRNet, as shown in Table 3.1.

CHAPTER 4: EXPERIMENTAL RESULTS

In this section, the method to evaluate scaling EfficientNet below the baseline through classification on the popular ImageNet [69] and CIFAR-100 [70] datasets is explored. Then, an exhaustive evaluation of five different EfficientHRNet models on the challenging COCO dataset is done and the results are compared to state-of-the-art methods. Results on EfficientHRNet Segmentation are also presented. In order to see the performance on edge devices, FPS numbers are reported for EfficientHRNet Pose Estimation and Segmentation. Finally, a qualitative evaluation of EfficientHRNet is presented, illustrating both where the models excel and where they fall short.

4.1 Classification for Compact EfficientNet

Dataset. ImageNet [69] has been a long time standard benchmark for object classification and detection thanks to its annual contest, the ImageNet Large Scale Visual Recognition Challenge, that debuted in 2010. The challenge uses a subset of the full dataset with over a million images spread out over 1000 object classes. For training, validating, and testing purposes, the trimmed ImageNet is divided into three sets: 800k images will be used for training the network, 150k will be used for validation after each epoch, and 50k will be used for testing the fully trained model. CIFAR-100 [70] consists of 100 object classes each with 500 images for training, and 100 for testing. This relatively small dataset helps illuminate the lightweight models, which start to struggle with the larger ImageNet as ϕ decreases, designed for resource constrained devices that might not need to classify as many object classes.

Training. Random rotation, random scale, and random aspect ratio is used to crop the input images to the desired resolutions based on the current EfficientNet model.

Color jitter was also used to randomly change the brightness, contrast, saturation, and hue of the RGB channels using principle component analysis [71]. The images are then normalized using per channel mean and standard deviation. Each model was trained using Stochastic Gradient Descent [72] with a weight decay of $1e-4$. The weights were initialized using the Xavier algorithm [73] and underwent five warm-up epochs with a learning rate of $1e-4$ that increased linearly until it reached 0.05. The networks were then trained for an additional 195 epochs and followed the step decay learning rate scheduler [74] that reduces the learning rate by a factor of 10 every 30 epochs.

Testing. The compact EfficientNet models were tested for accuracy based on their respective testsets. For a fair comparison, the number of ImageNet test samples were reduced to 10,000 to match the test set of CIFAR-100, where the batch size is set to 1. These results can be seen in Table 4.1.

Results on ImageNet and CIFAR-100. Looking at B_{-1} there is a 15% reduction in parameters and 25% reduction in operations, yet an accuracy drop of only 1.2% and 0.5% on ImageNet and CIFAR-100 respectively. More impressively, B_{-2} sees a 35-40% reduction in parameters and a 50% reduction in operations, yet only a 3.7% and 2.1% drop in accuracy on the two datasets. This minor accuracy loss is negligible compared to the massive reduction in model size and computation, allowing for much faster inference as well as deployment on low-power and resource constrained devices. In the most extreme, B_{-4} shows a parameter reduction of 68-75% and a 87.5% decrease in operations while having an accuracy drop of 9.4% and 7.6% on ImageNet and CIFAR-100. While the accuracy drop is a bit more significant here,

Table 4.1: Compact EfficientNet performance on ImageNet and CIFAR-100 datasets.

		ImageNet			CIFAR-100		
Model	Input size	# Params	FLOPS	Top-1	# Params	FLOPS	Top-1
B0 (ϕ = 0)	224	5.3M	0.4B	75	4.1M	0.4B	81.9
B_{-1} (ϕ = -1)	195	4.5M	0.3B	73.8	3.5M	0.3B	81.4
B_{-2} (ϕ = -2)	170	3.4M	0.2B	71.3	2.5M	0.2B	79.8
B_{-3} (ϕ = -3)	145	2.8M	0.1B	68.5	1.9M	0.1B	78.2
B_{-4} (ϕ = -4)	128	1.3M	0.05B	65.6	1.3M	0.05B	74.3

the massive reduction in computation allows for much more flexibility when it comes to deployment in systems where a lightweight approach is needed. This gives a solid foundation on which to build EfficientHRNet.

4.2 2D Human Pose Estimation for EfficientHRNet

Dataset. COCO dataset [2] has over 200,000 images with 250,000 person instances each labeled with 17 keypoints. The COCO dataset has three sets - *train* set with 57k images, *val* set with 5k images and *test*, which is divided into *test-dev* with 20k images and *test-challenge* with 20k images. The training is performed on all the training on *train* set, report results on *val* set, and compare with state-of-the-art methods on *test-dev* set for a fair comparison.

Evaluation. The COCO evaluation defines object keypoint similarity (OKS), and uses mean average precision (AP) over 10 OKS thresholds as the main evaluation metric[1]. The OKS is calculated from the scale of the person and the Euclidean distance between the GT and predicted points, similar to IoU in object detection. For the results, average precision and recall scores is reported: AP (mean of AP scores at OKS = 0.50, 0.55,$\cdots$, 0.90, 0.95), AP^{50} (AP at OKS = 0.50), AP^{75}, AP^{M} for medium objects, AP^{L} for large objects, and AR (mean of recall scores).

Training. Random rotation, random scale, and random translation for data augmentation is used to crop the input images to a fixed input resolution depending on

[1]http://cocodataset.org/#keypoints-eval

Table 4.2: Comparisons with bottom-up methods on COCO2017 val dataset

		single-scale			multi-scale				
Model	Input size	AP	AP^{50}	AP^{75}	AP	AP^{50}	AP^{75}	# Params	FLOPs
PersonLab	1401	66.5	86.2	71.9	-	-	-	68.7M	405.5B
HRNet	512	64.4	-	-	-	-	-	28.5M	38.9B
HigherHRNet	512	67.1	86.2	73.0	69.9	87.1	76.0	28.6M	47.9B
Lightweight OpenPose	368	42.8	-	-	-	-	-	4.1M	9.0B
H_0 (ϕ = 0)	512	64.8	85.3	70.7	68.1	87.0	74.1	23.3M	25.6B
H_{-1} (ϕ = -1)	480	59.2	82.6	64.0	63.2	84.3	68.6	16M	14.2B
H_{-2} (ϕ = -2)	448	52.9	80.5	59.1	56.4	82.2	63.4	10.3M	7.7B
H_{-3} (ϕ = -3)	416	44.8	76.7	48.2	46.4	76.6	50.8	6.9M	4.2B
H_{-4} (ϕ = -4)	384	35.7	69.6	33.7	40.3	73.0	41.9	3.7M	2.1B

the EfficientHRNet model. Following HigherHRNet [1], two ground truth heatmaps of different sizes, $\frac{1}{2}$ and $\frac{1}{4}$ of the original input size respectively are generated. Each EfficientHRNet model is trained using Adam optimizer [75] and weight decay of $1e-4$. All models from H_0 to H_{-4} are trained for a total of 300 epochs with a base learning rate of $1e-3$, decreasing to $1e-4$ and $1e-5$ at 200^{th} and 260^{th} epochs respectively. To maintain balance between heatmaps loss and grouping loss, the losses are weighted at 1 and $1e-3$ respectively.

Testing. For testing on COCO *val* and *test-dev* sets, the short side of test input image is resized to match our input resolution while preserving the aspect ratio. As in HigherHRNet [1], heatmap aggression is done by resizing the predicted heatmaps to the input resolution and taking the average. The models are tested using both single scale and multi-scale heatmaps, as is common. Following [29], the output detection heatmaps across different scales are averaged and the tags are concatenated into higher dimensional tags, making different objects and persons considerably more scale-invariant.

Results on COCO2017 *val*. The accuracy of EfficientHRNet H_0 to H_{-4} is reported on COCO *val* set along with parameters and FLOPs of the entire network and compare it with other bottom-up methods. As summarized in Table 4.2, the baseline H_0 model outperforms HRNet [1] with 0.4% more accuracy, 18% fewer parameters and 34% fewer FLOPs. H_{-2} and H_{-3} models outperform Lightweight OpenPose [5] in accuracy while having fewer FLOPs. H_{-4} has the worst accuracy of any model in Table 4.2, however it boast both the smallest model size and fewest number of operations, that later seeing an over 75% reduction from its lightest weight competitor. This makes our scaled-down models the new state-of-the-art for lightweight bottom-up human pose estimation.

Results on COCO2017 *test-dev*. Table 4.3 compares EfficientHRNet with other bottom-up pose estimation methods on COCO *test-dev* set. The baseline H_0 model

with single-scale testing serves as an efficient and accurate model for bottom-up methods as it is almost comparable to HRNet [1] in accuracy, losing by only 0.1%, while having a smaller model size and fewer FLOPs. H_0 outperforms Hourglass [13] in both single-scale and multi-scale testing by 7.4% and 1.6% respectively, with H_0 remarkably having about 10% the model size and number of FLOPs as Hourglass. In all cases where H_0 loses in accuracy, it more than makes up for it in a reduction in parameters and operations. Additionally, our H_{-1} model, with only 16M parameters and 14.2B FLOPs, outperforms both OpenPose [26, 27] and Hourglass [13], demonstrating EfficientHRNet's efficiency and suitability for low-power and resource constrained devices.

As the EfficientHRNet models are scaled down using the compound scaling method, somewhat minor drops in accuracy with significant drops parameters and FLOPs as compared to the baseline H_0 model are seen. H_{-1} has 31.3% less parameters and 44.5% less FLOPs as compared to H_0 while only being 4.9% less accurate. Similarly,

Table 4.3: Comparisons with state-of-the-art bottom-up methods on COCO2017 test-dev dataset.

Method	Backbone	Input size	# Params	FLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
w/o multi-scale test										
OpenPose	-	-	25.94M	160B	61.8	84.9	67.5	57.1	68.2	66.5
Hourglass	Hourglass	512	277.8M	206.9B	56.6	81.8	61.8	49.8	67.0	-
PersonLab	ResNet-152	1401	68.7M	405.5B	66.5	88.0	72.6	62.4	72.3	-
PifPaf	ResNet-152	-	-	-	66.7	-	-	62.4	72.9	-
HRNet	HRNet-W32	512	28.5M	38.9B	64.1	86.3	70.4	57.4	73.9	-
HigherHRNet	HRNet-W32	512	28.6M	47.9B	66.4	87.5	72.8	61.2	74.2	-
HigherHRNet	HRNet-W48	640	63.8M	154.3B	68.4	88.2	75.1	64.4	74.2	-
H_0 (ϕ=0)	$B0$	512	23.3M	25.6B	64.0	86.2	70.1	59.1	71.1	69.1
H_{-1} (ϕ=-1)	B_{-1}	480	16M	14.2B	59.1	83.9	66.4	54.6	65.7	64.7
H_{-2} (ϕ=-2)	B_{-2}	448	10.3M	7.7B	52.8	82.3	58.5	47.3	60.6	59.1
H_{-3} (ϕ=-3)	B_{-3}	416	6.9M	4.2B	44.5	78.0	47.6	39.8	51.0	51.8
H_{-4} (ϕ=-4)	B_{-4}	384	3.7M	2.1B	35.5	71.1	32.5	29.9	43.5	42.8
w/ multi-scale test										
Hourglass	Hourglass	512	277.8M	206.9B	63.0	85.7	68.9	58.0	70.4	-
Hourglass	Hourglass	512	277.8M	206.9B	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab	ResNet-152	1401	68.7M	405.5B	68.7	89.0	75.4	64.1	75.5	75.4
HigherHRNet	HRNet-W48	640	63.8M	154.3B	70.5	89.3	77.2	66.6	75.8	74.9
H_0 (ϕ=0)	$B0$	512	23.3M	25.6B	67.1	88.1	73.6	63.2	72.5	71.8
H_{-1} (ϕ=-1)	B_{-1}	480	16M	14.2B	62.3	85.1	67.8	58.2	67.8	67.4
H_{-2} (ϕ=-2)	B_{-2}	448	10.3M	7.7B	55.0	83.1	61.8	51.4	60.0	61.4
H_{-3} (ϕ=-3)	B_{-3}	416	6.9M	4.2B	45.5	78.2	49.4	42.7	48.9	53.1
H_{-4} (ϕ=-4)	B_{-4}	384	3.7M	2.1B	39.7	74.2	39.7	35.7	45.5	47.0

Table 4.4: EfficientHRNet Pose Estimation Performance Evaluation

Model	Input size	# Params	FLOPS	Accuracy	Runtime	FPS
HigherHRNet	512	28.6M	47.9B	67.1%	149.6ms	6.68
Lightweight OpenPose	512	10.3M	22.8B	43.1%	59.84ms	16.71
H_0	512	23.3M	25.6B	64.8%	62.68ms	15.95
H_{-1}	480	16M	14.2B	59.2%	76.01ms	13.15
H_{-2}	448	10.3M	7.7B	52.9%	56.76ms	17.61
H_{-3}	416	6.9M	4.2B	44.8%	44.74ms	22.34
H_{-4}	384	3.75M	2.1B	35.7%	27.15ms	36.82

the lightest model H_{-4} is 84% smaller and has 91.7% less FLOPs, with a less than 45% drop in accuracy. Interestingly, EfficientHRNet is the only bottom-up pose estimator that is able to provide such lightweight models while still having accuracies that are comparable to state-of-the-art bottom-up methods, as illustrated by both Table 4.3 and Fig 1.2. These results nicely show the validity of this approach to scalability and efficiency in EfficientHRNet, and its suitability as a lightweight human pose estimator. Table 4.4 provides the performance comparison of EfficientHRNet models with HigherHRNet [1] and Mobilenet v2 based Lightweight OpenPose [5]. While it as shown in the previous table 4.3, that the EfficientHRNet models were highly efficient than the state-of-the art models, here it is shown by inferencing the models on Nvidia Xavier embedded platform. While HigherHRNet [1] runs at 6fps, the H_0 model run at more than double speed and the smallest model runs 6x faster. Similarly, Lightweight OpenPose runs at 16 fps and it's near comparable model (based on parameters and FLOPs) run 5.6% faster. These results validates the efficiency and performance of EfficientHRNet models for pose estimation. Next, results of EfficientHRNet models for segmentation will be discussed.

4.3 Image Segmentation for EfficientHRNet

Dataset. Cityscapes dataset [68] is comprised of a large, diverse set of stereo video sequences recorded in streets from 50 different cities. 5000 of these images have high quality pixel-level annotations; 20000 additional images have coarse annotations to enable methods that leverage large volumes of weakly-labeled data. The finely an-

Table 4.5: EfficientHRNet Segmentation on Cityscapes dataset.

Model	Input size	Backbone	# modules	mIOU	# Params	FLOPS	FPS
HRNet	512x1024	HRNet-W48	1,4,3	79.2%	65.7M	692.2B	-
HRNet v1	512x1024	HRNet-W18	1,1,1	70.3%	1.5M	31.1B	5.92
HRNet v2	512x1024	HRNet-W18	1,3,2	76.2%	3.9M	71.6B	3.46
EfficientHRNet Seg	512	B_0	1,3,2	71%	19.3M	18.9B	$\sim$ 18
EfficientHRNet Seg	512x1024	B_0	1,4,3	74.3%	24.2M	70.5B	-

notated images are divided into 2,975/500/1,525 images for training, validation and testing. There are 30 classes out of which 19 are used for evaluation.

Evaluation. Mean Intersection-Over-Union (mIOU) is used for evaluation which is a common evaluation metric for semantic image segmentation, which first computes the IOU for each semantic class and then computes the average over classes.

Training. Data augmentation like random rotation, scaling and crop is used to get the input image into a fixed input resolution. The segmentation model is trained for a total of 450 epochs with a learning rate of 0.01. SGD optimzer [72] is used to backpropagate the classification loss.

Testing. For testing on Cityscapes dataset, the original image is resized to match the input resolution while preserving the aspect ratio. The segmentation model is evaluated on single-scale and the results are described below.

Results on Cityscapes. The accuracy of EfficientHRNet Segmentation is reported on Cityscapes test dataset along with parameters and FLOPs of the entire network and are compared with HRNet based segmentation [62]. As shown in Table 4.5, EfficientHRNet Segmentation model is much more efficient than its direct competitors based on HRNet backbone. For instance, HRNet has 692.2B FLOPS while EfficientHRNet has just 18.9B with only 8% drop in accuracy. This is significant for inference on edge devices. While HRNet, the biggest model does not even an inference on Nvidia Xavier, the EfficientHRNet model is expected to run at around 18 fps. This shows that EfficientHRNet models are much better for edge devices which will also be shown in the qualitative results discussed in the next section.

4.4 Qualitative Results

In this section, qualitative results for EfficientHRNet Pose estimation models are presented on COCO dataset. Also, the results on our own dataset is presented which shows pose estimation and segmentation models running in the pipeline along with tracking and person re-identification.

4.4.1 Qualitative Results of EfficientHRNet on COCO

To show how EfficientHRNet models perform on COCO dataset, qualitative results are presented on the *test-challenge* set. Fig. 4.1 shows simple, medium, and complex examples for all EfficientHRNet models from H_0 to H_{-4}. For the simple case, it can be seen that a single person pose is correctly formed for all our models, but as the smaller models are used for evaluation, the detected keypoints distort the pose due to the drop in accuracy. In the case of medium complexity examples, where there is an occlusion over multiple persons, it is observed that all our models are able to accurately detect different people in the image. However, smaller models, such as H_{-3} and H_{-4}, detect multiple keypoints for the same person, making the pose look clustered and creating noise. As far as complex examples are concerned, an image with more than 4 people are shown, all in different poses at different distances from the

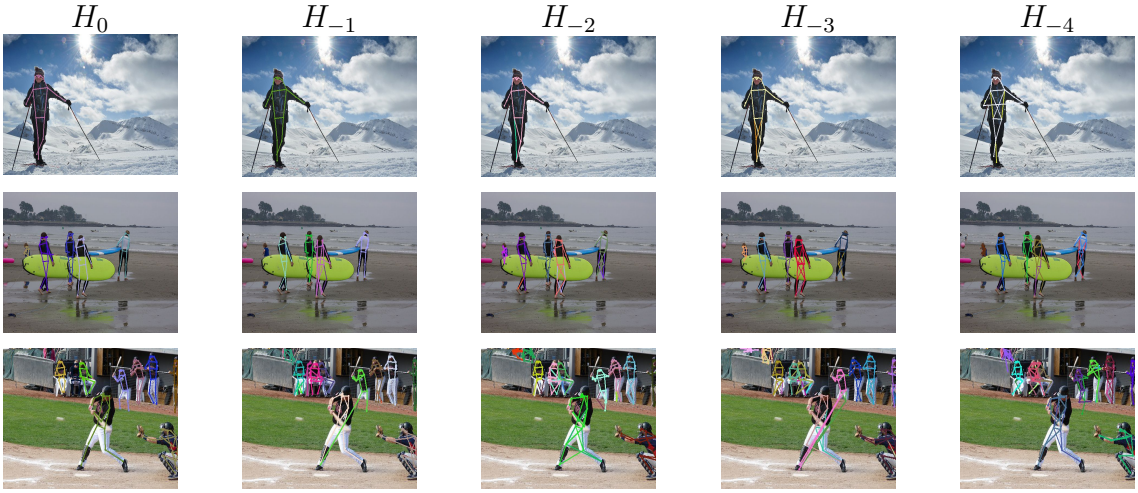

Figure 4.1: Qualitative Results for EfficientHRNet models on COCO2017 test dataset. Top to bottom: Simple, Medium and Complex examples.

camera. It is observed that the quality of poses is affected but all the models are still able to identify all the humans in the image. The smaller models particularly struggle with merging keypoints across multiple people and detecting multiple keypoints for the same person, just as in the medium example. These qualitative results nicely visualize the relationship between detected keypoints and model size, and show that the overall effect on accurate pose prediction is not significant in simple and medium examples when the evaluation is done on smaller models. However, complex scenes can still be a struggle for lightweight pose estimation, and additional post processing might be needed to redress the output for use in a real-world system.

4.4.2 Qualitative Results of EfficientHRNet in video analytics pipeline

Figure 4.2: EfficientHRNet Pose Estimation and Segmentation in Vision pipeline. Left: Pose Estimation, Right: Segmentation.

In order to demonstrate the quality of EfficientHRNet models, the pose estimation and segmentation H0 models were integrated into the vision pipeline - as a part of

National Science Foundation Smart City project. The dataset was created as a part of this research and was used with consent for this particular study. The entire vision pipeline includes applications like pose estimation, segmentation, tracking, person re-identification and action detection. As shown in Fig. 4.2, the leftmost images show EfficientHRNet pose estimation running on a single frame from a sequence of frames. The rightmost images show EfficientHRNet segmentation. These results demonstrate the applicability of our models on different datasets with people of small scale, occluded scenes, and overall a completely different application, surveillance system in this case.

4.4.3 Challenges in the current pipeline

The keypoint grouping as given in [29], the tags are compared from a given joint to the tags of the current pool of people, and try to determine the best matching between them. Two tags can only be matched if they fall within a specific threshold. In addition, this methods wants to prioritize matching of high detections and thus perform a maximum matching where the weighting is determined by both the tag distance and the detection score. If any new detection is not matched, it is used to start a new person instance. This accounts for cases where perhaps only a leg or hand is visible for a particular person. However, this does not deal with occlusion or when two people are very close to each other. This issue was solved by implementing a function which removes redundant poses. It filters out poses whose joints have, on average, a difference lower than 3 pixels. This is useful when grouping the joins together skeleton parts belonging to the same people (but then it does not remove redundant skeletons). The other challenge is people not detected at a far-away distance. This is because the training was on COCO which has objects close to the camera. Also, for segmentation, the classes like doors are not present and these two issues can be solved by creating own dataset or combining multiple datasets.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

5.1 Conclusion

In this paper, EfficientHRNet, a family of scalable networks for high-resolution and efficient bottom-up multi-person pose estimation and segmentation is presented, especially for low-power edge devices. The principles of state-of-the-art EfficientNet and HRNet are unified to create a network architecture for lightweight human pose estimation, and a new compound scaling method is proposed that jointly scales down the input resolution, backbone network, and high-resolution feature network. EfficientHRNet is not only more efficient than all other bottom-up human pose estimation methods, but it can maintain accuracy competitive with state-of-the-art models on the challenging COCO dataset. Along with efficiency, it can also be customized for other computer vision applications like segmentation and can be used for edge video analytics pipeline in order to run an end-to-end system. Remarkably, EfficientHRNet can achieve this near state-of-the-art accuracy with fewer parameters and less computational complexity than other bottom-up multi-person pose estimation networks. EfficientHRNet is also able to achieve high performance on edge devices and embedded platforms like Nvidia Xavier and Nano. Thus, this study provides models and networks which can highly benefit the end-to-end working of a system, like surveillance video system which runs multiple applications together and require high performance.

5.2 Future Work

There are several directions in which this work could be continued in the future. In this study, the focus was on creating lightweight human pose estimation and segmen-

tation and thus the network was scaled down from the baseline. However, there is nothing preventing EfficientHRNet from being scaled to a larger network in a fashion more similar to EfficentNet [65] or EfficientDet [67]. In addition, the following principles are shown in EfficientNet-lite, certain architectural elements of the EfficientNet backbone, mainly the squeeze and excite layers, as well as the swish activation, could be removed or replaced. An exploration of how this impacts inference on different hardware would be a potential avenue of research. Finally, while the focus was on 2D human pose estimation and segmentation, the principles and architecture of EfficientHRNet could be applied to numerous other computer vision tasks, such as object detection, and depth estimation, to name a few examples. The other important future of this thesis can be combining the pose estimation and segmentation model into one network i.e. sharing the backbone and potentially the High Resolution Network to reduce the redundant computations and enable knowledge distillation to learn from the stand-alone teacher models.

REFERENCES

[1] 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," 2019.

[2] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. DollÃ¡r, "Microsoft coco: Common objects in context," 2014.

[3] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *CoRR*, vol. abs/1902.09212, 2019.

[4] G. Papandreou, T. Zhu, L. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," *CoRR*, vol. abs/1803.08225, 2018.

[5] D. Osokin, "Real-time 2d multi-person pose estimation on CPU: lightweight openpose," *CoRR*, vol. abs/1811.12004, 2018.

[6] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR 2011*, pp. 1385–1392, 2011.

[7] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3041–3048, 2013.

[8] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR 2011*, pp. 1465–1472, 2011.

[9] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3674–3681, 06 2013.

[10] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik, "Articulated pose estimation using discriminative armlet classifiers," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR â13, (USA), p. 3342â3349, IEEE Computer Society, 2013.

[11] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *CoRR*, vol. abs/1312.4659, 2013.

[12] A. Jain, J. Tompson, M. Andriluka, G. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," 12 2013.

[13] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016.

[14] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," *CoRR*, vol. abs/1609.01743, 2016.

[15] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," *CoRR*, vol. abs/1605.02914, 2016.

[16] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *CoRR*, vol. abs/1602.00134, 2016.

[17] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," *CoRR*, vol. abs/1608.08526, 2016.

[18] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. P. Murphy, "Towards accurate multi-person pose estimation in the wild," *CoRR*, vol. abs/1701.01779, 2017.

[19] S. Huang, M. Gong, and D. Tao, "A coarse-fine network for keypoint localization," pp. 3047–3056, 10 2017.

[20] H. Fang, S. Xie, Y. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2353–2362, 2017.

[21] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," *CoRR*, vol. abs/1711.07319, 2017.

[22] B. Cheng, Y. Wei, H. Shi, R. S. Feris, J. Xiong, and T. S. Huang, "Decoupled classification refinement: Hard false positive suppression for object detection," *CoRR*, vol. abs/1810.04002, 2018.

[23] B. Cheng, Y. Wei, H. Shi, R. S. Feris, J. Xiong, and T. S. Huang, "Revisiting RCNN: on awakening the classification power of faster RCNN," *CoRR*, vol. abs/1803.06799, 2018.

[24] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016.

[25] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[26] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1611.08050, 2016.

[27] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018.

[28] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," *CoRR*, vol. abs/1903.06593, 2019.

[29] A. Newell and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *CoRR*, vol. abs/1611.05424, 2016.

[30] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," pp. 1014 – 1021, 07 2009.

[31] 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," *CoRR*, vol. abs/1511.06645, 2015.

[32] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," *CoRR*, vol. abs/1605.03170, 2016.

[33] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Articulated multi-person tracking in the wild," *CoRR*, vol. abs/1612.01465, 2016.

[34] M. Kocabas, S. Karagoz, and E. Akbas, "Multiposenet: Fast multi-person pose estimation using pose residual network," *CoRR*, vol. abs/1807.04067, 2018.

[35] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016.

[36] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," *CoRR*, vol. abs/1511.03339, 2015.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016.

[38] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016.

[39] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," *CoRR*, vol. abs/1708.01101, 2017.

[40] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," *CoRR*, vol. abs/1711.07319, 2017.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[43] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," *CoRR*, vol. abs/1505.04366, 2015.

[44] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," *CoRR*, vol. abs/1611.08323, 2016.

[45] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018.

[46] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," *CoRR*, vol. abs/1611.06612, 2016.

[47] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters - improve semantic segmentation by global convolutional network," *CoRR*, vol. abs/1703.02719, 2017.

[48] Z. Zhang, X. Zhang, C. Peng, D. Cheng, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," *CoRR*, vol. abs/1804.03821, 2018.

[49] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018.

[50] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," *CoRR*, vol. abs/1807.10221, 2018.

[51] J. Fu, J. Liu, Y. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *CoRR*, vol. abs/1708.04943, 2017.

[52] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," *CoRR*, vol. abs/1703.00862, 2017.

[53] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," *CoRR*, vol. abs/1702.07432, 2017.

[54] L. Ke, M. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," *CoRR*, vol. abs/1803.09894, 2018.

[55] W. Tang, P. Yu, and Y. Wu, *Deeply Learned Compositional Models for Human Pose Estimation: 15th European Conference, Munich, Germany, September 8â14, 2018, Proceedings, Part III*, pp. 197–214. 09 2018.

[56] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," *CoRR*, vol. abs/1804.06208, 2018.

[57] S. Saxena and J. Verbeek, "Convolutional neural fabrics," *CoRR*, vol. abs/1606.02492, 2016.

[58] Y. Zhou, X. Hu, and B. Zhang, "Interlinked convolutional neural networks for face parsing," *CoRR*, vol. abs/1806.02479, 2018.

[59] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. TremÃ©au, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," 01 2017.

[60] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense convolutional networks for efficient prediction," *CoRR*, vol. abs/1703.09844, 2017.

[61] J. Wang, S. ke, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," 08 2019.

[62] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *CoRR*, vol. abs/1904.04514, 2019.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[65] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019.

[66] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. DollÃ¡r, "Designing network design spaces," 2020.

[67] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," 2019.

[68] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016.

[69] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[70] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., 2009.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPSâ12, (Red Hook, NY, USA), p. 1097â1105, Curran Associates Inc., 2012.

[72] S. Ruder, "An overview of gradient descent optimization algorithms," *CoRR*, vol. abs/1609.04747, 2016.

[73] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 13–15 May 2010.

[74] R. Ge, S. M. Kakade, R. Kidambi, and P. Netrapalli, "The step decay schedule: A near optimal, geometrically decaying learning rate procedure," *CoRR*, vol. abs/1904.12838, 2019.

[75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.