EFFICIENT POSEVIT: EXPLORING EFFICIENT VISION TRANSFORMERS AND TRAINING STRATEGIES FOR FAST HUMAN POSE ESTIMATION

by

Diwas Lamsal

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Computer Science

Examination Committee: Prof. Matthew N. Dailey (Chairperson)
Dr. Mongkol Ekpanyapong
Dr. Chaklam Silpasuwanchai

Nationality: Nepalese
Previous Degree: Bachelor of Science in Computing in Software Engineering
University of Northampton
UK

Scholarship Donor: AIT Scholarships

Asian Institute of Technology
School of Engineering and Technology
Thailand
May 2023

AUTHOR'S DECLARATION

I, Diwas Lamsal, declare that the research work carried out for this thesis was in accordance with the regulations of the Asian Institute of Technology. The work presented in it are my own and has been generated by me as the result of my own original research, and if external sources were used, such sources have been cited. It is original and has not been submitted to any other institution to obtain another degree or qualification. This is a true copy of the thesis including final revisions.

Date: November 30, 2022

Name: Diwas Lamsal

Signature:

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Professor Matthew N. Dailey, for his continuous guidance, support, and encouragement throughout the course of my thesis. His expertise and insights have been invaluable in shaping my research and helping me grow as a researcher.

I would also like to extend my sincere appreciation to my committee members, Dr. Mongkol Ekpanyapong and Dr. Chaklam Silpasuwanchai, for their valuable suggestions and guidance in ensuring that my research stays on the right track. Their suggestions have been crucial in refining my research directions and enhancing the quality of my work.

I am grateful to the AIT AI Center for providing the necessary compute resources that enabled me to train and test the models. Their support was essential in facilitating the progress and completion of this work.

Lastly, I would like to thank Mr. Pramod for sharing his experience and suggestions and for his assistance in data collection.

ABSTRACT

Human pose estimation (HPE) serves as a critical component in various downstream applications, including activity recognition and human-computer interaction. However, it faces challenges such as occlusion, complex body poses, diverse clothing and environments, and low-quality images. State-of-the-art HPE methods adept at addressing these challenges often prove to be computationally expensive to run on low-cost edge devices. On the other hand, the efficient models designed to run on edge devices are not as accurate, especially when faced with occlusion. In this thesis, I explored the use of computationally efficient vision transformer (ViT) backbones for the task of HPE to produce an efficient model capable of running on edge devices while also addressing common HPE challenges. Specifically, I replaced the ViT encoder in ViTPose with efficient ViT variants, and performed a comprehensive comparison. The findings indicate that the EdgeNeXt-Base model offers an optimal speed-accuracy trade-off among the evaluated models. Through knowledge distillation and multi-task training strategies, the resulting models exhibit improved performance, particularly in handling occlusion. Additionally, I produced a wheelchair-based human pose estimation and activity recognition dataset aimed at developing a system that effectively monitors the activities of people sitting in wheelchairs.

CONTENTS

	Page
ACKNOWLEDGEMENTS	**iii**
ABSTRACT	**iv**
LIST OF TABLES	**vii**
LIST OF FIGURES	**viii**
CHAPTER 1 INTRODUCTION	**1**
1.1 Background of the Study	1
1.2 Statement of the Problem	2
1.3 Objectives of the Study	4
1.4 Limitations and Scope	5
CHAPTER 2 LITERATURE REVIEW	**6**
2.1 CNN-based HPE	6
2.2 Transformers in HPE	10
2.3 Vision Transformers	13
2.4 Adversarial Training and Data Augmentation	15
2.5 Video-Based HPE	16
2.6 Building Efficient Models	18
2.7 3D-HPE From Monocular Cameras	19
2.8 HPE Benchmarks and Datasets	20
2.9 2D Skeleton-Based Action Classification	21
2.10 Wheelchair-Based Pose Estimation and Activity Recognition	22
CHAPTER 3 METHODOLOGY	**25**
3.1 Research Methodology	25
3.2 Problem Formulation	27
3.3 ViTPose and Efficient ViTs	27
3.3.1 Pure Transformer-based ViT Variants	29
3.3.2 Hybrid ViT Variants	32
3.3.3 Models with best speed-accuracy trade-off on ImageNet	34
3.4 Training Strategies	35
3.4.1 Knowledge Distillation	35

3.4.2 Multi-task Training 36
3.4.3 GAN Setup 36
3.4.4 Additional Joint Vector Target 37
3.4.5 Other Techniques and Model Variants 37
3.5 Metrics for Evaluation 38
3.5.1 Object Keypoint Similarity 38
3.6 Dataset 39
3.7 Action Classification 41
CHAPTER 4 EXPERIMENTAL RESULTS 43
4.1 Pose Estimation Performance 43
4.1.1 Comparing Efficient ViT Variants 43
4.1.2 Comparison with state of the art 43
4.2 Knowledge Distillation Results 49
4.3 Improvements From Multi-Task Training 50
4.4 GAN Setup 51
4.5 Predicting Vector Fields as Additional Targets 51
4.6 Model Optimization and Other Techniques 52
4.6.1 Different Hyperparameters with T2TViT 52
4.6.2 Effect of Pretraining Distillation with TinyViT 52
4.7 Qualitative Analysis 53
4.8 Performance on the Wheelchair Dataset 53
4.8.1 HPE Performance on the Wheelchair Dataset 53
4.8.2 Action Classification Performance on the Wheelchair Dataset 57
CHAPTER 5 CONCLUSION AND RECOMMENDATIONS 59
5.1 Key Findings 59
5.2 Future Research Directions and Recommendations 60
REFERENCES 62

LIST OF TABLES

Tables **Page**

Table 3.1 Script used for data collection 40

Table 4.1 Pose Estimation Performance of Efficient ViT Variants 45

Table 4.2 Comparison with state of the art 46

Table 4.3 Inference Speed Comparison with State-of-the-art 47

Table 4.4 Improvements From Knowledge Distillation 49

Table 4.5 Improvements From Multi-Task Training 50

Table 4.6 Results of using GAN Setup 51

Table 4.7 Results from using Vector Fields as Additional Targets 51

Table 4.8 T2TViT-S Hyperparameters 52

Table 4.9 Testing Pretraining Distillation with TinyViT 53

Table 4.10 HPE Performance on the Wheelchair Dataset 55

Table 4.11 Action Classification Performance on the Wheelchair Dataset 56

Table 4.12 Performance of Propel and Fall on the Wheelchair Dataset 57

LIST OF FIGURES

Figures		Page
Figure 2.1	Hourglass Module	7
Figure 2.2	The HRNet Architecture	9
Figure 2.3	ViTPose Architecture	12
Figure 2.4	Vision Transformer Architecture	13
Figure 3.1	Research Methodology	26
Figure 3.2	Proposed Methodology	28
Figure 3.3	T2T Process in T2TViT	30
Figure 3.4	CrossViT Architecture	31
Figure 3.5	LeFF module of the CeiT architecture	32
Figure 3.6	LeViT architecture	33
Figure 3.7	Data Collection Camera Setup	41
Figure 3.8	Action Classification Framework	42
Figure 4.1	Pose Estimation Performance of Efficient ViT Variants	44
Figure 4.2	Comparison of Pose Estimation Performance with state of the art	45
Figure 4.3	CPU Inference Speed	48
Figure 4.4	GPU Inference Speed	48
Figure 4.5	Qualitative results on different datasets	54

CHAPTER 1
INTRODUCTION

In this chapter, I explore the fundamental and technical background of the thesis, with a primary focus on human pose estimation and a secondary emphasis on activity recognition for wheelchair users. Additionally, I outline the research questions, objectives, and the scope and limitations of the study, providing a foundation for the research methodology and subsequent chapters.

1.1 Background of the Study

With recent advances in deep learning and computer vision, problems such as classification, detection, and tracking have undergone tremendous improvements. Among these are the tasks of human pose estimation (HPE) and activity recognition. HPE attempts to estimate the human articulated posture from an image or video frame sequences. A common way to achieve this is to detect the location of different body joints (such as the elbow and knees) and connect them to form a 2D or 3D skeleton. HPE is commonly used in conjunction with classification techniques to recognize human activities (Du, Wang, & Wang, 2015a; Si, Chen, Wang, Wang, & Tan, 2019). HPE can also be used for applications beyond activity recognition, such as human tracking, movies, animation, human-computer interaction, surveillance, vision and language conversion, understanding of human behavior, autonomous driving, medical assistance and elder care (H. Chen et al., 2022; Y. Chen, Tian, & He, 2020).

In the medical domain, pose estimation has often been used in physical rehabilitation and therapy (Debnath, O'Brien, Yamaguchi, & Behera, 2021; Hagihara et al., 2020). Multiple attempts have been made to develop vision-based fall detection and activity recognition systems that involve some form of pose estimation (Feng, Yu, Naqvi, & Chambers, 2014; Nizam, Mohd, & Jamil, 2017). These systems are typically designed to help the elderly. It would be equally beneficial to assist people in wheelchairs using such activity recognition systems. Not only would this be helpful in automating smart healthcare and reducing costs (Postolache et al., 2014), but such a system can allow physicians to effectively monitor the general health conditions of people in wheelchairs (Ma, Li, Gravina, & Fortino, 2017). Most current systems for monitoring people in wheelchairs rely on sensors and smart wheelchairs (Ma et al., 2017), whereas vision-

based systems can be simpler, cheaper, and less intrusive. HPE methods enable the development of such vision-based systems to automate the task of monitoring people in wheelchairs.

Whether the individual is in a wheelchair or not, the key problem faced in pose estimation is the problem of occlusion and learning the local context in complex environments (Y. Chen et al., 2020). Although newer transformer architectures seem to be more robust to occlusion (Xu, Zhang, Zhang, & Tao, 2022) than previous approaches, they are expensive to train and expensive to run on low-cost edge devices. It is important to determine how best to deal with occlusion, as high accuracy, even under occlusion, is critical to the utility of a system in the medical domain. False positives will discourage people from using systems such as fall detection and activity recognition.

In this thesis, I explore ways to tackle the problem of partial occlusion in HPE while also focusing on an efficient vision-based pose and activity recognition system for people sitting in wheelchairs.

1.2 Statement of the Problem

A common way to apply HPE is to identify and label the body joint coordinates of a person and connect them to form 2D or 3D skeletons. Two-dimensional skeleton-based pose estimation methods are mainly divided into two categories: Top-down methods use a human detector to first detect and crop bounding boxes for each person; then a single-person pose estimation method is used to estimate 2D pose skeletons for each bounding box. Bottom-up methods, however, predict the location of each body joint in an image and connect them to form human skeletons. Increasing number of people in the image becomes computationally expensive with top-down methods, yet they provide much higher accuracy compared to bottom-up methods. These methods are suitable if we expect only a few people in an image and require better accuracy. The latter helps improve tasks such as action recognition (H. Chen et al., 2022).

There are several challenges in HPE. First, the flexibility of the human body allows complex poses that can cause self-occlusion. It is also common for objects or people to partially occlude the target person. Wheelchairs are likely to introduce occlusion whenever the person sitting does not directly face the camera. Similarly, people may appear in a variety of clothing, making it difficult to locate the keypoints. Humans may appear

close and entangled, making it difficult to connect the keypoints to obtain a skeleton (for bottom-up approaches) and for a person detector to detect individuals separately (for top-down approaches). Additionally, noise may be present in the images fed to the pose estimator as a result of blurring, incorrect exposure, or images taken with a low-quality camera. On the other hand, the most accurate models for pose estimation tend to be inefficient and almost impossible to run in real time (H. Chen et al., 2022; Y. Wang, Li, Cai, Chen, & Han, 2022). Crowded scenes further increase the computational complexity. Although efforts have been made to address these problems, it is challenging to find a perfect technique that works in every situation.

The current state-of-the-art method for HPE across multiple benchmarks uses a transformer-based architecture called ViTPose (Xu et al., 2022). It uses a plain vision transformer from Dosovitskiy et al. (2020) as an encoder and shows good performance even under occlusion. This is evident from its state-of-the-art performance on the OCHuman benchmark (S.-H. Zhang et al., 2019), which focuses primarily on occluded humans. The smallest model listed in the paper (ViTPose-B) beats the previous state-of-the-art methods and also provides a better speed-accuracy trade-off. My experiments show that multitask training, or the combination of multiple large-scale HPE datasets, is the primary driver behind this performance boost. Other transformer-based architectures, such as TransPose (S. Yang, Quan, Nie, & Yang, 2021) and HRFormer (Y. Yuan et al., 2021), have also shown improved performance with fewer parameters. This success may be attributed to the transformers' ability to capture long-range dependencies, which enables them to learn relationships between various body joints and human structures effectively.

Some studies have attempted to address the problem of occlusion using data augmentation. Peng, Tang, Yang, Feris, and Metaxas (2018) describe an augmentation generator network that learns alongside the pose estimator to augment the training data online. They jointly train the pose estimator and the data augmentation network in a generative adversarial network (GAN) setup to generate "hard" augmentations. Other studies have also used the GAN setup to train a pose estimation model (Y. Chen, Shen, Wei, Liu, & Yang, 2017; Chou, Chien, & Chen, 2018) and improve performance. The additional feedback from the discriminator network in these approaches may have helped the pose estimator produce better pose estimates. However, it is yet to be seen how this type of training setup works for transformer-based pose estimation techniques.

In addition to improving accuracy, the ability to run these models on low-cost edge devices is another important line of research. Some architectures have incorporated lightweight MobileNets and ShuffleNets (Debnath, O'Brien, Yamaguchi, & Behera, 2018; C. Yu et al., 2021) to reduce the model size. More recently, transformer-based architectures have shown better speed-accuracy trade-off in several computer vision tasks (Bao, Dong, & Wei, 2022; Z. Liu, Lin, et al., 2021) compared to traditional convolutional neural networks (CNNs). It would be appropriate to explore how these newer, more efficient vision transformers translate to HPE. Several studies have used knowledge distillation to improve the performance of lightweight models (Xu et al., 2022; F. Zhang, Zhu, & Ye, 2019). Recent studies have also focused on neural architecture search (NAS) to develop efficient models (McNally, Vats, Wong, & McPhee, 2021; Y. Wang et al., 2022).

To the best of my knowledge, only one study (Aihara et al., 2022) has attempted vision-based HPE for people in wheelchairs. Additionally, there is no publicly available dataset for this task. Therefore, it was necessary to collect a new dataset for HPE and activity recognition of people in wheelchairs. It is also important to explore techniques that are efficient and perform well even in highly occluded scenes, such as those caused by people sitting in wheelchairs. Given these circumstances, this thesis is driven by the following research questions:

- In human pose estimation, how can we deal with challenging occluded scenes, such as those that come from people sitting in wheelchairs?
- Can efficient transformer-based architectures outperform other methods in terms of the speed-accuracy trade-off while also being better at dealing with occlusion?
- Do tricks used in previous studies, such as knowledge distillation, multi-task training, GAN setup, data augmentation, and multi-objective training improve the performance of lightweight transformer-based HPE methods?

1.3 Objectives of the Study

The primary goal of this thesis is to find an efficient human pose estimation method that can run real-time on mobile devices and is also good at dealing with occluded scenes, especially those prevalent among people in wheelchairs. To achieve this, I performed the following specific steps:

- Develop efficient pose estimation models that work even under occlusion and provide a good speed-accuracy trade-off.

- Explore different training strategies from previous studies to further boost the performance of these efficient pose estimation techniques.
- Collect a new dataset for wheelchair-based pose estimation and activity recognition.
- Evaluate the pose estimation techniques and training strategies on state-of-the-art HPE benchmarks, as well as the wheelchair dataset.
- Evaluate action classification performance of the pose estimation techniques in a bench setting.

1.4 Limitations and Scope

While the work presented here addresses important challenges in HPE and activity recognition for people in wheelchairs, there are certain constraints and assumptions that must be acknowledged. The following points highlight these limitations and the scope of the study:

- The performance and generalizability of the models are largely dependent on the quality of data used to train them.
- Adequate exploration of hyperparameters, model architectures, and training strategies is limited by the availability of training resources.
- Run-time constraints limit the model architectures to being within a certain computational budget.
- The overall performance in the downstream tasks depends on the weakest module. For example, if the person detector fails, pose estimation and activity recognition are likely to fail, regardless of how well they may be under isolation.
- The volunteers used to produce the wheelchair dataset were not actual wheelchair users.

CHAPTER 2
LITERATURE REVIEW

In this section, I discuss the relevant literature on human pose estimation (HPE), vision transformers, and wheelchair-based activity recognition.

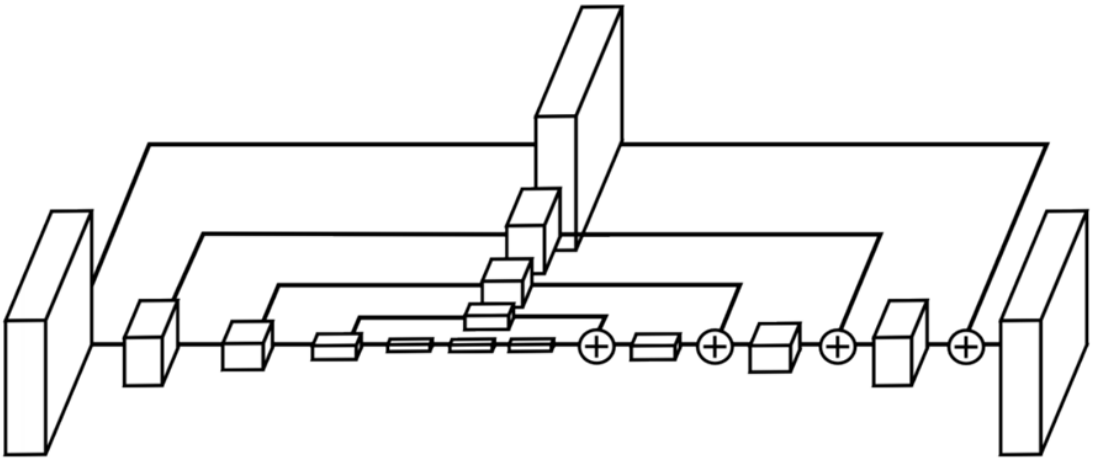
2.1 CNN-based HPE

Most classical approaches in HPE are based on pictorial structures (Felzenszwalb & Huttenlocher, 2005; Fischler & Elschlager, 1973). These classical methods use handcrafted features and exhibit a poor generalization performance (J. Tompson, Goroshin, Jain, LeCun, & Bregler, 2015). In recent years, approaches have shifted towards the use of convolutional neural networks (CNNs), inspired by the seminal work of Toshev and Szegedy (2014). Their DeepPose architecture was based on AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) and was the first to use CNNs for full-body HPE. They regressed the joint (x, y) coordinates after passing an image through the AlexNet backbone. However, J. J. Tompson, Jain, LeCun, and Bregler (2014) argue that this direct regression-based approach lacked efficiency and precision due to a non-linear and challenging-to-learn mapping. They propose generating heatmaps representing per-pixel likelihoods for joint locations and minimizing the mean squared error (MSE) between these target heatmaps (2D Gaussians centered at the ground-truth locations) and the predicted heatmaps from the model. This approach has become popular in recent HPE methods.

Here, I discuss some of the important studies on HPE. The stacked hourglass network from Newell et al. (2016) significantly outperformed previous methods in the FLIC (Sapp & Taskar, 2013) and MPII Human Pose (Andriluka, Pishchulin, Gehler, & Schiele, 2014) benchmarks (I discuss about these datasets in the HPE Benchmarks and Datasets section). This work emphasizes the need to capture information across different scales from an image and performs this using the hourglass module (Figure 2.1). The hourglass module is designed to pool down to a very low resolution, followed by upsampling to a higher resolution, and combining information across various scales. Additionally, they stack these hourglass modules with intermediate supervision, which helps in making the models deeper while tackling the problem of vanishing gradients, as shown by Wei, Ramakrishna, Kanade, and Sheikh (2016). Much of the recent work, such as that of Chu et al. (2017); Peng et al. (2018); Tang, Yu, and Wu (2018); W. Yang, Li, Ouyang, Li, and

Figure 2.1

Hourglass Module



A single hourglass module from the stacked hourglass networks architecture. Reprinted from Newell, Yang, and Deng (2016)

Wang (2017), refer to and improve on this work, which still remains a strong baseline in HPE.

J. Tompson et al. (2015) highlight the importance of max-pooling layers for promoting translation invariance while reducing computational complexity. However, these max-pooling layers also degrade spatial localization accuracy. This paper presents a coarse heatmap regression model that comprises max-pooling layers and a sliding window detector to produce coarse heatmaps for joint locations as well as a fine heatmap model that aims to recover the lost spatial accuracy due to the pooling. They jointly trained these networks to outperform the previous work in 2D HPE.

Convolutional pose machines (Wei et al., 2016) combine the advantages of the classical pose machine architecture (Ramakrishna, Munoz, Hebert, Andrew Bagnell, & Sheikh, 2014) with CNNs. In this work, Wei et al. (2016) developed a multi-stage CNN architecture to refine belief maps generated in previous stages. This study highlights the issue of vanishing gradients in very deep CNN architectures and demonstrates that intermediate supervision can mitigate the problem. They also show that this type of deep architecture can capture long-range dependencies (such as between the head and knee) owing to an increased receptive field.

Fang, Xie, Tai, and Lu (2017) introduce a regional multi-person pose estimation (RMPE)

framework, which aims to perform accurate HPE even when faced with inaccurate bounding boxes. To achieve this, the authors propose a symmetric spatial transformer network (SSTN) that extracts good single-person cropped regions from imprecise bounding boxes. Additionally, they develop novel non-maximum suppression and data augmentation techniques to improve performance.

Chu et al. (2017) employ stacked hourglass networks from Newell et al. (2016) to generate attention maps at various scales and multiple semantic levels. They create multi-resolution attention maps within each hourglass stage and replicate this process across the entire 8-stage hourglass network. This holistic attention model successfully captures full-body configurations and can handle occluded joints, as the receptive field of the model expands in deeper stages. Moreover, they design a hierarchical attention scheme for the final stages (5–8) to zoom into and refine heatmaps, enabling precise localization.

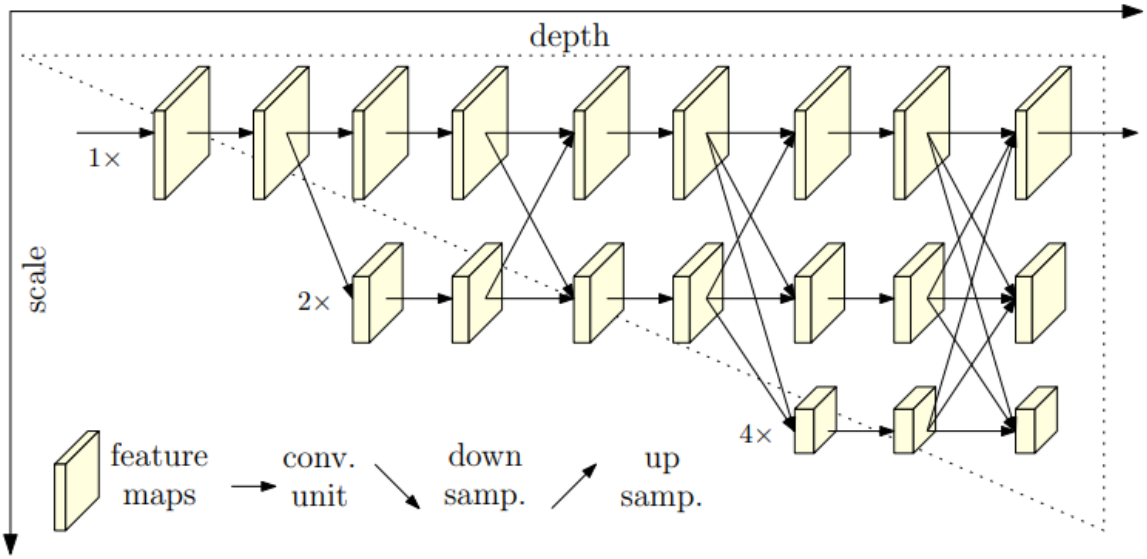
Cascaded pyramid networks (Y. Chen et al., 2018) are multi-stage networks consisting of two main stages. In the first stage, GlobalNet, a feature pyramid network, attempts to localize simple keypoints but may struggle with hard keypoints such as invisible or occluded ones. To address this issue, the second stage, RefineNet, explicitly focuses on localizing these challenging keypoints.

Xiao, Wu, and Wei (2018) present a simple baseline method for HPE. Despite significant differences in design among modern HPE methods, these approaches yield only minor variations in accuracy. This study demonstrates that simply adding a few deconvolution layers to a ResNet backbone for generating joint heatmaps can outperform previous methods in HPE. Furthermore, they adapt this implementation to tackle the problem of pose tracking using optical flow.

HRNet (Sun, Xiao, Liu, & Wang, 2019), initially introduced for HPE, has proven successful in other vision-based tasks such as semantic segmentation and object detection (J. Wang et al., 2020). The authors claim that maintaining high-resolution information throughout the whole process is crucial for precise localization, unlike previous methods that recover high-resolution information by upsampling. They design a model with a high-resolution subnetwork, progressively adding lower-resolution subnetworks and exchanging information between these subnetworks through multi-scale fusion. Despite requiring higher computational resources, the HRNet architecture has become a popular

Figure 2.2

The HRNet Architecture



The low-resolution subnetworks exist in parallel with the high-resolution subnetworks. Information is repeatedly exchanged between these subnetworks via multi-scale fusion. Reprinted from Sun et al. (2019)

baseline for numerous HPE methods (Cheng et al., 2020; C. Yu et al., 2021; Y. Yuan et al., 2021).

The aforementioned techniques represent some of the notable works in *top-down* HPE, which first requires a person bounding box detector to crop a person region before estimating the pose. In contrast, *bottom-up* approaches attempt to detect the body joints in an image and subsequently connect them into 2D skeletons. DeepCut from Pishchulin et al. (2016) was one of the earliest deep-learning-based bottom-up approaches. They adapted Fast RCNN (Girshick, 2015) to predict the joint locations and formulated the problem of connecting these body parts as an integer linear program (ILP) problem. Pishchulin and colleagues further improve this architecture in (Insafutdinov, Pishchulin, Andres, Andriluka, & Schiele, 2016) by adapting ResNet for part detection and introducing a novel image-conditioned pairwise terms to predict pairwise part-to-part probabilities. However, these models required several minutes to process an image during inference, rendering them unsuitable for real-time applications.

In this context, Cao, Simon, Wei, and Sheikh (2017) revolutionized bottom-up pose estimation by achieving a better balance between speed and accuracy than other methods.

It also became a suitable framework for real-time HPE on edge devices. This work introduced part affinity fields (PAFs) to represent pairwise relationships between body parts for each person. They predict a set of confidence maps for each body part location, along with a set of 2D vectors (PAFs), and perform bipartite matching to form a link between these body part candidates.

Other notable approaches include the HigherHRNet architecture developed by Cheng et al. (2020), which leverages high-resolution representations from HRNet to learn scale-aware representations, addressing the problem of scale variation in bottom-up HPE. Although less efficient during inference, this approach outperforms other bottom-up methods. Newell, Huang, and Deng (2017) introduce associative embedding, a technique that simultaneously predicts a tag along with each keypoint to group keypoints. This method aims to solve the problem of keypoint detection and grouping in a single stage. Furthermore, Kreiss, Bertoni, and Alahi (2019) present a novel two-head network called PifPaf, which focuses on improving HPE performance at lower resolutions. They emphasized the improvements their method offers over other bottom-up approaches in densely crowded scenes with occlusions.

Recently, model-agnostic methods like DARK (F. Zhang, Zhu, Dai, Ye, & Zhu, 2020) and UDP (Huang, Zhu, Guo, & Huang, 2020) have emerged, demonstrating their potential to boost the performance of existing HPE techniques. These approaches focus primarily on addressing issues or biases in the training process, making them adaptable to a wide range of models and architectures.

2.2 Transformers in HPE

Since the introduction of the transformer architecture for natural language processing in Vaswani et al. (2017), it has been adapted for vision-based tasks (Bao et al., 2022; Carion et al., 2020; Cheng, Schwing, & Kirillov, 2021; Dosovitskiy et al., 2020; Z. Liu, Lin, et al., 2021). TransPose from S. Yang et al. (2021) was among the first to apply a transformer network to HPE, aiming to overcome the limitations of CNNs in handling global spatial relationships. It uses the standard transformer encoder fed with flattened convolutional feature maps. Y. Li et al. (2021) argue that transformers naturally excel in handling global dependencies and propose a HPE method based on transformers, called TokenPose, for 2D HPE. Mao et al. (2021) leverage a transformer decoder to directly regress keypoint coordinates instead of heatmaps, achieving the best performance

among regression-based methods. Y. Yuan et al. (2021) build on the HRNet concept, using transformer blocks rather than CNN blocks in each multi-resolution stream, which surpasses the original HRNet in both inference speed and accuracy. SwinPose (Xiong, Wang, Li, Luo, & Cao, 2022) incorporates a swin transformer (Z. Liu, Lin, et al., 2021) backbone with a feature pyramid, showing improved accuracy, but emphasizing that the main challenge of employing transformer-based architectures in HPE is the large model sizes.

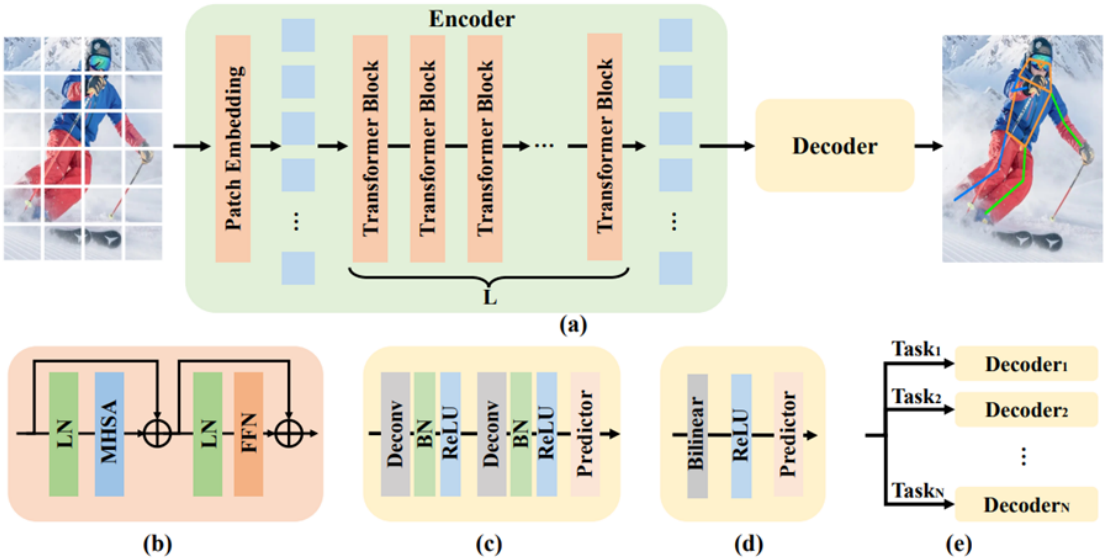
The current state-of-the-art model across multiple HPE benchmarks is the largest version of ViTPose (Xu et al., 2022). ViTPose-G achieves an AP of 81.1 on the MS-COCO dataset (with model ensemble) and 94.3 average PCKh on the MPII Human Pose dataset, which beats the second-best results by 0.6 AP and 0.2 PCKh respectively, while also being more efficient. The OCHuman (Occluded Human) dataset is a challenging benchmark with a special focus on occluded body joints (S.-H. Zhang et al., 2019). ViTPose-G achieves an AP of 93.3 on the test set, while the previous best method achieved an AP of just 74.1, which shows that this framework handles occlusion significantly better than previous methods.

The lighter ViTPose models show better computational efficiency, while either outperforming or closely matching the state-of-the-art methods in accuracy. For instance, the ViTPose-Base model, trained using a multi-task approach and an input resolution of 256×192, achieves an AP of 77.5 in the MS-COCO test-dev set, running at 944 fps on an Nvidia A-100 GPU. In comparison, a larger HRNet model, HRNet-W48, with the same input resolution of 256×192, achieves an AP of 75.1 and runs at 649 fps. This shows that a lighter transformer-based architecture could be a good choice for deployment with limited computing resources.

Unlike previous transformer-based frameworks that use a CNN to extract features in the initial stages, ViTPose directly extracts image features using plain vision transformers (Dosovitskiy et al., 2020). This is followed by a lightweight decoder for regressing keypoint heatmaps. The ViTPose model is flexible and scalable; different numbers of transformer layers can be stacked to control the model size. They also introduce a token-based knowledge distillation method for training smaller ViTPose models from larger ones, significantly enhancing the performance of their smallest ViTPose-Base model, which outperforms other popular HPE methods while remaining substantially faster during in-

Figure 2.3

ViTPose Architecture



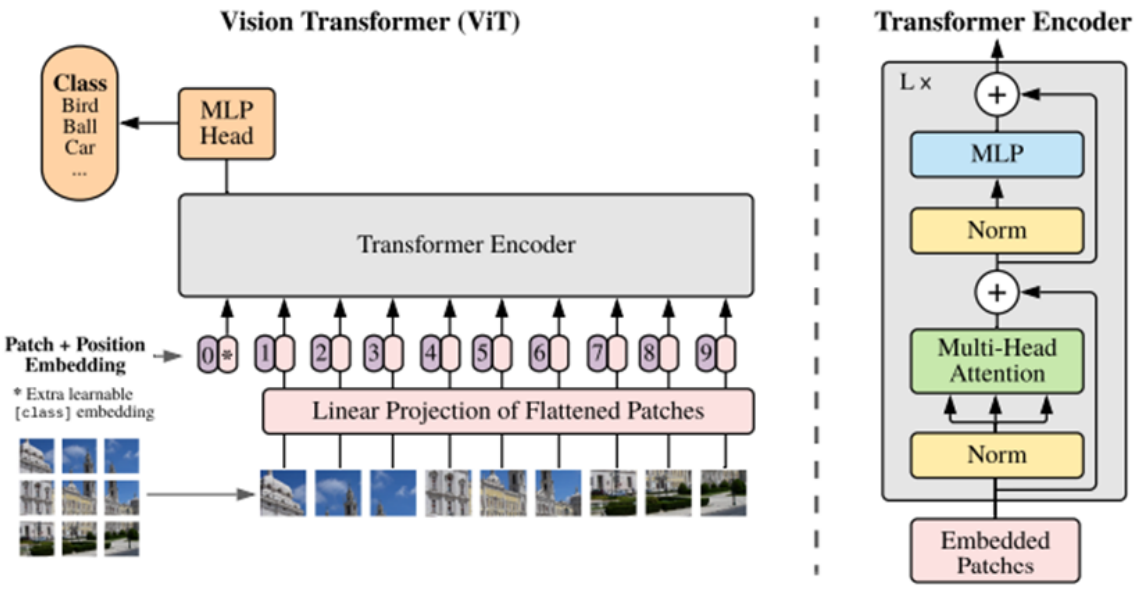
(a) The framework of ViTPose. It uses the plain Vision Transformer as the encoder to extract features. This is then fed to the decoder to regress joint heatmaps (b) The Transformer block (c) The classic decoder. (d) The simple decoder. (e) The decoders for multi-task training. Reprinted from (Xu et al., 2022)

ference. A multitask training setting is another technique that improves the performance of ViTPose models, which are trained with multiple decoders to handle multiple pose estimation tasks while sharing the same backbone decoder. Figure 2.3 depicts the structure of ViTPose. For a detailed overview of the ViTPose framework, refer to Xu et al. (2022).

These methods using transformers, especially ViTPose, show that transformer-based architectures are effective at addressing global spatial dependencies. This is particularly important for handling challenging cases such as occlusion. ViTPose considerably outperforms other methods on the challenging OCHuman dataset, demonstrating its robustness to occlusion. This demonstrates the potential of vision transformers in dealing with occlusion in 2D HPE. In my work, I explore replacing the Vanilla ViT Encoder in ViTPose with efficient ViT variants, aiming to achieve favorable speed-accuracy trade offs and enabling model deployment on edge devices.

Figure 2.4

Vision Transformer Architecture



Vision Transformer Architecture. The image is first split into fixed-size patches (usually 16x16). These patches are then linearly embedded, added with learnable position embeddings, and fed to a standard Transformer encoder. The classification token is added to perform classification. Reprinted from (Dosovitskiy et al., 2020)

2.3 Vision Transformers

The vision transformer (ViT) architecture introduced by Dosovitskiy et al. (2020) has been adapted to many vision-based tasks (Bao et al., 2022; Carion et al., 2020; Cheng et al., 2021; Z. Liu, Lin, et al., 2021). It was the first fully-transformer based architecture used for image classification. As depicted in Figure 2.4, it divides an image into patches of equal height and width (for example, $16 \times 16 \times 3$) and flattens these patches. These flattened patches are then projected to a linear embedding through a linear layer and concatenated with the position embeddings. This is fed to the original transformer encoder from Vaswani et al. (2017), which is simply followed with an MLP head for classification.

When pretrained on a large dataset with 300 million images, the original ViT outperforms CNN-based architectures in image classification. This was achieved without the inherent inductive biases such as translation equivariance and locality present in CNNs, forcing ViTs to independently learn these biases. Without sufficient training data, these

models fail to generalize as effectively as CNNs. Touvron, Cord, Douze, et al. (2021) introduce a novel knowledge distillation strategy called token distillation. Combined with strong data augmentation and regularization, they show that ViTs can be directly trained on ImageNet-1K without the large 300-M dataset and achieve performance comparable to state-of-the-art CNN-based methods. Recent studies have focused on the efficiency and data requirements of ViTs and to make them suitable for downstream tasks (Z. Liu, Lin, et al., 2021; Touvron, Cord, Sablayrolles, Synnaeve, & J'egou, 2021; W. Wang et al., 2021; L. Yuan et al., 2021).

Here I discuss some of the notable efforts in improving ViTs by making them efficient. T2T-ViT from L. Yuan et al. (2021) iteratively reduces the token length by aggregating neighboring tokens to better capture the local structure. The paper further proposes a ResNet-like deep-narrow structure (He, Zhang, Ren, & Sun, 2016) that reduces the model size while improving performance. The pyramid vision transformer (PVT) initially uses high resolution (4x4) image patches and progressively shrinks the sequence length as the depth increases (W. Wang et al., 2021). Z. Liu, Lin, et al. (2021) merge image patches in deeper layers and builds hierarchical feature maps. It also uses a shifted window approach to compute attention within these windows. PiT introduces a novel pooling strategy for ViTs and improves efficiency (Heo et al., 2021). Touvron, Cord, Sablayrolles, et al. (2021) increase the depth of ViTs with their new strategy and also insert a CLS token later in the transformer to separate self-attention from classification. C.-F. R. Chen, Fan, and Panda (2021) use a dual branch architecture to process image patches of varying sizes and a token fusion module to exchange information between these two branches.

The architectures discussed above are purely based on transformers. Another line of work considers combining ViTs with CNNs to include the benefits of capturing long-range dependencies from ViTs and the efficiency and inductive biases from CNNs. Most of these architectures produce patch embeddings using the features processed by a CNN (H. Wu et al., 2021; K. Yuan et al., 2021). CeiT (K. Yuan et al., 2021) further proposes replacing the feedforward layer with a locally enhanced feedforward network to promote the correlation between neighboring tokens. Graham et al. (2021) suggest a hybrid architecture comprising CNN layers initially, followed by ViT layers. They incorporate additional changes, such as using batch normalization instead of layer nor-

malization, reducing the spatial resolution while increasing the number of channels, as done in CNN-based architectures, and replacing the positional embedding with attention bias. This architecture significantly improves the speed-accuracy trade off and shows good inference throughput even on when running on a CPU. Most recent methods, such as MobileViT and its variants (Mehta & Rastegari, 2022a, 2022b; Wadekar & Chaurasia, 2022), EdgeNext (Maaz et al., 2022), and TinyViT (K. Wu et al., 2022) substantially push the boundaries of the speed-accuracy trade off in vision-based tasks and outperform pure CNN-based architectures.

This study examines the benefits of these efficient ViT variants in a pose estimation framework similar to Xu et al. (2022) by using them as backbone feature extractors. The goal is to determine whether the enhanced efficiency of these ViT variants can be successfully translated into improved efficiency in the context of pose estimation without sacrificing accuracy.

2.4 Adversarial Training and Data Augmentation

Some HPE methods use an adversarial training setup for data augmentation or for directly improving HPE performance. Adversarial PoseNet (Y. Chen et al., 2017) implicitly teaches the human body structure to a pose estimator network using a generative adversarial network (GAN) setup. In this setup, the generator aims to create pose and occlusion heatmaps, while the discriminator seeks to determine the plausibility of the generated pose estimates. A qualitative evaluation shows this architecture's proficiency in predicting occlusion heatmaps and producing reasonable 2D skeletons, even under occlusion. Chou et al. (2018) introduce similar work aimed at teaching the generator about plausible human body configurations. Exploring whether additional feedback from the discriminator improves pose estimation performance can be a valuable research avenue.

Data augmentation techniques also play a crucial role in improving the performance of data-driven deep learning techniques. In addition to common data augmentation techniques, such as random rotation, flipping, and scaling, some methods have looked at learnable data augmentation (H. Chen et al., 2022). Peng et al. (2018) propose a GAN setup where the generator aims to generate difficult data augmentations. This approach differs from those mentioned above, as it utilizes a discriminator for pose estimation rather than the generator. The generator explores the weaknesses of the discriminator and generates hard examples online during training to increase the discriminator's

loss. These hard examples are generated through adversarial distributions from which augmentations such as rotation, scaling, and occlusion are sampled. Since augmentation operations are not differentiable, they have a reward penalty policy to create online ground truths for the generator. The process involves creating random data augmentations, comparing their difficulty, and subsequently rewarding or penalizing the augmentation network accordingly. By incorporating this data augmentation framework, they demonstrate a significant performance improvement in the stacked hourglass network (Newell et al., 2016).

Bin et al. (2020) employ a similar setup, using the generator to create human body parts and overlaying them on existing images to occlude body parts and generate challenging examples. AdvMix, introduced in J. Wang et al. (2021), is a model-agnostic technique that combines adversarial data augmentation and knowledge distillation to enhance the robustness of existing HPE techniques. Adversarial data augmentation methods, such as those discussed above, appear to be effective in guiding pose estimation models to handle occluded body parts.

Ronchi and Perona (2017) demonstrate that state-of-the-art HPE methods exhibit similar error distributions. Moon, Chang, and Lee (2019) attempt to generate synthetic poses based on these error statistics to train their model-agnostic pose refinement model. This model called PoseFix takes pose estimates from a different model and refines them without requiring architectural knowledge.

2.5 Video-Based HPE

The primary challenges in video-based HPE come from camera shift, rapid movement, defocus, and motion blur (H. Chen et al., 2022). An ideal real-time video processing HPE method should be able to model the geometric consistency and temporal dependency between video frames (Luo et al., 2018) and address such challenges. Unlike single-frame HPE methods, video-based methods aim to capture these temporal cues for better performance in videos.

Previous studies using temporal information to estimate 2D poses were based on optical flow (Jain, Tompson, LeCun, & Bregler, 2014; Pfister, Charles, & Zisserman, 2015). Pfister et al. (2015) use optical flow fields to align features using information from multiple frames. In J. Song, Wang, Van Gool, and Hilliges (2017), the authors propose a

thin-slicing network to incorporate spatial and temporal information into an end-to-end architecture. Despite their ability to capture temporal information, optical flow-based models are easily affected by noise such as defocus and motion blur(H. Chen et al., 2022).

Recurrent neural network (RNN)-based methods can also capture the temporal contexts in videos. LSTM (long-short-term memory) pose machines (Luo et al., 2018) derived their idea from convolutional pose machines (CPM) introduced in Wei et al. (2016). The multi-stage refinement strategy of a single image in CPM is replaced by recurrent LSTM units, which are fed consecutive video frames in the subsequent stages. While this framework works remarkably well for single-person pose estimation, it is difficult to apply it to multi-person HPE in videos (H. Chen et al., 2022; Z. Liu, Chen, et al., 2021).

UniPose (Artacho & Savakis, 2020) is a state-of-the-art method for video-based HPE. It comprises a one-stage network with a waterfall atrous spatial pooling (WASP) module as the main component, which generates an efficient multiscale representation. This is then used by the decoder to predict keypoint heatmaps. The architecture provides one of the best results in single-frame HPE and is extended by applying a linear sequential LSTM to capture temporal information.

DeciWatch (Zeng et al., 2022) is an efficiency-focused method and is applied to both 2D and 3D HPE. The authors propose a method to get pose estimates through sparsely sampled video frames, specifically less than 10% of the video frames, using a single-frame HPE technique. Then they use an efficient transformer-based DenoiseNet to denoise these pose estimates. They further use another transformer-based RecoverNet to recover the pose estimates in remaining frames. They achieve a higher mean PCKh in the sub-JHMDB dataset (Jhuang, Gall, Zuffi, Schmid, & Black, 2013) compared to LSTM Pose Machines (Luo et al., 2018) at a fraction (5%) of the FLOPs.

Training models to use temporal information typically requires a densely annotated video dataset. However, due to the difficulty in producing them, only a few publicly available datasets of this kind exist (Andriluka et al., 2018; Jhuang et al., 2013; W. Zhang, Zhu, & Derpanis, 2013). These datasets are discussed in the HPE Benchmarks and Datasets section.

2.6 Building Efficient Models

In addition to accuracy, efficiency plays a crucial role in HPE. While large models such as ViTPose-G (Xu et al., 2022) demonstrate exceptional performance and handle occlusion more effectively than smaller models, real-time execution on mobile edge devices remains a significant challenge. Therefore, developing efficient models that provide a good balance between speed and accuracy is an important area of research.

CNN-based models, such as MobileNet (Howard et al., 2017) and ShuffleNet (X. Zhang, Zhou, Lin, & Sun, 2018), are known for their efficiency, as they are specifically designed for deployment on mobile devices with minimal compromise in accuracy. Certain HPE methods (Debnath et al., 2018; C. Yu et al., 2021) explore the application of these models to enhance efficiency.

LiteHRNet (C. Yu et al., 2021) combines ShuffleNet (X. Zhang et al., 2018) with HRNet (Sun et al., 2019) to improve the speed-accuracy trade off. The model employs conditional channel weighting as a substitute for pointwise convolutions in ShuffleNet, resulting in a further improvement in performance.

LitePose (Y. Wang et al., 2022), which is based on the MobileNetV2 architecture (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018), demonstrates through a gradual shrinking experiment that the high resolution branches of HRNet (Sun et al., 2019) are not only redundant, but also degrade both performance and accuracy in the low-precision region. To better address scale variation, LitePose increases the receptive field size of its relatively shallow model by using a 7×7 convolution kernel. Furthermore, instead of manually designing their CNN architecture, the authors utilized neural architecture search (NAS) to generate four LitePose models, each targeting different computational budgets. Consequently, this model approaches OpenPose's (Cao et al., 2017) performance with substantially fewer parameters.

EvoPose2D (McNally et al., 2021) leverages NAS, specifically neuroevolution, to enhance efficiency and yield one of the best-performing models in terms of speed-accuracy trade off.

Lately, knowledge distillation (Hinton, Vinyals, & Dean, 2015), a technique that emphasizes transferring knowledge from a larger teacher network to a more compact student

network, has been effective in enhancing the performance of smaller, lightweight models. In Xu et al. (2022), the authors use a modified knowledge distillation technique to train their smallest model (ViTPose-B) using a larger model (ViTPose-L) as a teacher, resulting in a 0.6 AP increase in accuracy with only a marginal addition to the memory footprint. Fast pose distillation (FPD) (F. Zhang et al., 2019), a method proposed to train a 4-stage Hourglass network using an 8-stage hourglass network, achieves a 0.8 mean PCKh boost on the MPII dataset (Andriluka et al., 2014).

This paper aims to identify efficient models that also show robustness against occlusion. To achieve this, knowledge distillation methods are also explored to find the ideal combination to develop such models. Additionally, several training strategies are explored to find the best way to train small and efficient models for improved accuracy and robustness to occlusion.

2.7 3D-HPE From Monocular Cameras

Estimating the 3D information from multiple 2D images is achievable through traditional techniques such as structure from motion (SfM) and simultaneous localization and mapping (SLAM). Recently, deep learning techniques are also used for 3D reconstruction using single 2D images (B. Li, Dai, & He, 2018; B. Li, Shen, Dai, van den Hengel, & He, 2015). As a result, 3D HPE from monocular cameras is now feasible, although predicting depth information presents an additional challenge. S. Li and Chan (2014) were among the first to train 3D point regression and joint detection tasks simultaneously. In Pavlakos, Zhou, Derpanis, and Daniilidis (2017), instead of directly regressing 3D joint coordinates, they first predict voxel likelihoods for each joint and then apply a coarse-to-fine prediction scheme for refinement. Some methods directly estimate 3D human poses from images, while others predict an intermediate 2D pose (C. Li & Lee, 2019; Tekin, Márquez-Neila, Salzmann, & Fua, 2016). Research focusing exclusively on multi-person 3D-HPE with monocular cameras is relatively scarce, with advances in multi-person 3D HPE primarily stemming from single-person methods (Y. Chen et al., 2020).

Although 3D HPE techniques offer more detailed information for downstream tasks such as action classification, accurately predicting depth information from a monocular camera presents a challenge. This typically requires additional computational resources, which can pose difficulties in running the models on edge devices. Furthermore, training

datasets for monocular 3D HPE are limited and available only for constrained environments, restricting the generalization capabilities of these models (Y. Chen et al., 2020; W. Liu, Bao, Sun, & Mei, 2022). Monocular 3D HPE is a broad and extensively studied area; however, providing a comprehensive literature review falls outside the scope of this work. For a more in-depth review, the reader is referred to recent surveys such as the ones by W. Liu et al. (2022).

2.8 HPE Benchmarks and Datasets

The current popular benchmarks for HPE are the MS COCO Keypoint Detection (Lin et al., 2014) and MPII Human Pose (Andriluka et al., 2014) datasets. The COCO dataset was initially created for large-scale object detection and later extended to include keypoint detection for 2D HPE. For this task, it features 200,000 images with 250,000 annotations, which comprise 17 joints for each person, along with visibility information. The MPII Human Pose dataset, sourced from online videos, contains approximately 25,000 images with 40,000 annotations. The annotations cover 16 body joints and include occlusion data. This dataset spans about 410 human activities and can also be used for activity recognition.

Introduced in S.-H. Zhang et al. (2019), the OCHuman (Occluded Human) dataset primarily focuses on heavily occluded humans, with annotations consisting of instance masks, bounding boxes, and 2D skeletons. The dataset comprises 5,081 images and 13,360 annotation instances. Similarly, the CrowdPose (J. Li et al., 2019) dataset contains crowded scenes and features around 20,000 images with 80,000 person instances. The AI Challenger Human Keypoint Detection dataset (J. Wu et al., 2017) includes approximately 300,000 images and 700,000 annotations, making it the largest dataset for HPE to date. ViTPose beats previous methods in all of the aforementioned datasets (Xu et al., 2022).

The Leeds Sports Poses (LSP) dataset (X. Yu, Zhou, & Chandraker, 2016) was a common benchmark for early HPE methods. It features approximately 10,000 images (expanded from the original 2,000) with 14 joint annotations. The FLIC dataset (Sapp & Taskar, 2013), which includes around 5,000, was also used frequently.

Certain datasets include videos and can be used to train HPE methods to leverage temporal information. PoseTrack (Andriluka et al., 2018) consists of around 514 videos,

with the training set containing 30 annotated frames from the center for each video. The Penn Action dataset (W. Zhang et al., 2013), primarily intended for action classification, includes about 2,326 videos with human joints annotated across each video frame, along with visibility data. The Joint-annotated Human Motion Data Base (JHMDB) dataset (Jhuang et al., 2013) was designed for both action classification and HPE. The subset used for HPE encompasses 316 densely annotated video clips.

The wheelchair dataset produced at AIT as part of this work is another dataset that can be used for action classification as well as an HPE benchmark.

2.9 2D Skeleton-Based Action Classification

A sequence of human pose skeletons can provide an estimate of the action being performed by a human. These pose skeleton estimates are not affected by noise factors such as background variation and varying lighting conditions (Duan et al., 2022). Their compact nature, combined with these advantages, makes them suitable for human action classification (Fanuel, Yuan, Nam Kim, Qingge, & Roy, 2021).

Action classification using a sequence of pose skeletons can be treated as a time series problem (Du, Wang, & Wang, 2015b). RNNs have historically been widely used in time series problems, including skeleton-based action classification. Du, Wang, and Wang (2015b) divide the pose skeletons into five parts, feeding them into separate Bidirectional RNNs (BRNNs) and hierarchically fusing them before classifying the action in the softmax layer. Si et al. (2019) integrate graph convolutional networks (GCNs) and LSTMs to concurrently learn spatiotemporal features, as they argue that pose skeleton sequences exhibit a co-occurrence relationship between the spatial and temporal domains. Various other studies have used RNNs, particularly LSTM networks, in action classification (Lee, Kim, Kang, & Lee, 2017; S. Song, Lan, Xing, Zeng, & Liu, 2017; Zhu et al., 2016).

In contrast, Du, Fu, and Wang (2015) utilize a CNN rather than an RNN. They represent the sequence of skeletons as a matrix by concatenating and arranging them in chronological order, then feeding the matrix to the CNN as an image. C. Li, Zhong, Xie, and Pu (2018) and Ke, Bennamoun, An, Sohel, and Boussaïd (2017) similarly adopt a CNN architecture for action classification.

Sequential pose estimates can intuitively be viewed as graphs connected in a series (Fanuel et al., 2021). This perspective has enabled the application of GCNs to the problem, with representative works including H. Yang, Gu, Zhu, Hu, and Zhang (2020); H. Yang et al. (2022); X. Zhang, Xu, and Tao (2020).

These 2D skeleton-based action classification methods heavily rely on the quality of 2D skeletons. Inaccurate joint locations, for instance, due to occlusion, might result in false classifications. Therefore, it is crucial to prioritize the quality of such 2D skeletons. In some cases, it has been observed that 2D skeletons may not capture enough information to correctly classify similar actions (Du, Wang, & Wang, 2015a). However, the compactness of a 2D skeleton-based model allows for efficient inference in action classification, and, if accurate, can serve as an effective method for real-time human action classification.

2.10 Wheelchair-Based Pose Estimation and Activity Recognition

One of the common problems faced by individuals in wheelchairs with a spinal cord injury (SCI) is pain and damage to the shoulders (de Vries et al., 2022; Ding, Hiremath, Chung, & Cooper, 2011). This problem is commonly linked to prolonged dependence on upper limbs for demanding tasks like lifting heavy objects and propelling the wheelchair (Ding et al., 2011). Not only does this reduce the patients' quality of life, but it also limits their daily activities (de Vries et al., 2022). To avoid such problems, guidelines suggest minimizing strain on these limbs.

On the other hand, lack of physical exercise for people in wheelchairs is associated with various health issues, including chronic diseases (Booth, Roberts, & Laye, 2012; Marco-Ahulló et al., 2021). Obesity (Hiremath, Intille, Kelleher, Cooper, & Ding, 2015), hypertension, hyperlipidemia, and diabetes (Popp et al., 2018) are some of the prevalent conditions among this population. In fact, the leading cause of death for wheelchair users with an SCI is related to cardiovascular diseases (Popp et al., 2018). Moreover, extended pressure on a particular area of the body may result in pressure sores, which is an additional factor contributing to premature death among wheelchair users (Jaffery et al., 2022).

Among wheelchair users are also elderly individuals who may require monitoring (Ma, Li, Gravina, & Fortino, 2016). Falls are common for wheelchair users when they attempt

to transfer from the wheelchair without assistance. The elderly population is particularly susceptible to falls, and they pose a serious health risk for all wheelchair users (Yousuf & Jilani, 2021).

Keeping track of the physical activities of people in wheelchairs and presenting this information to medical professionals would be highly beneficial. This system can enable professionals to effectively monitor the general health conditions and activities of wheelchair users (Ma et al., 2017), and help them suggest behavioral modifications for a healthier lifestyle (Hiremath, Ding, Farringdon, Vyas, & Cooper, 2013). Moreover, such a system would save significant time and cost in healthcare (Postolache et al., 2014).

Here, I review previous work on monitoring the activities and posture of people in wheelchairs, which typically relies on wearable sensors such as Shimmer IMUs (de Vries et al., 2022) and accelerometers (Marco-Ahulló et al., 2021) or smart wheelchairs (Jaffery et al., 2022). To the best of my knowledge, only one paper has ever attempted to perform pose estimation for people in wheelchairs using computer vision (Aihara et al., 2022). Consequently, there is no publicly available dataset for wheelchair-based pose estimation or activity recognition. It is necessary to consider the different types of postures or activities that previous studies have chosen with proper scientific reasoning before creating a dataset for vision-based pose estimation.

Most previous work has attempted to classify similar activities with slight variations. de Vries et al. (2022) extract the most relevant activities from the previous literature and through discussions with clinical professionals from a specialized SCI center in Switzerland. The activities include *desk work, picking and placing weights, sitting still, transferring from the wheelchair, intermittent and regular wheelchair propulsion, arm cranking*, and *weight relief lift*. They use Shimmer IMU sensors to collect electromyography signals and pass the signals through an LSTM network for classification. Goodwin et al. (2021) include activities such as *reaching counter height, overhead reaching, lifting a 6.8 kg backpack, transfer between the wheelchair and a plinth*, and *propulsion*.

Marco-Ahulló et al. (2021) rely on a smartphone sensor to collect accelerometer data. The activities they collect include *lying down, watching TV, working on a computer, moving items, mopping the floor, cleaning the window, ironing, exercising using an arm ergometer*, and *slow and fast propulsion*. García-Massó et al. (2015) use the same ac-

tivities, while Ding et al. (2011) additionally include *being propelled by a person, doing laundry, preparing meals*, and *transferring from the wheelchair*.

Several other studies have investigated similar activities with minor differences, such as *playing wheelchair basketball*, and *propelling at a given pace* (Hiremath et al., 2015), as well as *washing dishes* and *playing table tennis* (Popp et al., 2018). Some papers such as those by Ma et al. (2017) and Ma et al. (2019) focus directly on sitting postures such as *proper sitting, leaning left, leaning right, leaning forward*, and *leaning backward*. These methods utilize smart wheelchairs rather than wearable sensors.

The need to collect a new dataset for vision-based HPE and activity recognition arises from the lack of publicly available datasets for people in wheelchairs. As part of this thesis, by reviewing the sensor-based papers and discussing with experts, I narrowed down a list of activities for data collection. The collected activities include *sit, propel the wheelchair around the room, pick and place weights on a desk, deskwork, phone call, eat, drink, bend, exercise, sit still, transfer from the wheelchair to bed, transfer from the bed to wheelchair, rest on the bed, fall while transferring from the wheelchair to bed, fall while transferring from bed to wheelchair*, and *get propelled by someone else*.

CHAPTER 3
METHODOLOGY

In this section, I propose a methodology for creating a computationally efficient top-down human pose estimation (HPE) framework based on transformers, that is also expected to deal better with occlusion. I also explore different training strategies to further boost the performance of the small and efficient methods. My framework is particularly aimed towards action classification for people sitting in wheelchairs. To this end, I collected a new dataset specifically for training pose estimation and action classification models that works with people in wheelchairs. My proposed pose estimation methods and training strategies are assessed based on multiple factors, such as pose estimation performance on benchmark datasets, speed-accuracy trade-off, and action classification performance on the new wheelchair dataset.

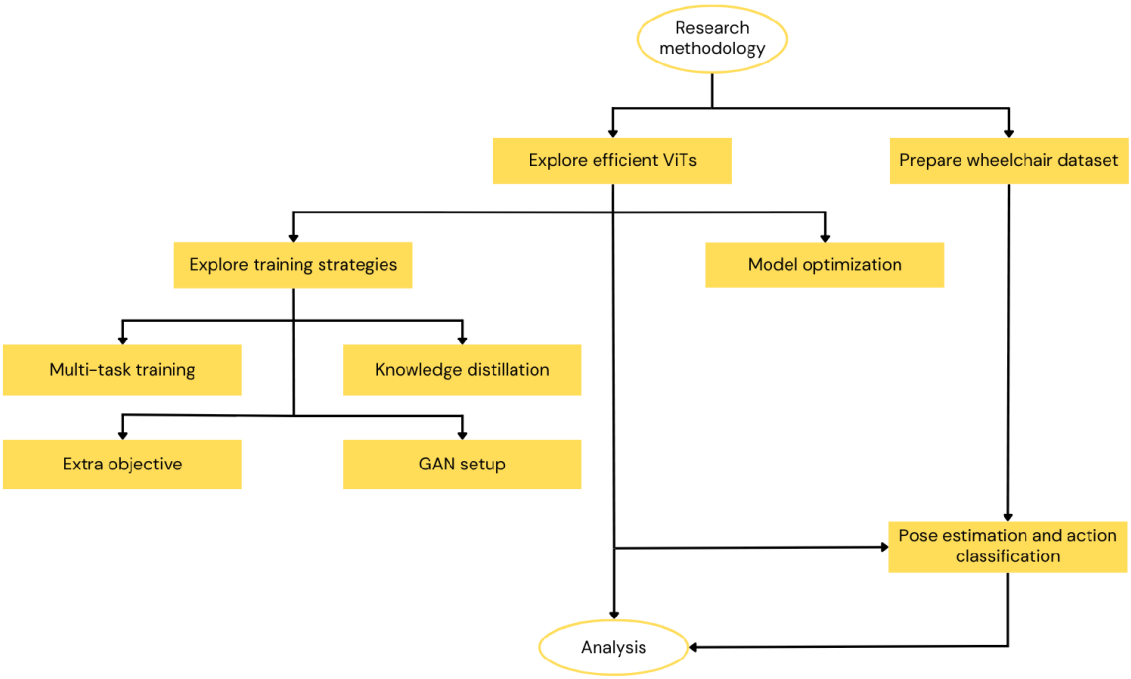
3.1 Research Methodology

The research methodology (Figure 3.1) I adopted in this study involves a systematic approach to explore the applicability of efficient ViT architectures for HPE tasks and to address the challenges associated with running them on low-cost edge devices. The methodology is organized into the following steps:

1. **Efficient ViT Architectures:** Investigate the potential of computationally efficient vision transformer backbones for HPE tasks, specifically by replacing the ViT encoder in ViTPose with various efficient ViT variants.
2. **Knowledge Distillation:** Employ knowledge distillation strategies to improve model performance by transferring knowledge from a larger, more accurate teacher model to a smaller, more efficient student model.
3. **Multi-Task Training:** Implement multi-task training strategies that allow models to learn from multiple heterogeneous datasets.
4. **GAN Setup:** Leverage generative adversarial networks (GANs) to improve the plausibility of the model's predictions.
5. **Extra Objective:** Introduce an additional objective to the model (such as vectors formed between connected joints) to guide its learning towards more effective representations and improve overall performance.
6. **Model Optimization:** Fine-tune the model's hyperparameters and architecture to

Figure 3.1

Research Methodology



I initially explored computationally efficient ViT variants to find models that provide good speed-accuracy trade-off. I then explored several training strategies that further improve performance of these models. In parallel, I also worked on preparing a wheelchair dataset for pose estimation and action classification. The models were analyzed based on their pose estimation performance on the COCO dataset, OCHuman dataset, and wheelchair dataset, action classification performance on the wheelchair dataset, and inference speed.

maximize its performance.

7. **Wheelchair Dataset:** Create a wheelchair-based HPE and activity recognition dataset to facilitate the development of a system that effectively monitors the activities of people using wheelchairs.
8. **Action Classification:** Implement an action classification module on top of the optimized HPE model to recognize and classify various activities performed by individuals in wheelchairs.

3.2 Problem Formulation

HPE is focused on locating different body parts and predicting a human structure or pose. In this study, I concentrate on 2D-skeleton-based HPE. My primary focus is on top-down methods, which utilize cropped bounding box regions generated by an object detector. These cropped regions are typically passed through a deep learning model to predict the body part locations.

The training process for a pose estimation model involves generating heatmaps that closely match ground-truth heatmaps in the training set. Typically, a 2D-Gaussian is created around ground-truth keypoint locations to create ground-truth heatmaps. The mean-squared difference between these target heatmaps and the model-predicted heatmaps is used as the loss function to drive the model's learning. Given a set of ground-truth heatmaps C={C_1 ... C_N}, $C_i \in \mathbb{R}^{h \times w}$, $i \in \{1...N\}$, the objective of the model is to produce corresponding heatmaps $\hat{C}$={$\hat{C}_1$ ... $\hat{C}_N$}, $\hat{C}_i \in \mathbb{R}^{h \times w}$, $i \in \{1...N\}$, that minimize the objective

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \|C_i - \hat{C}_i\|_F^2, \tag{3.1}$$

where N denotes the number of keypoints, and C_i and $\hat{C}_i$ denote the ground-truth and the generated heatmaps, respectively.

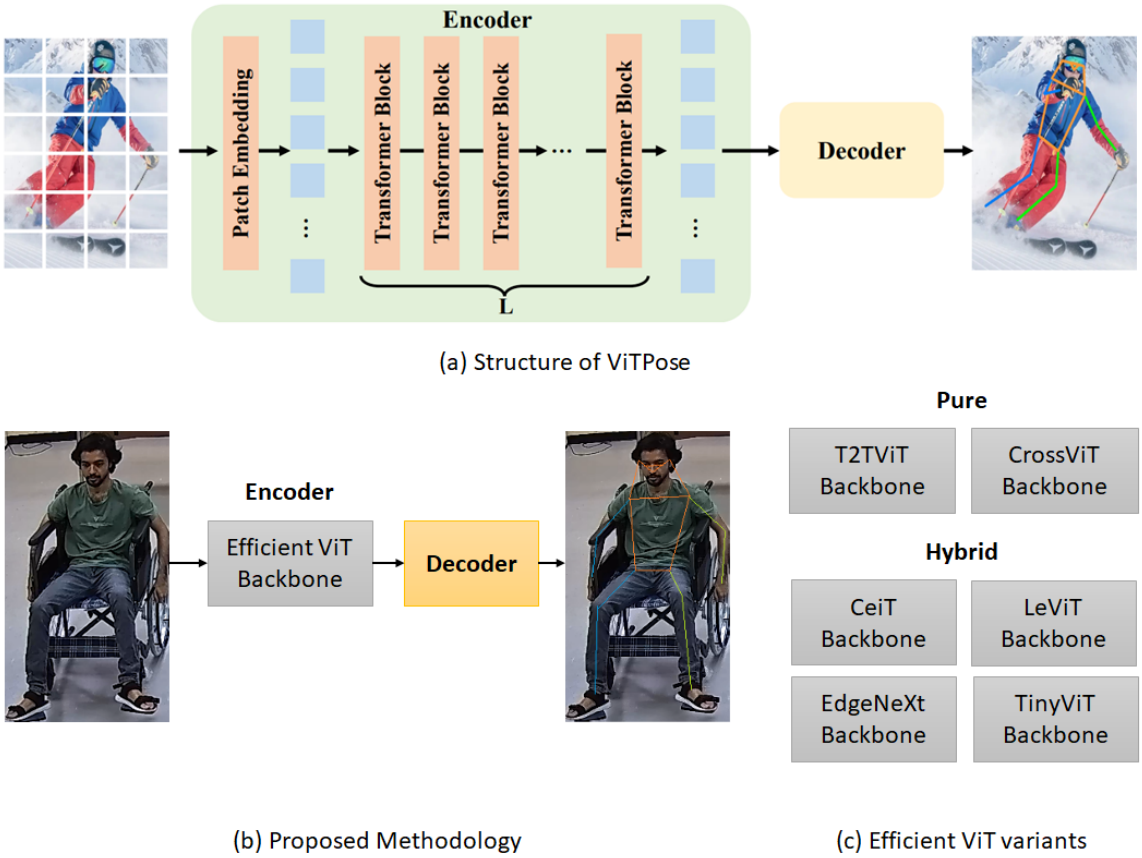
3.3 ViTPose and Efficient ViTs

ViTPose (Xu et al., 2022) uses a plain vision transformer (Dosovitskiy et al., 2020) as an encoder. The features extracted from an image are passed through a decoder, usually composed of one or more deconvolution layers, which generates the predicted heatmaps. Typically, the size of these heatmaps is $N \times 64 \times 48$.

From the literature review, I concluded that ViTPose is at the time of writing the state-

Figure 3.2

Proposed Methodology



(a) Framework of ViTPose from Xu et al. (2022). (b) Proposed Methodology. I replace the ViT encoder in ViTPose with computationally efficient ViT variants. The input is a cropped bounding box region of a person that comes from a person detector and the output is a set of keypoint heatmaps for each body joint. The decoder remains the same as in ViTPose. (c) The ViT variants used. Initially, I compare four ViT variants, with two being purely transformer-based and the other two incorporating CNNs into their architecture. Later, I also test EdgeNeXt and TinyViT.

of-the-art method in HPE across multiple benchmarks. Notably, it outperforms other methods in the challenging OCHuman dataset, demonstrating its ability to handle occlusion effectively. In my work, I replace the vanilla ViT backbone of ViTPose with more computationally efficient ViT backbones developed in more recent work. I compare their performance against each other, as well as other HPE methods, with the goal of finding the best speed-accuracy trade-off while maintaining the capability of dealing with occlusion. The proposed methodology is depicted in Figure 3.2

3.3.1 Pure Transformer-based ViT Variants

ViTPose has proved that a purely transformer-based architecture can provide superior performance across multiple HPE benchmarks and can robustly deal with occlusion. However, it remains to be seen whether this holds true as the ViT backbone is made more efficient and whether computationally efficient and purely transformer-based architectures can surpass other architectures. This section describes the pure transformer variants that I used.

Tokens-to-Token ViT

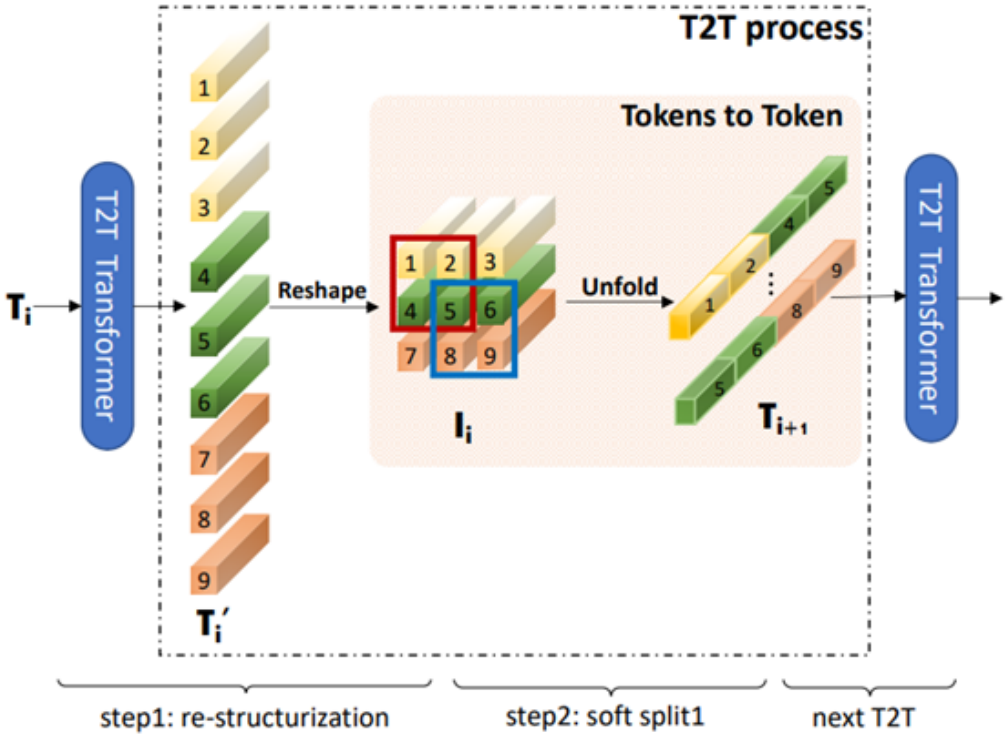
T2TViT proposes an aggregation of neighboring tokens into a single token to capture local structure (L. Yuan et al., 2021).The T2T module, as shown in Figure 3.3, can be applied iteratively to reduce the number of tokens. The authors further explore CNN-based architectures and determine that a deep-narrow architecture offers optimal performance and efficiency for vision transformers.

CrossViT

Sun et al. (2019) show that capturing information at multiple scales improves HPE performance. The CrossViT (C.-F. R. Chen et al., 2021) architecture uses a dual-branch transformer to learn feature representations at multiple scale. Each branch processes both small and large patches concurrently, which are then fused multiple times using the efficient cross-attention module. The architecture is depicted in Figure 3.4.

Figure 3.3

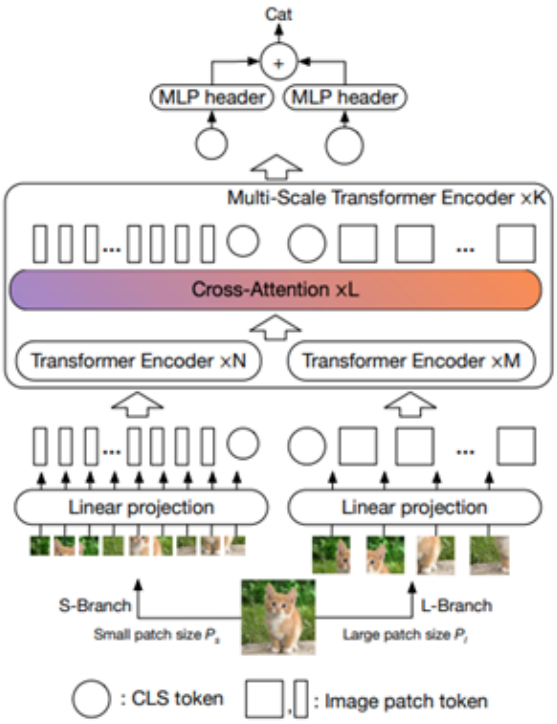
T2T Process in T2TViT



The T2T process in the T2TViT architecture. The tokens are restructured as an image and then split again with overlap.This approach reduces the number of tokens while simultaneously capturing local information. Reprinted from L. Yuan et al. (2021).

Figure 3.4

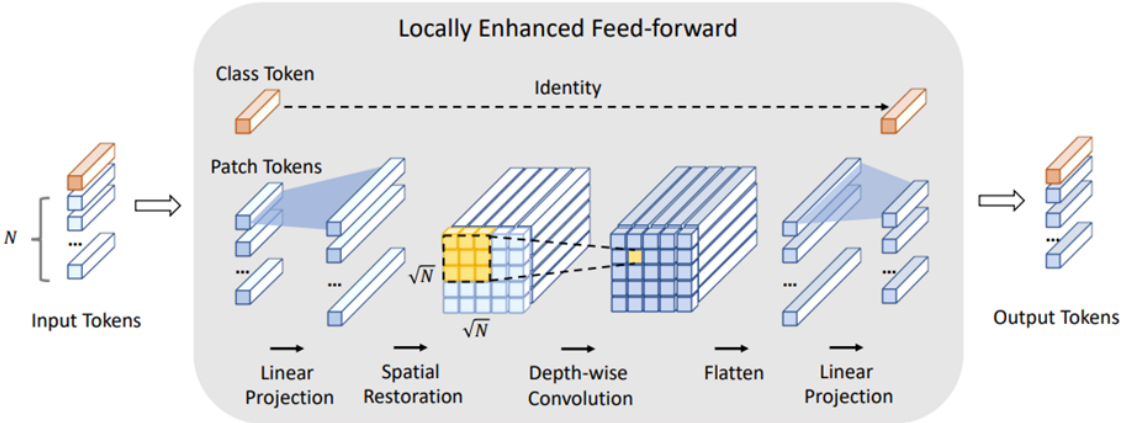
CrossViT Architecture



CrossViT uses two branches (L and S) to process image tokens of different sizes. The tokens are fused using cross attention with the CLS tokens. Reprinted from C.-F. R. Chen et al. (2021).

Figure 3.5

LeFF module of the CeiT architecture



LeFF module of the Incorporating Convolutions Design into Vision Transformers architecture. Instead of the regular feed-forward layer in transformers, the architecture uses the LeFF module to capture local information. Reprinted from K. Yuan et al. (2021).

3.3.2 Hybrid ViT Variants

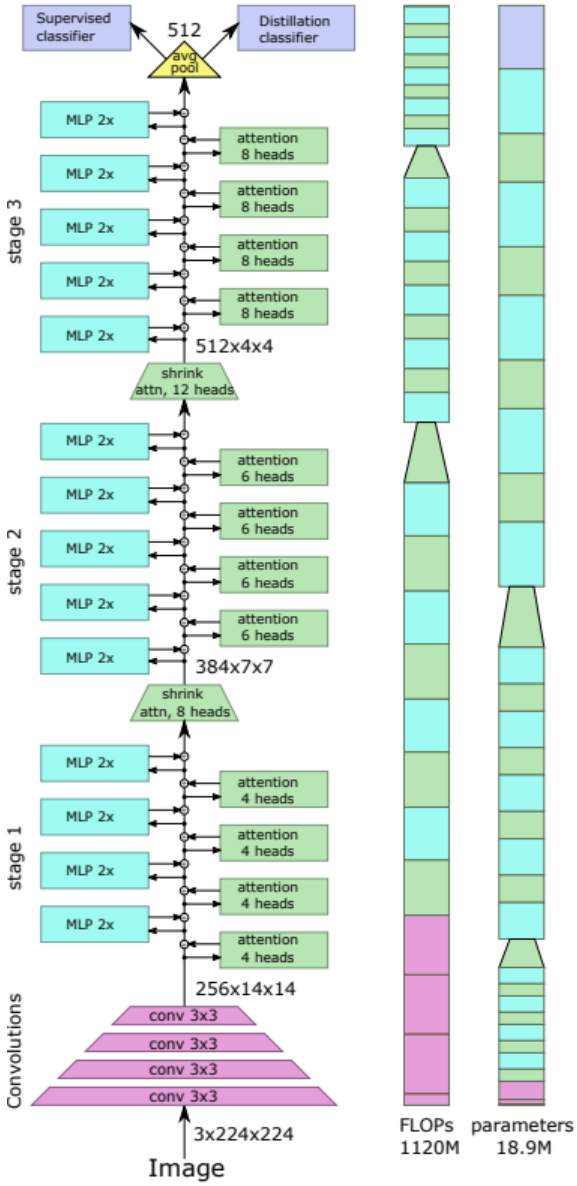
Although ViTPose is a purely transformer-based architecture and showcases impressive performance, hybrid transformer architectures, which combine both CNNs and transformers, have displayed superior computational efficiency in image classification tasks. These hybrid architectures may potentially excel in low-computation scenarios for HPE as well. To investigate this, I evaluate hybrid ViT variants, comparing their performance against each other and the pure variants. This section outlines the architectures of the hybrid transformer variants that I used in this study.

Incorporating Convolutions Design into Vision Transformers (CeiT)

The CeiT architecture (K. Yuan et al., 2021) uses low-level CNN features to generate patch embeddings. It further introduces a locally enhanced feed-forward (LeFF) layer designed to capture local information through depth-wise convolutions. The self-attention module remains unchanged, but the feed-forward layer is modified to incorporate this LeFF layer.

Figure 3.6

LeViT architecture



The transformer stages in LeViT begin after four 3x3 convolutions. The channels are increased and the resolution is reduced by half after each stage, similar to CNN architectures. Reprinted from Graham et al. (2021).

LeViT

Graham et al. (2021) propose to replace transformer components with convolution-like components. This involves several changes, including applying four 3×3 convolutions with a stride of 2 to reduce reduce the number of features in the initial layers, replacing layer normalization with batch normalization, and substituting GELU with Hardswish non-linearity. Additionally, they use a shrinking attention block that reduces the spatial resolution while increasing the number of channels like in CNNs. Moreover, they replace positional embeddings with attention bias, and MLP blocks with 1×1 convolutions.

3.3.3 Models with best speed-accuracy trade-off on ImageNet

While the aforementioned ViT variants showcase impressive speed-accuracy trade-offs, they do not represent the absolute best models in this regard. To further investigate how the image classification performance on ImageNet transfers to pose estimation, I examine additional hybrid architectures that currently excel in terms of their speed-accuracy trade-offs. These models achieve higher accuracy with a similar number of parameters and within a similar computational budget.

EdgeNeXt

EdgeNeXt (Maaz et al., 2022) is a multi-stage architecture that comprises four stages, beginning with downsampling layers that reduce the spatial resolution of the image. These layers are followed by adaptive convolutional encoders, in which the kernel size is increased after each stage. A subsequent split depth-wise transpose attention (SDTA) block utilizes a transposed query and key attention mechanism. Unlike the conventional transformer dot product attention, which involves the dot product between Q and K^T, the SDTA block applies the dot product between Q^T and K, which applies attention weights to channel dimensions instead of the spatial dimensions. This reduces the cost of dot product operation in the attention block to have a linear complexity with respect to the channel dimensions.

TinyViT

TinyViT (K. Wu et al., 2022) is also a multi-stage architecture. It scales down larger models using a progressive model contraction approach. This method involves contracting factors such as embedding dimensions, the number of blocks, window size, and the number of attention heads. Starting with a large model, each step generates a set of smaller models. The model that satisfies the given constraints and achieves the best validation accuracy is chosen before proceeding to the next step. The primary focus of the TinyViT paper is on knowledge distillation. The authors also emphasize that knowledge distillation on ImageNet can help boost performance across various downstream tasks. However, it is yet to be seen whether it helps in tasks such as human pose estimation.

3.4 Training Strategies

In addition to a model's architecture, training strategies can also significantly impact the final performance of the model. As observed in the literature, various training strategies have been investigated to enhance the performance of different types of models. These strategies encompass techniques such as data augmentation, knowledge distillation, adversarial training, multi-task training, and the use of appropriate alternative loss functions. In this section, I discuss the training strategies I explored to further improve the performance of efficient pose estimation methods, seeking optimal solutions to achieve the desired speed-accuracy trade-offs while maintaining a strong capability to deal with occlusion.

3.4.1 Knowledge Distillation

F. Zhang et al. (2019) demonstrate that a much smaller student pose estimation model can achieve significantly higher accuracy when learning from a larger teacher model. In their paper, they explore the optimal balance between learning teacher predictions and learning from ground truth to achieve the best result. They found that dividing the loss equally between the ground truth and teacher-generated labels yielded the highest accuracy for their 4-stack hourglass network. Their loss function is defined as

$$\mathcal{L}_{\text{fpd}} = \alpha \mathcal{L}_{\text{pd}} + (1 - \alpha) \mathcal{L}_{\text{mse}} \,, \tag{3.2}$$

where $\mathcal{L}_{\text{fpd}}$ is the total loss, $\mathcal{L}_{\text{pd}}$ is the loss from teacher, and $\mathcal{L}_{\text{mse}}$ is the loss obtained from Equation 3.1. The α parameter balances between the two loss terms.

In this case, I investigate how knowledge distillation from a larger ViTPose model to more efficient pose estimation methods works with efficient ViTs as backbones. Furthermore, I examine the ideal balance between the ground truth and teacher-predicted losses by testing with different values of the α parameter, aiming to maximize the performance of these efficient pose estimation methods.

3.4.2 Multi-task Training

Xu et al. (2022) combine multiple pose estimation datasets such as the AI Challenger, COCO, and MPII Human Pose datasets and jointly train a pose estimation model. As the output (target) structure is different in these datasets, different heads have to be used to produce outputs with varying number of channels. For example, the COCO dataset contains 17 keypoints per person, and the ground truth target size is $64 \times 48 \times 17$. However, the AI Challenger dataset and the MPII dataset contain 14 and 16 keypoints, respectively. Hence, instead of merging these datasets together, the pose estimation models are trained with separate heads by sampling data from different datasets simultaneously. Each sample is sent to the respective head with respective targets. The loss is propagated through all of these heads and is used to update the backbone encoder. The total loss is obtained through the equation

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{COCO}} + \mathcal{L}_{\text{AIC}} + \mathcal{L}_{\text{MPII}}, \tag{3.3}$$

This multi-task training approach has proven to be the primary factor in improving performance under occlusion, as highlighted by its state-of-the-art performance in the OCHuman dataset. In this work, I train some of my models that demonstrate the best speed-accuracy trade-offs using a multi-task training setup, aiming to boost their performance on the COCO validation set and also make them significantly better at dealing with occlusion.

3.4.3 GAN Setup

Y. Chen et al. (2017) and Chou et al. (2018) show that using an adversarial setup to train a pose estimation model improves performance. They use the generator as a pose estimator, and the discriminator distinguishes the fake poses generated by the generator from the real poses (ground-truth).

I also explore an adversarial setup for training pose estimation networks. I explore a traditional GAN setup in which the discriminator distinguishes between real and fake

samples. For the discriminator, I use a CNN with several convolution layers followed by linear layers. The optimization process for the discriminator in this case is the same as that in a traditional GAN setup: it takes an input heatmap x and outputs a score $D(x)$ indicating whether the input came from the pose estimator or from the ground truth. The primary loss for the generator is the same as Equation 3.1. The secondary adversarial loss (L_{adv}) comes from the generator's ability to fool the discriminator. The equation can be written as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{G}} \mathcal{L}_{\text{adv}}, \tag{3.4}$$

where λ_{G} is a constant used to give a weight to the adversarial loss.

3.4.4 Additional Joint Vector Target

Humans inherently possess the ability to estimate the locations of invisible or occluded joints. We use information such as the connections between joints, knowledge of human body structure, and estimated angles and lengths between connections formed by different body joints to predict such invisible joints. This prompts the question: can we use information such as lengths and angles between different body parts as an additional objective for the pose estimation model to improve its performance?

To investigate this, I explore incorporating an additional objective to train the pose estimation model. This extra target involves predicting vector fields, that model the relationship between body joints. These vector fields capture the spatial relationships between different body parts by representing them as directed line segments with associated directions. Although the use of such vector fields was originally introduced by Cao et al. (2017) as part affinity fields (PAFs) for bottom-up pose estimation, they could also be beneficial in a top-down framework if the model successfully learns the human body structure, leading to improved performance under occlusion. While they use an arbitrary width to define the vector field, I use a straight line segment. By integrating vector field prediction as an additional objective, the pose estimation model may better understand human body structure and be more robust in challenging situations, such as occlusions and complex poses.

3.4.5 Other Techniques and Model Variants

Apart from the approaches discussed above, I test other techniques to improve performance. CutOut (DeVries & Taylor, 2017) data augmentation randomly removes square

patches from the image during training. This technique is tested to assess whether it helps the model in dealing with occlusion by forcing the model to learn the relationship between the hidden joints and visible joints.

For each model used, I test two different model variants are based on their performance in ImageNet classification (Deng et al., 2009). The smallest models that achieve accuracy scores of 75% and 80% on ImageNet are used as backbone encoders for comparison with other models.T2TViT-S and EdgeNeXt-Small are used to explore the effect of different hyperparameters.

3.5 Metrics for Evaluation

To comprehensively evaluate the performance of the proposed pose estimation models and training strategies, a combination of quantitative and qualitative metrics is used:

- **Pose estimation performance**: Model performance is assessed on the COCO validation set, the the OCHuman dataset, and the wheelchair dataset, comparing them with similar lightweight models as well as larger accuracy-focused models.
- **Action classification performance**: Models are also assessed based on the action classification performance in the manual wheelchair dataset to evaluate their usefulness in this task.
- **Inference performance**: Inference performance is measured to compare the models' effectiveness in situations requiring fast inference time.
- **Qualitative Analysis**: A qualitative analysis is conducted to gain insight about the models' performance under various scenarios.

These evaluation metrics provide a comprehensive understanding of the models' performance, efficiency, and applicability in different use cases.

3.5.1 Object Keypoint Similarity

The object keypoint similarity (OKS) metric was introduced in the COCO dataset (Lin et al., 2014) as a way to evaluate HPE performance. It essentially measures how close the predicted joint locations are to the ground-truth joint locations. The OKS metric is defined as

$$OKS = \frac{\sum_i [\exp(-\frac{d_i^2}{2s^2\kappa_i^2})\delta(v_i > 0)]}{\sum_i [\delta(v_i > 0)]}, \tag{3.5}$$

where d_i gives the distance between the predicted and ground-truth keypoint at index i, v_i is the visibility flag of the keypoint, s gives the scale of the person in the image, and

k_i is a constant used for normalization or controlling fall-off. Using this equation, the OKS is calculated between a set of predicted and ground-truth keypoints. The prediction is considered correct if the OKS value is above the given threshold (for example, 0.5) and incorrect otherwise. This result is then used to calculate average precision (AP) and average recall (AR). The final AP and AR values are calculated using the mean of the precision and recall values respectively over recall or precision values between 0.5 and 0.95 with an interval of 0.05. This is the metric used to compare the performance of models in the wheelchair dataset, the OCHuman dataset, and the COCO dataset in this study.

3.6 Dataset

As discussed in Section 2, to my knowledge, there is not as of yet any publicly available dataset for wheelchair-based HPE and activity recognition. Hence, I collected my own dataset. To come up with the list of activities to collect, I prepared a list of commonly considered activities in the literature on sensor-based activity recognition for wheelchair users. The final set of activities were picked from the list with the help of experts. Table 3.1 shows the script used for data collection and the classes of data collected. The participants were asked to mimic wheelchair users as much as possible. For example, they were instructed to rely as little as possible on their legs while transferring to and from the wheelchair, in order to simulate the experience of actual wheelchair users more accurately.

Eight cameras were installed on the ceiling of a room to capture the recordings from different angles. A manual wheelchair was used to perform wheelchair-based activities. For transferring between the wheelchair and another medium, a couch was used. To minimize potential injury during voluntary falls, a mattress was placed beneath the couch. Figure 3.7 illustrates the camera setup and provides an overview of the laboratory environment. This was followed by labeling the timestamps at which each activity starts and ends for each participant.

I ran Yolo-X Large (Ge, Liu, Wang, Li, & Sun, 2021) and ViTPose-H (Xu et al., 2022) on extracted frames to generate initial annotations in the COCO format. Then I used COCO Annotator (Brooks, 2019) to manually refine the bounding boxes and pose estimates. Currently, about 1500 images with a total of about 1700 person instances have been annotated in the COCO format. These annotated images can be used to fine-tune an

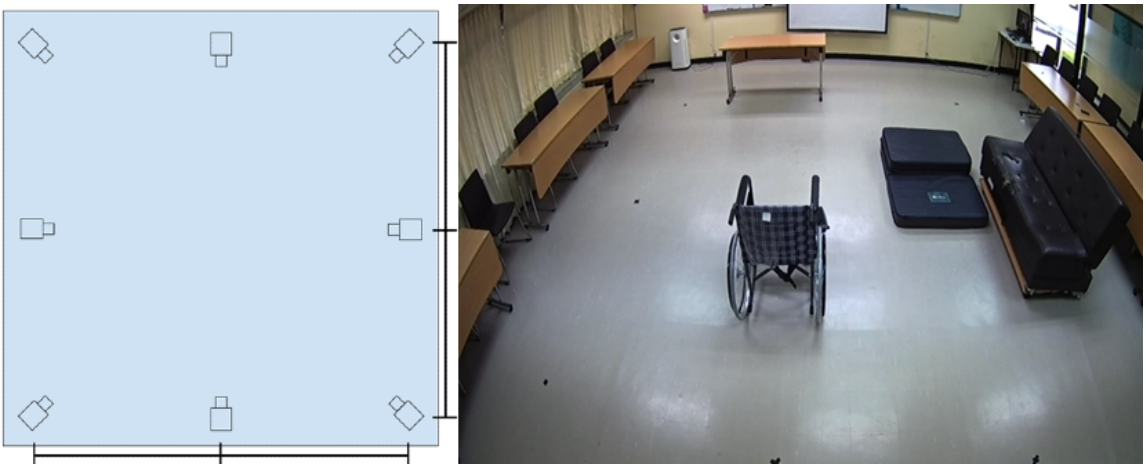
Table 3.1

Script used for data collection

Activity/Instruction	Time/Number of Times
Propel the wheelchair	1 time around the room
Reach the desk	1 time
Pick left weight from the floor and place it on the desk	1 time
Pick right weight from the floor and place it on the desk	1 time
Pick left weight from the desk and place it on the floor	1 time
Pick right weight from the desk and place it on the floor	1 time
Deskwork (Use computer)	10 seconds
Phone call, drink, eat	10 seconds each
Move to the middle of the room	1 time
Bend (alternate between trying to touch the floor)	3 times
Exercise	10 seconds
Rest with hands on the wheels	5 seconds
Rest with hands on the lap	5 seconds
Move towards the bed/couch	1 time
Transfer from the wheelchair to the couch	1 time
Rest (Lie on the couch)	10 seconds
Transfer from the couch to the wheelchair	1 time
Go back to the wheelchair	1 time
Fall while transferring from the wheelchair to the couch	1 time
Go back to the couch	1 time
Fall while transferring from the couch to the wheelchair	1 time
Go back to the middle of the room	1 time
Get propelled by someone else	1 time around the room

Figure 3.7

Data Collection Camera Setup



(Left) The camera setup used for data collection. Eight cameras were used to capture recordings at different angles. (Right) The view from Camera 4. This picture shows the initial wheelchair position and the locations of the desk, mattress, and the couch.

object detection model or test the performance of pose estimation models on this dataset.

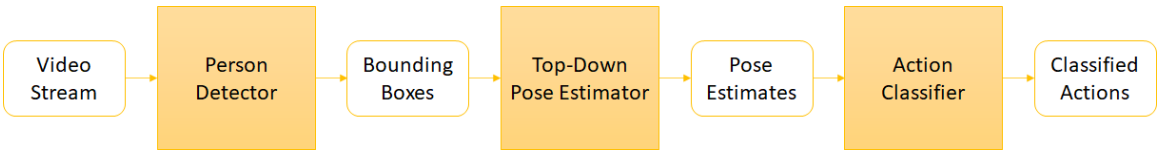
3.7 Action Classification

For action classification, a pose estimation method such as ViTPose is used to first extract keypoints, which are then used to train an action classification model. A sequence of pose estimates from 28 frames is taken and fed into a 1D-CNN model, along with a label, to train it for action classification. Upon completion of action classification training, the entire framework for action classification starts by running an object detection model to obtain person detections and cropped bounding box regions. These are then fed to the pose estimator to acquire the pose estimates, and finally, a sequence of 28 pose estimates is input to the action classification model to determine the person's action. One of the primary objectives of identifying an efficient pose estimation technique is to achieve a better speed-accuracy trade-off for action classification. Therefore, performance in action classification is also an important factor to consider.

To evaluate action classification performance, a held-out test set is used. This test set contains videos from a participant not used in the training set. A sliding window is applied over all video samples in the test set, extracting the ground-truth action over a batch of 28 frames. For each batch of these 28 frames, the action that appears in the highest number of frames is selected as the ground-truth action for that batch. Simul-

Figure 3.8

Action Classification Framework



The video stream is first fed to the person detector, which returns the bounding box for each person in the image. These are fed to a pose estimator such as ViTPose, which produces pose estimates for each person. A sequence of such pose estimates is then fed to the action classification model, which finally predicts the action.

taneously, these batches are passed through the action classification pipeline mentioned earlier and depicted in Figure 3.8 to extract the predicted actions. These data are used to calculate the accuracy, average precision, average recall, and the precision and recall for each class. Different pose estimation models are tested in the framework to compare their performance in action classification.

CHAPTER 4
EXPERIMENTAL RESULTS

This chapter discusses the results I obtained when testing the efficient ViT variants, training strategies, and action classification performance on the wheelchair dataset as outlined in the methodology section.

4.1 Pose Estimation Performance

In this section, I compare the human pose estimation (HPE) performance of the efficient ViT variants with each other, as well as with state-of-the-art methods. The goal is to identify models with a good balance between accuracy and computational efficiency, which will be useful for action classification.

4.1.1 Comparing Efficient ViT Variants

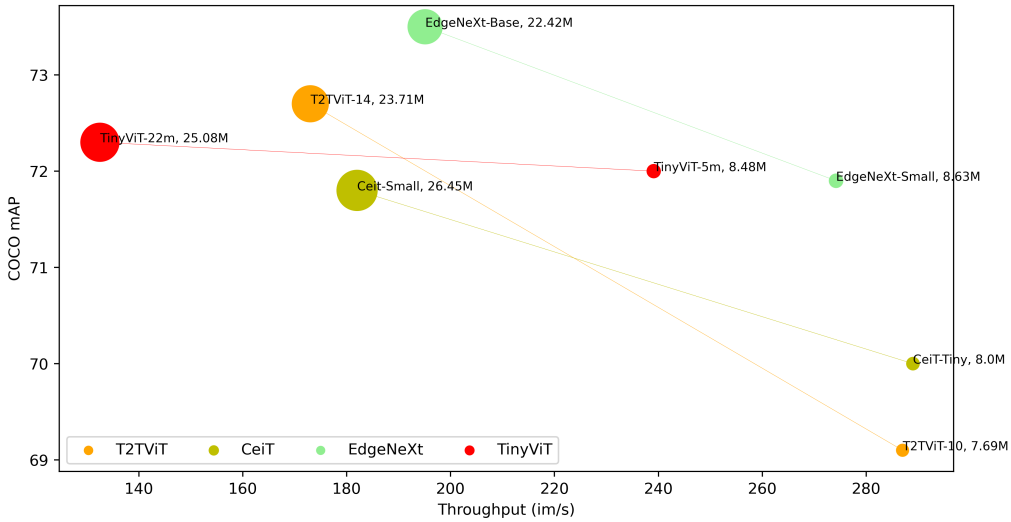
Table 4.1 compares various computationally efficient ViT variants based on their number of parameters, throughput speed while running an epoch on the validation set in images per second (im/s), and accuracy score on the COCO dataset. CrossViT and LeViT backbones struggle to achieve the same level of accuracy as other models. However, the smaller variants of the remaining models demonstrate a good speed-accuracy trade-off. Considering the larger models that also provide good accuracy, EdgeNeXt-Base stands out, with the best AP score on the COCO dataset at 73.5 and a throughput of 195.1 im/s. Despite these positive results, none of the models excel at handling occlusion, as evidenced by their poor performance in the OCHuman dataset.

4.1.2 Comparison with state of the art

In Table 4.2 and Figure 4.2, I compare the performance of efficient ViT variants (CeiT, T2TViT, and EdgeNeXt) with modern HPE techniques. As in Section 4.1.1, the throughput speed is the of a validation epoch on the COCO validation set. Among the lightweight models, efficient ViT variants demonstrate substantial improvements over LiteHRNet. For example, T2TViT-10 achieves a COCO AP of 69.1 with a throughput of 287 im/s, while LiteHRNet-18 achieves a COCO AP of 64.3 with a similar throughput of 294.5 im/s. CeiT and EdgeNeXt similarly outperform LiteHRNet in this regard. This demonstrates the superior performance of efficient ViT variants in terms of both accuracy and computational efficiency.

Figure 4.1

Pose Estimation Performance of Efficient ViT Variants



Pose estimation performance of efficient ViT variants, illustrating the trade-off between accuracy and throughput. T2TViT-S, CeiT-Tiny, EdgeNeXt-Small, and TinyViT-S seem suitable if speed is a priority, as they provide a relatively high throughput. Among the larger variants, the EdgeNeXt-Base model achieves the highest accuracy, requires the fewest number of parameters, and also offers the highest throughput.

Table 4.1

Pose Estimation Performance of Efficient ViT Variants

Model	#Params(M)	im/s	ImageNet Accuracy	COCO AP	OCHuman AP
T2TViT-10	**7.69**	287	75.2	69.1	52.4
T2TViT-14	23.71	173	81.5	72.7	56.8
CrossViT-9	10.66	299	73.9	62.5	44.4
CrossViT-S	26.7	168	81.0	68.8	49.3
CeiT-Tiny	8.0	289	76.4	70.0	53.2
Ceit-Small	26.45	182	82.0	71.8	56.1
LeViT-128S	9.8	**368.5**	76.6	60.6	44.5
LeViT-192	13.0	314	80.0	65.7	50.4
EdgeNeXt-S	8.63	274.2	79.4	71.5	54.2
EdgeNeXt-B	22.42	195.1	82.5	**73.5**	**56.9**
TinyViT-5	8.48	239.1	80.7	72	55.6
TinyViT-22	25.08	132.5	**84.8**	72.3	54.7

Figure 4.2

Comparison of Pose Estimation Performance with state of the art

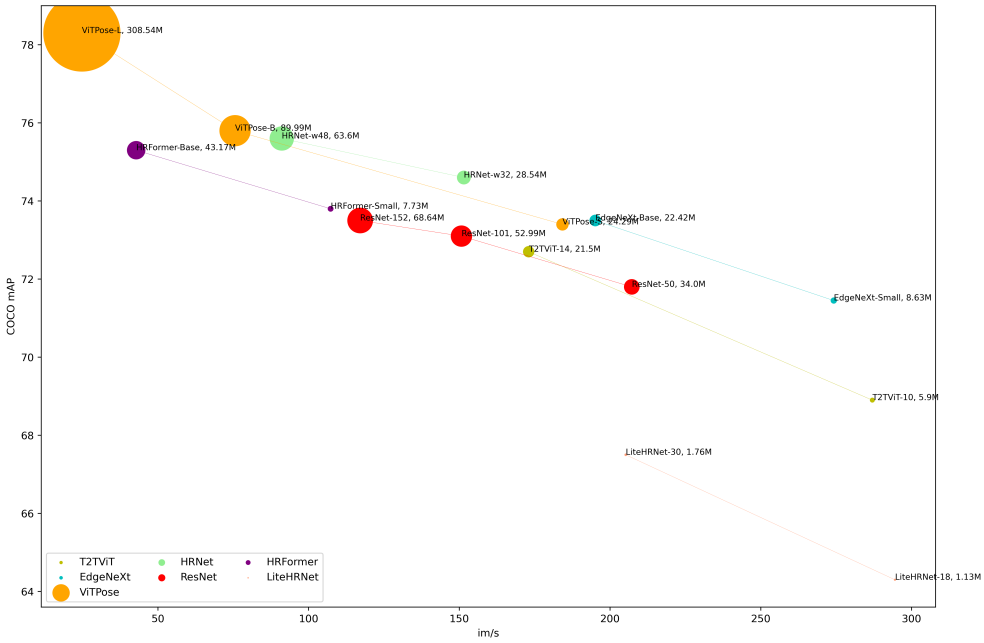


Table 4.2

Comparison with state of the art

Model	Parameters(m)	im/s	COCO AP
T2TViT-10	7.69	287	69.1
T2TViT-14	23.71	173	72.7
CeiT-Tiny	8.0	289	70.0
Ceit-Small	26.45	182	71.8
EdgeNeXt-Small	8.63	274.2	71.5
EdgeNeXt-Base	22.42	195.1	73.5
ViTPose-B	89.99	75.6	75.8
ViTPose-L	308.54	24.8	**78.3**
HRNet-w32	28.54	151.5	74.6
HRNet-w48	63.6	91.1	75.6
ResNet-50	34	207.2	71.8
ResNet-101	52.99	150.7	73.1
ResNet-152	68.64	117.1	73.5
HRFormer-Small	7.73	107.3	73.8
HRFormer-Base	43.17	42.8	75.3
LiteHRNet-18	**1.13**	**294.5**	64.3
LiteHRNet-30	1.76	205.2	67.5

Table 4.3

Inference Speed Comparison with State-of-the-art

Model	COCO AP	Params	CPU (BS 1) im/s	GPU (BS 4) im/s	GPU (BS 8) im/s
T2TViT-S	69.1	5.9	11.65	**55.52**	26.96
T2TViT-L	72.7	21.5	6.66	37.71	18.9
EdgeNeXt-S	71.9	8.63	12.05	49.62	**27.67**
EdgeNeXt-B	73.5	22.42	8.34	41.17	19.93
TinyViT-S	72	8.48	10.07	46.67	23.14
TinyViT-L	72.3	25.08	6.01	30.47	15.29
ViTPose-B	75.8	89.99	2.45	21.25	9.51
ViTPose-L	78.3	308.54	0.82	7.27	3.15
ViTPose-H	**79.1**	637.21	0.41	3.81	1.67
HRNet-w32	74.6	28.54	8.62	17.85	13.5
HRNet-w48	75.6	63.6	5.4	17.12	8.1
ResNet-50	71.8	34	10.94	45.98	22.39
ResNet-101	73.1	52.99	7.79	29.55	15.61
ResNet-152	73.5	68.64	6.96	21.64	11.7
LiteHRNet-18	64.3	**1.13**	**15.7**	18.61	13.79
LiteHRNet-30	67.5	1.76	9.85	11.13	9.02

Even considering larger models, EdgeNeXt-Base stands out as a competitive alternative to state-of-the-art models. With 22.42 million parameters and a throughput of 195.1 im/s, EdgeNeXt-Base attains a COCO AP of 73.5, which is comparable to ResNet-152-S's AP of 73.5, achieved with 68.64 million parameters and a throughput of 117.1 im/s. EdgeNeXt-Base also compares favorably to HRNet-w32, which achieves a COCO AP of 74.6 with a larger number of parameters (28.54 million) and a lower throughput (151.5 im/s).

In Table 4.3, Figure 4.3, and Figure 4.4, I compare the inference performance of the efficient ViT variants with state-of-the-art methods using a Ryzen 7 5800X CPU and an Nvidia RTX 3060Ti GPU. EdgeNeXt-S and T2TViT-S consistently show the highest speed among ViT variants. Among CNN-based architectures, ResNet-50 offers a good speed and accuracy trade-off, while the remaining models appear less efficient. LiteHR-

Figure 4.3

CPU Inference Speed

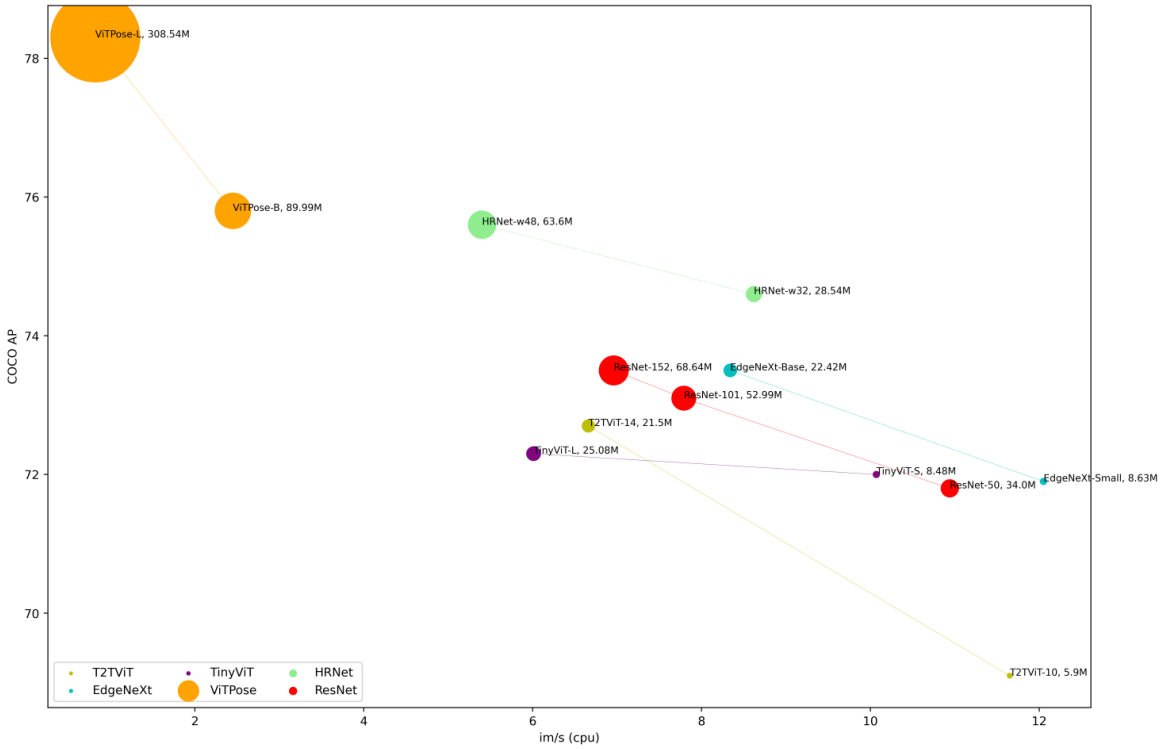


Figure 4.4

GPU Inference Speed

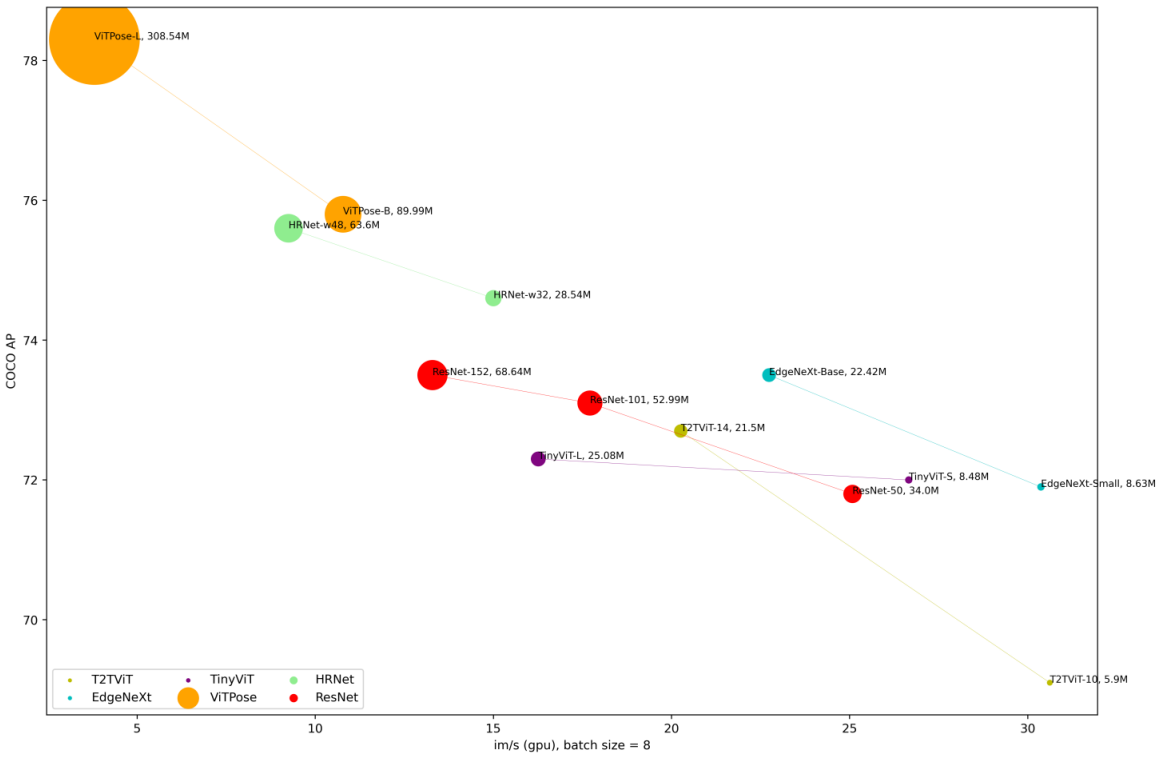


Table 4.4

Improvements From Knowledge Distillation

Model	Distillation	α	Batch Size	COCO AP	OCHuman AP
T2TViT-S	No	0	64	68.8	51.9
T2TViT-S	Yes	0.5	64	69.2	52.5
T2TViT-S	Yes	1	64	69.3	53.2
T2TViT-S	No	0	128	69.1	53.2
T2TViT-S	Yes	1	128	69.6	53.6
EdgeNeXt-S	No	0	64	71.5	54.2
EdgeNeXt-S	Yes	0.5	64	71.9	55.23
EdgeNeXt-S	No	0	128	71.9	54.4
EdgeNeXt-S	Yes	1	128	72.4	57.0

Net models provide the fastest CPU inference, but compromise on accuracy. On the other hand, ViTPose-B, ViTPose-L, and ViTPose-H achieve high accuracy at the cost of slower inference speeds. In summary, T2TViT-S and EdgeNeXt-S stand out as the most efficient models, providing good accuracy with a fast inference speed.

4.2 Knowledge Distillation Results

Table 4.4 presents the results of applying knowledge distillation to the T2TViT-S and EdgeNeXt-S models. α balances the weight between the teacher model and ground-truth during training, with a value of 1 indicating that the model learns solely from the teacher, without ground-truth data being involved in the loss calculation (refer to Equation 3.2). The teacher model used in this study, ViTPose-B, achieves an average precision (AP) of 77.51 on the COCO dataset and 88.0 on the OCHuman dataset.

The results indicate that the application of knowledge distillation using a large teacher model yields significant improvements in the performance of these efficient models compared to training with ground-truth data alone. The improvements are consistent across different models and batch sizes. For example, the T2TViT-S attains AP scores of 69.25 and 69.3 when trained with α values of 0.5 and 1, respectively, and a batch size of 64. The results suggest that relying entirely on the teacher model is more effective than using any ground-truth data, which contrasts with the findings of F. Zhang et al. (2019), where an optimal α value of 0.5 was reported.

Table 4.5

Improvements From Multi-Task Training

Model	Size	Epochs	Train	COCO	OCHuman
ViTPose	B	210	COCO*	75.8	59.6
ViTPose	B	210	Multi*	77.5	88.3
T2TViT	S	210	COCO	68.8	51.9
T2TViT	S	110	Multi	68.4	70.5
T2TViT	S	210	Multi	69.3	73.3
EdgeNeXt	S	210	COCO	71.5	56.9
EdgeNeXt	S	110	Multi	71.6	74.0

*: Models taken from the original repository.

However, while the performance on the OCHuman dataset does show some improvement, the models seem unreliable in handling occlusion, as their scores remain considerably lower than that achieved by ViTPose-B. This highlights that knowledge distillation alone cannot provide these models with the ability to deal with occlusion.

4.3 Improvements From Multi-Task Training

The multi-task training strategy discussed in Section 3.4.2 uses multiple datasets to train the backbone encoder. In this case, the COCO dataset is used as the primary dataset, with the AIC dataset and the MPII Human Pose dataset as the secondary datasets. Table 4.5 presents the improvements obtained from this multi-task training on ViTPose, T2TViT, and EdgeNeXt models.

ViTPose-B trained in the multi-task setting achieves a higher COCO AP (77.5) and a substantially higher OCHuman AP (88.3) compared to training just on COCO. For T2TViT-S, multi-task training for 110 and 210 epochs leads to higher OCHuman scores of 70.48 and 73.3, respectively, compared to training on COCO for 210 epochs (OCHuman AP of 51.9). The COCO AP remains comparable across different training settings. EdgeNeXt-S shows similar improvements with multi-task training for 110 epochs.

These results demonstrate the benefits of multi-task training for improving the generalization capabilities of pose estimation models, especially on more challenging datasets like OCHuman that primarily focus on occluded humans. Multi-task training is the pri-

Table 4.6

Results of using GAN Setup

Model	Target	disc_lr	λ_G	COCO AP
T2TViT-S	None	-	-	69.1
T2TViT-S	Heatmap	2.00E-05	0.0001	61.44
T2TViT-S	Keypoints	2.00E-04	0.001	66.17

Table 4.7

Results from using Vector Fields as Additional Targets

Model	Extra Target	Loss Weight	COCO AP	OCHuman AP
T2TViT-S	No	-	69.1	53.2
T2TViT-S	Vector Field	0.1	69.3	52.1
T2TViT-S	Vector Field	1.0	68.9	52.1

mary factor to consider if there is a need to deal with occlusion.

4.4 GAN Setup

In Table 4.6, I show the results of training the pose estimation models under a GAN setup. I tested with two different targets: heatmaps and keypoints for the discriminator. For the heatmaps, the predicted heatmaps and the ground-truth heatmaps were fed to the discriminator to distinguish between them. For the keypoints, the keypoint coordinates were used instead by extracting them from the respective heatmaps.

The results show that this setup was unable to improve the performance of the T2TViT-S model that I tested. Instead, the accuracy dropped substantially. This is possibly because the model was trying to produce outputs that deviate from the actual ground-truth in trying to fool the discriminator.

4.5 Predicting Vector Fields as Additional Targets

Table 4.7 presents the results of using an additional objective during training. This additional objective involves predicting vector fields that show the connection between different body joints. Although the model appeared to perform better when the loss from the vector prediction was weighed down to 0.1, it was unable to achieve the same level of accuracy when this was set to 1. However, this method requires additional exploration with hyperparameters such as the learning rate, learning rate schedule, number

Table 4.8

T2TViT-S Hyperparameters

Batch Size	Drop Path	Layer Decay	QKV bias	COCO AP	OCHuman AP
64	0.3	0.75	No	68.3	51.2
64	0.3	0.8	No	68.8	51.5
64	0.3	0.9	No	68.6	50.9
64	0.4	0.8	No	68.5	51.3
64	0.3	0.8	Yes	68.8	51.9
128	0.3	0.8	Yes	69.1	53.4
256	0.3	0.8	Yes	68.8	51.8

of epochs, etc. before we draw any conclusions.

4.6 Model Optimization and Other Techniques

This section discusses the various hyperparameter settings and other techniques tested to improve performance.

4.6.1 Different Hyperparameters with T2TViT

Table 4.8 shows the results of the T2TViT-S model's performance under various hyperparameter configurations, such as batch size, drop path rate, layer decay, and QKV bias. The results indicate that a layer decay rate of 0.8 and the inclusion of QKV bias contribute to better performance. The optimal drop path rate is 0.3. Additionally, increasing the batch size to 128 further improves the performance, while going beyond that appears to have negative effects.

Additionally, using CutOut data augmentation with the optimal hyperparameter configurations led to a COCO AP of 68.5 and an OCHuman AP of 51.4. Although I expected CutOut data augmentation to improve the model's performance, particularly in handling occlusion, it appears to have had a negative effect. The negative impact of CutOut data augmentation might be due to the failure of the model to learn effectively from the occluded regions introduced by CutOut.

4.6.2 Effect of Pretraining Distillation with TinyViT

The results in Table 4.9 indicate that pretraining distillation with TinyViT does not significantly impact its performance in HPE tasks, despite improving its accuracy on the

Table 4.9

Testing Pretraining Distillation with TinyViT

Distillation	Pretraining Dataset	ImageNet-1K Accuracy	COCO AP
No	1K	79.1	72.0
Yes	1K distil	80.7	72.0
Yes	21K to 1K distil	80.7	71.9

ImageNet-1K dataset. This contrasts with K. Wu et al. (2022)'s findings about pretraining distillation improving the performance for downstream tasks. Although they demonstrate that pretraining TinyViT models on ImageNet using knowledge distillation helps in some downstream tasks such as object detection and segmentation, it seems that there is not much improvement in pose estimation performance.

4.7 Qualitative Analysis

This section presents some qualitative results of the EdgeNeXt-Small-Multi model on images from various datasets, including COCO, MPII, CrowdPose, OCHuman, and the wheelchair dataset (Figure 4.5). The model performs well under a variety of conditions such as blur, objects at various scales, occlusion by objects or humans, complex poses, and diverse clothing.

4.8 Performance on the Wheelchair Dataset

This section discusses the performance of various HPE methods both in pose estimation and action classification on the wheelchair dataset.

4.8.1 HPE Performance on the Wheelchair Dataset

Table 4.10 demonstrates the performance of efficient ViT-based HPE methods, ViTPose variants, and other common HPE methods on the wheelchair dataset. The results highlight that ViTPose models, especially those using multi-task training, achieve the highest scores. Additionally, the HRNet and ResNet backbones also perform well. However, the efficient ViT variants show competitive performance while being substantially more compact and faster, as demonstrated by the number of parameters and throughput speed in Table 4.2.

Consistent with results from other datasets, EdgeNeXt emerges as the best model in terms of the speed-accuracy trade-off. In particular, when trained with multi-task train-

Figure 4.5

Qualitative results on different datasets

Qualitative results (from top: COCO, CrowdPose, MPII, OCHuman, wheelchair datasets) using EdgeNeXt-Small-Multi. Other models show similar results.

Table 4.10

HPE Performance on the Wheelchair Dataset

Model	Parameters(m)	COCO AP	Wheelchair AP
T2TViT-S	7.69	69.1	74.24
T2TViT-S-Multi	7.69	69.3	77.26
T2TViT-L	23.71	72.7	78.76
CeiT-Tiny	8.0	70.0	76.3
Ceit-Small	26.45	71.8	78.44
EdgeNeXt-Small	8.63	71.5	77.2
EdgeNeXt-Small-Multi	8.63	71.6	80.3
EdgeNeXt-Base	22.42	73.5	80.2
ViTPose-B	89.99	75.8	81.96
ViTPose-B-Multi	89.99	77.1	85.34
ViTPose-L	308.54	78.3	86.20
ViTPose-L-Multi	308.54	78.7	88.20
ViTPose-H	637.21	79.1	87.23
ViTPose-H-Multi	637.21	**79.5**	**88.79**
HRNet-w32	28.54	74.6	81.89
HRNet-w48	63.6	75.6	82.90
ResNet-50	34	71.8	79.11
ResNet-101	52.99	73.1	80.22
ResNet-152	68.64	73.5	81.76
HRFormer-Small	7.73	73.8	70.41
HRFormer-Base	43.17	75.3	73.31
LiteHRNet-18	**1.13**	64.3	67.32
LiteHRNet-30	1.76	67.5	72.03

Table 4.11

Action Classification Performance on the Wheelchair Dataset

Model	Accuracy	AP	AR
T2TViT-S	73.72	72.11	67.82
T2TViT-S-Multi	75.56	75.03	70.46
T2TViT-L	77.41	74.92	71.24
EdgeNeXt-S	75.71	74.46	70
EdgeNeXt-S-Multi	75.92	73.9	70.51
EdgeNeXt-B	77.55	76.51	71.58
ViTPose-B	78.67	79.44	72.8
ViTPose-B-Multi	**81.38**	**80.52**	**74.98**
HRNet-w32	78.17	76.61	71.17
HRNet-w48	79.83	79.38	72.8
ResNet-50	75.8	73.78	69.11
ResNet-101	75.78	74.71	69.29
ResNet-152	78.85	78.73	73.17
LiteHRNet-18	64.68	64.53	59.12
LiteHRNet-30	68.77	68.57	62.72

ing, EdgeNeXt-Small appears to be the most suitable choice for deployment in resource-constrained environments. Selecting such a lightweight model would also provide more compute resources for other tasks in the action classification pipeline, such as object detection and action classification.

In Section 4.3, we observed a substantial improvement in performance on the OCHuman dataset with multi-task training. The results on the wheelchair dataset also show a substantial boost in performance for models trained using multi-task training. This approach enables the smaller model variants to achieve performance levels close to their larger counterparts, and in some cases, to even outperform them. The results agree with the initial hypothesis that the ability to handle occlusion would be useful for improving HPE performance for people in wheelchairs.

Table 4.12

Performance of Propel and Fall on the Wheelchair Dataset

	Propel		Fall	
Model	**Precision**	**Recall**	**Precision**	**Recall**
T2TViT-S	90.69	89.83	34.69	35.11
T2TViT-S-Multi	90.73	89.74	48.08	36.74
T2TViT-L	91.18	95	34.71	29.24
EdgeNeXt-S	91.68	95.22	45.01	**37.28**
EdgeNeXt-S-Multi	89.62	91.15	34.73	32.28
EdgeNeXt-B	90.35	94.65	45.44	34.13
ViTPose-B	91.1	94.22	**58.2**	30.87
ViTPose-B-Multi	**91.34**	**97.45**	54.16	27.61
HRNet-w32	90.4	95.8	34.39	32.93
HRNet-w48	90.31	94.87	47.1	29.02

4.8.2 Action Classification Performance on the Wheelchair Dataset

Table 4.11 displays the action classification performance of various pose estimation models on the wheelchair dataset. The table reports the accuracy, average precision (AP), and average recall (AR) for all classes. In addition, the precision and recall scores for the *fall* and *propel* classes are included in Table 4.12.

Most of the models exhibit similar performance in terms of accuracy, AP, and AR, with ViTPose-B-Multi (trained with multi-task training) achieving the highest scores. Among other methods, HRNet variants also perform well, consistent with the results of Table 4.10. The efficient ViT-based methods demonstrate competitive performance, despite being lightweight. In contrast, the LiteHRNet models cannot match the performance of other models, which aligns with their performance in previous comparisons.

Examining the precision and recall scores for the *fall* class reveals inconsistencies compared to the other classes. Falls occur rapidly and are present only in a small number of frames in the dataset, making the training data for the *fall* class limited. As a result, the models struggle to perform well in this class. This suggests the need for an expanded dataset with more fall instances to improve precision and recall scores. In contrast, the scores for the *propel* class are much higher across all models, implying that this action

might be relatively easier for the models to predict.

CHAPTER 5
CONCLUSION AND RECOMMENDATIONS

In this chapter, I summarize the key findings of the thesis and make recommendations for the future.

5.1 Key Findings

In conclusion, my work explored various strategies to improve the performance of efficient ViT backbones in human pose estimation (HPE), particularly in the presence of occlusion.

I replaced the original ViT backbone in ViTPose with six efficient ViT variants, namely T2TViT, CrossViT, CeiT, LeViT, EdgeNeXt, and TinyViT. I tested the hypothesis that purely transformer-based architectures, such as T2TViT and CrossViT, would be better at dealing with occlusion and result in superior performance. However, the results indicated that these pure transformer-based architectures had no advantage over hybrid variants that incorporate convolutional layers in their design. Among the efficient ViT variants, CrossViT and LeViT did not provide satisfactory accuracy, while others demonstrated better speed-accuracy trade-offs compared to state-of-the-art methods in HPE.

I tested knowledge distillation to further improve the performance of efficient models. Using a ViTPose-B model as a teacher network, efficient variants improved consistently, showing on average an increase of 0.5 in the AP score in the COCO data set.

I also explored multi-task training (used by Xu et al. (2022)) in this work, which involved combining multiple data sets: COCO, AI Challenger, and MPII human pose datasets for training. This strategy yielded significant improvements on the OCHuman dataset, indicating that multi-task training enables the model to learn to handle occlusion more effectively. The ability to deal with occlusion is crucial for accurate HPE, especially in challenging scenarios with overlapping body parts and complex backgrounds. The results show that the efficient ViT backbones can benefit from multi-task training to improve their performance in HPE tasks without adding computational complexity.

The additional training target, which involved regressing vector fields formed between connected joints in the human skeleton, was incorporated into the training process with

the hypothesis that it would enable the model to learn plausible human body configurations. Although this strategy resulted in minor performance improvements, it did not significantly impact performance in the challenging OCHuman dataset, suggesting that the additional training target may not be sufficient to fully capture complex body configurations in occluded scenes.

I prepared a wheelchair-based HPE and activity recognition dataset to facilitate the development of a system that effectively monitors the activities of people using wheelchairs. I compared the HPE and action classification performance among several techniques. For pose estimation on the wheelchair dataset, efficient ViT variants demonstrated a favorable speed-accuracy trade-off. Additionally, multi-task training improved the performance on the wheelchair dataset, emphasizing that the ability to deal with occlusion is crucial for accurate pose estimation of people sitting in wheelchairs.

Comparing the performance of action classification, most of the models showed similar results in terms of accuracy, AP, and AR. However, the ViTPose-B-Multi model, which was trained with multi-task training, achieved the highest scores. One challenge faced by the models was limited training data for the fall class, as falls occur rapidly and are present across only a few number of frames in the dataset. Consequently, the models struggled to perform well in this class, suggesting the need for an expanded dataset with more fall instances to improve precision and recall scores for better action classification performance. In summary, the results highlight the potential of efficient ViT variants and training strategies to improve HPE performance in resource-constrained environments.

5.2 Future Research Directions and Recommendations

The following future directions are recommended to improve the performance of efficient ViT variants in HPE:

- Exploration of more efficient ViT architectures: Since the efficient ViT variants tested in this work do not cover all possible architectures, exploring additional variants or combining ideas from existing ones could lead to the discovery of architectures that are more naturally suited for HPE tasks.
- Investigating different teacher architectures for knowledge distillation: Touvron, Cord, Douze, et al. (2021) demonstrate that knowledge distillation from a CNN architecture could result in better performance compared to distillation from a transformer-based model. Examining different teacher architectures and their im-

pact on pose estimation performance could reveal new insights and improve model performance.

- Understanding the impact of multi-task training: While the multi-task training strategy significantly boosts performance, especially on the OCHuman dataset, perhaps as a type of regularization, the exact reasons behind this improvement are not yet clear. Gaining a deeper understanding of the factors contributing to this performance boost could enable optimization of the training process, making it more resource efficient and compatible with other techniques. For instance, combining it with knowledge distillation would either require a teacher trained under the same multi-task setting, or separate teachers for each dataset. The latter would require substantially more compute resources during training, as well as teacher models trained for every dataset beforehand.

Expanding the wheelchair dataset to include more diverse and realistic scenarios is crucial for the development of more robust and practical human pose estimation and action classification systems. This can be achieved by considering the following:

- Gathering data in real-world environments: Instead of relying solely on laboratory environments, it is important to collect data from various real-world settings, such as homes, offices, and outdoor locations. This would enable the model to generalize better and handle a wider range of situations that wheelchair users may encounter.
- Including more fall instances: Performance in the *fall* class highlights the need for more fall instances in the dataset. I collected and added more fall data to the original dataset, which improved the precision and recall scores of the fall class by a substantial margin. This shows that the fall class in particular requires more samples than other classes.
- Involving real wheelchair users: To ensure that the dataset is representative and meets the needs of wheelchair users, it may be necessary to involve real wheelchair users in data collection. This would help capture more authentic scenarios in the way people use wheelchairs, making the system more applicable to real-life situations.
- Diversifying the dataset: Expanding the dataset to cover different types of wheelchair, users with varying physical abilities, and a range of daily activities would make the system more versatile and adaptable to different user needs.

REFERENCES

Aihara, S., Sakai, T., Shibata, R., Matsubara, T., Mizukami, R., Yoshida, Y., & Shionoya, A. (2022). Deep learning model for integrated estimation of wheelchair and human poses using camera images. In *Isea 2022 – the engineering of sport 14*. Purdue University. Retrieved from `https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1070&context=resec-isea`

Andriluka, M., Iqbal, U., Milan, A., Insafutdinov, E., Pishchulin, L., Gall, J., & Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5167-5176.

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014, June). 2d human pose estimation: New benchmark and state of the art analysis. In *Ieee conference on computer vision and pattern recognition (cvpr)*.

Artacho, B., & Savakis, A. (2020). Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7035–7044).

Bao, H., Dong, L., & Wei, F. (2022). Beit: Bert pre-training of image transformers. *ArXiv*, *abs/2106.08254*.

Bin, Y., Cao, X., Chen, X., Ge, Y., Tai, Y., Wang, C., ... Sang, N. (2020). Adversarial semantic data augmentation for human pose estimation. In *Eccv*.

Booth, F., Roberts, C., & Laye, M. (2012, 04). Lack of exercise is a major cause of chronic diseases. *Comprehensive Physiology*, *2*, 1143-1211. doi: 10.1002/cphy.c110025

Brooks, J. (2019). *COCO Annotator*. `https://github.com/jsbroks/coco-annotator/`.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7291–7299).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *ArXiv*, *abs/2005.12872*.

Chen, C.-F. R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. , 357–366.

Chen, H., Feng, R., Wu, S., Xu, H., Zhou, F., & Liu, Z. (2022). 2d human pose estimation: A survey. *arXiv preprint arXiv:2204.07370*.

Chen, Y., Shen, C., Wei, X.-S., Liu, L., & Yang, J. (2017). Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the ieee international conference on computer vision* (pp. 1212–1221).

Chen, Y., Tian, Y., & He, M. (2020). Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, *192*, 102897.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., & Sun, J. (2018). Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7103–7112).

Cheng, B., Schwing, A. G., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. In *Neurips.*

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5386–5395).

Chou, C.-J., Chien, J.-T., & Chen, H.-T. (2018). Self adversarial training for human pose estimation. In *2018 asia-pacific signal and information processing association annual summit and conference (apsipa asc)* (pp. 17–30).

Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A. L., & Wang, X. (2017). Multi-context attention for human pose estimation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5669-5678.

Debnath, B., O'Brien, M., Yamaguchi, M., & Behera, A. (2018). Adapting mobilenets for mobile based upper body pose estimation. In *2018 15th ieee international conference on advanced video and signal based surveillance (avss)* (p. 1-6). doi: 10.1109/AVSS.2018.8639378

Debnath, B., O'Brien, M., Yamaguchi, M., & Behera, A. (2021). A review of computer vision-based approaches for physical rehabilitation and assessment. *Multimedia Systems*, 1–31.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

de Vries, W. H. K., Amrein, S., Arnet, U., Mayrhuber, L., Ehrmann, C., & Veeger, H. E. J. (2022). Classification of wheelchair related shoulder loading activities from wearable sensor data: A machine learning approach. *Sensors*, *22*(19). Retrieved from `https://www.mdpi.com/1424-8220/22/19/7404` doi: 10.3390/s22197404

Ding, D., Hiremath, S., Chung, Y., & Cooper, R. (2011). Detection of wheelchair user activities using wearable sensors. In C. Stephanidis (Ed.), *Universal access in human-computer interaction. context diversity* (pp. 145–152). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y., Fu, Y., & Wang, L. (2015). Skeleton based action recognition with convolutional neural network. In *2015 3rd iapr asian conference on pattern recognition (acpr)* (p. 579-583). doi: 10.1109/ACPR.2015.7486569

Du, Y., Wang, W., & Wang, L. (2015a). Hierarchical recurrent neural network for skeleton based action recognition. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 1110-1118). doi: 10.1109/CVPR.2015.7298714

Du, Y., Wang, W., & Wang, L. (2015b). Hierarchical recurrent neural network for skeleton based action recognition. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 1110-1118). doi: 10.1109/CVPR.2015.7298714

Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2959-2968.

Fang, H.-S., Xie, S., Tai, Y.-W., & Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In *Proceedings of the ieee international conference on computer vision* (pp. 2334–2343).

Fanuel, M., Yuan, X., Nam Kim, H., Qingge, L., & Roy, K. (2021). A survey on skeleton-based activity recognition using graph convolutional networks (gcn). In *2021 12th international symposium on image and signal processing and analysis (ispa)* (p. 177-182). doi: 10.1109/ISPA52656.2021.9552064

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International journal of computer vision*, *61*(1), 55–79.

Feng, P., Yu, M., Naqvi, S. M., & Chambers, J. A. (2014). Deep learning for posture analysis in fall detection. In *2014 19th international conference on digital signal processing* (p. 12-17). doi: 10.1109/ICDSP.2014.6900806

Fischler, M., & Elschlager, R. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, *C-22*(1), 67-92. doi: 10.1109/T-C.1973.223602

García-Massó, X., Serra-Añó, P., González, L., Ye Lin, Y., Prats-Boluda, G., & Garcia-Casado, J. (2015, 05). Identifying physical activity type in manual wheelchair users with spinal cord injury by means of accelerometers. *Spinal cord*, *53*. doi: 10.1038/sc.2015.81

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Girshick, R. B. (2015). Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440-1448.

Goodwin, B. M., Jahanian, O., Van Straaten, M. G., Fortune, E., Madansingh, S. I., Cloud-Biebl, B. A., … Morrow, M. M. (2021). Application and reliability of accelerometer-based arm use intensities in the free-living environment for manual wheelchair users and able-bodied individuals. *Sensors*, *21*(4), 1236.

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., J'egou, H., & Douze, M. (2021). Levit: a vision transformer in convnet's clothing for faster inference. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 12239-12249.

Hagihara, H., Ienaga, N., Enomoto, D., Takahata, S., Ishihara, H., Noda, H., … Terayama, K. (2020, 04). Computer vision-based approach for quantifying occupational therapists' qualitative evaluations of postural control. *Occupational Therapy International*, *2020*, 1-9. doi: 10.1155/2020/8542191

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Heo, B., Yun, S., Han, D., Chun, S., Choe, J., & Oh, S. J. (2021). Rethinking spatial dimensions of vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11916-11925.

Hinton, G. E., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *ArXiv*, *abs/1503.02531*.

Hiremath, S. V., Ding, D., Farringdon, J., Vyas, N. J., & Cooper, R. A. (2013). Physical activity classification utilizing sensewear activity monitor in manual wheelchair users with spinal cord injury. *Spinal Cord*, *51*, 705-709.

Hiremath, S. V., Intille, S. S., Kelleher, A., Cooper, R. A., & Ding, D. (2015). Detection of physical activities using a physical activity monitor system for wheelchair users. *Medical Engineering Physics*, *37*(1), 68-76. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1350453314002793` doi: https://doi.org/10.1016/j.medengphy.2014.10.009

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, *abs/1704.04861*.

Huang, J., Zhu, Z., Guo, F., & Huang, G. (2020). The devil is in the details: Delving into unbiased data processing for human pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5699-5708.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Eccv.*

Jaffery, M., Ashraf, M., Almogren, A., Mahnoor, H., Arshad, J., Khan, J., ... Hussen, S. (2022, 05). Fsr-based smart system for detection of wheelchair sitting postures using machine learning algorithms and techniques. *Journal of Sensors*, *2022*. doi: 10.1155/2022/1901058

Jain, A., Tompson, J., LeCun, Y., & Bregler, C. (2014). Modeep: A deep learning framework using motion features for human pose estimation. In *Accv.*

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., & Black, M. J. (2013, December). Towards understanding action recognition. In *International conf. on computer vision (iccv)* (p. 3192-3199).

Ke, Q., Bennamoun, An, S., Sohel, F., & Boussaïd, F. (2017). A new representation of skeleton sequences for 3d action recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4570-4579.

Kreiss, S., Bertoni, L., & Alahi, A. (2019). Pifpaf: Composite fields for human pose estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11969-11978.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Wein-

berger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`

Lee, I., Kim, D., Kang, S., & Lee, S. (2017). Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *2017 ieee international conference on computer vision (iccv)* (p. 1012-1020). doi: 10.1109/ICCV.2017.115

Li, B., Dai, Y., & He, M. (2018). Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *ArXiv*, *abs/1708.02287*.

Li, B., Shen, C., Dai, Y., van den Hengel, A., & He, M. (2015). Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 1119-1127). doi: 10.1109/CVPR.2015.7298715

Li, C., & Lee, G. H. (2019). Generating multiple hypotheses for 3d human pose estimation with mixture density network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9879-9887.

Li, C., Zhong, Q., Xie, D., & Pu, S. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Ijcai*.

Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H., & Lu, C. (2019). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10855-10864.

Li, S., & Chan, A. (2014, 11). 3d human pose estimation from monocular images with deep convolutional neural network. In (Vol. 9004, p. 332-347). doi: 10.1007/978-3-319-16808-1_23

Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S., & Zhou, E. (2021). Tokenpose: Learning keypoint tokens for human pose estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11293-11302.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., … Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, *abs/1405.0312*. Retrieved from `http://arxiv.org/abs/1405.0312`

Liu, W., Bao, Q., Sun, Y., & Mei, T. (2022). Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys*

(CSUR).

Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S., Yang, B., & Wang, X. (2021). Deep dual consecutive network for human pose estimation. In *2021 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 525–534).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., … Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992-10002.

Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., … Lin, L. (2018). Lstm pose machines. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5207–5215).

Ma, C., Li, W., Cao, J., Du, J., Li, Q., & Gravina, R. (2019, 06). Adaptive sliding window based activity recognition for assisted livings. *Information Fusion*, *53*. doi: 10.1016/j.inffus.2019.06.013

Ma, C., Li, W., Gravina, R., & Fortino, G. (2016). Activity recognition and monitoring for smart wheelchair users. In *2016 ieee 20th international conference on computer supported cooperative work in design (cscwd)* (p. 664-669). doi: 10.1109/CSCWD.2016.7566068

Ma, C., Li, W., Gravina, R., & Fortino, G. (2017, 03). Posture detection based on smart cushion for wheelchair users. *Sensors*, *17*, 719. doi: 10.3390/s17040719

Maaz, M., Shaker, A. M., Cholakkal, H., Khan, S., Zamir, S. W., Anwer, R. M., & Khan, F. S. (2022). Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. *ArXiv*, *abs/2206.10589*.

Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., & Wang, Z. (2021). Tfpose: Direct human pose estimation with transformers. *ArXiv*, *abs/2103.15320*.

Marco-Ahulló, A., Montesinos-Magraner, L., Gonzalez, L.-M., Llorens, R., Segura-Navarro, X., & García-Massó, X. (2021). Validation of using smartphone built-in accelerometers to estimate the active energy expenditures of full-time manual wheelchair users with spinal cord injury. *Sensors*, *21*(4). Retrieved from `https://www.mdpi.com/1424-8220/21/4/1498` doi: 10.3390/s21041498

McNally, W., Vats, K., Wong, A., & McPhee, J. (2021). Evopose2d: Pushing the boundaries of 2d human pose estimation using accelerated neuroevolution with weight transfer. *IEEE Access*, *9*, 139403–139414.

Mehta, S., & Rastegari, M. (2022a). Mobilevit: Light-weight, general-purpose, and

mobile-friendly vision transformer. *ArXiv*, *abs/2110.02178*.

Mehta, S., & Rastegari, M. (2022b). Separable self-attention for mobile vision transformers. *ArXiv*, *abs/2206.02680*.

Moon, G., Chang, J. Y., & Lee, K. M. (2019). Posefix: Model-agnostic general human pose refinement network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7765-7773.

Newell, A., Huang, Z., & Deng, J. (2017). Associative embedding: End-to-end learning for joint detection and grouping. *ArXiv*, *abs/1611.05424*.

Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision* (pp. 483–499).

Nizam, Y., Mohd, M. N. H., & Jamil, M. M. A. (2017). Human fall detection from depth images using position and velocity of subject. *Procedia Computer Science*, *105*, 131-137. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1877050917302090` (2016 IEEE International Symposium on Robotics and Intelligent Sensors, IRIS 2016, 17-20 December 2016, Tokyo, Japan) doi: https://doi.org/10.1016/j.procs.2017.01.191

Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3d human pose. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1263-1272.

Peng, X., Tang, Z., Yang, F., Feris, R. S., & Metaxas, D. (2018). Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2226–2234).

Pfister, T., Charles, J., & Zisserman, A. (2015). Flowing convnets for human pose estimation in videos. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1913-1921.

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., & Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4929–4937).

Popp, W. L., Richner, L., Brogioli, M., Wilms, B., Spengler, C. M., Curt, A., … Gassert, R. (2018). Estimation of energy expenditure in wheelchair-bound spinal cord injured individuals using inertial measurement units. *Frontiers in Neurology*, *9*.

Postolache, O., Viegas, V., Dias Pereira, J. M., Vinhas, D., Girão, P. S., & Postolache, G. (2014). Toward developing a smart wheelchair for user physiological stress and physical activity monitoring. In *2014 ieee international symposium on medical measurements and applications (memea)* (p. 1-6). doi: 10.1109/MeMeA.2014 .6860097

Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J., & Sheikh, Y. (2014). Pose machines: Articulated pose estimation via inference machines. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – eccv 2014* (pp. 33–47). Cham: Springer International Publishing.

Ronchi, M. R., & Perona, P. (2017). Benchmarking and error diagnosis in multi-instance pose estimation. In *2017 ieee international conference on computer vision (iccv)* (p. 369-378). doi: 10.1109/ICCV.2017.48

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510-4520.

Sapp, B., & Taskar, B. (2013). Modec: Multimodal decomposable models for human pose estimation. In *In proc. cvpr.*

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 1227–1236).

Song, J., Wang, L., Van Gool, L., & Hilliges, O. (2017). Thin-slicing network: A deep structured model for pose estimation in videos. In *2017 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 5563–5572).

Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Aaai.*

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5693–5703).

Tang, W., Yu, P., & Wu, Y. (2018). Deeply learned compositional models for human pose estimation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision – eccv 2018* (pp. 197–214). Cham: Springer International Publishing.

Tekin, B., Márquez-Neila, P., Salzmann, M., & Fua, P. (2016). Fusing 2d uncertainty and 3d cues for monocular body pose estimation. *ArXiv*, *abs/1611.05708*.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 648–656).

Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, *27*.

Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1653–1660).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & J'egou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & J'egou, H. (2021). Going deeper with image transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 32-42.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *ArXiv*, *abs/1706.03762*.

Wadekar, S. N., & Chaurasia, A. (2022). Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *ArXiv*, *abs/2209.15159*.

Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., & Luo, P. (2021). When human pose estimation meets robustness: Adversarial algorithms and benchmarks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11850-11859.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., … others (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3349–3364.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., … Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 548-558.

Wang, Y., Li, M., Cai, H., Chen, W.-M., & Han, S. (2022). Lite pose: Efficient ar-

chitecture design for 2d human pose estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 13126–13136).

Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4724–4732).

Wu, H., Xiao, B., Codella, N. C. F., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 22-31.

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., … Wang, Y. (2017). Ai challenger : A large-scale dataset for going deeper in image understanding. *ArXiv*, *abs/1711.06475*.

Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., & Yuan, L. (2022). Tinyvit: Fast pretraining distillation for small vision transformers. *ArXiv*, *abs/2207.10666*.

Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *Proceedings of the european conference on computer vision (eccv)* (pp. 466–481).

Xiong, Z., Wang, C., Li, Y., Luo, Y., & Cao, Y. (2022). Swin-pose: Swin transformer based human pose estimation. *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 228-233.

Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*.

Yang, H., Gu, Y., Zhu, J., Hu, K., & Zhang, X. (2020). Pgcn-tca: Pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition. *IEEE Access*, *8*, 10040-10047. doi: 10.1109/ACCESS.2020.2964115

Yang, H., Yan, D., Zhang, L., Li, D., Sun, Y., You, S., & Maybank, S. J. (2022). Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, *31*, 164-175.

Yang, S., Quan, Z., Nie, M., & Yang, W. (2021). Transpose: Keypoint localization via transformer. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 11802–11812).

Yang, W., Li, S., Ouyang, W., Li, H., & Wang, X. (2017). Learning feature pyramids for human pose estimation. In *proceedings of the ieee international conference*

on computer vision (pp. 1281–1290).

Yousuf, S., & Jilani, M. T. (2021, 04). A ubiquitous wheelchair fall detection system using low-cost embedded inertial sensors and unsupervised one-class svm. *Journal of Ambient Intelligence and Humanized Computing*, 1-16. doi: 10.1007/s12652-021-03279-6

Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., & Wang, J. (2021). Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10440–10450).

Yu, X., Zhou, F., & Chandraker, M. (2016). Deep deformation network for object landmark localization. *ArXiv*, *abs/1605.01014*.

Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., & Wu, W. (2021). Incorporating convolution designs into visual transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 559-568.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F. E. H., … Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 538-547.

Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., & Wang, J. (2021). Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, *34*, 7281–7293.

Zeng, A., Ju, X., Yang, L., Gao, R., Zhu, X., Dai, B., & Xu, Q. (2022). Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation. In *Eccv.*

Zhang, F., Zhu, X., Dai, H., Ye, M., & Zhu, C. (2020). Distribution-aware coordinate representation for human pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7091-7100.

Zhang, F., Zhu, X., & Ye, M. (2019). Fast human pose estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3517–3526).

Zhang, S.-H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., … Hu, S.-M. (2019). Pose2seg: Detection free human instance segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 889–898).

Zhang, W., Zhu, M., & Derpanis, K. G. (2013). From actemes to action: A strongly-supervised representation for detailed action understanding. In *2013 ieee international conference on computer vision* (p. 2248-2255). doi: 10.1109/ICCV.2013.280

Zhang, X., Xu, C., & Tao, D. (2020). Context aware graph convolution for skeleton-based action recognition. In *2020 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 14321-14330). doi: 10.1109/CVPR42600.2020.01434

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6848-6856.

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., & Xie, X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Aaai.*