3D Human pose estimation

Using geometric self-supervision with temporal methods

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Nandi Bau
1064787

Supervised by Prof. Richard Klein

September 10, 2024

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

Abstract

This dissertation explores the enhancement of 3D human pose estimation (HPE) through self-supervised learning methods that reduce reliance on heavily annotated datasets. Recognising the limitations of data acquired in controlled lab settings, the research investigates the potential of geometric self-supervision combined with temporal information to improve model performance in real-world scenarios. A Temporal Dilated Convolutional Network (TDCN) model, employing Kalman filter post-processing, is proposed and evaluated on both ground-truth and in-the-wild data from the Human3.6M dataset. The results demonstrate a competitive Mean Per Joint Position Error (MPJPE) of 62.09mm on unseen data, indicating a promising direction for self-supervised learning in 3D HPE and suggesting a viable pathway towards reducing the gap with fully-supervised methods. This study underscores the value of self-supervised temporal dynamics in advancing pose estimation techniques, potentially making them more accessible and broadly applicable in real-world applications.

Declaration

I, Nandi Bau, hereby declare the contents of this dissertation to be my work. This dissertation is submitted to fulfil the requirements for a Master of Science in Computer Science (by Dissertation), at the University of the Witwatersrand. This work has not been submitted to any other university, nor for any other degree.

Acknowledgements

It is with great pleasure that I express my heartfelt appreciation to Prof. Richard Klein. I am truly grateful for your warm guidance and inspirational teaching style, as well as for generously dedicating your time and giving me this invaluable opportunity.

I would like to extend my warmest gratitude to my mother, Dr Hairong Wang, and my father, Prof. Sheng Bau, for their unwavering support and nurturing throughout the years.

I am deeply thankful to my partner, Siddarth Singh, whose constant encouragement and unwavering support have been instrumental on this journey.

Thank you to my sisters, Christine and Marilyn, for being precious beacons of happiness throughout.

I thank the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa, for funding this research. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the CoE-MaSS.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Being able to accurately locate, estimate, and predict human body poses is a crucial computer vision problem. Human pose estimation (HPE) has been extensively studied, and ongoing research aims to develop increasingly accurate deep learning models that can approach real-time estimation and generalise to in-the-wild images and videos.

The importance of studying and improving human pose estimation lies in its wide range of potential applications. HPE enables higher-level reasoning of the human body in the context of activity recognition and human-computer interaction. Specific applications include pedestrian recognition and motion prediction in self-driving cars, automated clinical analysis of pose and gait pathology, markerless motion capture (Mo-Cap) technologies, augmented or virtual reality interactions, sports prediction and analysis, and robot training.

There are decades worth of cumulative research in the field of human pose estimation which has resulted in specialised open-access datasets for both two-dimensional (2D) and three-dimensional (3D) human pose estimation and a set of specific, standardised metrics in measuring pose estimation performance. While there have been significant advancements and popularity in the field of human pose estimation, there are still unresolved challenges and gaps in research and application. In particular, 3D human pose estimation suffers from a lack of diverse and large-scale labeled datasets compared to 2D pose estimation datasets.

Despite the great developments and popularity of the human pose estimation field, there remain unresolved challenges and gaps in research and application. Specifically in 3D human pose estimation, there is a lack of diverse and large scale labelled data compared to the 2D pose estimation datasets.

This scarcity of data necessitates the development of more robust unsupervised, semi-supervised, or self-supervised methods to compensate for the limited availability of diversified large-scale datasets.

This research proposes using temporal methods to improve on geometric self-supervised methods in 3D HPE. Experiments are framed around several potential approaches: temporal convolutional neural networks using dilated convolutions, using Kalman filters to better optimise joint predictions and using optical flow methods to try and improve the prediction accuracy of a geometric self-supervised model.

This research proposes the use of temporal methods to enhance geometric self-supervised methods in 3D HPE. The experiments are framed around several poten-

tial approaches, including temporal dilated convolutional neural networks, Kalman filters post-processing to improve the prediction accuracy of a geometric self-supervised model.

This dissertation is structured to systematically address the research topic, with each chapter building upon the knowledge established in the preceding ones.

Chapter 2 lays the foundation by presenting a comprehensive overview of the structure, frameworks and challenges prevalent in the field of both 2D and 3D pose estimation. This includes an exploration of the datasets, metrics, and backbone techniques that are integral to understanding and addressing the problem at hand. Following this, chapter 3 delves deeper into related work in existing literature, focusing on alternative supervision paradigms and temporal methods that have been applied in pose estimation. This chapter critically analyses the current methodologies and highlights the gaps in the research that this dissertation aims to address. In chapter 4, the dissertation presents the core research hypothesis and questions that guide the study. This chapter outlines the proposed methodology, detailing the experimental design, dataset details, and analytical techniques employed in the pursuit of answering the research questions. Chapter 5 documents the results and discussions of the empirical research. Chapter 6 proposes avenues for further research. This serves as a bridge between the current study and future work, suggesting enhancements and areas that warrant additional investigation. Finally, we present a summary in chapter 7.

Chapter 2

Background and literature review

2.1 Introduction

In computer vision, human pose estimation (HPE) is the study of systems that assess the pose of an articulated human body using image inputs. Recognition and perception of human beings and their activities in different environments and situations is an important capability of digital systems [Chen *et al.* 2020]. With various applications that would allow people to better interact with computers, the HPE problem area is well defined and intensively studied in computer vision [Chen *et al.* 2020; Dang *et al.* 2019; Gong *et al.* 2016; Guo *et al.* 2020; Ji *et al.* 2020].

The current methods used in the pose estimation problem are rich and varied, including a plethora of deep learning methods studied and developed over the past few decades. Significant advancements have been made in the accuracy, efficiency and generalisability of pose estimation models. In this chapter, we will present an overview of this research landscape that leads to and motivates the work presented in this dissertation. We will present the robust development of two-dimensional (2D) human pose estimation methods that lead into the foray towards three-dimensional pose estimation methods. Methods for 3D human pose estimation and the challenges encountered for these are studied. Due to the diversity and variation in the human pose estimation field, there are several broad categories that human pose estimation can be organised for a general understanding of the state of research, such as categories based on whether a 2D or 3D pose is measured. These categories will be used to help gain a good overview of the notable research methods and how different methods developed relate to each other.

The adult human body has 244 degrees of freedom consisting of 230 joints [Dsouza *et al.* 2020].

The structure of presenting the methods in our literature review is presented below.

First we define the general problem space of human pose estimation, and present logic-based algorithmic research image processing-based methods before deep-learning methods became a standard.

The pose estimation problem space can be partitioned on a high level based on whether the pose being predicted is two-dimensional or three-dimensional. There are many further classifications of methods among each. Within both 2D and 3D pose

estimation, there are distinct studies for detecting both single-person and multi-person poses.

Within 2D single-person HPE, methods are generally classified based on whether or not they are regression-based or detection-based. In 2D, multi-person HPE methods may fall into either a top-down or bottom-up estimation approach [Chen *et al.* 2020]. The importance of discussing 2D pose estimation methods for our research is due to the fact that 2D pose estimation is used as the first step in many two-step 3D HPE methods, where 2D poses are obtained first for facilitating 3D poses in the next step. Our research uses self-supervision data in the 3D pose context by first estimating 2D poses as a first step using a 2D pose estimation network. 2D human pose estimation is an extensively studied field, where great accuracy and speed is achieved with estimating poses in 2D space. There are also extensive and diverse datasets that are available that contain large samples of training in many different environments, lighting conditions and activities.

Current research trends beyond 2D HPE to have accurate predictions for human poses in 3 dimensions. Three dimensions are a more natural representation of a human body within an environment. As with 2D HPE, 3D pose estimation solutions can also classify on a higher level as being either single-person or multi-person. Multi-person pose estimation is generally more complex than single-person estimation because of the need to detect and distinguish individual humans in an image as a first step. Methods in multi-person human pose estimation can be either a top-down or bottom-up approach. Top-down methods first locate the potential number of people within a given image, then infer the location of each body part belonging to the person. Bottom-up methods detect and estimate the position of body parts first, then associate a set of body parts into distinct persons.

We will also introduce some of the conventions used in human pose estimation, such as the type of abstraction used for the human body, and which error metrics are common. Error metrics and pose abstractions are closely related to individual datasets available in the pose estimation research area. We will focus on the most common type of pose abstraction, and the well-established public datasets for both 2D and 3D human pose estimation.

We will introduce a survey of the datasets used in the general human pose estimation task, based on whether the dataset is used for 2D or 3D pose estimation tasks. Where appropriate, we will introduce the common conventions and the loss criteria associated with and defined by each dataset. The backbone of 2D and 3D pose estimation is the data sets that offer annotations and data diversity. The biggest difference between a 2D pose dataset and a 3D pose dataset is that 2D pose datasets can be lifted from many sources of existing image sources and annotated by hand. 3D pose datasets can not be annotated by hand and rely on lab equipment which is costly, and data collection in a strict lab environment (see Ionescu *et al.* [2014]) often reduces the diversity and amount of the data. Self-supervised, unsupervised and semi-supervised methods may be able to alleviate some of these problems inherent to the 3D pose datasets that exist now. This offers some insight into why we chose to explore the research that we present in the later chapters of this dissertation.

Temporal methods and geometric self-supervision methods are related to to our work and will be discussed in detail and presented in Chapter 3, leading up to why we chose to study self-supervised and temporal methods with our research.

Although a mature and diverse field, the solutions that are being developed within the human pose estimation field is constantly evolving. In the final section of this chapter, other works of interest that are contemporary in the field are introduced for comparison.

2.2 Human body models

Human bodies are complex structures, which contribute to the complexity of the human pose estimation problem. The biological make-up of a non-disabled adult human body consists of 230 joints with 244 degrees of freedom [Dsouza *et al.* 2020]. The range and diversity of this structure is a part of what makes pose estimation a challenging task to perform to great detail and accuracy. This challenge necessitates abstractions and simplified models to represent human bodies for general application and research.

In order to study and estimate the locations of articulated human poses, it is important to abstract the human body into various forms of digital representations. Articulated human poses refer to systems of configurations for a human body in a specific position, where the specific position is further defined by angles and positions of joints in the body [Pishchulin *et al.* 2012]. Many applications and practical research further simplifies the number of joints defined, such as the standard of using 17 or 18 joints representing the positions of major body parts, with 17 joints defined by the well-known MPII Human Pose dataset [Andriluka *et al.* 2014b]. This is done to reduce complexity of data collection and computation for most general applications such as in different pose estimation datasets [Andriluka *et al.* 2014b; Ionescu *et al.* 2014].

The prevalent types of pose representation models are grouped in the following section 2.2. The purposes, advantages and disadvantages of each type is discussed with a brief description of how each of the pose structure types are generally collected. Comparisons of the suitable contexts for which each body abstraction type is used is also discussed.

Skeleton-based pose representation

One of the most commonly-used pose representation structures is a skeleton-based, also referred to as the kinematic or stick-figure model. This model of the human body is joint-based, meaning that each major body part useful for positioning articulated pose is represented as a point with a location in 3D space. The body itself is represented as a set of joints (typically between ten to thirty joint locations). Some relational structure may be pre-defined, with relational connections between certain joints such as what is expected between a left hand and a left elbow.

The skeleton model may be described as a graph where vertices indicate joints and edges encode connections or prior constraints between the joints defined. This human body topology is simple and highly flexible and used widely in both 2D and 3D human pose estimation methods [Chen *et al.* 2020; Ji *et al.* 2020; Liu *et al.* 2015b; Gong *et al.* 2016], and defined along with many pose datasets [Ionescu *et al.* 2014; Andriluka *et al.* 2014b]. This type of body abstraction is able to simplify positions of body joints very well, but on its own it is not able to represent texture or volumetric information for

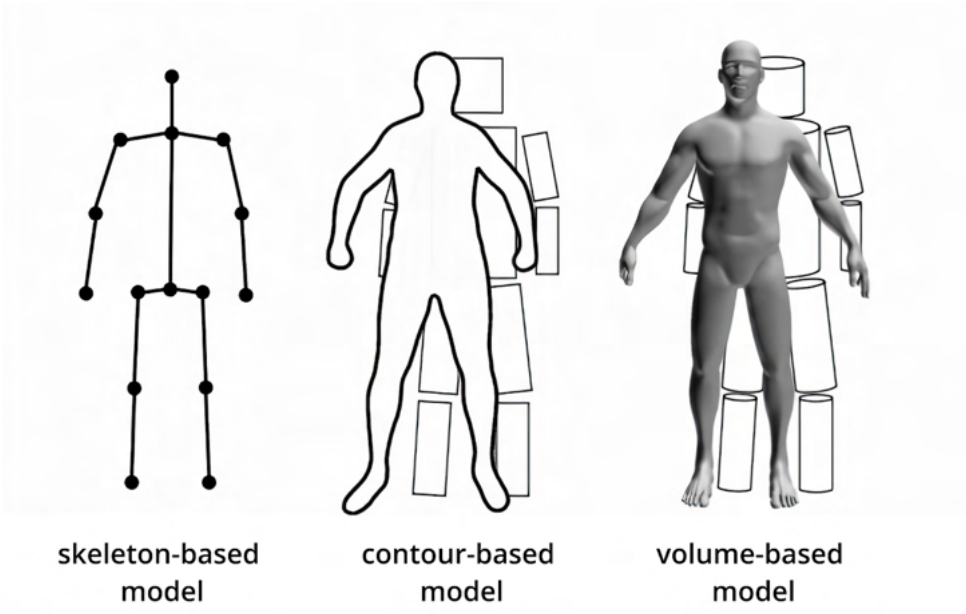


Figure 2.1: Human body model abstractions used in pose estimation

problems that may require more detailed analysis of human bodies that require these properties [Chen *et al.* 2020].

Contour-based pose representation

The contour-based pose representation is an area-based representation where the human body or body parts are encoded as areas within an image. This can serve to represent the contour information of bodies and limbs, along with approximate widths and lengths of joints and bodies. There are contour-based pose estimation methods where rectangles or person-boundaries are used to denote the space that articulated poses occupy within an image. Representative methods of contour-based body models include cardboard models [Ju *et al.* 1996] and Active Shape Models [Cootes *et al.* 1995]. The strengths of this type of pose representation is that there is information for the rough boundaries and shape of the human body, especially in 2D estimation problems [Dang *et al.* 2019; Chen *et al.* 2020]. The main weaknesses are that the precise locations of joints are not clearly represented, and the model is generally harder to superimpose onto something like a virtual body [Chen *et al.* 2020].

Volume-based pose representation

Geometric meshes are used to represent articulated poses in volume-based pose models. Early methods used simple geometric shapes to represent the human body, such as cylinders and cones [Sidenbladh *et al.* 2000]. More modern volume-based representations use a mesh to capture the human body. Meshes are widely used in computer graphics and modelling as a collection of vertices and edges that define the outer surface of a physical object representation [Sorkine 2006]. Meshes provide a way to rep-

resent complex 3D forms in a computationally effective way, however, they are difficult and expensive to capture human pose datasets for [Huang *et al.* 2022].

Widely-used volume-based representations in pose estimation include Skinned Multi-Person Linear Model (SMPL) [Loper *et al.* 2015], Shape Completion and Animation of People (SCAPE) [Anguelov *et al.* 2005], and unified deformation model [Joo *et al.* 2018].

2.3 Sensor and image types for pose data

This section serves to provide a background into the common sensors and image types used within the human pose estimation field, due to the traditionally heavy reliance on well-annotated large datasets for training models. As shown in section 2.4, large datasets have been a requirement to solve the human pose estimation problem since even before the rise in popularity of deep learning methods. Thus it is important to give a briefing and catalogue of the image types, collection mechanisms, and sensors used to collect the datasets used to benchmark modern pose estimation methods. This section focuses solely on providing introduction to the technicalities of collecting data for pose estimation, instead of the specific datasets used to benchmark pose estimation methods.

2.3.1 Common sensor types for pose estimation

The most common type of sensor used for pose data collection is red-green-blue (RGB) imaging, which is collected and taken with RGB cameras. RGB cameras take colour images of the scene, with each image composed of a matrix of pixels that include colour information. RGB cameras are widely available and ubiquitous in contemporary society, making RGB image datasets for pose estimation cheaper and easier to collect than other types. RGB cameras are a popular choice for image collection and is the primary method of data collection for many vision tasks and pose estimation tasks [Ionescu *et al.* 2014; Sigal *et al.* 2010]. Apart from RGB images which are widely used, other colour formats may also be used including gray-scale images [Jain *et al.* 2011].

Depth sensors and depth cameras (one example being Microsoft's Kinect [Karl-Bridge-Microsoft 2012]) capture depth information using depth sensors in addition to color information similar to RGB images. Depth information can be particularly useful to train models that are able to handle occlusions and complex poses better than others. However, pose datasets that contain pose data are more difficult to collect and often require lab environments to collect, resulting in less diverse appearances in the dataset, less lighting variability and background or environment variability [Ionescu *et al.* 2014].

Inertial measurement units (IMUs) are sensors that measure acceleration and rotation rates of the body. Existing commercial IMUs are wearable sensors that are dense and complex and require time-consuming setup and calibration. Sensors such as these are costly and also need to be worn in lab environments and cause environment and lighting invariability in the dataset similar to depth sensor cameras. The sensors also can cause intrusions to subject movement and hindrance in full-range motions. IMUs

and IMU data is primarily experimented with towards reducing occlusion errors when training models, where it is able to effectively overcome viewpoint limitation of accessible optical sensors and be able to produce the rotation information of joints [Liao *et al.* 2023].

The last common type of sensor used for pose estimation are cameras with infrared (IR) sensors. Infrared sensors use infrared thermography to capture images using infrared radiation, which can be used to detect body heat and motion. The advantages of using IR sensors is invariance to lighting and weather conditions, along with emphasising the shapes and boundaries of a pose being tracked. Disadvantages to using IR sensors are fewer texture features available, poor signal-to-noise ratio and blurred edge effects [Guo *et al.* 2022].

2.3.2 Common image types for pose estimation

Along with different sensor types being used for the HPE task, it is also important to introduce and contextualise the different image types used in the field. This is important to contextualise the constraints and the potential improvements towards different research methods in the field in future. Both sensor and image types constrain the availability of features within human pose datasets, which are typically expensive and time-consuming to create, especially in the 3D HPE problem space [Ionescu *et al.* 2014; Sigal *et al.* 2010; Joo *et al.* 2015; Mehta *et al.* 2017a; Trumble *et al.* 2017].

Single images are a common type of training input for human pose estimation. They can be captured using a wide variety of sensor types including RGB cameras, depth sensors, and thermal cameras. However, single-image datasets have limitations such as not providing any temporal information, which makes it challenging to estimate poses in dynamic scenes by leveraging consecutive image frames. There are many single-image datasets for the 2D pose estimation problem, where the problem space can benefit from the diversity and dynamism of the images and easier annotation process [Lin *et al.* 2014; Andriluka *et al.* 2014a; Sigal *et al.* 2010; Rostamzadeh *et al.* 2018].

Image sequences such as videos or image streams provide information over time [Hussein *et al.* 2013], which are useful for estimating human poses in dynamic scenes or complex scenes where the subject is in motion or conducting a specific type of activity. Unlike single images which only capture a single frame of a subject pose, image sequences provide a temporal context that helps disambiguate ambiguous poses and improve the accuracy of pose estimation in challenging situations such as handling self-occlusions and complex poses. Datasets that use image sequences are typically collected in video format and have a clear sequencing of each image frame [Ionescu *et al.* 2014].

A point cloud is a discrete set of points in a space. A point cloud can be used to approximate the space or volume that a subject is occupying [Guo *et al.* 2020], and thus used for the pose estimation problem by defining either a 2D point cloud to estimate the space a pose is occupying, or a 3D point cloud to estimate the volume of a pose. Point clouds data is generally captured using depth sensors such as the Microsoft Kinect [Karl-Bridge-Microsoft 2012] or Intel RealSense [Intel 2023].

Synthetic images are generated images created using computer graphics software, and represent an alternative way to generate large datasets with ground-truth anno-

tations without relying on sensors. This type of data is especially advantageous to 3D pose estimation, where 3D poses are much more difficult to annotate or create crowd-sourced annotations for [Chen *et al.* 2016]. The lack of accessible 3D annotations hinders exploration of end-to-end solutions for 3D pose estimation problems, and the main strength of synthesized pose datasets is accessibility. A weakness of synthesized datasets is that they may not accurately represent natural humans in diverse habitats, however, this is often a trade-off between having a dataset taken in a lab environment versus having a very customise-able but not necessarily representative dataset to train on [Ionescu *et al.* 2014; Chen *et al.* 2016]. Some methods leverage both real and synthetic datasets to try and generalise models to more diverse appearances and lighting [Pavlakos *et al.* 2018; Zhang *et al.* 2018].

2.4 Non-deep learning based approaches

Before the widespread use of deep learning methods in human pose estimation, many non-deep learning-based methods were developed and used for articulated pose estimation, due to the long-time saliency of the problem in computer vision [Liu *et al.* 2015b]. These methods are briefly introduced and discussed in the following section to establish a background of the methods and direction before the impact of deep learning in the computer vision field in general. Most non-deep learning methods are used for 2D HPE, at the time being the salient problem of articulated pose estimation. These non-deep learning-based approaches can be roughly categorized according to the approach taken when developing the methods, and represent some of the well-known and well-studied methods in earlier human pose estimation papers.

We introduce the major methods and advances in non-deep learning based pose estimation by examining some of the prominent methods, and attempting to break down the methods into categories. These are split into two taxonomies, model-based methods and data-driven methods. These non-deep learning (or traditional) HPE methods involved hand-crafted feature extraction methods and body models to obtain pose structure estimations, data driven methods using traditional machine learning techniques, and hybrid methods that may use a combination of techniques from both categories [Liu *et al.* 2015b; Dantone *et al.* 2013; Chen and Yuille 2014].

2.4.1 Model-based pose estimation methods

One of the main types of non-deep learning articulated pose estimation methods is the model-based method, which generally use a predefined model of the human body similar to those introduced in section 2.2. These human body models would have a-priori knowledge of the human body, typically based on geometric, physical or anatomical information of the human body [Liu *et al.* 2015b; Dantone *et al.* 2013; Eichner and Ferrari 2010].

Model-based methods include methods such as pictorial structure models, where a graphical structure is used to encode the spatial relationships between body parts and joints. A set of appearance models are used to learn possible appearances of each body part using annotated datasets [Liu *et al.* 2015b]. The graphical structure of the human

model typically uses a tree or graph representation [Wang and Li 2013; Felzenszwalb and Huttenlocher 2005; Yang and Ramanan 2012; Andriluka *et al.* 2014a]. The appearance models used to account for differences in body part appearance variation typically uses traditional vision and image processing markers such as edges, textures, orientations and colours [Felzenszwalb and Huttenlocher 2005; Yang and Ramanan 2012], trained using traditional machine learning techniques such as support vector machines (SVMs) and principal component analysis (PCA) [Bourdev and Malik 2009; Ramanan *et al.* 2006].

2.4.2 Data-based pose estimation methods

The other general category of pose estimation methods before deep learning methods became widespread were what could be considered data-driven methods for human pose estimation. This category of methods relies mainly on large datasets to learn relationships between image features and pose. These methods typically use 'traditional' machine learning methods such as SVMs [Johnson and Everingham 2010] and random forests [Shotton *et al.* 2011]. Data-based models are usually more scalable and faster than model-based systems, but are generally more prone to errors and overfitting. Data-based methods generally do not handle human pose estimation challenges such as occlusion and complex poses very well [Wei *et al.* 2016; Yang and Ramanan 2012].

Exemplary data-based hand-crafted feature techniques using Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) [Lowe 1999] rely on domain-specific knowledge and expertise to design features that are relevant to the pose estimation task [Dalal and Triggs 2005; Shotton *et al.* 2011]. HOG is a feature extraction image processing method that computes the distribution of gradient orientations in an image. HOG is a widely used method in object detection and human pose estimation, as it is able to capture the local shape and texture information of an object [Dalal and Triggs 2005; Lowe 1999]. SIFT detects keypoints in an image that are stable under different transformations, and then computing a descriptor for each keypoint based on the local image gradient orientations [Lowe 1999 2004]. SIFT has been widely used in object recognition and image matching, as it is able to capture the distinctive local features of an object [Lowe 1999], its invariance to scaling, rotation and changes in lighting make it a powerful image processing technique [Lowe 2004; Pishchulin *et al.* 2012]. While both data-based and model-based human pose estimation methods have been very successful in the past, they have largely been replaced by deep learning methods in recent years, with the ability to learn features automatically from large datasets.

2.5 Deep 2D human pose estimation

Two-dimensional (2D) pose estimation is a computer vision task that involves detecting or localising human body joints and parts with 2D image inputs. The main goal is to estimate the 2D position, orientation and articulation of the human body or bodies in the image. It is very important to note that the categories of methods in the taxonomy

provided in this section does not mean that methods within each are mutually exclusive. Studies and methods may be using multiple of the categories defined here. The taxonomic survey thus serves more to distinguish between method purposes, rather than a strict catalogue for models.

In pursuit of increased accuracy and generaliseability of pose estimation methods, development of deep-learning model architectures such as convolutional neural networks (CNNs) [Simonyan and Zisserman 2014; Krizhevsky *et al.* 2012; He *et al.* 2016] and recurrent neural networks (RNNs) [Graves and Graves 2012; Sutskever *et al.* 2014] provided the foundation for learning complex mappings between input data and pose outputs. Deep-learning methods, like many other salient fields in contemporary machine learning problems, became the standard of development and study. Large-scale annotated datasets were created, exemplars being the MPII Human Pose [Andriluka *et al.* 2014a] and COCO datasets [Lin *et al.* 2014] which enabled deep-learning models to learn from large datasets. Development of computer hardware such as graphics processing units (GPUs), new hardware architectures, and cloud computing services accelerated the uptake and availability of deep-learning.

In the first step of training a two-step network 3D pose estimator, a 2D pose estimator is deployed to provide 2D estimations from which the 3D positions will be estimated [Zhou *et al.* 2017; Tome *et al.* 2017; Véges *et al.* 2019]. Because of this, it is important to provide an overview of methodology-based taxonomy for understanding the foundations for two-step 3D human pose estimation. In our methodology we leverage the accuracy and speed achievements of a two-step pose estimator to enable self-supervised 3D pose estimation methods. Thus, we feel it is important to provide a background and brief general taxonomy of methods for 2D pose estimation and 2D pose estimation methods.

The challenges and solutions that occur in 2D pose estimation inform much of the development and methodology in 3D pose estimation. It is also important to provide the context within which our methodology and experimentation exists within.

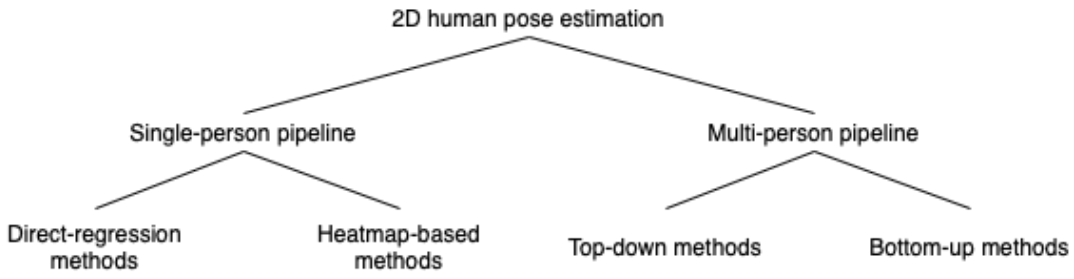


Figure 2.2: Method-based taxonomy of this section

2.5.1 Single-person pipelines

The first distinction for 2D human pose estimation methods is whether they are single-person or multi-person pose estimation pipelines. This section will introduce single-person methods [Toshev and Szegedy 2014; Wei *et al.* 2016] based on the taxonomy diagram shown in 2.2. Multi-person methods are introduced in 2.5.2 below.

Single-person methods typically assume that the input images will only contain one person, and they localize body joint positions of a single person in the input image. Single-person pose estimation essentially solves a regression problem in which the number of keypoints is explicitly given (in the case of using a skeleton-based human body abstraction). This simplifies the problem and can improve accuracy [Wei *et al.* 2016; Ji *et al.* 2020]. Often in single-person pose estimation problems, the number of keypoints is implicitly given, and the task is to regress the 2D locations (x and y positions relative to the image) of joint keypoints.

By understanding the differences between single-person and multi-person pose estimation methods, researchers can choose the appropriate method for their specific application and improve the accuracy of their results.

Direct regression methods

The next level of categorisation for single-person human pose estimation methods is whether it is a regression-based or heatmap-based approach. Regression-based methods use a pipeline where output feature maps are directly used to regress the joint locations of the human body [Toshev and Szegedy 2014; Carreira *et al.* 2016; Sun *et al.* 2017]. Toshev and Szegedy [2014] experimented with a cascaded DNN regressor to directly predict joint keypoints. Due to the difficulty in learning keypoint locations without other procedures, an iterative self-correcting model is developed [Carreira *et al.* 2016]. Progress in accuracy is made by including structural awareness in 'compositional pose regression' methods [Sun *et al.* 2017]. This method re-parameterises pose representation using bones instead of joints, which can be seen as a modification to the skeleton-based pose structure introduced in section 2.2. This method represents long-range interactions between bones as a compositional loss function, which proved more stable learning poses compared to earlier methods.

Heatmap-based methods

Instead of directly regressing keypoints or body parts, heatmap-based methods first calculate a heatmap of representing the probability field of joint locations in the image and then uses the heatmap to produce the most likely location for the joint [Chen and Yuille 2014; Chen *et al.* 2017; Wei *et al.* 2016; Newell *et al.* 2016]. A graphical model DCNN that learns parts types and pairwise parts relations is used in [Chen and Yuille 2014]. Training strategies similar to conditional Generative Adversarial Networks (GANs) are used to incorporate and learn the priors of human bodies in some works such as Adversarial PoseNet [Chen *et al.* 2017]. Convolutional Pose Machines (CPM) introduced by [Wei *et al.* 2016] regress heatmaps in multiple stages and use intermediate supervision techniques to overcome the vanishing gradient method. The stacked hourglass novel network structure is designed to leverage intermediate supervision with repeated bottom-up and top-down network processing to improve pose detection performance [Newell *et al.* 2016]. Many models have then used the stacked hourglass network as a baseline, an example being building a multi-context attention mechanism to make the model more robust and accurate [Chu *et al.* 2017].

Comparison of direct regression and heatmap-based methods

Direct regression methods discussed above [Toshev and Szegedy 2014; Carreira *et al.* 2016; Sun *et al.* 2017] are non-linear and have a tendency to have difficulty in learning mappings. Direct regression methods cannot be applied to multi-person pose estimation where a bottom-up approach or a bounding box containing more than one subject. In contrast, heatmap-based frameworks [Chen and Yuille 2014; Chen *et al.* 2017; Wei *et al.* 2016; Newell *et al.* 2016] may be used to visualise and enhance understanding of probability fields in images that keypoints and points of interest may occupy in any input. Direct regression methods are usually more lightweight and applied to 3D pose estimation scenarios without too many modifications. Heatmap resolution in heatmap-based methods may result in a high memory requirement for devices used to train and run the models, making direct-regression based methods a better choice for mobile applications in many situations [Dang *et al.* 2019].

It is important for our research to note that the relationship between 2D pose estimation and 3D pose estimation is explored [Martinez *et al.* 2017], following the logic of wanting to use deep neural networks to directly predict 3D keypoints on 2D data. Some experiments revealed that the 2D pose estimation model is one of the main causes of errors in 3D human pose estimation. This informs that robust 2D pose estimation methods and novelty can contribute greatly to robustness and performance of 3D human pose estimation [Martinez *et al.* 2017].

2.5.2 Multi-person pipelines

Multi-person methods must detect poses of multiple human bodies in the image input [Cao *et al.* 2017; Newell *et al.* 2017; Kocabas *et al.* 2018], as opposed to only accounting for single-subject poses in single-person approaches in section 2.5.1. Multi-person pose estimation methods [Cao *et al.* 2017; Kocabas *et al.* 2018] are more challenging to address due to the unconstrained problem of having any number of people in the input images. Two types of pipelines have been proposed: top-down and bottom-up approaches.

Top-down approach

Top-down approaches define a type of multi-person pose estimation method by first detecting the number of people in an input image, defined using bounding boxes. Any single-person pose estimation method may then be used on each subject pose in each person bounding box. Aside from person detection, top-down approaches have more process involved, which may employ information context from the whole input image.

The first deep learning top-down approach on multi-person inputs is studied by Toshev and Szegedy [2014]. They use a face-based detector to detect entire subject bodies demarcated by bounding boxes, where a multi-stage cascade DNN is used to directly regress joint positions. Other examples of notable works using the top-down approach include Faster R-CNN [Ren *et al.* 2015], Mask R-CNN [He *et al.* 2017]. Top-down approaches leverage the assumptions to help faster estimation of poses [Ren *et al.* 2015; He *et al.* 2017]. Mask R-CNN simultaneously predicts human bounding boxes

and human keypoints, which makes the detection and pipeline faster by sharing the features of the CNN between keypoint estimation and human detection.

An earlier, notable proof of concept substantiating our research is introduced by Radosavovic *et al.* [2018], which used all labeled and unlabeled available data in a dataset to train the model. The authors explain traditional supervised learning methods are limited by the availability of labeled data, which can be expensive and time-consuming to obtain. To address this issue, the authors propose a method called "data distillation". "Data distillation" under their "omni-supervised" learning (a technique of semi-supervision introduced by Radosavovic *et al.* [2018]) involves using ensemble predictions from multiple transformations of unlabeled data using a single model to automatically generate new training annotations. The network is then retrained on the pseudo-labels, which helps to improve its performance. The authors demonstrate the effectiveness of their approach on several computer vision tasks, including object detection, semantic segmentation, and human pose estimation. They show that data distillation can significantly improve the performance of deep neural networks, especially when labeled data is limited. This method proves the concept that state-of-the-art human pose detectors being accurate enough to apply self-training techniques to challenging real-world data.

Bottom-up approach

In contrast to the top-down approach, bottom-up methods first detect all the body parts in the image and then associate sets of body parts with each instance of a human detected in the image (such as using clustering techniques) to form the pose of the human body. Inference time for bottom-up methods is likely to be faster, since the methods do not need to perform a pose estimation pipeline for each person in the image.

Pishchulin *et al.* [2016] presented Deepcut, the first multi-person bottom-up approach for deep pose estimation. They effectively formulate the problem of multi-person pose estimation as a minimum-cost multi-cut problem, modelling part candidates as vertices and relations as edges. Modifications and improvements were made with Deepcut as a baseline in Deepercut [Insafutdinov *et al.* 2016]. Deepercut made improvements using three areas of modification. The network used is deeper than work done in Pishchulin *et al.* [2016], which generates more effective body parts proposals. A novel image-conditioned pairwise term enables variable number of human instances in input images. The third modification is that the optimisation strategy is changed, allowing for improvements to accuracy and faster inference.

A notable development by Cao *et al.* [2017] uses a non-parametric representation called Part Affinity Fields (PAFs) with heatmaps. A greedy algorithm is used to generate person instances based on the PAF and heatmap combinations. Zhu *et al.* [2017] builds on Cao *et al.* [2017] and performs several modifications to achieve better results. These include experimenting with deeper networks and using reductant PAFs to help connect broken child joint connections to a parent link.

Associative embedding is another group of bottom-up approaches introduced by Newell *et al.* [2017]. Associative embedding is a method to supervise CNNs for detection and grouping, able to simultaneously predict part and tagging heatmaps. Tagging

heatmaps are used for person detection, and are similar for the same person while indicative of dissimilarity for different persons.

Comparison of top-down and bottom-up approaches

Multiple aspects of real-world applications are important to consider when weighing the advantages and disadvantages between top-down and bottom-up approaches. Accuracy and speed are two crucial factors [Dang *et al.* 2019] to consider against each approach, as they have been identified to be important factors when it comes to the human pose problem.

One advantage of the top-down approach is that it can have better accuracy depending on the quality of the human detector used as the first step. The current top-performing model for the COCO [Lin *et al.* 2014] dataset is a top-down model, DWPose by Yang *et al.* [2023] based on RTMPose [Jiang *et al.* 2023]. One of the disadvantages of bottom-up approaches at the moment is that the scale variety of persons bring difficulties due to the networks not being able to learn robust and consistent features from the scale variant images, at least not to the extent that the top-down approaches are able to perform.

On the other hand, bottom-up approaches are likely to have faster inference time since the methods do not need to perform a pose estimation pipeline for each person in the image. Each person in the top-down pipeline is likely going to be estimated one by one, which consumes linear time with the addition of each subject present in the image. In contrast, in the bottom-up pipeline all keypoints are detected first which means the image only gets one layer of inference for all the body parts all at once. Forward propagation will take relatively long as neural network architectures tend to be deep in contemporary literature. Bottom-up methods may achieve faster inference speed due to the relatively fast grouping methods being used compared to speeds taken for network forward propagation.

Overall, both top-down and bottom-up approaches have their strengths and weaknesses, and the choice of approach may depend on the specific requirements of the application and interest in experimentation and improving on methods that exist.

2.6 Datasets for 2D pose estimation

2D human pose datasets serve as a precursor to 3D human pose datasets, and are generally more available compared to 3D pose datasets. 2D pose datasets are important in relation to 3D pose datasets because they are easier to annotate and collect and can be used for transfer learning to improve the performance of 3D pose estimation models. Importantly, many real-world applications of pose estimation rely on 2D pose estimation, making 2D pose datasets essential for developing and evaluating pose estimation models for these applications.

Pose estimation datasets are important for providing the metrics and benchmarks for the problem space as they often provide a standardized way to evaluate the performance of different solutions and models. This means that the datasets themselves are foundational to tracking the progress of the field over time, similar to other application-

based datasets in other problem areas. They serve as a way to evaluate the generalisation ability of models and assess the practical relevance of models for real-world applications.

In this section, we introduce the common 2D pose estimation datasets as a way to both understand the foundations of the pose estimation field and be able to track the evolution of the important metrics over time. Introducing 2D pose estimations data lays the foundational work of a piece of our research decisions introduced later.

There are many well-established datasets for deep learning in 2D pose estimation. We will introduce Microsoft COCO [Lin *et al.* 2014], MPII Human Pose dataset [Andriluka *et al.* 2014b], Leeds Sports Pose (LSP) [Johnson and Everingham 2010], Frames Labelled in Cinema (FLIC) [Sapp and Taskar 2013], AI Challenger Human Keypoint Detection (AIC-HKD) [Tang and Wu 2019] and PoseTrack [Doering *et al.* 2022].

2.6.1 Microsoft Common Objects in Context (COCO) dataset

The Microsoft Common Objects in Context (COCO) dataset is originally proposed for common object detection in natural contexts or environments. Objects are labelled using per-instance segmentations. The dataset identifies 91 object types over 2.5 million labelled instances in 328 thousand images [Lin *et al.* 2014]. Despite being intended for an object identification dataset for natural environments, it has since been extended in the COCO Keypoints 2017 dataset with more than 200 thousand images and 250 thousand labelled person instances [Lin *et al.* 2014]. The COCO keypoints dataset defines a standard for using 17 keypoints in a skeleton-based human body model to abstract and represent the human body. The images are in RGB format, one of the standout possible data augmentation methods used for the COCO keypoints dataset is image mosaicing, where several training images can be combined into one image input for the model training.

Figure 2.3 shows some examples of the diversity of environments and visual representations of the annotations used for the dataset. The dataset supports both single-person and multi-person estimation.

The foundational metric defined for the Microsoft COCO dataset is the Object Keypoint Similarity (OKS). Ground truth keypoints have the form $[x_1, y_1, v_1, ..., x_k, y_k, v_k]$, where x, y are the keypoint locations. The variable v is a visibility flag defined as $v = 0$ not labeled, $v = 1$ labeled but not visible, and $v = 2$ labeled and visible. Each ground truth object also has a scale s which is defined as the square root of the object segment area.

Predicted keypoints have the same form as the ground truth $[x_1, y_1, v_1, ..., x_k, y_k, v_k]$. However, the detector's predicted v_i are not currently used during evaluation, meaning the keypoint detector is not required to predict per-keypoint visibility or confidences. The OKS metric is formally defined as follows:

$$OKS = \frac{\Sigma_i \left[\exp(\frac{-d_i^2}{2s^2\kappa_i^2})\delta(v_i > 0)\right]}{\Sigma_i \left[\delta(v_i > 0)\right]} \tag{2.1}$$

where d_i are the Euclidean distances between each corresponding ground truth and detected keypoint. v_i are visibility flags of the ground truth (predicted v_i are excluded).

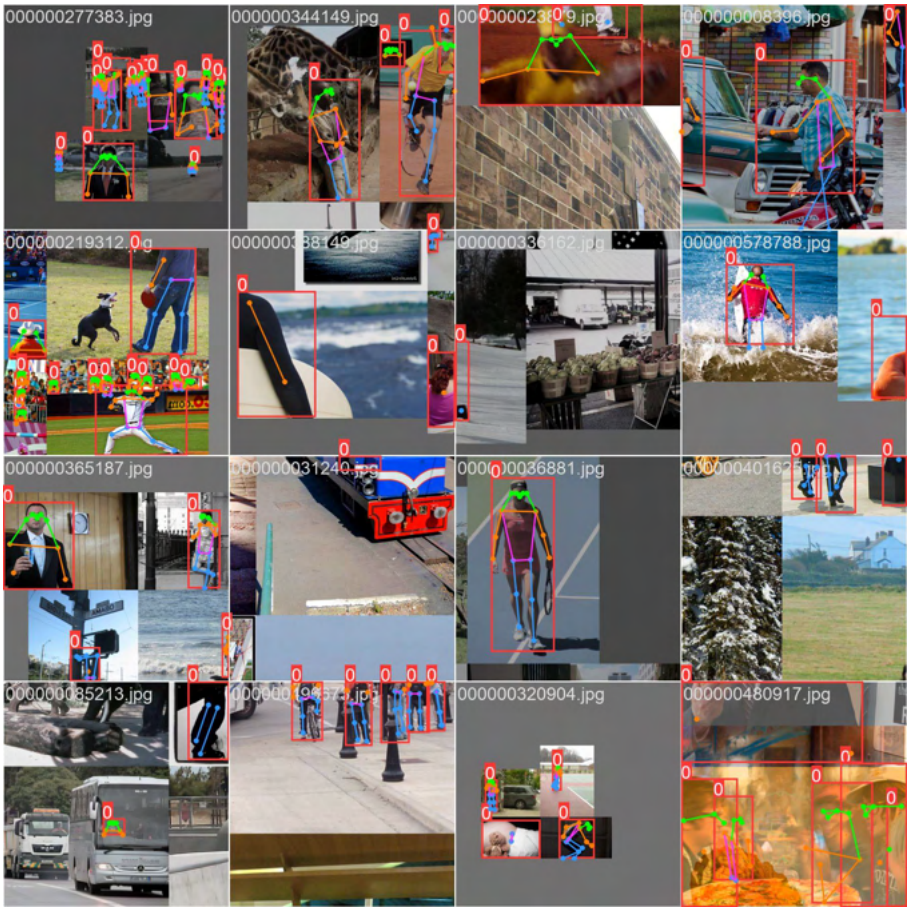


Figure 2.3: Example range of images found in the Microsoft COCO keypoints dataset, showing some examples of image mosaicing [Lin *et al.* 2014]

To compute OKS, d_i is passed through an un-normalised Gaussian with standard deviation $s\kappa_i$. s is the object scale and κ_i is a per-keypoint constant controlling falloff. For each keypoint prediction, this yields a keypoint similarity that ranges between 0 and 1. These similarities are averaged over all labeled keypoints (keypoints for which $v_i > 0$). Predicted keypoints that are not labeled ($v_i = 0$) do not affect the OKS. Perfect predictions will have $OKS = 1$ and predictions for which all keypoints are off by more than a few standard deviations $s\kappa_i$ will have $OKS \sim 0$. The OKS is analogous to Intersection over Union (IoU) in the object detection problem space. Given the OKS, the other metrics and their derivatives can be calculated.

Having defined the OKS then helps define Average Precision (AP), Average Recall (AR) and their derivatives. The Microsoft COCO keypoints dataset uses 10 derivative metrics defined using OKS (2.1) to benchmark and compare model performances.

MS COCO Average Precision (AP) represents the predicted keypoint locations at different recall levels. AP is calculated by computing the precision and recall values for different thresholds of the OKS metric, and then computing the area under the precision-recall curve (AUC). The AP metric ranges from 0 to 1, where a higher value

indicates a better precision-recall trade-off. The AP formulation is as follows:

$$AP = \sum p_i \Delta r_i \tag{2.2}$$

where p_i is the precision at i, Δr_i is the change in recall at i. This effectively computes the area under the precision-recall curve.

Average Precision (AP) and its derivatives for the Microsoft COCO dataset is defined as follows:

$$AP \text{ is } AP \text{ at } OKS = [0.50 : 0.05 : 0.95] \tag{2.3a}$$

$$AP^{OKS=0.50} \text{ is } AP \text{ at } OKS = [0.50] \tag{2.3b}$$

$$AP^{OKS=0.75} \text{ is } AP \text{ at } OKS = [0.75] \tag{2.3c}$$

AP across scales is calculated by computing the average precision of the predicted keypoint locations at different scales, where the scale is defined as the ratio of the object size to the median object size in the dataset. The AP across scales metric ranges from 0 to 1, where a higher value indicates a better precision-recall trade-off across different subject image scaling.

$$AP^{medium} = AP \text{ for medium subject size with } 32^2 < \text{ segment area } < 96^2 \tag{2.4a}$$

$$AP^{large} = AP \text{ for large subjects with segment area } > 96^2 \tag{2.4b}$$

AR for the Microsoft COCO keypoints dataset measures recall of the predicted keypoint locations at different precision levels. AR is calculated by computing the precision and recall values for different thresholds of the OKS metric, and then computing the area under the recall-precision curve. The AR metric ranges from 0 to 1, where a higher value indicates a better recall-precision trade-off. The Average Recall (AR) and its derivatives defined for the Microsoft COCO dataset is as follows:

$$AR \text{ is } AP \text{ at } OKS = [0.50 : 0.05 : 0.95] \tag{2.5a}$$

$$AR^{OKS=.50} \text{ is } AR \text{ at } OKS = [0.50] \tag{2.5b}$$

$$AR^{OKS=.75} = \text{ is } AR \text{ at } OKS = [0.75] \tag{2.5c}$$

AR across scales is similar to AP across scales, measuring the average recall of the predicted keypoint locations at different scales, same as how the AP across scales is measured. The AR across scales metric ranges from 0 to 1, where a higher value indicates a better recall-precision trade-off across different scales.

$$AR^{medium} = AR \text{ for medium subject size with } 32^2 < \text{ segment area } < 96^2 \tag{2.6a}$$

$$AR^{large} = \text{ for large subjects with segment area } > 96^2 \tag{2.6b}$$

For the MS COCO dataset, AP defined in (2.3a) is the primary (or default) metric used. The standard for the metrics in this dataset is that AP and AR are averaged over multiple OKS values $(0.50 : 0.05 : 0.95)$ unless otherwise specified. $\kappa i = 2\sigma i$ for each keypoint type i. For person subjects, the σ's are pre-defined as $.026$, $.025$, $.035$, $.079$, $.072$, $.062$, $.107$, $.087$ and $.089$ for the nose, eyes, ears, shoulders, elbows, wrists,

hips, knees, and ankles respectively. The single most important metric in the MS COCO dataset that is used to benchmark SoTA methods and models is the AP averaged across all 10 OKS thresholds. In the MS COCO dataset, small subjects and objects (for the object detection use case of the MS COCO dataset) under the segment size of 32^2 do not contain keypoint annotations.

Note that Frames per Second (FPS), a measurement of graphics processing speed per second is also considered in some MS COCO dataset-based solutions and papers [Fang *et al.* 2017]. FPS is a way to benchmark and compare the processing speed of many computer vision models, and is a very generalised metric that may apply to most datasets. Frames per second, also known as frame rate, has a very simple calculation and interpretation and is as follows:

$$FPS = f/t \tag{2.7}$$

where f is the number of still image frames processed (can be variable resolution and size), and t represents time elapsed in one second. Average FPS is used to give a better representation of model speed performance over time, which is simply that the FPS is measured over a period of seconds and the average over the seconds is calculated.

2.6.2 Max Planck Institute for Informatics human pose dataset

The Max Planck Institute for Informatics (MPII) human pose dataset is one of the SoTA benchmarks specifically for evaluating articulated 2D pose estimation [Andriluka *et al.* 2014a]. The dataset contains around $25,000$ images containing over $40,000$ distinct people with annotated body joints. The images in MPII are systematically collected using an established taxonomy of every day human activities, overall the dataset covers 491 distinct human activities. The images in the MPII dataset are collected from YouTube, frames are sampled along with their preceding and following frames. These are then sampled and annotated by hand. Visibility and left-or-right labels for joints are labelled. Frames are sampled from $3,913$ videos.

The MPII dataset uses a standard of 16 keypoints in a skeleton-based human body representation, images are collected in RGB format. The dataset is good for estimating both single- and multi-person poses.

The strength of the MPII dataset is that images in the MPII dataset are diverse and varied with many in-the-wild data examples for more general pose estimation and activity recognition. The MPII dataset is challenging in terms of diversity on subjects, poses, challenges, camera quality, views, lighting and environments [Andriluka *et al.* 2014a].

A secondary contribution of this diverse and natural dataset is recognising the most common issues with pose estimation on images that holds across both the 2D and 3D pose estimation domains. Figure 2.6.2 shows examples of these difficulties.

Two main metrics are used to benchmark performance in the MPII dataset (along with generic ones such as FPS defined in (2.7)). These are Average Precision (AP), and head-normalised Percentage Correct Keypoints (PCKh), which is based on Percentage Correct Keypoints (PCK).

Percentage of Correct Keypoints (PCK) measures the distance between the predicted joint location and the true joint location. The PCK evaluation metric measures specific

Figure 2.4: Challenging example scenarios in HPE found in the MPII dataset [Andriluka *et al.* 2014a]

body joint localization accuracy. PCK uses a threshold defined by the maximum side length of the external rectangle bounding box of ground truth body joints. Thus, detecting a joint is considered correct if the distance between the predicted joint and the true joint is within a certain fraction of the specified threshold. The higher the PCK, the better the model, with PCK ranging between 0 and 1. PCK can be interpreted using the following:

$$PCK_i^k = \frac{\sum_p \delta(\frac{d_{pi}}{d_p^{df}} \leq T_k)}{\sum_p 1} \tag{2.8}$$

where p represents the p_{th} subject (if performing multi-person pose analysis), i is the i_{th} joint of the skeleton body model, d represents the distances and T_k is the threshold value.

PCKh is a modified version of PCK. Head-normalisation means that the matching threshold of PCKh is a percentage of the head segment length. The most common threshold used is 50% of the head portion length. PCKh can be interpreted as the head-normalized probability of the correct keypoint metric. In PCKh, joint detection is considered correct if the predicted joint location is with a certain threshold from the true joint location. But the threshold should be adaptive based on the apparent size of the subject image. It should fall within αl pixels of the ground-truth position, where α is a constant and l is the head size that corresponds to 60% of the diagonal length of the ground-truth head bounding box. The $PCKh@0.5(\alpha = 0.5)$ score is most commonly used for reporting performance of MPII results and methods. PCKh is used to make the metric articulation independent.

Average Precision (AP) for the MPII dataset is the same metric as the AP used for the MS COCO dataset (refer to (2.3)). The values at which the AP may vary is different for different datasets [Andriluka *et al.* 2014a].

2.6.3 Leeds Sports Pose dataset

The Leeds Sport Pose dataset and its extension LSP Ext, is a collection of images and annotations for human pose estimation in sports [Johnson and Everingham 2010]. It contains a combination of $12,000$ images ($2,000$ from the base LSP dataset and a further $10,000$ from the extended dataset) of athletes in various poses and activities, with annotations for 14 body joints in the skeleton body model. These images are mostly from sports activities images sampled from Flickr which make it very challenging in their pose appearance. Most subjects in the dataset are also scaled to roughly 150 pixels in length.

The Leeds Sport Pose dataset has a specific focus on sports-specific poses and activities. This makes the LSP dataset well-suited for pose estimation in sports applications, such as motion analysis or sports performance tracking. The annotations are collected manually using a custom-built annotation tool, resulting in some high-quality sports and complex pose annotations. Expert annotation is used to ensure a high level of accuracy and consistency in the labeling.

LSP includes images from a variety of sports, including football, basketball, and tennis, among others. This makes it a good representation of the types of poses and activities that are common in sports, and allows researchers to evaluate the performance of their algorithms across different sports and activities.

Figure 2.5: Example of sports variety in the Leeds Sports Pose (LSP)dataset [Johnson and Everingham 2010]

The main metric used to evaluate performance for the LSP dataset is the Percentage Correct Keypoints (PCK) metric, defined previously in (2.8).

2.6.4 Frames Labelled in Cinema dataset

The Frames Labeled in Cinema (FLIC) dataset is a 2D human pose benchmark dataset. It consists of just over $5,000$ frames from videos of people performing various activities, such as dancing, playing sports, and doing gymnastics. The FLIC dataset is unique in that it was created using frames from 30 popular Hollywood movies using a person detector model by Bourdev and Malik [2009] on every tenth frame of the movie input. The subjects in the dataset are labelled with 10 skeleton-model body joints including head, shoulders, elbows, wrists, hips, knees, and ankles and supports both single-person and multi-person use models.

The images in the FLIC dataset contain individuals in diverse poses with diverse clothing. This makes it a good representation of real-world scenarios, where people are

often captured in motion and in a variety of cinematic and staged lighting conditions [Sapp and Taskar 2013].

The FLIC dataset has gained an expanded super-set (sometimes called FLIC full) containing more challenging examples with $20,928$ images.

Figure 2.6: Example appearances in the Frames Labelled in Cinema (FLIC) dataset [Bourdev and Malik 2009]

The FLIC dataset is sometimes used to train elbow and wrist detection specifically, due to cinematic techniques resulting in cameras capturing and focusing on upper-body poses more often [Newell *et al.* 2016]. The FLIC dataset at large most commonly uses the PCK metric defined in equation (2.8), with the elbow and wrists detection metric threshold at $PCK@0.2$.

2.6.5 AI Challenger Human Keypoint Detection dataset

The AI Challenger (AIC) dataset by Wu *et al.* [2017] has three parts: human keypoint detection (HKD), large-scale attribute dataset (LAD) and image Chinese captioning (ICC). We only focus on introducing the relevant subset, the HKD part. The HKD dataset contains $300,000$ images with a large amount of labelled person subjects ($700,000$). The data is labelled using 14 keypoints in the skeletal body model. Bounding box annotations are also provided with the use of a person detector model [Wu *et al.* 2017].

The primary evaluation metric provided by AIC is AP (mAP) based on OKS, similar to the AP and OKS metrics defined in equations (2.3a) and (2.1) (using the same thresholds).

One of the features of the AIC dataset is that the HKD set has additional metadata annotations such as scenarios and environments (refer to figure 2.7 below).

Figure 2.7: Example appearances in the AI Challenger (AIC) HKD dataset [Wu *et al.* 2017]

2.6.6 PoseTrack dataset

The PoseTrack dataset is specifically a multi-person pose estimation dataset with video data (sequential image frames), requiring temporal tracking across frames [Andriluka *et al.* 2018]. The dataset is composed of 514 videos including $66,374$ stills in total. The train-test-validation split is organised by videos, with 300 for training, 50 for validation and 208 for testing. For the training videos, 30 frames from the center of the video is annotated. In the validation and test videos, besides 30 frames from the center, every fourth frame is also annotated for evaluating long range articulated tracking. The annotations include 15 body keypoints location, a unique person identifier and a head bounding box for each person instance.

There are some complexities involved in having sequence image data. Using bounding box IoU(Intersection-over-Union) as the similarity metric (SBbox) to link instances is sometimes problematic when a person instance moves fast. The bounding boxes do not overlap. An additional problem scenario is in crowded scenes where boxes may not have total corresponding relationship with instances. A more fine-grained metric is defined by the PoseTrack dataset benchmark [Andriluka *et al.* 2018]. A flow-based pose similarity metric S_{pose} is introduced, calculating the body joints distance between two instances using Object Keypoint Similarity as in (2.1).

Given one instance J_i^k in frame I^k and one instance J_j^l in frame I^l, the flow-based pose similarity metric is represented as:

$$S_{\text{flow}}(J_i^k, J_j^l) = OKS(\hat{J}_i^l, J_j^l), \tag{2.9}$$

where OKS represents calculating the Object Keypoint Similarity (OKS) in equation (2.1) between prediction and annotation human poses, and J^l represents the propagated joints for J^k from frame I^k to I^l using optical flow field $F^k \rightarrow l$.

The following figure 2.8 is an example sequence of images in the PoseTrack dataset.

Figure 2.8: Example data sequence in the PoseTrack dataset [Andriluka *et al.* 2018]

2.7 3D human pose estimation

In the following section we will introduce and discuss the background supporting and leading up to the methodologies and experiments used in our research. This is separate from the more pertinently related works introduced later in chapter 3.

3D pose estimation is generally more challenging compared to 2D pose estimation for a multitude of reasons. Firstly is that 3D pose estimation needs to account for the additional dimension of estimation. Secondly, despite the advent of depth sensors such as those introduced by Karl-Bridge-Microsoft [2012], large datasets for 3D pose estimation are much more difficult to annotate and contain compared to 2D pose datasets [Ionescu *et al.* 2014]. 3D datasets often necessitate the use of motion capture technologies, which also constrains data collection to lab environments and subject persons needing to wear motion capture markers and special equipment used to measure either locations or speed and angular motion. Many 3D human pose datasets that shape the advances of the field and determine the quality of the benchmarks (in the same way introduced in 2D pose estimation) is less variant compared to 2D pose datasets in terms of lighting, activity, environment, subject and pose variability.

Interest in the 3D human pose estimation problem area arises in computer vision, graphics and robotics where the pose or transformation of an object can be used for alignment of computer-aided design models, identification, grasping, or manipulation of the object. 3D HPE has been an active area of research for several decades, and originates from before the popularisation and advent of deep-learning [Sarafianos *et al.* 2016]. Model-based methods such as those in section 2.4 were developed for the 3D problem space such as a particle filtering methods making use of a probabilistic model Liu *et al.* [2015a]; Sarafianos *et al.* [2016]. However, such methods were constrained by the common challenges faced by non-deep learning models such as reliance on a priori information resulting in bad generalisability, and high complexity in computations.

The recent advances in the field of 3D human pose estimation hinges strongly on extensions of the methodology involved in 2D human pose estimation such as in section 2.5. Many of the methods and logic of methods involved in 2D pose estimation may be extended to help study 3D human pose estimation. As such, many of the metrics and

benchmarks used in 3D methods are direct extensions of, or inspired by metrics and benchmarks used in 2D methods. Thus, similarly to 2D pose estimation introduced in section 2.5, the human body abstractions and image types used are included in sections 2.2 and 2.3.2.

The research landscape of 3D pose estimation methods that are deep-learning based are introduced in this section. Classical or non-deep learning approaches such as exemplar-based approaches and pictorial models will not be included in the scope of this section. The following figure 2.9 is a framework for the taxonomy of this section.

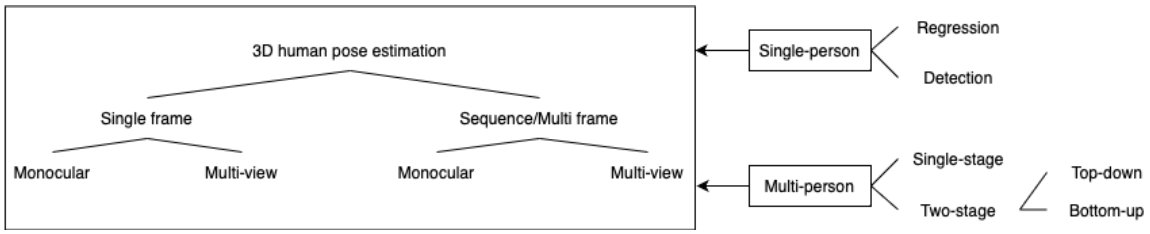


Figure 2.9: Method taxonomy of 3D human pose estimation based on deep learning

With reference to human body models used in the wider pose estimation field in section 2.2, 3D HPE does not make use of the contour-based models since contours are essentially representative of a grouping of flat area pixels, and is not informative nor intuitive to use to represent three dimensional data. Only skeleton-based and volume-based abstractions of the human body are used for 3D pose estimation, prominent examples being the skeletal model used for the MPI-INF-3DHP dataset [Mehta *et al.* 2017b], the SMPL model which is a triangle polygon human body mesh [Loper *et al.* 2015], and DensePose-COCO which is a human body model that is able to create a detailed mapping between individual pixels in an image and a three-dimensional representation of the human form's surface [Güler *et al.* 2018].

2.7.1 Single-person methods

We will discuss here an overview of existing methods using only single-frame information without taking into account the relations between frames. These methods therefore does not account for any temporal information, and can be either monocular or multi-view (calibrated images taken from multiple camera angles at the same time).

Direct end-to-end networks are a straightforward way to estimate a 3D coordinates of poses. These direct networks may be classified by whether or not they are a detection or regression-based method following the framework laid out in figure 2.9. Representations based on heatmaps encounter the traditional limitation of the process of selecting the peak values from probability heatmaps. This step is non-differentiable, which hinders the ability to train the system in a seamless end-to-end manner. Regression-based approaches have the capability to accomplish end-to-end learning and generate continuous outputs by considering expectations instead of relying on maxima.

Detection-based methods

Methods based on detection initially generate a probability heatmap for every joint, and the position of the joint is ascertained by identifying the point of highest probability on the heatmap. Pavlakos *et al.* [2019] employ a volumetric approach, training a convolutional neural network (CNN) to forecast the likelihood of each joint within the body's three-dimensional space on a per-voxel basis. They utilise a hierarchical prediction strategy, enhanced by intermediate supervisory steps in an iterative module for estimation, which progressively refines the predictive accuracy within the supervisory volume.

Luvizon *et al.* [2018] implement a framework that simultaneously processes action recognition and the estimation of both 2D and 3D human poses. By integrating 2D and 3D poses with volumetric models of the human body, they unify the pose representations. Models based on detection necessitate extra processing to transform the probability heatmaps into specific joint locations. Typically, this transformation is performed using the argmax function, which lacks differentiability and thus interferes with the learning process of neural networks, representing a potential area for improvement. Moreover, the accuracy of joint location predictions in detection-based networks is contingent on the resolution of the heatmap. A higher resolution generally translates to greater model precision. Consequently, detection-based models may experience a deficiency in spatial generalisation capabilities. To attain high precision, the model often requires heatmaps of substantial resolution, which in turn leads to a quadratic rise in computational demands and memory usage [Wang *et al.* 2021].

Regression-based methods

Regression-based methods are another commonly used method category for 3D single-frame pose estimation. Both 2D and 3D human pose estimation are essentially regression problems directly reconstructing the relative positions of joints in relation to a pre-defined root joint position. Li *et al.* [2014] present an early iteration of a deep neural network featuring a two-branch design that concurrently detects root joint and regresses relative positions of the remaining joints. Zhou *et al.* [2016b] integrate prior knowledge by encoding geometric structures inherent in human bodies. They use a kinematic object model comprised of various bones and joints, where the bones are fixed in length and are able to rotate around their adjoining joints. They ascertain that the unchanging bone lengths do not adequately represent the variation present in human body sizes, thereby limiting the generalisability of their model. Sun *et al.* [2017] expand on these findings by experimenting with the regression of bones as opposed to joints. They discovered that bone representations are not only simpler to learn but also more accurately reflect the geometric constraints and exhibit greater stability. To address the issue of the independence of L2 loss, they encoded long-range interactions among the bones using an innovative compositional loss function. This function imposes penalties for deviations from the anticipated compositional structure of the human body and motivates the network to learn a representation that aligns with the body's intrinsic structure. This approach necessitates the conversion of pose data into a relative bone-centric format [Sun *et al.* 2017]. A soft-argmax function is proposed by Luvizon *et al.* [2019] that enables converting of feature maps into joint co-ordinates,

which enables a fully-differentiable framework. In a similar vein, Nibali *et al.* [2019] introduce a new differential spatial layer that numerically transforms data to maintain differentiability for the model end-to-end. This also conserves spatial generalisation capabilities of the model.

2.7.2 Multi-person methods

Multi-person 3D human pose estimation is generally considered a more difficult problem than single-person 3D human pose estimation. This is primarily due to the more challenging state space existing in the data. The larger state space comes from the dimensional increase of possible rotations and translations of human body parts in 3D space. This comes with the increased challenges of occlusions and cross-view ambiguities. This section presents representative work and progression made in multi-view 3D pose estimation only. We present this background research in a similar manner introduced in the single-person problem space (section 2.7.1) following the taxonomy introduced in figure 2.9.

Similarly to the challenges faced in 2D multi-person pose estimation (section 2.5.2), there are challenges in distinguishing joints and body parts that belong to a specific person. There may be many persons in the image that are standing in close distance to each other or mutually occluding in complex pose positions. There may often be scaling issues that occur due to what resolution data is available at and at which resolutions the data is being processed at. A second challenge unique to the 3D multi-person pose estimation problem is that absolute pose position needs to be estimated in a global 3D space instead of a root joint-relative position (in 2D multi-person pose estimation).

Single-stage methods

Single-stage methods in multi-person 3D HPE aim to directly estimate the 3D pose of all individuals in the image without relying on intermediate representations. These methods typically use a single neural network to predict the 3D joint locations of all individuals in the image simultaneously, then apply grouping methods to try and group the joints into specific persons. This is the same logic as bottom-up approaches for multi-person 2D pose estimation. Just like the previous sections, we follow the pattern here of presenting representative works that offer insight into notable works and direction of research in each category of methodology in the field.

For single frame multi-person 3D pose estimation, Nie *et al.* [2019] claims the first Single-stage multi-person Pose Machine (SPM) approach (model architecture shown in figure 2.10 and figure 2.11), which predicts the 2D joint locations of all individuals in the image simultaneously using a single neural network. A novel Structured Pose Representation (SPR) that unifies person instance and body joint position representations is proposed to support the SPM. This multi-task learning framework jointly optimises the detection and pose estimation tasks. This method is proposed to address the challenge of two-stage methods often suffering low efficiency, as they require multiple passes through the network and post-processing steps to obtain the final pose estimates. The SPM approach is shown to achieve state-of-the-art performance on several benchmark datasets, while also being more efficient than previous methods.

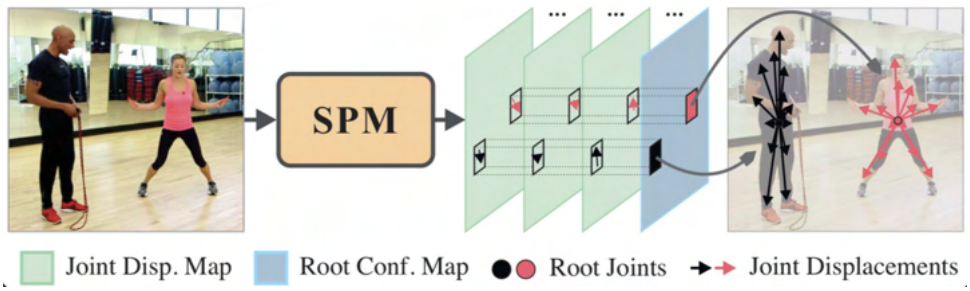


Figure 2.10: Single-stage multi-person pose machine (SPM) [Nie *et al.* 2019]

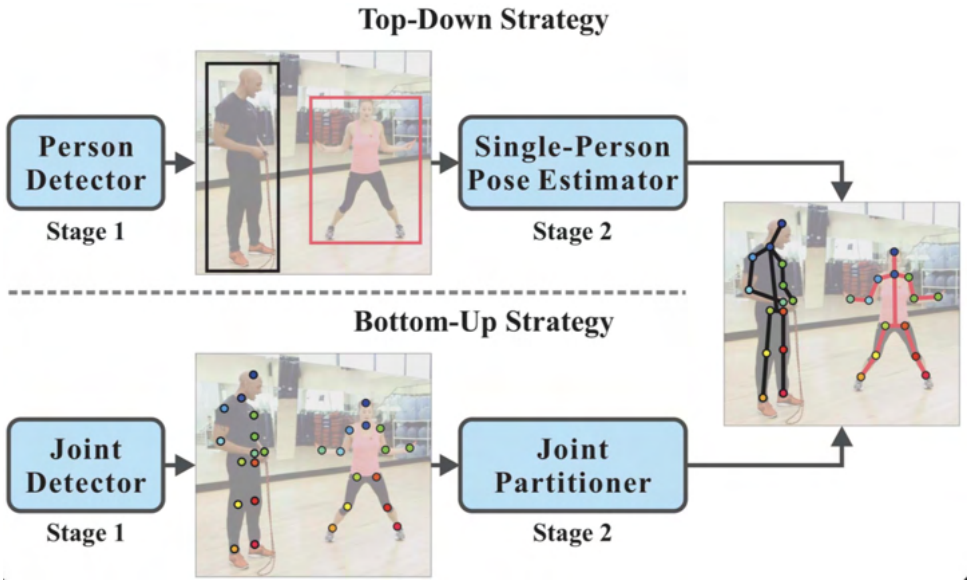


Figure 2.11: Two-stage pose machine [Nie *et al.* 2019]

A recent paper introduced by Wang *et al.* [2022] proposes a novel Distribution-Aware Single-stage (DAS) 3D human pose estimation model capable of multi-person estimation. The authors introduce a new distribution-aware loss function that is able to take into account distributions of the 3D pose annotations in training data. This loss function encourages the model to predict poses that are more likely to occur in the training data, which leads to better generalisation on unseen data. The method also incorporates a novel multi-person pose estimation module that is able to estimate the poses of multiple people in an image simultaneously. This module uses a set of learnable keypoints to represent the locations of many people in the image, and uses a graph convolutional network to estimate the poses of each person. The authors evaluate their method on several benchmark datasets and show that it outperforms state-of-the-art methods in terms of both accuracy and efficiency. They also demonstrate that their method is able to generalise well to images with varying numbers of people and different poses. Figure 2.12 illustrates the general pipeline of the DAS model.

Single-stage methods have the advantage of being computationally efficient and can handle a variable number of individuals in the image. Single-stage methods often encounter challenges accurately localizing joints in crowded scenes or when persons are occluding each other.

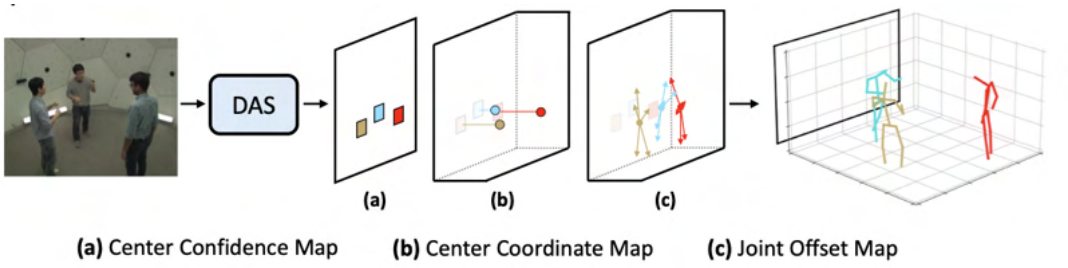


Figure 2.12: Distribution-Aware Single-stage model [Wang *et al.* 2022]

Two-stage methods

Two-stage methods first detect individuals in the image and then estimate their 3D pose. These methods typically use a person detector to identify individuals in the image and then use a separate neural network to predict the 3D joint locations for each person. In the following we will introduce three exemplary two-stage methods and illustrate the advancements and general ideas behind each.

The paper presents a novel approach for both 2D and 3D multi-person pose detection in in-the-wild images, called LCR-Net++. The proposed method builds upon the previous LCR-Net architecture, which uses a two-stage approach to detect human poses. The first stage generates a set of candidate poses, while the second stage refines the poses and assigns them to specific individuals. Rogez *et al.* [2019] finds allowing the model to estimate both 3D and 2D poses of multiple individuals in tandem can enhance the precision and efficiency of pose estimation. This improvement is attributed to obviating the need for initial localisation of each person. Initially, a pose generator creates multiple pose candidates throughout the image. Subsequently, a classifier evaluates these poses, and a regression head is employed to fine-tune the proposals for both 2D and 3D poses. Rogez *et al.* [2019] conduct end-to-end training of a location-classification-regression network, LCR-Net++, which enhances the original LCR-Net architecture through several significant alterations. Firstly, it adopts a multi-stage architecture that facilitates more precise pose estimation. Secondly, the network features an innovative multi-scale feature fusion module that amalgamates features across various scales, thereby refining the accuracy of pose estimation. Thirdly, the introduction of a new 3D pose estimation module utilises a novel volumetric representation of the human form to ascertain 3D poses.

A 3D multi-person framework adopting a Mask R-CNN is proposed to identify human bounding boxes and compute image characteristics [Moon *et al.* 2019]. RootNet is a model that estimates the 3D coordinates of the root joint (usually the pelvis) of a human skeleton, given a 2D image of the person. It uses a convolutional neural network (CNN) to extract features from the image, and then applies a regression network to reconstruct the three-dimensional coordinates of the root joint [Moon *et al.* 2019]. PoseNet, on the other hand, estimates the 3D coordinates of all the joints in a human skeleton, given a 2D image of the person. It also uses a CNN to extract features from the image, but then applies a multi-stage regression network predicting 3D coordinates of all joints [Kendall *et al.* 2015].

Mehta *et al.* [2018] present innovative occlusion-robust pose-maps (ORPM), which produce a fixed number of maps encoding 3D postures of each individual in a provided image. This approach draws inspiration from the 2D pose estimation technique using Part Affinity Fields (PAFs), as employed by Cao *et al.* [2017]. It leverages associations between body parts, enabling the deduction of poses for an indeterminate number of individuals without relying on detection bounding boxes.

Two-stage methods have the advantage of being able to more accurately localise joints even in crowded scenes or when persons are occluding each other. However, they may be more computationally expensive.

2.7.3 Single-frame 3D pose estimation

Single-frame pose estimation is a grouping of 3D pose estimation methods that are trained on single frames of input data, without explicitly designing or accounting for any temporal relations between the frames, such as frame sequences from videos or multi-view perspectives. This approach has several advantages, including reduced computational complexity and the ability to handle large datasets with diverse camera viewpoints and lighting conditions [Wang *et al.* 2021].

Constructing a 3D pose from one image is an attractive approach due to its low-resource image and labelled data requirements. Nevertheless, this poses a challenge because distinct 3D poses may correspond to an identical 2D image, rendering the problem ill-posed. Additionally, in certain settings, using temporal or multi-view data to diminish ambiguity in the recovery process may not be feasible. Consequently, extensive research has been undertaken, and a variety of methods have been devised to tackle these challenges.

In this section, we present and discuss some exemplary studies of three principal approaches to single-frame 3D pose recovery: direct prediction of 3D poses, elevation from 2D poses, and techniques based on the Skinned Multi-Person Linear model (SMPL). These approaches have been formulated to resolve the issue of uncertainty in deducing 3D poses from single images.

Monocular methods

The most direct method for estimating 3D human poses involves creating a network that is end-to-end to estimate the 3D joint locations of poses. Approaches mapping input images directly to 3D joint locations fall into two broad categories of methods: detection-based and regression-based. Efforts have been made to integrate heatmap representations with joint regression. These methodologies and their overarching categories are summarised in section 2.7.1.

Detection-based approaches generate a probability heatmap of each keypoint, with the location of the joint identified by selecting the point of highest likelihood on the heatmaps. Pavlakos *et al.* [2017a] represents a 3D pose using a volumetric approach and train a convolutional neural network (CNN) that predicts the likelihood for each joint at the voxel level within the volume, significantly enhancing direct regression of keypoint coordinates. They create a coarse-to-fine approach that is used to predict pose

that incorporates intermediate supervision and an iterative estimation module, which gradually enhances resolutions of supervision volumes.

Luvizon *et al.* [2018] introduce a multitasking model framework that simultaneously performs the estimation of 2D/3D poses and action recognition, unifying the estimation of 2D and 3D poses through the use of volumetric heatmaps. Nonetheless, these methods depend on supplementary steps to translate heatmaps into joint locations, typically employing the argmax function. Since argmax is non-differentiable, it can interfere with learning processes of neural networks. Furthermore, the accuracy of the predicted keypoints is dependent on the resolution of the heatmap, which does not naturally offer spatial generalisation. For high-precision outcomes, the heatmaps need to be of sufficient spatial resolution, leading to a quadratic increase in both computational costs and memory requirements.

Estimating human poses is fundamentally a regression task that aims to directly determine the positions of joints in relation to the location of the root joint. A straightforward and efficacious neural network is developed by Li and Chan [2015] with a dual-branch system that concurrently identifies the root joint's coordinate and regresses relative positions of the non-root keypoints. To integrate priors about the human body's geometric structure, Zhou *et al.* [2016b] proposed a kinematic object model composed of multiple bones along with joints, with the bones being of constant length and capable of rotating around the connecting joint. Nevertheless, the invariant bone lengths fail to capture the diverse nature of a real human skeleton, thus constraining their model capacity for generalisation.

Sun *et al.* [2017] suggested that for pose estimation, regressing bones rather than joints is more practical, given that representations of bones are simpler to learn, more accurately embody geometric constraints, and are inherently more stable. In order to address the problem of L2 loss for bones being localised and independent, they utilised a compositional loss function in their research, which accounts for the long-range interactions among the bones. Nevertheless, this approach necessitates the transformation of pose data into a format that is relative to the bones.

Nibali *et al.* [2018] presented a novel network layer termed the Differentiable Spatial to Numerical Transform (DSNT), which maintains the model's differentiability end-to-end and enhances its spatial generalisation. The objective of these methods is to augment the model's ability to generalise and to retain the end-to-end differentiability of the entire framework. In more recent work, Luvizon *et al.* [2019] introduced the soft-argmax function as a means to transform feature maps into joint coordinates, thereby also achieving a fully differentiable framework.

An additional prominent category of methods in 3D multi-person single-frame pose estimation involves the elevation of 2D poses into 3D space. Martinez *et al.* [2017] introduce a neural network-based method that is both simple and effective for lifting 2D poses to 3D, spurring further research in this area. Various techniques have been developed within this domain, which are concisely reviewed and discussed in this section. Among these are techniques aimed at diminishing ambiguity by merging 2D joint heatmaps with image features that provide depth cues [Habibie *et al.* 2019; Park *et al.* 2016; Tekin *et al.* 2017; Zhou *et al.* 2019].

Park *et al.* [2016] eliminate implausible 3D human joint positions by combining the results of 2D pose estimation with image features. Habibie *et al.* [2019] explicitly

represent 2D poses with heatmaps and encode implicit depth information to deduce viewpoints and 3D pose parameters, which allows for the projection of the generated 3D pose onto the 2D plane using these viewpoint parameters. Tekin *et al.* [2017] create a fusion stream that incorporates 2D joint heatmaps and 3D image cues within a trainable framework, circumventing the necessity for manually designed fusing techniques. Zhou *et al.* [2019] explore usage of part-centric heatmap triplets (HEMlets), which segment body skeletons into 14 components and capture the local ordinal information of subsections. Each HEMlet comprises 3 distinct map: the positive polarity map, the zero polarity map and a negative polarity heatmap. Through this methodology, the relative depth information serves as a link between 3D and 2D poses, bridging the dimensional gap.

There are a group of methods using spatial configurations between human pose joints in the development of algorithms for elevating from 2D to 3D. Hochreiter and Schmidhuber [1997] and Nie *et al.* [2017] use long short-term memory (LSTM) to process body part images with 2D poses. They express the structure of the skeleton body model as a tree graph structure in order to propagate contextual geometric information. Moreno-Noguer [2017] diminishes ambiguity by computing the pairwise distances between body joints, which are encapsulated in a Euclidean distance matrix (EDM). This matrix encodes data informing the structure of the body and capture the pairwise dependencies and correlations between joints. Zhao *et al.* [2019] introduce a semantic graph convolutional network (GCN), which discerns the edge weights within the skeleton graph on a per-channel basis, deriving 3D pose from 2D joint coordinates. Wang *et al.* [2019] directly estimate 3D keypoints progressing from a lower degree of freedom (DOF) to a higher DOF. Ci *et al.* [2019] propose the locally connected network (LCN), which employs distinct feature extraction filters instead of a shared filter, which is common in GCNs. This approach allows for the natural skeleton adjacency matrix to be learnable, echoing the methodology of Zhao *et al.* [2019].

Additional methods are researched to ensure consistency between estimated 3D pose from 2D pose lifting. Tome *et al.* [2017] use a multi-stage approach that concurrently takes into account both 2D and 3D poses, projecting the synthesised 3D pose back to 2D in order to create 2D belief maps. These maps are then integrated with belief maps generated by a standard 2D pose estimator. Jack *et al.* [2019] adopt a learning approach for the loss function through the use of an energy-based loss function. This energy loss function evaluates the congruence between a suggested 3D pose and the ground truth using a dense two-layer network, as well as assessing the plausibility of the proposed pose in real-world scenarios. Habibie *et al.* [2019] project the synthetic 3D pose onto the 2D plane by employing viewpoint parameters obtained through estimation. Chen *et al.* [2019b] use an unsupervised rotation-invariant approach with the generated 3D pose by lifting the 2D pose into the 3D space, then rotating and projecting it back to 2D, subsequently lifting this new 2D pose back to 3D, and finally computing the Euclidean loss between the two 3D poses as well as between the restored 2D pose and the original 2D input. Novotny *et al.* [2019] introduce a canonicalisation network designed to map members of an equivalence class to a canonical reconstruction, thereby regularising the outcomes. They reconstruct the 3D shape using a factorisation network capable of separating viewpoint information from object deformations. The training of the reconstruction branch is conducted by minimising a re-projection loss.

Adversarial learning methods such as Generative Adversarial Networks (GANs) are studied in HPE to produce more realistic 3D human pose [Fish Tung *et al.* 2017; Wandt and Rosenhahn 2019; Yang *et al.* 2018]. Fish Tung *et al.* [2017] propose adversarial inverse graphics networks (AIGNs), which consist of three components: a generator, a renderer, and a discriminator. First, a generator component generates heatmaps for 2D pose and subsequently infers the shape and camera parameters from these heatmaps. The renderer acts as the re-projection function, and the discriminator evaluates the plausibility of the 3D pose generated by the network. Yang *et al.* [2018] train a multi-source discriminator that processes several types of input: the original image, a geometric descriptor that details the correspondence between the image and the pose, the 2D pose, and depth heatmaps. Wandt and Rosenhahn [2019] experiment with a Generative Adversarial Network (GAN) to discern whether a 3D pose produced by the network appears realistic, as opposed to assessing the 2D pose. Additionally, they introduce a network specifically designed to estimate camera parameters. These parameters are then used to project the generated 3D pose back into the 2D space, where the loss is calculated in comparison with the original 2D pose.

Another notable class of single-frame techniques employs methods based on the Skinned Multi-Person Linear (SMPL) model. The SCAPE body model, which was introduced in earlier research [Anguelov *et al.* 2005], is adapted to images through the use of manually annotated keypoints and outlines. In contrast, more contemporary studies have adopted the SMPL model, as demonstrated in Loper *et al.* [2015], where the model is automatically adjusted. This adjustment is accomplished by either resolving an optimisation challenge to align the model with the data, as seen in Bogo *et al.* [2016], or by directly inferring the model weights through a comprehensive deep learning approach that converts pixel information into SMPL model parameters [Kanazawa *et al.* 2018]. The approach by Kanazawa *et al.* [2018] also involves reducing the re-projection error and training a discriminator with an extensive mesh database to assess the plausibility of the generated shape and posture parameters. The SMPL model integrates inherent understanding of human anatomy, enabling it to be accurately adjusted with minimal data. A variety of optimisation techniques have been introduced for estimating 3D human poses. One such technique is employed by Bogo *et al.* [2016] in their SMPLify algorithm, which initially predicts 2D keypoints through the use of DeepCut [Pishchulin *et al.* 2016], followed by the fitting of the SMPL model to these identified keypoints. The fitting process is directed by aligning the 2D keypoints projected from the SMPL model with the detected 2D keypoints. Enhancements to SMPLify have been introduced recently in Lassner *et al.* [2017], this involves an additional step of matching the image silhouette with the shape silhouette generated by projecting the SMPL model. This shape encompasses all the pixels that make up the projection of the body. Tan *et al.* [2017] developed an encoder-decoder framework to lessen the reliance on datasets containing body poses with labelled ground truth data. The decoder takes in SMPL parameters and generates a corresponding shape silhouette. After establishing the decoder, the entire network undergoes end-to-end training using real images and their associated silhouettes. Pavlakos *et al.* [2018] devised a network known as PosePrior, which processes heatmaps to deduce the pose attributes of a SMPL model. Concurrently, a separate model is used to estimate the shape parameters from the silhouette. Subsequently, the pose projection is aligned with annotated keypoints and masks in

3D. Meanwhile, Omran *et al.* [2018] employ a semantic segmentation CNN to divide the image into twelve semantic regions, then transform each semantic part probability map into SMPL attributes. This transformation utilises the semantic segmentation of body parts in conjunction with the anatomical constraints inherent in the SMPL model. Regression-based SMPL approaches often exhibit sub-optimal image-model congruence due to their reliance on single-step predictions and the necessity for extensive data. To overcome this limitation, Kolotouros *et al.* [2019] introduced the SMPL 'oPtimization IN the loop' (SPIN) technique, which amalgamates the strengths of both optimisation and regression methodologies. Within SPIN, the outcomes of regression serve as the basis for subsequent optimisation, and a pixel-precise optimisation phase is employed to further refine the supervision signal.

Multi-view single-frame methods

The next categorisation of methods is multi-view single-frame methods. This refers to models that use a single pose in time (frame) as input to the model, but take into account multiple camera views. The obvious advantage here is to reduce pose ambiguity by leveraging the additional data from different camera perspectives. One drawback is that the data for this category of methods is difficult to collect and requires a lab environment setup with synchronised cameras taking pictures of likely a smaller group of subject persons [Ionescu *et al.* 2014]. Here we will introduce a research background for multi-view single-frame methods that show general research direction and challenges in the area. Common approaches involve combining or integrating 2D multi-view heatmaps [Qiu *et al.* 2019; Tome *et al.* 2018; Pavlakos *et al.* 2017b], geometric methods such as triangulation [Kocabas *et al.* 2019], developing multi-view consistency [Rhodin *et al.* 2018b] and SMPL model-based human body representation [Liang and Lin 2019].

For techniques designed to integrate 2D heatmaps from multiple camera views, Pavlakos *et al.* [2017b] employ a pictorial structures model to combine the 2D heatmaps from each pose derived from each camera view. The heatmaps are back-projected into a shared discretised three-dimensional space, where pose priors are applied by imposing constraints on the proportional lengths of limbs and the likelihood represented by the heatmaps. The estimation of the pose is achieved by calculating the average of the marginal distribution for each joint. Multi-stage frameworks are designed such as Tome *et al.* [2018] that use iterative refinement on 3D poses from images captured from multiple viewpoints, informed by 3D priors. At every stage, the CNN receives inputs that include images from various camera angles and the pose heatmaps generated in the preceding stage. The final 3D poses are obtained by optimising the latent 3D pose prior space that aligns with the 2D poses inferred from 2D heatmaps derived from all views. This 3D pose is simultaneously reprojected onto the 2D plane for each camera viewpoint to form 2D heatmaps. This is fused together with the 2D heatmaps obtained using the 2D estimator. Works such as Qiu *et al.* [2019] feeds multi-view images into a CNN model to merge information from other views to a focus view. They also use a pictorial structures model recursively to optimise 3D poses, and demonstrate better results by showing a progressive reduction in the quantisation error.

Multi-view consistency-focused methods try to use the assumption that the model should predict equal poses from all camera views on the pose frame. This assumption is used to obtain more accurate model results. Rhodin *et al.* [2018b] implements a weakly-supervised model by using the assumption of consistency by forcing the model to predict the same (or similar with threshold) pose all views during training. It is found that this approach reduces the volume requirement for labelled data and that the model can be applied to traditionally difficult 3D data environments such as sports. Rhodin *et al.* [2018a] uses a semi-supervised encoder-decoder network that learns a body representation that is geometry-aware. First, unlabelled multi-view images of the same single-frame pose are used to get body representation information, then weak supervision in the form of small amounts of training data is used to learn a geometry-aware representation to 3D pose mapping. Chen *et al.* [2019b] encodes 2D pose geometries to latent 3D pose space using a constraint for representation consistency. The encoder undergoes training using pairs of multi-view images. The latent geometry representation of one image is scaled by the relative rotation matrix that transforms it to the pair image. The rotated representation is then used as input to the decoder which tries to output the image pair pose.

Geometric methods (such as triangulation) is another method category for pose construction. A baseline approach is proposed by Iskakov *et al.* [2019] which first obtains a 3D pose using an algebraic triangulation module with the 2D joint positions and confidences of joints. The limitation of this model is that it processes images from separate viewpoints in isolation. In the same paper they propose a more advanced triangulation procedure that aggregates feature maps into a CNN to produce 3D heatmaps. Separately, Kocabas *et al.* [2019] creates the EpipolarPose model pipeline which leverages stereo/epipolar geometry to reconstruct pose from multi-view 2D poses. By using stereo-geometry inherent in the nature of two or more viewpoints, they are able to generate a self-supervised signal to train a 3D pose estimation model. Geometric methods to produce and reinforce 3D poses prove to be a resource-efficient and fast way to obtain prerequisite information used to train models to predict 3D poses.

The final general group of methods for multi-view single-frame methods is using human-body models, most prominently SMPL. Existing shape and pose estimation methods that rely on silhouette or shape supervision cannot be directly applied to subjects wearing loose-fitting clothing, which is part of the shape-awareness problem in HPE [Wang *et al.* 2021].

Liang and Lin [2019] create a large synthetic dataset with variant and diverse views, clothing and shapes to train models to be more shape-aware. The model pipeline and architecture consist of multiple stages that estimate parameters on a view-by-view basis. This pipeline accepts encoded image features and the previous body pose as input, and uses camera estimates to determine corrective values. These parameters encompass view-specific camera attributes and shared human body parameters across all views.

Single-frame pose estimation is a promising approach to 3D pose estimation that offers several advantages over traditional methods. However, it also presents several challenges that may be addressed through the development of new techniques and algorithms. Heatmap representations have practical drawbacks, the most prominent of which is that the maximum-taking operation is not differentiable, which makes end-to-end model training difficult [Luvizon *et al.* 2018; Pavlakos *et al.* 2017a]. Regression

approaches, on the other hand, allow for holistic model learning which produces continuous outputs by leveraging taking expectations instead of maxima [Luvizon *et al.* 2019; Nibali *et al.* 2018; Sun *et al.* 2017; Zhou *et al.* 2016b; Li and Chan 2015]. However, regression models are not as effective as detection-based models. To combine the benefits of both methods, Sun *et al.* [2018] introduce an integral regression approach that estimates keypoints by calculating the probability-weighted average of all joint coordinates within a heatmap. This approach enables end-to-end training, while using less computational resources. In the next subsection, we will look at 3D pose estimation using multi-frame input as opposed to single-frame input, where the main advantage is leveraging temporal information to garner better insights on poses.

2.7.4 Multi-frame 3D pose estimation

Multi-frame 3D pose estimation is a type of pose estimation that uses multiple frames, or image sequences for training, as opposed to single-frame methods that are trained on individual frames of input data. Video data is an efficient way to collect large amounts of pose data, so extracting 3D poses from sequential or temporal data holds appeal. By incorporating temporal information from multiple frames, multi-frame methods can improve the accuracy and robustness of pose estimation, particularly in challenging scenarios such as occlusions, motion blur, and complex camera viewpoints [Ionescu *et al.* 2014]. Some methods use real-time systems to process video sequences, while others use entire videos as input and output pose sequences.

One of the main advantages of multi-frame 3D pose estimation is the ability to capture the temporal dynamics of human motion, which can provide important contextual pose information. For example, by tracking the movements of individuals over time, multi-frame methods can better distinguish between different poses and actions, and can more accurately estimate joint locations and orientations.

However, multi-frame 3D pose estimation also presents several challenges, such as the need to handle missing or noisy data, and the increased computational complexity of processing multiple frames. To address these challenges, researchers have developed various techniques for multi-frame pose estimation, including recurrent neural networks (RNNs) that can model temporal dependencies between frames, and graph-based methods that can incorporate both spatial and temporal information. Occlusion, lighting and background variation, movement in the camera, fast motion, loose clothing and diverse subjects may result in significant variations in the shape and appearance of individuals, posing challenges to the generaliseability of the model.

Multi-frame 3D pose estimation has numerous applications, including sports analysis, motion capture, and virtual reality. For example, it can be used to analyze the movements of athletes during training or competition, or to create realistic animations of human motion in virtual environments.

Monocular image sequences

Performing 3D pose estimation through monocular image sequences means that only one camera viewpoint is used for each sequence of image data. Taking pose from monocular images in general suffers inherent depth ambiguity. Image sequencing or

temporal data is thus used in many methods to overcome depth ambiguity. A main intuitive advantage of this group of methods is that given a mostly stationary camera, properties such as shape and bone lengths of the same subjects are constant, while the person's movements may be regular. Temporal relations in video inputs can be learned using temporal-enabled models such as LSTMs [Lee *et al.* 2018; Coskun *et al.* 2017; Lin *et al.* 2017], CNNs [Dabral *et al.* 2018; Tekin *et al.* 2016], Temporal Convolutional Networks (TCNs) [Pavllo *et al.* 2019] and Graph Convolutional Networks (GCNs) [Cai *et al.* 2019]. Temporally smooth poses are obtained by some works by penalising pose-related parameters, such as limb length priors about the body [Du *et al.* 2016; Mehta *et al.* 2017c]. Some research incorporates trajectory optimisation [Li *et al.* 2019; Arnab *et al.* 2019]. We briefly look into these methods as a background to understand the overall picture of the research landscape of the 3D HPE problem space.

Generally sequence-to-sequence models such as LSTMs are used to model temporal sequences and capture patterns and information of data changing over time. An LSTM Kalman filter (LSTM-KF) model is proposed by Coskun *et al.* [2017]. This model takes advantage of motion and noise models through combining an LSTM and a Kalman filter in order to learn temporal information. Lin *et al.* [2017] try to capture long-range dependencies using a recurrent 3D pose sequence machine. These also enforce the temporal consistency of predicted pose sequences. Multi-stage refinement is used for the reconstructed 3D poses. Lee *et al.* [2018] proposed a method for 3D human pose estimation using a Propagating LSTM (p-LSTM) network that models interdependency between joints. The model propagates information across the temporal dimension and using the interdependency between joints to refine the predictions. The p-LSTM network consists of a top-down pathway that propagates high-level information to lower-level features, and a bottom-up pathway that extracts features from the input sequence.

CNN-based neural networks have also been experimented with to learn temporal sequences. A method proposed by Tekin *et al.* [2016] use a CNN to directly predict 3D poses using motion-compensated sequence data. They use a CNN to learn a mapping between the input sequence and the corresponding 3D pose. Essentially, the model directly predicts 3D poses by regressing spatial-temporal volume of bounding boxes centered on a target frame. They observe that extracting such features directly using CNNs does not significantly enhance performance compared to using only spatial CNNs.

This generated interest and further investigation by Dabral *et al.* [2018] with Temporal PoseNet, a low-resource fully connected two-layer network that uses rectified linear units (ReLUs). This relatively simple network is able to take a constant number of time-adjacent pose frames as input and is able to learn complex motion and structure cues. Pavllo *et al.* [2019] propose a model for estimating 3D poses from 2D keypoint sequences using temporal dilated convolutions. Their approach is a type of lifting network that constructs 3D poses using 2D poses obtained by a 2D keypoint detector. This model is able captures long-term information by using dilated temporal convolutions. This results in a computationally efficient model with a large temporal window.

Due to the graph nature of skeleton pose models, researchers such as Cai *et al.* [2019] represent pose sequences as graphs and design Graph Convolutional Networks (GCNs) that exploit spatial-temporal relationships. This is a local-to-global model that uses 2D pose sequences to generate corresponding 3D pose sequences. This network

demonstrated its capability to learn multi-scale features and incorporate a temporal constraint for target human body pose sequences.

Temporally consistent poses through image sequences can also be obtained by setting pose-related parameters based on structural priors of the human body and training models on these. In works by Du *et al.* [2016], a markerless 3D human motion capture method is used to produce a dataset with monocular image sequences and height-maps. 2D joint locations are first estimated for each subject in each frame of the image sequence using a CNN. They impose constraints on limb length and ensure a temporal constraint on 3D poses when estimating 3D motions from the predicted 2D pose sequences. Mehta *et al.* [2017c] create a model called VNECT (Volumetric Non-rigid Multi-Person 3D Reconstruction). Their method focuses on estimating the 3D pose and shape of multi-person data from a single RGB image in a real-time manner. The method uses a volumetric representation of the human body to estimate the 3D pose and shape of each person in the image. The network takes as input a set of 2D joint heatmaps and a binary mask segmenting each person in the image. The estimated 3D joint locations and the binary mask are then used to generate a volumetric representation of each person used to estimate the 3D shape of the person. The model estimates a consistent temporal global 3D pose using temporal information.

Some models are developed to use trajectory and tracking motion to exploit temporal information and consistency, while attempting to shift towards semi-supervised or weakly-supervised settings. For example, Li *et al.* [2019] use unlabelled monocular videos in a two-stage configuration. In the initial phase, preliminary predictions are derived from a pose estimation model using weakly-supervised data. Matrix completion methods are used in the second phase to supervise further training of the pose estimation network by matrix completion methods applied on 3D joint trajectories. This model also takes into account the geometric pose prior constraints of the body. Arnab *et al.* [2019] propose an in-the-wild method that inputs a set of 2D joint heatmaps and a set of temporal features that encode a temporal context of a pose sequence. They design and experiment with a multi-frame bundle adjustment method for capturing temporal information. To generate a dataset as a weak supervision signal, bundle adjustment is applied to real-world unlabelled YouTube video data [Arnab *et al.* 2019]. Additionally, the authors propose a 3D pose and mesh reconstruction algorithm to mitigate estimation ambiguities.

Monocular image sequence methods are often studied due to the improved availability of data, particularly in weakly- and semi-supervised settings. The above methods offer a look into the research direction and prominent progress of work in the field in recent years. Next we will look at exploiting sequential data using multi-view images for training.

Sequence of multi-view images

Using multi-view image sequences for training has often resulted in better accuracy performance in models. However as mentioned previously, they require expensive synchronised camera setups and controlled lab environments. This may contribute to datasets being more visually homogeneous and less subjects mean models may not generalise to diverse human appearances and poses.

The idea of using multi-view image sequences to provide robust 3D pose predictions has generated interest for a relatively long time in the machine learning space. Stoll *et al.* [2011] and Rhodin *et al.* [2015] use 2D joint positions obtained from multi-view sequences to guide model initialisation for 3D pose estimation for more effective and robust training. Rhodin *et al.* [2015] proposes a versatile scene model for generative pose estimation that takes into account the visibility of body parts over time from multiple camera views. The model uses a differentiable visibility function that allows for efficient optimisation of the pose parameters, which they prove robust against occlusions and self-occlusions.

There is research into using multi-view temporal data to construct volumetric human bodies. Trumble *et al.* [2018] use a symmetric convolutional autoencoder to estimate body shape from a sparse set of wide-baseline camera views. Their autoencoder uses dual loss to enforce learning latent representational encoding of skeletal joint positions, at the same time learning a deep representation of volumetric body shape. The resultant volumetric representation, which has a coarse resolution, is up-scaled using tricubic interpolation, which is used to learn a deep representation by passing to a convolutional autoencoder. As a result, the latent representation of pose sequences is fed into an LSTM (Long Short-Term Memory) network to estimate human body pose in 3D and ensure temporal consistency.

Pavlakos *et al.* [2019] proposes a method for human body mesh estimation that leverages textural consistency in a multi-camera view video. The method, called TexturePose, uses a network that estimates both the shape and 3D pose. The model leverages multi-view sequential appearance consistency and provides evidence that this approach is effective for model-based pose estimation. This concept revolves around the idea that the alterations in the texture appearance of subjects in images are not drastic between consecutive frames. This means that consistency in texture can be used as additional information cue for reconstructing a body model. Schwarcz and Pollard [2018] presents a method for estimating 3D human pose from multi-view 2D poses. The proposed method uses a convolutional neural network (CNN) to predict 2D pose from each view, and then combines the 2D poses in a network to reconstruct the 3D pose. The model is trained using a combination of supervised and unsupervised learning techniques, including a novel self-supervised loss function that encourages the network to learn temporally consistent 3D pose across multiple views. Shuai *et al.* [2022] create a unified Multi-view Temporal Fusing Transformer (MTF-Transformer) framework that is capable of adaptive handling of varying view cameras and video input lengths. The framework uses a transformer-based architecture that can effectively fuse information from multiple views and temporal frames. It also includes an adaptive attention mechanism that can dynamically adjust the importance of different views and frames based on their relevance to the current pose estimation task.

These methods prove a good research area for taking advantage of additional data provided in multi-view datasets to create models that are able to run inference on monocular inputs. Typically, multi-view 3D HPE that takes advantage of temporal information provides robust pose predictions and proves an interesting area of research.

2.8 Datasets for 3D human pose estimation

Work in the HPE domain is highly dependent on available data sets and annotations. By providing annotated datasets of human poses, researchers can train and evaluate algorithms that can accurately estimate the 3D pose of a human from 2D images and videos. Datasets for 3D HPE are often, but not always, obtained using expensive calibrated synchronised cameras and Mo-Cap systems in constrained indoor lab environments [Ionescu *et al.* 2014; Sigal *et al.* 2010]. Due to this, diverse 3D HPE datasets are more difficult to obtain than 2D HPE datasets which can be annotated manually on any source of image. 3D pose estimation datasets, like the 2D datasets introduced in section 2.6, are important for benchmarking and comparing different algorithms. In this section, we will introduce and discuss most of the prominent datasets available for public research. Generally, the HumanEva dataset and the Human3.6M [Sigal *et al.* 2010; Ionescu *et al.* 2014] datasets are current salient standard benchmarks. Referring back to the datasets introduced for 2D pose estimation in section 2.6, the error metrics used for training and benchmarking 3D datasets are mostly extensions of the types of distance and thresholding metrics used in 2D benchmarking from a 2D space to a 3D one.

2.8.1 Human3.6M

The Human3.6M (H36M) dataset is currently the largest and most prevalent dataset for 3D HPE [Ionescu *et al.* 2014]. The dataset comprises at least 3.6 million image sequences and their corresponding hardware-collected 3D human pose annotations. The H36M dataset is captured in a laboratory environment using Mo-Cap on 11 subjects, consisting of 5 female and 6 male persons. The images are captured in the format of video sequences and are classified into 17 distinct activities. This dataset is multi-view, captured using 4 calibrated synchronised cameras at $50Hz$. The annotations are pixel-level, split into 24 body part labels. Options such as pre-made person bounding boxes and background subtraction are included.

This is the sole dataset for the experiments proposed in this document. Properties of the data provided in Human3.6M are a good fit due to the size and diversity of the activities provided. The dataset is in video format, meaning that temporal methods can be leveraged effectively. The dataset is multi-view, enabling the use of geometric self-supervision. We have dedicated a section in our Methodology chapter 4.4 to the dataset due to its importance in our experimentation and benchmarking, where we delve deeper into the specifics of the dataset.

The Human3.6M dataset is a larger dataset having the advantage of more diverse sets of actions and subjects, which can improve the generalisability of 3D pose estimation algorithms. However, the dataset is more complex and requires more computational resources to process, compared to a dataset such as the HumanEva dataset, introduced in the next subsection.

2.8.2 HumanEva-I&II

The HumanEva dataset has two versions offered by the Max Planck Institute for Intelligent Systems, HumanEva-I and HumanEva-II [Sigal *et al.* 2010]. Similarly to the Human3.6M dataset, a lab environment with hardware setup is used to capture the data. Synchronised multi-view camera setups with motion capture hardware is used to obtain video and 3D ground truth data. There is $40,000$ frames of synchronised motion capture available in this dataset, collected at a camera frequency of $60Hz$ with an additional $37,000$ time instances of pure motion capture data. This dataset has 4 subjects performing 6 types of annotated actions, across 7 calibrated videos (3 colour and 4 grey-scale) that are synchronised with bodies in 3D. The HumanEva-I dataset is more commonly used.

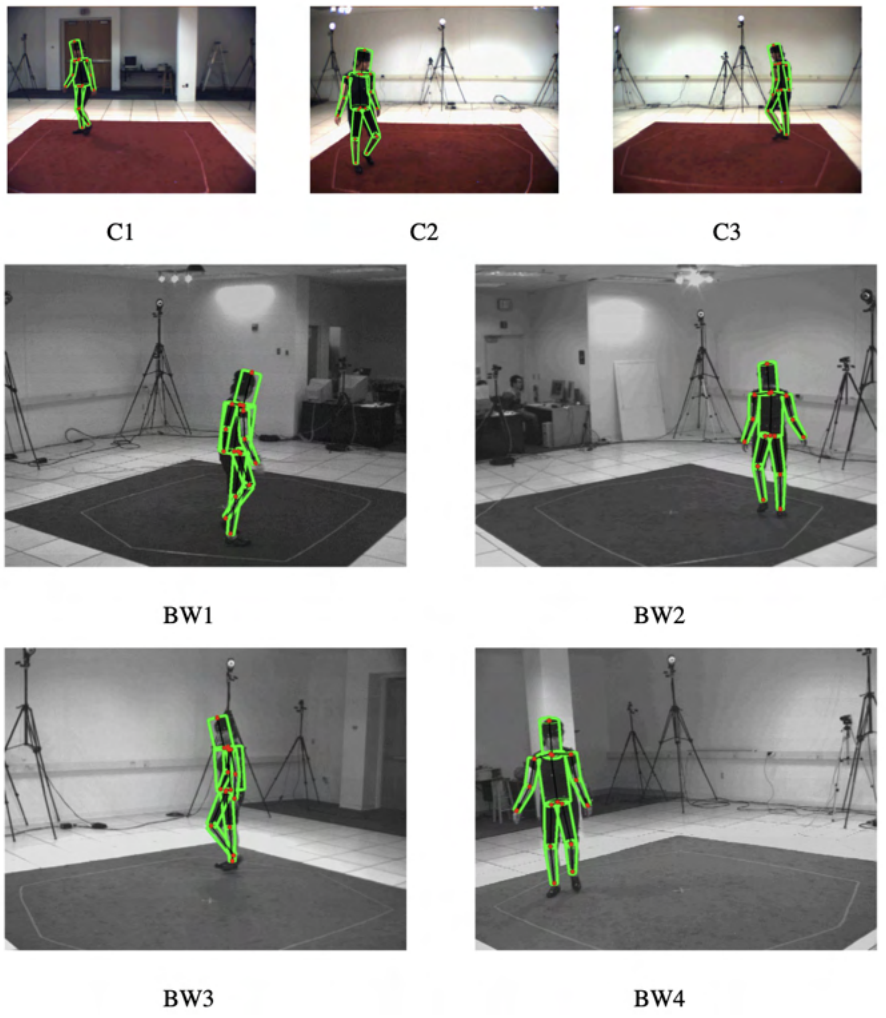


Figure 2.13: Example images and some annotations found in the HumanEva dataset [Sigal *et al.* 2010]

Similarly to other datasets introduced by the background of this dissertation, this dataset has its own suite of error measures defined for benchmarking models. The HumanEva dataset uses an Euclidean distance-based error metric based on 15 joints

markers as follows:

$$D(x, \hat{x}, \hat{\Delta}) = \frac{1}{\Sigma_{j=1}^{M} \hat{\delta}_j} \sum_{i=1}^{M} ||\hat{\delta}_i m_i(x) - m_i(\hat{x})||, \tag{2.10}$$

x represents pose, M is a set of 15 virtual markers $\{m_i(x)\}$, where $i = 1...M$. $m_i(x) \in \mathbb{R}^3$. When pose is in 2D space the markers are in $\mathbb{R}^2$. The error between estimated pose $\hat{x}$ and ground-truth pose x is expressed as the average Euclidean distance between individual virtual markers. The binary selection variable Δ is added to accommodate researchers wanting to use different numbers of parts or joints in their models [Sigal *et al.* 2010]. $\hat{\Delta}$ is a set of binary selection variables (where $\hat{\delta}_i = 1$ if the algorithm is able to estimate marker i, 0 otherwise) for each pre-defined marker in the dataset available for estimation, where $\hat{\Delta} = \{\hat{\delta}_1, \hat{\delta}_2, ..., \hat{\delta}_M\}$.

For a sequence of T frames, HumanEva datasets offers an average performance metric defined as the following:

$$\mu_{seq} = \frac{1}{T} \sum_{t=1}^{T} D(x, \hat{x}, \hat{\Delta}). \tag{2.11}$$

The advantage of the HumanEva dataset is that it provides accurate ground-truth annotations, which can be valuable for assessing the performance of 3D pose estimation algorithms. However, the dataset is relatively small and highlights the difficulties of collecting a dataset and creating benchmarks for the 3D pose estimation problem space. is limited in terms of the number of actions and subjects, which may limit its generaliseability to other scenarios.

2.8.3 MPI-INF-3DHP

Mehta *et al.* [2017a] introduces the MPI-INF-3DHP dataset. The dataset consists of over 1.3 million frames captured from video footage with corresponding Mo-Cap pose annotations from 8 actors, 4 male and 4 female, performing 8 activities. The dataset is multi-view, with annotations captured from a total of 14 cameras. The dataset is captured with a commercial motion capture setup that does not need special mo-cap markers or suits, allowing the persons to wear everyday apparel. The MPI-INF-3DHP dataset is captured in a lab environment with green screens to allow automatic image segmentation and augmentation of different types of textures and variable images to the background. The main part of the dataset allowing for 3D HPE is the subset of $500,000$ images captured using a 5-camera multi-view setup at around chest height. Both 3D annotations and skeletal annotations are made available in this dataset. Some visual examples of the data are included in figure 2.14.

Although the dataset has more clothing variation than other datasets due to the data capture and augmentation techniques, the appearance variation is not comparable to true in-the-wild datasets [Mehta *et al.* 2017a]. Figure 2.14 shows a visual example of the augmentation techniques used to simulate true environmental variability found in-the-wild.

Figure 2.14: Example data in the MPI-INF-3DHP dataset, showing actors in the green screen studio environment (far left), with segmentation masks for different regions (center left) and augmentation on the footage by compositing different textures to the different picture regions, independently (right) [Mehta *et al.* 2017a].

The MPI-INF-3DHP dataset uses the standard 3D MPJPE metric, along with some specialised alternate metrics for the dataset. Mehta *et al.* [2017a] suggest a 3D extension of the Percentage Correct Keypoints (PCK) metric from 2D PCK metrics such as those found in equation (2.8) [Andriluka *et al.* 2014a]. A 3D extension of the AUC metric [Lin *et al.* 2014] similar to that of equation (2.3) is also computed for the ranges of PCK thresholds.

The MPI-INF-3DHP dataset contributes in-the-wild images for the 3D HPE problem, which could be useful for models with better generalisability in in-the-wild applications. The dataset provides greater diversity in poses, human appearances, clothing, occlusion and viewpoints.

2.9 Conclusion

This chapter serves to introduce and show the research landscape and direction around pose estimation. We have explored various methods and techniques for estimating human pose from sensor and image data. We discuss different human body models and pose representations, including skeleton-based, contour-based, and volume-based representations. We examined common sensor and image types used for pose estimation, as well as non-deep learning based approaches such as model-based and data-based methods.

We offer a method-based taxonomy into deep 2D human pose estimation in order to show the background that led to 3D HPE. We introduce both single-person and multi-person pipelines, as well as direct regression and heatmap-based methods. We also compared top-down and bottom-up approaches for multi-person pipelines. We concluded this section by discussing popular datasets for 2D pose estimation, including COCO, MPII, Leeds Sports Pose, Frames Labelled in Cinema, AI Challenger, and PoseTrack.

We also do a method-based taxonomy on 3D human pose estimation, including single-person and multi-person methods, as well as single-frame and multi-frame meth-

ods. We discussed monocular and multi-view approaches for single-frame 3D pose estimation, as well as sequence-based methods for multi-frame 3D pose estimation. Finally, we highlighted popular datasets for 3D pose estimation, including Human3.6M, HumanEva, and MPI-INF-3DHP. Introducing datasets offers insight into the benchmarking and metrics of the problem area.

This chapter has provided a comprehensive overview of the state-of-the-art methods and datasets for human pose estimation. By understanding the strengths and limitations of different approaches, we show the context and the direction that we take in our research area and methods. In the next chapter, we will take a closer look at the related works and methods to the research areas that we have selected to study.

Chapter 3

Related Work

3.1 Introduction

In this chapter we take a look at the related work to our research in greater detail. This will help place the spaces that our research occurs in, the topics and methods that are most relevant to our work. We will take a deeper look into some self-supervised, weakly-supervised and unsupervised training methods in 3D HPE, with a particular focus on geometric self-supervision. We will have a look at other salient works on gathering temporal information from 3D human pose data.

In our background chapter 2, we identify that 3D human pose estimation datasets are expensive and confined to laboratory environments for collection. This is mostly due to the fact that with our current technologies, researchers are only able to collect annotated pose data using specialised equipment, which most often may only be used in limited spaces in laboratories. Part of our research is attempting to prove the validity of being able to generate self-supervision pose annotations using the rich diversity of 2D pose datasets, so we pay particular attention to other supervision paradigms enabling the use of less annotation-intensive data. We will study some of the current proven work in the field that we may experiment with.

3.2 Weak supervision paradigms in 3D human pose estimation

Using other supervision paradigms including weakly supervised, self-supervised, and unsupervised methods in 3D HPE has some advantages over traditional supervised methods. Weak supervision requires less labelled data, which is generally expensive and time consuming to obtain [Ionescu *et al.* 2014; Sigal *et al.* 2010; Mehta *et al.* 2017a]. Weak supervision techniques can leverage a variety of alternative sources, such as rules and heuristics, to generate pose annotation data. Weak supervision methods may also be able to improve generalisability of models to unseen data instances. As mentioned in the introduction, there is potential to research leveraging the superior variability and appearance diversity from 2D pose datasets for 3D human pose estimation.

3.2.1 Weakly-supervised methods

A model called the Classification Regression Network (LCR-Net) uses three modules to perform weakly-supervised 3D pose detection [Rogez *et al.* 2017]. A pose proposal generator is created by the first module, followed by a classifier and regressor modules. On the image input, the model first extracts candidate regions and obtains pose proposals in the localisation module. The proposals are then scored by a classifier and regressed into a pose through a regressor. The outcome contains both 2D and 3D poses per image, collected by aggregating similar pose proposals, in terms of localisation and 3D pose. Pseudo ground-truth 3D poses are inferred from 2D pose annotations using a nearest neighbor search performed on the annotated joints.

The authors create an extension of the network in updated works in Rogez *et al.* [2019], where they create LCR-Net++ by improving the performance of the method in some new ways. They augment their weakly-supervised data by generating rendered image data by using a SMPL 3D model, along with using a ResNet backbone [He *et al.* 2016] which proves an effective model backbone for pose estimation. Further, Rogez *et al.* [2019] use an improved alignment scheme and an iterative process to refine the classification and regression results during training.

A two-stage approach is studied by Dabral *et al.* [2019] that uses Regions of Interests (RoI) to estimate keypoints in 2D, then performs lifting process on the joints into 3D space. They adopt an augmented Mask-RCNN with a shallow hourglass network to output 2D heatmaps, which are then passed to a 3D pose module to regress the root-relative 3D joint coordinates in the second stage. This method proves to not require any multi-person datasets for training, and performs inference on multi-person subjects in-the-wild. This proves the weakly-supervised class of models are able to avoid the costly 3D annotations for multi-person in-the-wild setting data.

3.2.2 Unsupervised methods

Many unsupervised methods in 3D HPE uses geometry-aware body representations or some form of body prior to model human bodies for unsupervised learning. Rhodin *et al.* [2018a] for example, develop a representations that are geometry-aware for the human body to facilitate unsupervised learning (figure 3.1). This method leverages multi-view datasets to learn a latent representation that captures the 3D geometry of the human body. The process for learning the latent representation does not rely on any 2D or 3D labels and annotations. An encoder-decoder is used to predict an image seen in one viewpoint from another viewpoint of a synchronised camera. Rhodin *et al.* [2018a] prove in their paper that their latent representation is able to adequately capture 3D geometry so that the mapping to 3D pose is simpler and may be learned using few examples compared to existing solutions that rely on multi-view supervision.

One of the recent prominent unsupervised methods for 3D pose estimation is Kundu *et al.* [2020]. The authors note that though weakly-supervised models can improve model generalisability compared to supervised methods, performance of weakly-supervised models rely on availability of supervision on related tasks, such as multi-view camera setups or 2D pose. Kundu *et al.* [2020] propose a novel kinematic-structure-preserved unsupervised 3D pose estimation framework. The model makes use of body-structure

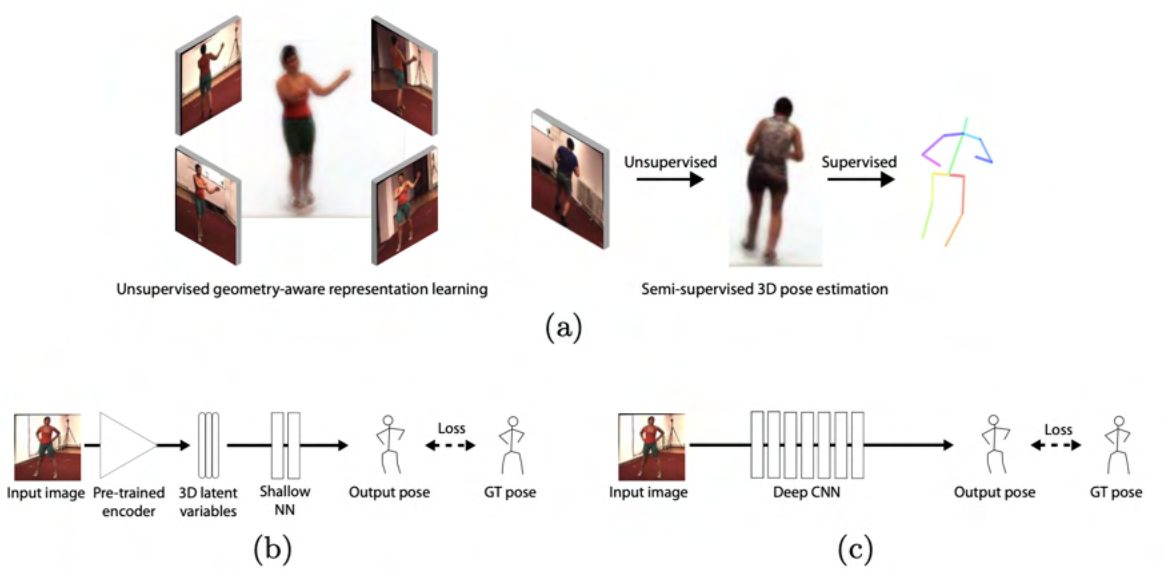


Figure 3.1: Geometry-aware pose representation for unsupervised pose learning [Rhodin *et al.* 2018b]

priors defining a 3D kinematic structure of a human body. Bone-length ratios and skeletal joint connectivity information is encoded in a fixed canonical scale. The model uses three differentiable forward-kinematics transformations, camera projection and spatial-map transformations. This acts as a suitable bottleneck simulating effective pose disentanglement, and provides latent pose representations. This is shown to be interpretable by Kundu *et al.* [2020] and they avoid training an explicit mapper from latent embedding to pose. An energy-based loss is designed by Kundu *et al.* [2020] to enable learning from in-the-wild images. The authors prove the model performance and flexibility on both Human3.6M and the MPI-INF-3DHP datasets.

3.2.3 Self-supervised methods

It is interesting to note that many self-supervised methods in contemporary 3D HPE research space often rely on off-the-shelf 2D pose estimators, which (referring back to the background chapter 2) tend to be fairly robust. Many other methods involve leveraging the multi-view property of the large 3D pose datasets introduced in section 2.8, such as geometry-based methods where the pose is identified using a pre-trained off-the-shelf 2D pose estimator, then using some form of multi-view triangulation to obtain a 3D pose from data to be used as the pseudo-ground truth. We focus a bit on more on this area since leveraging the multi-view nature of many datasets to infer on monocular frames is an efficient way to train models.

In order to leverage the previous success of Convolutional Networks (ConvNets) without the intensive need of ConvNets for labelled training data, Pavlakos *et al.* [2017b] propose a self-supervised model that represents an early work that makes use of a geometry-driven approach.

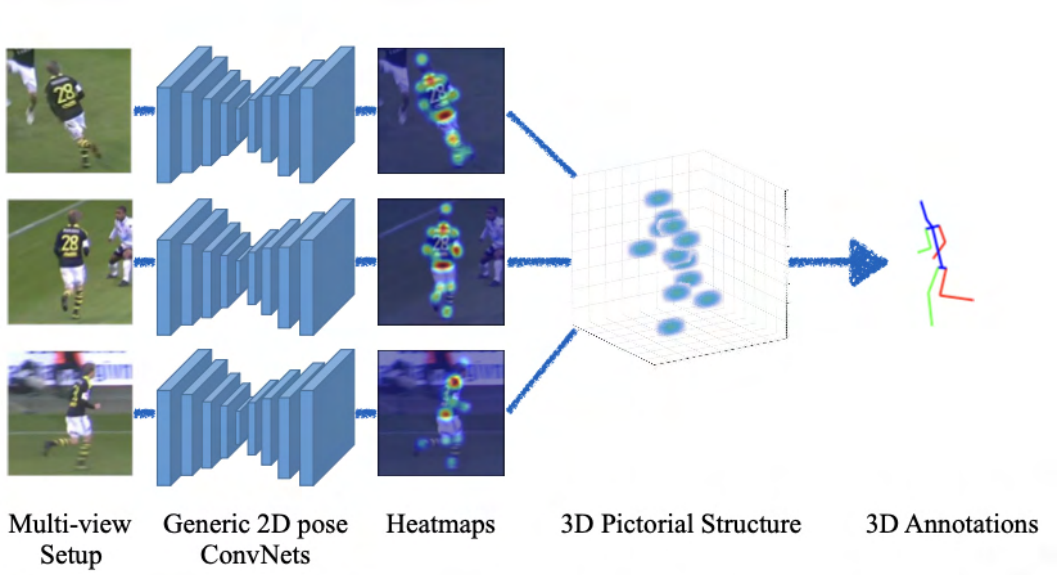


Figure 3.2: Heatmap-based architecture used by the pictorial structures model [Pavlakos *et al.* 2017b]

A pictorial structures model is used by Pavlakos *et al.* [2017b] (shown in figure 3.2). This is a graphical model that represents a pose as a collection of parts, each of which has a specific appearance and spatial relationship to the other parts. For inference, the model estimates the pose of the object in a new image by searching for the configuration of parts that best matches the observed image features. This is typically done using an iterative optimization algorithm, such as belief propagation or gradient descent.

Kocabas *et al.* [2019], as introduced in 2.7.3, uses epipolar geometry to obtain a pseudo-annotation in 3D for poses. Epipolar geometry is based on the geometries involved in stereo vision. Given a pair of 2D cameras, each camera views a 3D scene from two distinct positions. There are fundamental geometric relations between the 3D points and their projections onto the 2D camera images that lead to constraints between the image points.

Along with using self-supervision in the form of epipolar geometry, the authors also propose a new metric to be used along with MPJPE and PCK. This metric is the Pose Structure Score (PSS), and is developed to address the issue of MPJPE (3D) not accounting for structural issues in an estimated pose. PSS is not used as a loss function, only a performance metric/score that computes a scale-invariant index for the structural plausibility of a pose with respect to the pose annotations. PSS is specifically computed by first modelling the natural distribution of ground-truth poses, p being a predicted pose for an image where the corresponding annotation is q. Unsupervised clustering is used on pose distributions, and cluster centers are found closest to p and q. If both are assigned to the same cluster center, the PSS for the joint is 1, otherwise 0. More specifically, PSS is described as such:

$$PSS(p,q) = \delta(C(p), C(q))$$
$$\text{where } C(p) = \arg\min_k ||p - \mu_k||_2^2, \text{ and } \delta(i,j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{3.1}$$

In the equation 3.1 above, given a ground-truth set composed of n poses q_i, $i \in \{1, ..., n\}$, each pose vector is normalised as $\hat{q}_i = \frac{q_i}{||q_i||}$. Then k cluster centres $\mu_j, j \in \{1, ..., k\}$ are computed using k-means clustering.

The authors then use PSS for each pose to calculate a mean Pose Structure Score (mPSS), over a set of poses averaged over their individual scores in equation 3.1.

The model architecture used in EpipolarPose is based on the model developed by Sun *et al.* [2018] mentioned already in 2.7.3. This itself is a heatmap-based method using a ResNet [He *et al.* 2016] backbone.

3.3 Temporal methods

In this section we will study some of the salient methods involved in temporal methods. As shown in chapter 2, temporal methods leveraging video-format data proves to be effective in allowing models to be able to deal with vision problems and challenges such as occlusion, self-occlusion, and jittery predictions. We will take a closer look at some work representing progress using LSTMs, Optical Flow, Temporal Convolutional Networks and attention mechanisms.

3.3.1 Optical flow

Optical flow in computer vision is a representation of the apparent motion or pattern of movement of objects of interest. By tracking the apparent velocities of pixel movement in a sequence of images, temporal information for the movement of more abstract objects in an image is extracted. The main idea behind optical flow used in human pose tracking is that the change in optical flow between consecutive frames largely depends on the change in the human pose behind them. However, optical flow requires data pre-processing which can be computationally expensive to obtain. There have been many methods to gather optical flow information without needing to incur the pre-processing step because of this problem.

For two consecutive video frames, optical flow is a vector field that matches a point in the first frame to the same displaced point in the second frame. Even though it is calculated in the image plane, optical flow contains 3D information as it can be interpreted as the projection of a 3D scene flow. The observed optical flow is caused by relative movement between the object or person and the camera observer.

The work by Alldieck *et al.* [2017] exploits optical flow properties in a sequence of images to enforce temporal coherence and resolve pose ambiguities of either joint-based or silhouette-based approaches. An energy minimisation system is used to infer model parameters that minimise the distance between observed flow between two input frames. This work exploits the property that optical flow is able to extract rigid boundaries in images and applies the inferences to a Skinned Multi-Person Linear Model

(SMPL), which is a volume-based human pose representation. A common weakness of Optical flow methods is that large occlusions and reappearances of previously-occluded body parts can lead to large tracking errors.

The work by Li *et al.* [2020b] proposes the use of a novel temporal consistency exploration (TCE) module, using feature maps of neighbouring frames as input. Enhanced feature maps are obtained using the input feature maps, which are then fed into a fully convolutional network (FCN) to predict precise joint heatmaps.

3.3.2 Long short-term memory networks

Long short-term memory (LSTM) networks are a type of recurrent neural network (RNN) that is able to track long-term dependencies in sequential data. LSTMs are a significant advancement for models to be able to learn longer-term insights and dependencies in temporal data. Like many other temporal methods, it was able to be applied and experimented with in the 3D HPE problem space due to the availability of video data in large and prominent 3D human pose datasets (such as introduced in section 2.8).

Lee *et al.* [2018] present a method using a modified LSTM called propagating LSTM networks (p-LSTM). Lee *et al.* [2018] focus on using joint interdependency (JI). JI incorporates body part-based structural connectivity of joints to learn a spatial correlation of human poses. A series of LSTM are connected sequentially to reconstruct 3D depth from a centroid to edge joints through learning some intrinsic JI. Figure 3.3 shows the general idea and network structure proposed [Lee *et al.* 2018].

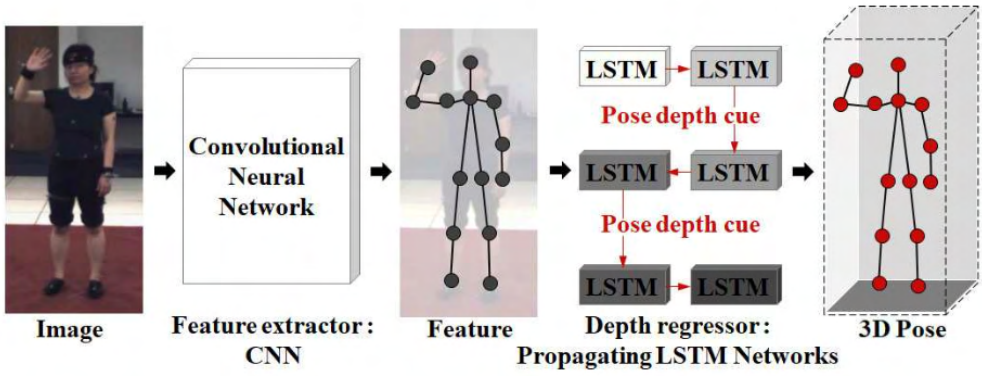


Figure 3.3: Network architecture of p-LSTM [Lee *et al.* 2018]

3.3.3 Attention mechanisms

Research in RNNs eventually led to the predecessor for the relatively new attention mechanisms in machine learning [Hafiz *et al.* 2021]. Generally, attention mechanisms intuitively simulate cognitive attention. Tang *et al.* [2023] is a recent method that proposes a transformer-based solution for the task of 3D human pose estimation. Attention mechanisms used in transformer networks have been proven successful in many tasks,

including 3D human pose estimation. In their paper, Tang *et al.* [2023] propose a novel spatio-temporal criss-cross attention (STC) block. Figure 3.4 shows the general model architecture.

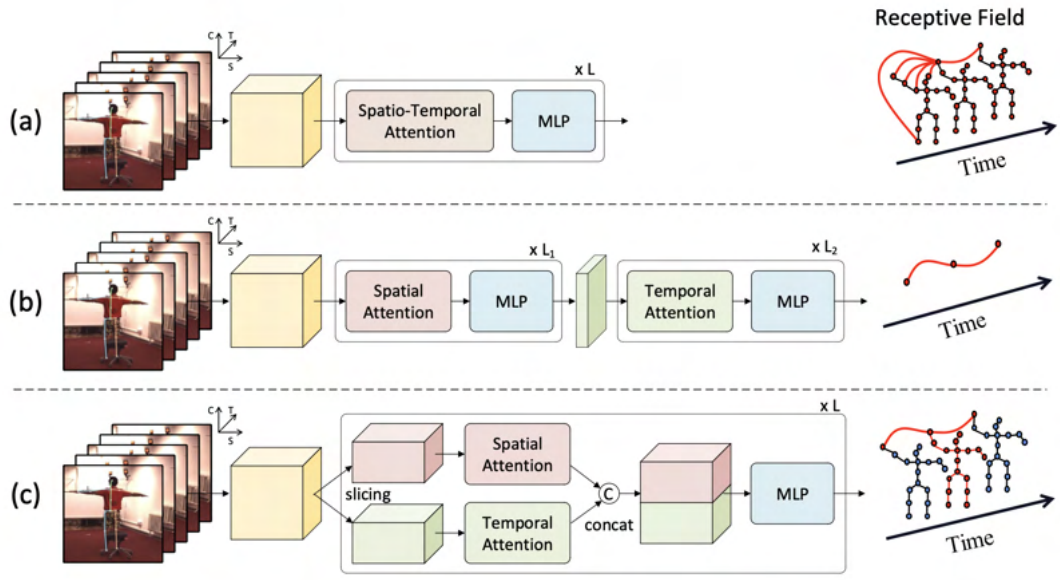


Figure 3.4: Network architecture of STCFormer [Tang *et al.* 2023]

Despite the usefulness of joint-to-joint affinity matrices in pose estimation, its computational cost increases quadratically with the number of joints. This drawback is particularly problematic for video sequences, which require spatio-temporal correlation across the entire video. The STC block is able to decompose correlation learning into space and time. Specifically, STC partitions its input feature evenly along the channel dimension and applies spatial and temporal attention to each partition. By concatenating the outputs from attention layers, STC models the interactions between joints in an identical frame and joints in an identical trajectory simultaneously. Tang *et al.* [2023] then introduce STCFormer, which stacks multiple STC blocks and integrates a new Structure-enhanced Positional Embedding (SPE) that takes the structure of the human body into consideration. The SPE consists of two components: spatio-temporal convolution around neighboring joints to capture local structure, and part-aware embedding to indicate which part each joint belongs to. They conduct extensive experiments on Human3.6M and MPI-INF-3DHP benchmarks, and report state-of-the-art results.

3.4 Conclusion

This chapter provides a more detailed look at some of the methods close to our research area and methodology. There is much potential and research into the various supervision paradigms outside of fully-supervised learning, in order to leverage the fact that the most prominent 3D human pose estimation datasets are multi-view and contain sequential image formats. The motivations are to make sure that the models

created for this problem space are less reliant on large amounts of expensive and constrained labelled data. Traditional fully-supervised methods require a large amount of labelled data to train models, which can be expensive and time-consuming to obtain. Other supervision paradigms, on the other hand, are able to use less or unlabelled data by defining a proxy task that does not require explicit annotations. By leveraging the inherent structure and patterns in the data, self-supervised methods can learn useful representations that can be transferred to downstream tasks such as 3D human pose estimation. This approach has the potential to significantly reduce the amount of labelled data required for training, making it more accessible and cost-effective for researchers, practitioners and development of applications.

Likewise, temporal methods are well studied in the problem space to leverage temporal information present in video sequences to improve the accuracy of pose estimation. By considering the motion of the human body over time, temporal methods can better capture the dynamics and variability of human movement, which is particularly important for tasks such as action recognition and motion tracking. Additionally, temporal methods can help address the issue of occlusion, where certain parts of the body may be obscured in a single frame but can be inferred from the motion in adjacent frames. Furthermore, temporal methods can help to improve the robustness of pose estimation in challenging scenarios such as low-light conditions or noisy environments.

Chapter 4

Experiment Methodology

4.1 Introduction

In this chapter, we will focus on the target area of our research and outline methodologies and experiments to explore this area further.

In the previous chapters, we have established the significance of improving human pose estimation through other supervision paradigms and temporal methods. We discuss the intensive resources required to collect annotated and multi-view 3D human pose datasets. Often the constrained lab environments means extensive camera and equipment setup, which leads to less variability in subject, pose, lighting, clothing, appearance and environment. Due to this, we want to experiment with proving the concept of leveraging self-supervised learning to generate pseudo ground truth annotations. To this end, we are inspired by Kocabas *et al.* [2019] and other geometric self-supervised methods that use geometric rules to generate annotations for supervising methods. This way, we are able to not only measure how well a model performs using geometric self-supervision but also how close the generated labels are compared to the dataset ground-truth.

Another area of research interest is that of temporal modelling and methods. Existing large and established 3D pose estimation datasets and benchmarks conveniently offer sequential data in the form of video. We see in the background (chapter 2) and related works (chapter 3) chapters that temporal methods are effective at overcoming some issues present with single-frame estimation methods such as jittery inference and difficulties handling occlusion.

By presenting our hypotheses and methodologies around combining the two research areas of self-supervision and temporal methods, we aim to gain a deeper understanding of the effectiveness of self-supervised and temporal methods in improving human pose estimation accuracy. This knowledge can inform the development of more robust and accurate models that are more generalisable and less dependent on large-scale, fully-labelled datasets.

4.2 Hypothesis

Recent geometric self-supervision methods are able to leverage spatial relations to generate pseudo-label annotations. These can then be used to train models to estimate 3D human poses using 2D image information. It requires multi-view data and provides a simple, yet computationally effective method to obtain pseudo ground-truth poses.

Temporal methods have been applied with success in many deep learning problems where there is time-series data available and modelling inter-dependencies and relations over time. In the 3D HPE space where video data is available tracking poses of people, it has shown to be able to improve model efficiencies and track pose over time. It is also used to tackle issues with occlusion, self-occlusion and complex poses.

By experimenting with and integrating temporal methods with geometry-based self-supervised pose estimation, we hope to see some improvements to accuracy and better model response to common pose estimation problems such as occluded body parts.

Hypothesis: Using temporal methods to enhance self-supervised geometric 3D human pose estimation will improve the accuracy and temporal consistency of model predictions on unseen monocular inference data.

4.3 Research questions

The following research questions have been established using the above hypothesis to inform all subsequent methodologies, experiments, and analysis:

- How will the use of self-supervision in the form of epipolar geometry affect the accuracy of a baseline fully-supervised model? Will the self-supervised results be robust enough to justify further research?
- To what extent does the use of a temporal convolutional network (TCN) improve or affect the accuracy of the model?
- How will a TCN perform when experimenting with temporal ranges with dilated TCNS?
- Can we implement some form of Kalman filter post-processing method for improved temporal accuracy?
- How much better do models that use temporal convolutions and Kalman filters perform compared to a baseline model that only uses naïve temporal information?
- Can combining the different methods result in a model that can potentially improve efficiency of the model and act as a proof of concept for further research into using these methods to alleviate the reliance on expensive training data?
- Can the model achieve competitive results when inferring only on monocular data, despite being trained on multi-view data?

4.4 Dataset: Human3.6M

As shown in chapter 2, datasets are paramount in defining benchmarks, metrics, loss functions, and constraints for training in the general HPE problem space. They offer different settings, variability, and activity types for pose estimation. While there are some publicly available datasets for 2D pose estimation research, such as the COCO dataset and the MPII Human Pose dataset [Lin *et al.* 2014; Andriluka *et al.* 2014a], 3D pose estimation datasets as a whole may be prohibitively difficult and expensive for individuals and small researchers to establish and obtain.

To address this challenge, we have selected the Human3.6M dataset by Ionescu *et al.* [2014] for use in our research. This dataset is one of the largest motion-capture 3D pose estimation datasets, consisting of over 3.6 million 3D human poses. It was collected by recording performances of 5 female and 6 male subjects in a lab environment using 4 synchronised cameras on different viewpoints. The purpose of the dataset is to train realistic human sensing systems and evaluate HPE models and algorithms. It includes a diverse set of motions and activity annotations, such as walking, posing, greeting, and eating, as well as time-of-flight (depth) data and person bounding boxes.

In this section, we will discuss the Human3.6M dataset in depth, including how we set up our experiments and methodologies to leverage its multi-view pose information and temporal data to test our hypothesis (section 4.2) and research questions (section 4.3).

4.4.1 Overview

As of the writing of this document, the Human3.6M dataset remains the largest 3D HPE dataset by volume of frames. Created by Ionescu *et al.* [2014], the dataset was several orders of magnitude larger than the former state-of-the-art at the time of its release. Sampling the human kinematic space regularly and densely is very challenging due to its large size. To address this challenge, Ionescu *et al.* [2014] chose to collect data focused on a set of poses that are likely to be of interest in an urban and office scene. The dataset offers 17 distinct categories of actions performed by professional actors with the direction that they should behave in a variety of ways. For example, the subjects will perform walking actions with a hand in their pocket or carrying a backpack. It is very unlikely that people are found in 'perfect' walking poses that are highly regulated, and the Human3.6M dataset tries to account for this by allowing their actors ample room for improvisation. The 17 scenarios are: directions, discussion, eating, activities while seated, greeting, taking photo, posing, making purchases, smoking, waiting, walking, sitting on chair, talking on the phone, walking dog, walking together.

Human3.6M contains additional synchronised image, human motion capture and time of flight depth data. Ionescu *et al.* [2014] also offer accurate 3D body scans of the subjects involved. Using the 3D body scans, controlled mixed reality evaluation scenarios are created where the body scans are used to create 3D human animated models inserted using 3D geometry into complex real environments. This mixed reality data is also able to be viewed with a moving camera and with augmented occlusion. Some examples of the 3D scans can be viewed in figure 4.2.

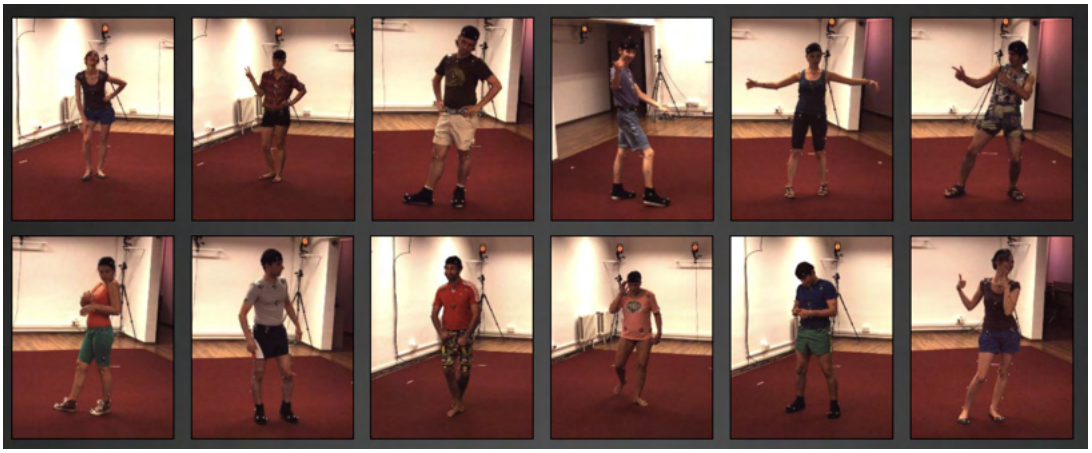


Figure 4.1: Subjects and pose examples found in Ionescu *et al.* [2014]

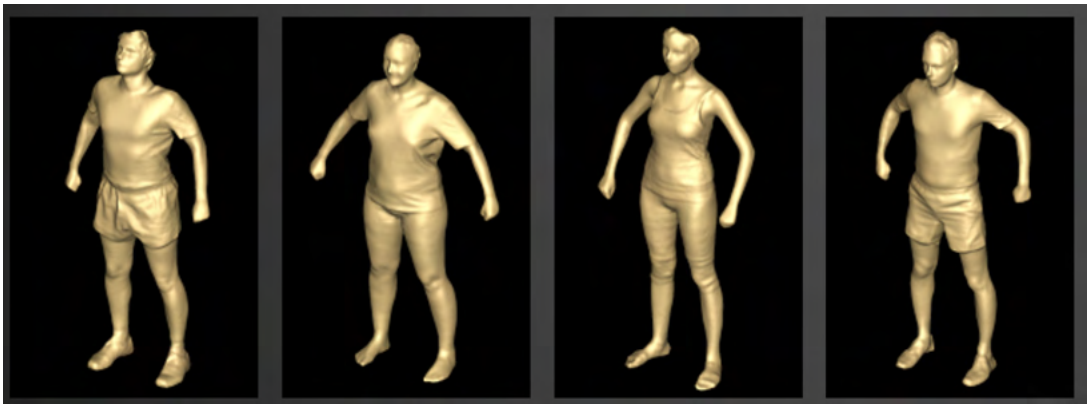


Figure 4.2: Subject 3D scan examples found in Ionescu *et al.* [2014]

The dataset is collected in a lab environment in the configuration shown in figure 4.3. Here we can see some of the potential model generalisability issues often encountered in 3D HPE. The entire dataset is captured within the confines of a 3 metre by 4 metre space, with very little variability in lighting and background.

For the Human3.6M dataset, pose is represented using a skeleton body model in section 2.2. The dataset uses a total of 32 skeletal body joints for all parameterisations. The joints are reduced to include only the relevant ones, for example leaving one joint to represent each hand or foot instead of several. Most often models will use 15 or 17 joints instead of all 32 included in a dataset. Common pose parameterisations are used, including relative 3D joint positions and kinematic representations. Human3.6M itself also provides 2D joint positions in case of methods wanting to use the 2D ground-truth annotations.

For relative 3D joint positions, joint positions in 3D space are provided. These are provided by a forward kinematics fitting procedure on the subject skeleton. Relative 3D joint positions are challenging to apply on a dataset since it is difficult to estimate the size of a person in an image. Limb length consistency is used for all frames that each subject appears in by applying the constraint of the same skeleton limb lengths. A root joint is defined for enforcing relative joint positions (roughly corresponding to the

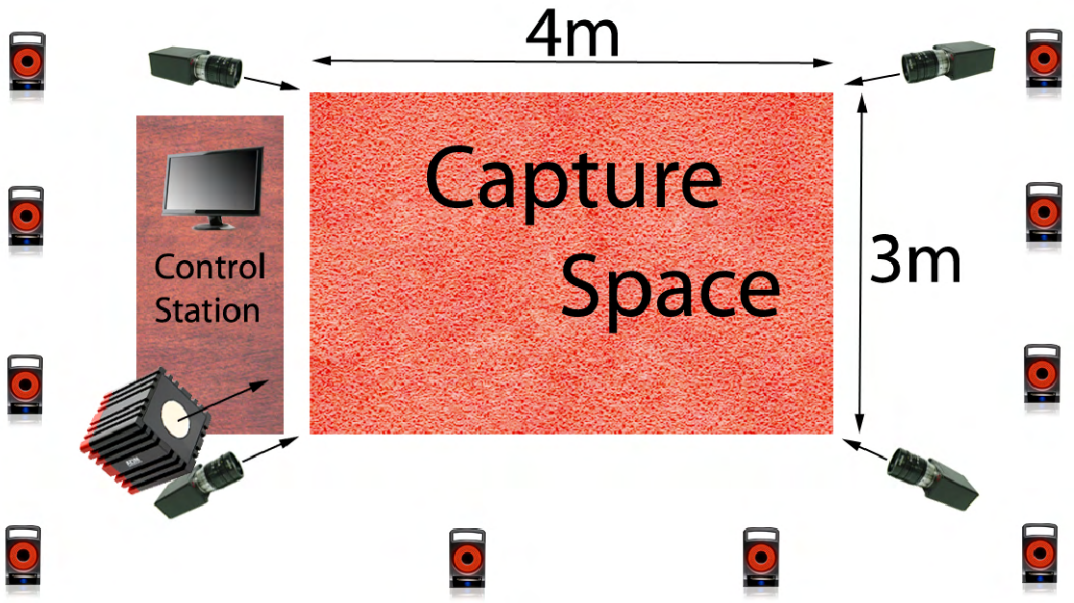


Figure 4.3: Lab space where data is collected for Human3.6M [Ionescu *et al.* 2014]

pelvis bone position). This is taken as the center of the prediction coordinate system and all the other joints are estimated relatively to it.

Kinematic representation in the Human3.6M dataset considers the relative joint angles between limbs. This is invariant to both scale and body proportions. The dependencies between variables are much more complex, thus making estimation more difficult. The process of obtaining joint angle values involves a complex constrained non-linear optimization process. Both the relative 3D joint positions and the kinematic representations can be used directly in monocular predictions or in multi-camera settings (which we are using).

The dataset is organised by subject, and there are pre-defined protocols for training and testing. We will detail some of this information as it is pertinent to our experiment setup. Following, we will look at the metrics and benchmarking available with the dataset.

4.4.2 Evaluation metrics

In the following we will detail and discuss the metrics tied to the benchmarks and evaluation protocols of the Human3.6M dataset [Ionescu *et al.* 2014]. Some of which are used in our research to train and evaluate our models. The most important of these is the 3D mean per-joint position error (MPJPE) which is not only a benchmark but also the primary loss function used to train our models. We also look at percentage correct keypoints (PCK).

Mean per joint position error

Mean per joint position error (MPJPE) in 3D space is measured in millimetres. For an image frame S and a skeleton system S consisting of i joints, MPJPE error E_{MPJPE} is computed as:

$$E_{MPJPE}(f,S) = \frac{1}{N_S}\sum_{i=1}^{N_S} ||m_{p,S}^{(f)}(i) - m_{gt,S}^{(f)}(i)||_2, \tag{4.1}$$

where $m_{p,S}^{(f)}(i)$ is a function that returns coordinates of the i_{th} joint of a skeleton S at frame f, from a pose estimator p. N_S is the total number of joints in the skeleton system.

Similarly, $m_{gt,S}^{(f)}(i)$ is the ground truth (gt) coordinates of joint i of skeleton S corresponding to frame i [Ionescu *et al.* 2014]. In summary, MPJPE calculates an Euclidean distance from the estimated joints to their ground truth position in millimetres, averaged over the number of joints present. If there are multiple frames present, the MPJPE is averaged over the frames.

It is important to note that for different protocols in the Human3.6M dataset, different data post-processing of the joints occurs before computing MPJPE. These are standard to each pose dataset and protocol and must be observed [Ionescu *et al.* 2014; Sigal *et al.* 2010; Mehta *et al.* 2017a].

An example of these modified metrics includes the normalised mean per joint position error (N-MPJPE) for protocol $\#1$ of Human3.6M where the MPJPE is calculated after aligning the depths of the root joints. The dataset also includes the Procrustes mean per joint position error (P-MPJPE), which applies a Procrustes analysis that aligns the predicted 3D pose with the ground truth before calculating the mean per joint position error. This alignment process involves translation rotation and scaling steps [Ionescu *et al.* 2014].

This metric is currently the most widely accepted error metric in 3D HPE and is proposed as the primary metric and loss function for the experiments proposed in this document.

Mean per joint angle error

Mean per-joint angle error (MPJAE) is similar to MPJPE but applied to joint angles. This represents a a different approach to compare poses using the angles between joints of the skeleton. These angles are computed in 3D and is defined as the following:

$$E_{MPJAE}(f,S) = \frac{1}{3N_S}\sum_{i=1}^{3N_S} |m_{p,S}^{(f)}(i) - m_{gt,S}^{(f)}(i)| \bmod \pm 180, \tag{4.2}$$

Percentage correct keypoints

Percentage of Correct Keypoints (PCK) is introduced by Mehta *et al.* [2017a] for 3D pose estimation similar to the PCK metric in 2D pose estimation used in the 2D MPII Human Pose Dataset (MPII) [Yang and Ramanan 2012]. PCK in the 3D domain is also

called 3DPCK. PCK counts the percentage of points that fall in a threshold distance of the ground truth joint. The threshold is often determined by the dataset and instance of use.

A candidate keypoint is defined to be correctly labelled if it falls within $\alpha \cdot max(h, w)$ pixels of the ground truth annotation, where h and w are the height and width of the person bounding box, respectively. The parameter α controls the relative threshold for considering correctness. Common values for α are $\alpha = 0.1$, $\alpha = 0.2$ and should be clearly notated.

4.5 Methodology

The main goal of our research is to compare the accuracy of geometric self-supervision to a fully-supervised model, and improve the accuracy of geometric self-supervision in the 3D human pose estimation task, by experimenting with various prominent temporal methods. In the interest of this, different methodologies have been identified and detailed below.

The results for each methodology is presented and summarised in chapter 5.

For evaluation of the final results, the primary metric being analysed is the average MPJPE error across all protocols and activities in the baseline model, over multiple runs. For each experiment, random seeds for each run and different holdouts of data will be used to ensure that the results are statistically significant.

4.5.1 Epipolar geometric self-supervision

Geometric self-supervision in the 3D HPE problem is reliant on multi-view datasets. EpiPolar Pose, presented by Kocabas *et al.* [2019], is the most prominent self-supervision method for 3D HPE to date. Many current state-of-the-art models use it as a backbone. The model requires multi-view data during training only and runs inferences on monocular images. The model requires at least 2 camera views to train and relies on training using 3D data obtained from triangulating two or more 2D poses from different simultaneous viewing positions and angles.

The EpiPolar Pose model (as shown in figure 4.4) takes in at least a pair of images (I_i, I_{i+1}) simultaneously taken by two separate cameras or views. The figure shows only two camera inputs for simplicity's sake, but more than two views can be used using the same principle. The 3D pose V, generated using the 2D pose estimator after triangulation T (epipolar geometry), is used as a training signal for the CNN in the 3D pose estimator. The blue background in the upper half shows the inference pipeline, which is monocular and takes in a single image I_i. L represents $L1$ loss.

The CNNs in the 2D and 3D estimation branches in the architecture shown in figure 4.4 can be pre-trained using the MPII dataset [Yang and Ramanan 2012]. During training, the 2D pose estimation network needs to be kept frozen to avoid a degenerate solution where the keypoints collapse on a single location. This is done due to many multi-view models having the same problem [Rhodin *et al.* 2018b; Suwajanakorn *et al.* 2018]. Rhodin *et al.* [2018b] solved the problem by using a small set of ground-truth samples. However, to maintain the integrity of a fully-supervised model, especially

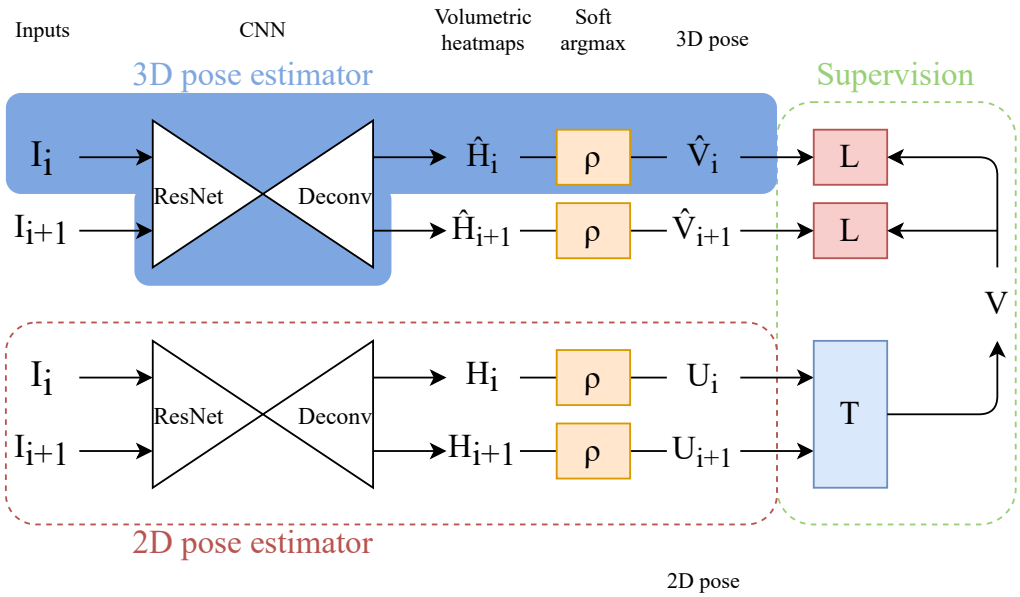


Figure 4.4: Figure showing the general architecture of the EpiPolar Pose network

when annotations are challenging to acquire for in-the-wild scenarios, it is advisable to keep the 2D pose estimation network frozen during training to prevent the degeneration of solutions where keypoints might collapse into a single location.

Epipolar pose shows that triangulation results on the Human3.6M dataset is close in error when compared to the projected 2D ground truth annotations [Kocabas *et al.* 2019]. When given the 2D ground truth annotations and the camera geometry, triangulation gives $4.3mm$ error.

The model achieves an average MPJPE error of $76.6mm$ across all results. This result is significant for being one of the best performers in 3D HPE self-supervised methods.

There are currently a few 3D HPE methods that are self-supervised without requiring any 3D supervision or camera extrinsics [Kocabas *et al.* 2019; Li *et al.* 2020a; Chen *et al.* 2019a; Drover *et al.* 2018]. By implementing a model using the main concepts of the aforementioned papers, a robust baseline may be obtained. The baseline will require an accurate and efficient 2D pose estimator for finding 2D keypoints for obtaining ground-truth 3D keypoint locations. This will then be used to train a 3D pose CNN for inferring 3D joint positions.

A prerequisite of this phase is to find and prepare the relevant datasets on which to experiment. The most prominent dataset being used in the 3D human pose estimation problem is the Human3.6M dataset [Ionescu *et al.* 2014]. This dataset is well-suited for our research, being multi-view and sufficiently large, consisting of over 3.6 million frames of captured pose in video format well suited to temporal methods.

The Epipolar Pose network in section 3.2 will be used as the baseline for the research [Kocabas *et al.* 2019]. The model will follow the architecture shown in figure 4.4, and the starting set-up will use two camera views for obtaining the self-supervised annotations. The 2D pose estimation network in the model will be pre-trained on the MPII 2D human pose dataset, based on the model by Andriluka *et al.* [2014b]. The output of the 2D pose estimation branch for each camera view will be used to

triangulate the 3D human pose in the global coordinate frame. The aim is to recreate the model and the results, which will serve as a baseline to the experiments from the next 4 phases.

4.5.2 Naïve temporal network

Along with the baseline outlined in Phase 1, some naïve temporal methods can be used to establish a baseline temporal experiment to compare with more robust methods that will be implemented in later phases.

The setup is by incorporating a naïve method with the baseline model explained in Phase 1. The training can simply take into account keypoint information from the previous few frames in a video sequence. When frame I_t at time t is input into the model, we also take into consideration a certain number of consecutive previous frames I_{t-1} and I_{t-2}. The purpose of using this simple naïve approach is to study the preliminary viability of using temporal methods. Any positive results serve as an affirmation of the more involved methods investigated in the next three phases. Investigations can be made about how varying the amount and configuration of the frames being sampled affects the results.

4.5.3 Temporal dilated convolutional network

A convolution of two discrete signals f of length N and zero-centered h of length $2M+1$ is the following:

$$(f * g)[n] = \sum_{m=-M}^{M} f[n-m]h[m] \tag{4.3}$$

Dilated convolutions are a particular form of convolution with a sparse structure where kernel points are spaced uniformly so that there is no need to explicitly apply kernel sparsity. A discrete dilated convolution is:

$$(f * h_D)[n] = \sum_{m=-M}^{M} f[D(n-m)]h[m], \tag{4.4}$$

where D is the dilation factor that is used to control the sparsity of the resultant filter. The larger the dilation factor, the sparser the filter.

By using a dilated convolution over the input signals, the receptive field of the model can be increased without increasing the computational cost. Computations are skipped over zero points.

By using dilated temporal convolutions, the information flow of the model can be controlled. The convolutions can be either symmetric or causal. Given a sequence of frames, symmetric convolutions use information from future frames along with past and present frames. Casual convolutions do not make use of future frames and can potentially be used where live data is concerned.

We propose to use dilated temporal convolutions with a self-supervised method in Pavllo *et al.* [2019] as a baseline. Using the 2D keypoints obtained from self-supervision

as dilated inputs to the convolutional neural network, the temporal receptive field can be controlled and varied without significantly increasing the computational cost.

Temporal convolutions are used with success in CNN-based models, created by Pavllo *et al.* [2019] and Shin *et al.* [2020]. Similarly, we propose using temporal convolutions to efficiently capture temporal information across a variable receptive field in our self-supervised model. A potential advantage of this method in conjunction with the baseline model is that it allows the use of efficient 1D convolutions over keypoint coordinates, which should theoretically decrease the computational complexity compared to using temporal convolutions with 2D heatmaps over time.

Dilated convolutions are proposed to model long-term dependencies by increasing the receptive field that the model samples from the input video sequence. The size of the receptive field is varied to capture long-term video information.

By experimenting with the receptive field parameter and ways to improve the adaptiveness of this method, the goal is to improve the accuracy of the baseline model. Temporal convolution filters should provide significant inference improvements with occluded parts.

4.5.4 Kalman filters

Kalman filtering algorithms provide estimates of unknown variables given previous measurements observed over time. Kalman filters (KF) are state estimators under the assumptions of linearity and Gaussian noise.

In the context of 3D HPE, Kalman filters are used by Coskun *et al.* [2017] to leverage temporal information in video sequence datasets. Their Kalman filter model assumes that a state at a given time t evolves from the state at $(t-1)$ following:

$$\begin{aligned} y_t &= Ay_{t-1} + w_t, w \sim N(0, Q), \\ z_t &= Hy_t + v_t, v \sim N(0, R), \end{aligned} \tag{4.5}$$

where the state is y_t and the measurement is z_t.

An iterative feedback loop with an update and a prediction step is used. In the prediction step used the pose model, the mean and covariance of the current state are measured regardless of the current measurements. $\hat{y}_t$ is the best estimate in terms of sum-of-squares error.

$$\begin{aligned} \hat{y}_t^{'} &= A\hat{y}_{t-1} \\ \hat{P}_t^{'} &= A\hat{P}_{t-1}A^T + Q \end{aligned} \tag{4.6}$$

The Kalman gain K_t is calculated in the update step, and used along with the observed measurement z_t to estimate the mean and covariance of y_t:

$$\begin{aligned} K_t &= \hat{P}_t^{'} H^T (H\hat{P}_t H^T + R)^{-1} \\ \hat{y}_t &= \hat{y}_t^{'} + K_t(\hat{z}_t - H\hat{y}_t^{'}) \\ \hat{P}_t &= \hat{P}_t^{'}(I - K_t H_t) \end{aligned} \tag{4.7}$$

The goal of Coskun *et al.* [2017] developing this particular Kalman filter for temporal regularisation pose estimation is to avoid specifying the linear transition function A or the fixed process and measurement covariance matrices Q and R. Kalman filtering is an algorithm that produces estimates of localisations based on a series of measurements over time. Since our baseline model infers using the 3D locations of joints in a skeleton, Kalman filters may provide a way to reduce noise in the estimations. By incorporating and experimenting with Kalman filters in the other phases where possible, the aim is to demonstrate increased accuracy that requires relatively small computational power.

Using the LSTM-KF model by Coskun *et al.* [2017] as inspiration, we can integrate the Kalman filters with the other phases and investigate the accuracy and stability of the results. Experiments are run on a baseline CNN model with a Kalman filter, temporal convolutions and Kalman filter, LSTM with Kalman filter. Using Kalman filters incorporated with an LSTM network, all the prior frames are used instead of only a single previous state in a normal Kalman Filter model.

4.6 Conclusion

The primary objective of our research is to evaluate and enhance the accuracy of geometric self-supervision in the 3D human pose estimation (HPE) task, comparing it against fully-supervised models. We use the average MPJPE error across multiple runs using dataset-standard protocols as our evaluation metric, ensuring statistical significance through the use of different random seeds.

Our first methodology explores Epipolar Pose, a prominent self-supervision method that relies on multi-view datasets. This model requires at least two camera views for training and uses 3D data obtained from triangulating 2D poses from different views. To avoid degenerate solutions, the 2D pose estimation network is pre-trained and kept frozen during training. We use the Human3.6M dataset, which is well-suited for our research due to its multi-view and large-scale nature. The Epipolar Pose network serves as our baseline.

In addition to the baseline, we incorporate a naive temporal method to establish a preliminary temporal experiment. This method considers keypoint information from previous frames in a video sequence to study the viability of temporal methods. By examining how varying the amount and configuration of sampled frames affects the results, we aim to provide initial insights into the potential benefits of temporal methods.

We also explore the use of temporal dilated convolutions, which increase the receptive field without increasing computational cost. By using dilated temporal convolutions, our model can capture long-term dependencies in video sequences. This method aims to improve the accuracy of the baseline model by varying the receptive field parameter, allowing for efficient capture of temporal information across a variable receptive field.

Finally, we investigate Kalman filters as a means to reduce noise in 3D joint estimations. These filters provide estimates of unknown variables given previous measurements over time. Inspired by the LSTM-KF model, we integrate Kalman filters with other phases to investigate accuracy and stability improvements. By incorporating Kalman filters with the baseline CNN model, temporal convolutions, and LSTM net-

works, we aim to demonstrate increased accuracy with relatively small computational power.

Each methodology is evaluated and summarised in the results chapter 5, ensuring that our findings are statistically significant and provide a comprehensive comparison of the different approaches to improving geometric self-supervision in 3D HPE.

Chapter 5

Results

5.1 Introduction

In this chapter, we present the results and interpretations derived from the methodologies and experimental details outlined in chapter 4. The findings are crucial for addressing the research questions posed at the outset of this study, providing insights into the effectiveness of our proposed techniques.

The chapter is structured to reflect the sequence of methods applied in our research. We begin by discussing the results obtained from epipolar self-supervision in section 5.2, which sets the foundation for subsequent analyses. This is followed by an examination of the naive temporal results in section 5.3, which offers a baseline for comparison with more advanced approaches. We employ our dilated Temporal Convolutional Network (TCN) in section 5.3, demonstrating the robustness of the methods above. The application of Kalman filter post-processing is then explored in section 5.4.

Finally, section 5.5 showcases a selection of visual results that provide an intuitive understanding of model results and support the quantitative findings. These visualisations demonstrate and convey the practical implications of our research in an easily accessible manner.

5.2 Epipolar self-supervision

The first result we obtain for using our epipolar geometric self-supervision is the distance error between our (calculated) pseudo-ground truth from the true ground-truth annotations provided by the dataset. Table 5.1 shows the errors obtained between epipolar geometry in the ground truth annotations themselves and the outcomes of the off-the-shelf 2D estimation model introduced in section 4.5.1 that is trained on both Human3.6M and MPII used in our model pipeline.

The table illustrates that the pseudo-ground truth, derived from the 2D ground truth of the Human3.6M dataset, exhibits a Mean Per Joint Position Error (MPJPE) of 5.29mm. This indicates an average discrepancy of approximately 5mm between the estimated geometric pose and the actual 2D ground truth. In contrast, the MPJPE for the off-the-shelf 2D pose estimator (based on Andriluka *et al.* [2014b]) trained on Human3.6M data is significantly higher, at 30.64mm, translating to an average error of

Data	MPJPE (mm)
Ground Truth 2D	5.29
H36M 2D	30.64
MPII 2D	47.38

Table 5.1: MPJPE Error between self-supervision data and various ground truth annotations.

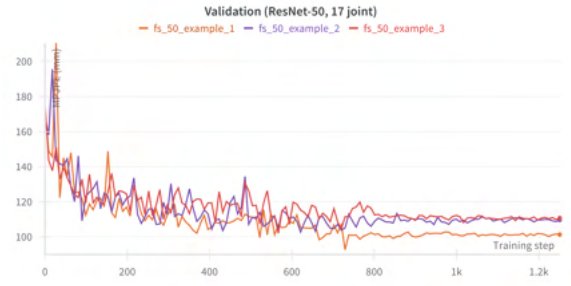
about 3cm across all joints in the world space. The estimator trained on the MPII 2D ground truth data fares worse, with an average MPJPE of 47.38mm. Despite this higher error rate, the MPII-based estimator's output remains sufficiently accurate to serve as a basis for further model training. These results show the effects of different keypoint sources on the geometric self-supervision data.

We then examine the outcomes of models trained under consistent conditions, with the only variable being the mode of supervision. A comparative analysis of the training on a suitable model architecture is conducted to compare the disparities in model performance when trained in a fully supervised environment against a self-supervised setting. This investigation seeks to determine the impact of supervision mode on the accuracy and reliability of the model's output, providing insights into the efficacy of self-supervised learning techniques in relation to the fully supervised counterpart.

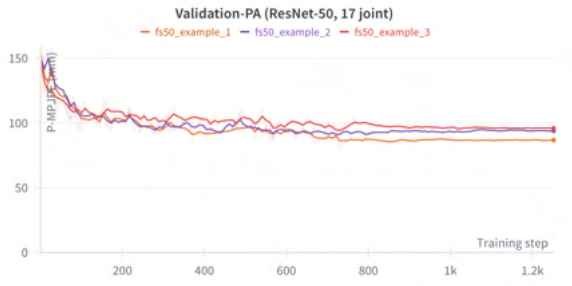
In the following graphs, we present the model training outcomes of self-supervision using epipolar geometry and normal fully-supervised training on the available annotations in the H36M dataset. We not only vary the supervision data source of the model, but experiment with the model size. We demonstrate the most effective results in terms of training times and outcomes on a network with a ResNet-50 backbone, while showing some results from other network sizes.

For our fully-supervised model using the ResNet-50 backbone, we achieve an average MPJPE of 107.238mm over 3 runs shown in figure 5.1. The average P-MPJPE is 92.283mm, and 93.255mm for N-MPJPE. The P-MPJPE result suggests that, after Procrustes alignment, the predicted joint positions closely match the scale, rotation, and translation of the ground truth, indicating that the model captures the overall pose structure well. In contrast, the N-MPJPE, which is typically scale and translation invariant, implies that the model's joint predictions are consistent across different individuals and poses after normalising for individual body sizes and other scale factors. Taking the MPJPE error, this intuitively represents around 10cm of average error across all joints across all frames using our fully-supervised baseline model. For comparison, figure 5.2 shows the average MPJPE of 3 self-supervised runs being 118.306mm, which is a 9.35% decrease in MPJPE performance relative to the average performance of the fully-supervised model.

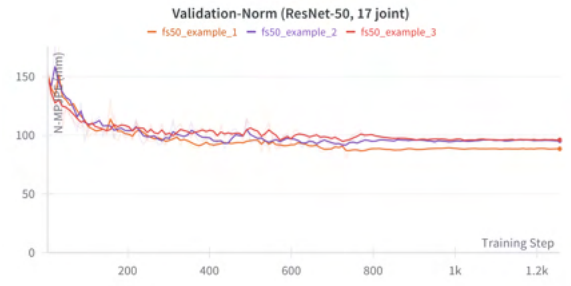
Despite performing with more error, the decrease in results is not significantly lower (less than 10%) considering this is around a 2cm average increase in error across all joints over all frames. The P-MPJPE for our self-supervised models are 102.036mm, indicating that even without explicit supervision, the self-supervised approach is capable of learning pose structures that align well with the true poses after Procrustes analysis. The N-MPJPE average is 112.726mm, showing that the self-supervised model maintains a degree of consistency across normalized poses. We also show the standard



(a) Fully-supervised (FS) validations (MPJPE (mm))



(b) FS validations (P-MPJPE (mm))



(c) FS validations (P-MPJPE (mm))

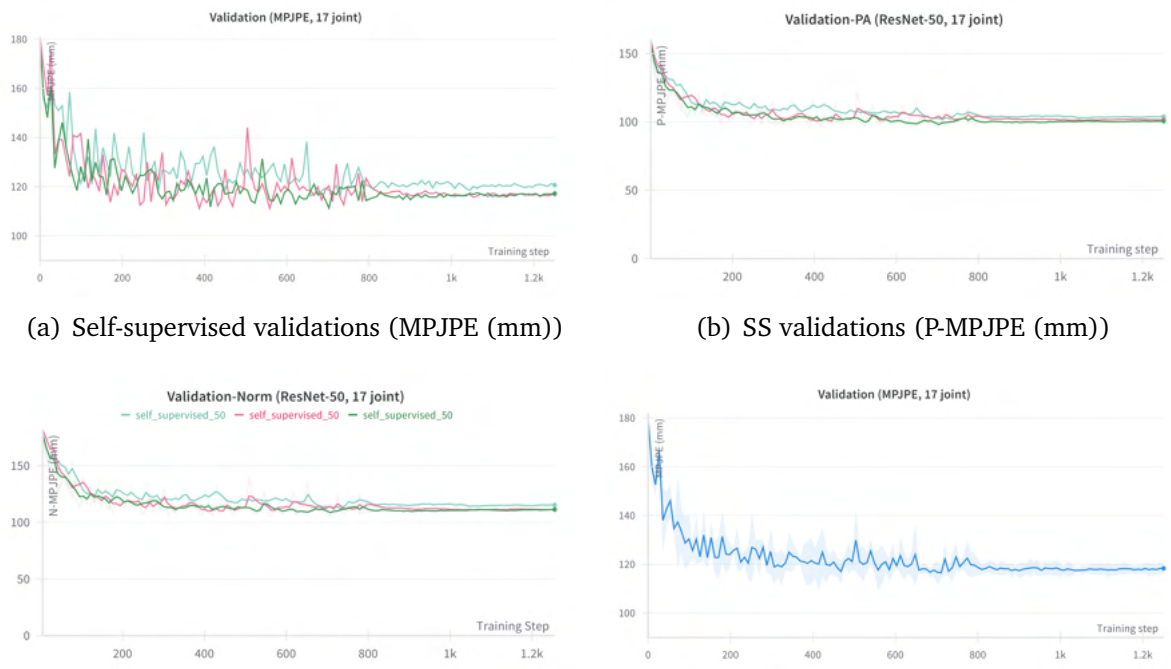
(d) MPJPE (mm) validation std. deviation

Figure 5.1: Validation graphs for ResNet-50 fully-supervised model runs

deviation of the model runs in figures 5.1(d) and 5.2(d) for the fully-supervised and self-supervised model runs, respectively, which provide insights into the variability of the model's performance.

The comparative analysis of fully-supervised and self-supervised models using ResNet-50 backbones reveals compelling evidence in favor of self-supervised learning approaches. Despite the absence of explicitly labeled 3D data, the self-supervised models demonstrate a remarkable ability to approximate human pose estimation with an average MPJPE of 118.306mm, which is only marginally greater than that of their fully-supervised counterparts. Notably, the self-supervised models achieve a P-MPJPE of 102.036mm, suggesting that they are proficient at capturing the intrinsic geometric structure of the human body after alignment.

The N-MPJPE value of 112.726mm demonstrates a high level of uniformity across normalized poses, which is important for practical applications that require scale and translation invariance. The slight decrease in performance, less than 10% compared to fully-supervised models, is a small compromise given the substantial reduction in annotation expenses and the potential for scalability provided by self-supervised methods. These results highlight the effectiveness of self-supervised learning as a reliable alternative to conventional fully-supervised approaches, especially in situations where obtaining large annotated datasets is difficult or not feasible.



(a) Self-supervised validations (MPJPE (mm))

(b) SS validations (P-MPJPE (mm))

(c) SS validations (P-MPJPE (mm))

(d) MPJPE (mm) validation std. deviation

Figure 5.2: Validation graphs for ResNet-50 self-supervised model runs

5.3 Temporal convolutional network

In this section, we present the results of experimentation with the temporal convolutional neural network on the fully-supervised and self-supervised training paradigms. We will discuss these results in contrast with our fully-supervised and self-supervised models from section 5.2 as a baseline.

First we experiment with the receptive field of the temporal dilated convolutional neural network (TDCN) under a fully-supervised setting. We use symmetric dilation following the methodology outlined in section 4.5.3 as the setup. Figure 5.3 shows the validation error of the model averaged over three runs. From this figure we can see good consistency of the results between different receptive field sizes of 9, 27, 81. From the figure, the receptive field of 81 achieved the best results with an average 49.084mm of error, while a receptive field size of 27 achieved an average of 50.277mm. The receptive field size 9 has average final validation loss of 51.321mm. Compare this to the baseline ResNet model in section 5.2. This shows that the best temporal experiment model performed on average, 54.23% better than the best fully-supervised baseline, which is significantly more accurate. This demonstrates the remarkable improvement in accuracy over a non-temporal model that the TDCN can achieve. Also note the fair consistency and expected results from the variation of the receptive field. Having a larger receptive field should improve the accuracy of the model. However, there is also expectation that a receptive field that is too large will result in the model losing temporal context and detail. In our improvements section in section 6.1, we discuss the potential for experimenting to find the receptive field size that will cause a drop-off of

accuracy for the model. These results indicate that the higher average final error of the model is due to the acquisition of temporal information from the training data.

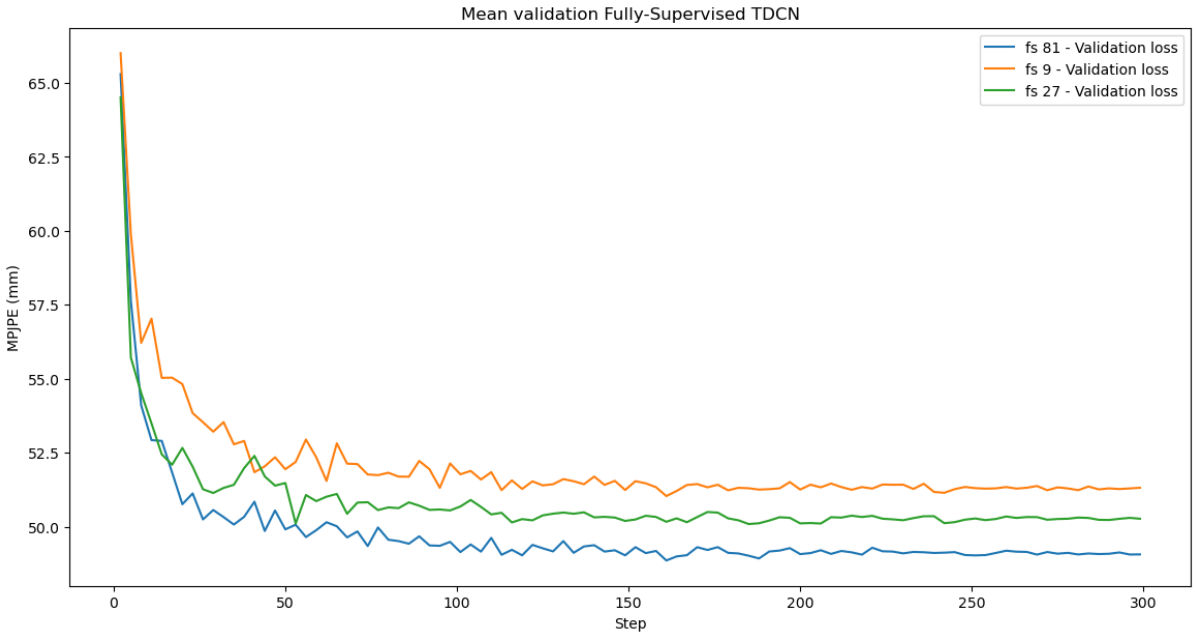


Figure 5.3: Experimentation on receptive field sizes in fully-supervised temporal dilated convolutional neural network

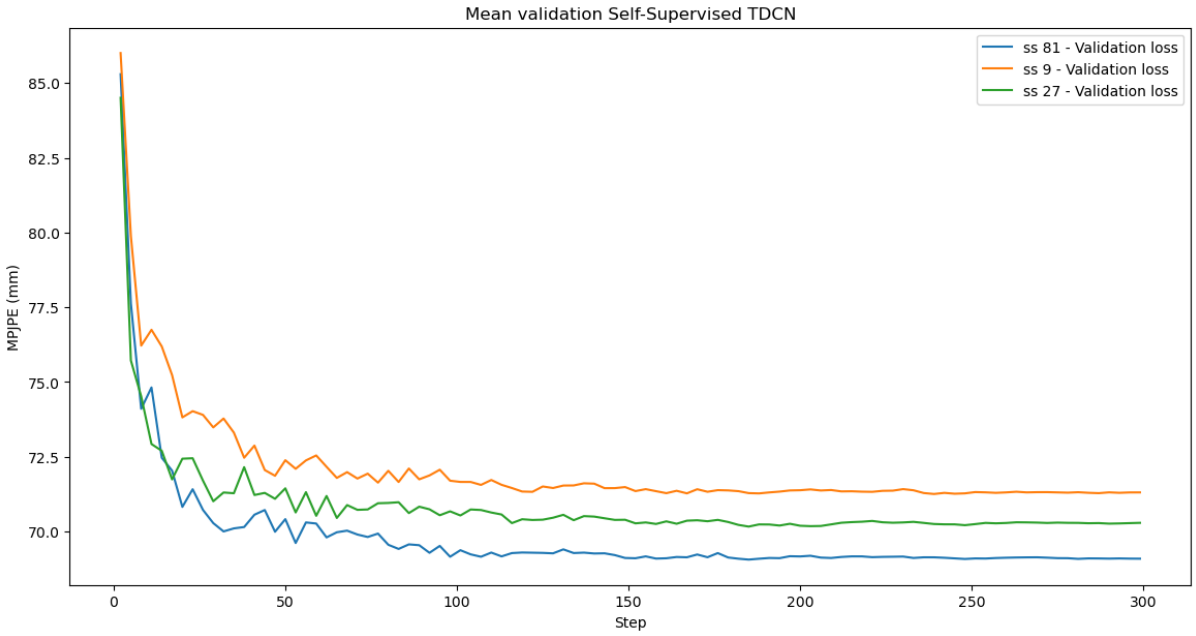


Figure 5.4: Experimentation on receptive field sizes in self-supervised temporal dilated convolutional neural network

Next, we evaluate the combination of Temporal Dilated Convolutional Network (TDCN) with epipolar self-supervision for its efficacy in self-supervised learning for 3D

human pose estimation. The results, presented in figure 5.4, indicate that the model's performance varies with different receptive field sizes. Specifically, the model with a receptive field size of 81 registers an MPJPE of 69.088mm. For receptive field sizes of 27 and 9, the average errors are 70.287mm MPJPE and 71.305mm MPJPE, respectively.

When these results are juxtaposed with those from the fully-supervised model, as shown in figure 5.3, there is a consistent increase in MPJPE, approximately 20mm across the different models. This increment in error can be attributed to the reliance on the accuracy of epipolar geometry within the self-supervised data. Despite a 40% reduction in accuracy when comparing the best self-supervised model to the fully-supervised counterpart, the self-supervised approach still demonstrates a 35% improvement over the baseline fully-supervised model, as referenced in section 5.2. These findings underscore the potential of self-supervised methods in leveraging temporal information, even in the absence of fully annotated training data.

5.4 Kalman filter post-processing

For our Kalman filter post-processing step introduced in section 4.5.4, we present here a summary result applied to the self-supervised TDCN at various receptive fields. We already show the effectiveness of temporal methods combined with self-supervision in section 5.3, and here we will discuss the results of further applying a temporal post-processing step to the network.

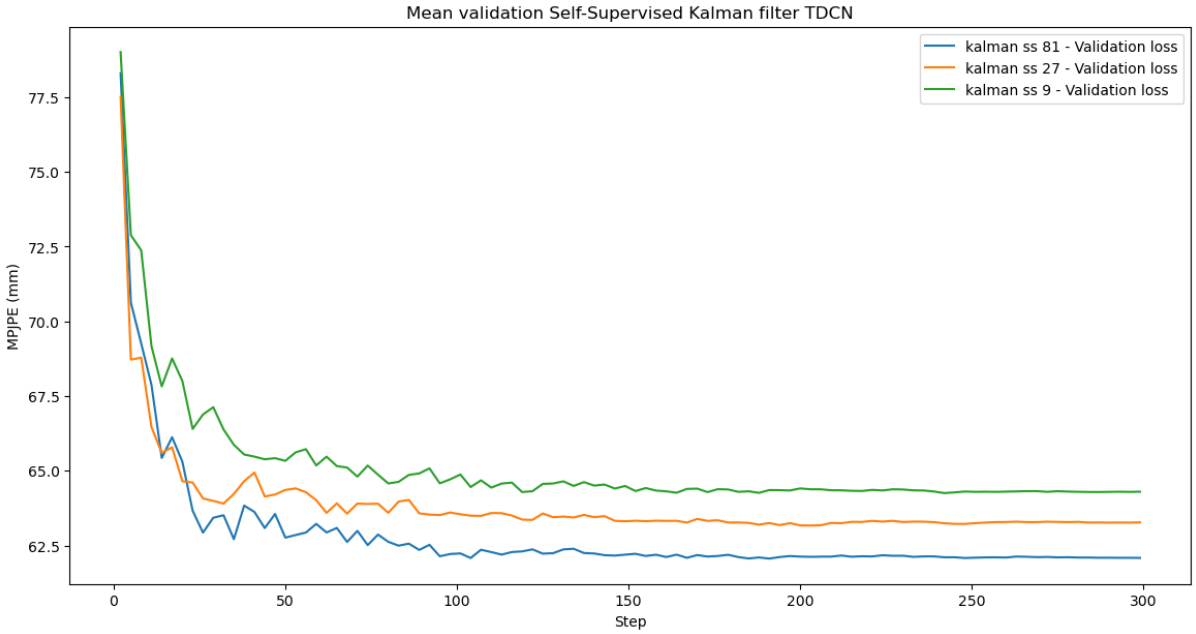


Figure 5.5: Kalman filter validation performance on self-supervised TDCN with varying receptive fields

Figure 5.4 shows average performance over 3 runs of the model with Kalman filter post-processing for receptive fields of sizes 81, 27 and 9. We find that there is a consistent but relatively small increase in model performance with the Kalman filter

post-processing step in place, averaging approximately 7mm MPJPE. This represents a 10.13% increase in performance compared to the equivalent self-supervised TDCN without a post-processing step. Whilst small, this is still an improvement on the model and further illustrates the effectiveness of temporal methods when it comes to improving model performances. The best performing self-supervised TDCN model with a Kalman filter post-processing step performs at a final accuracy of 62.088mm MPJPE on the unseen test set of Human3.6M.

5.5 Visual results

For an intuitive and empirical understanding of how our model works, we create a visualiser tool for rendering and viewing the result 3D poses in short sample data segments. The visuals are able to be rendered in video format, and we have extracted random frames here for qualitatively reviewing model performance. These figures 5.6 and 5.7 show the resultant pose from the model, superimposed on the sample data, and also isolated in a 3D axis.

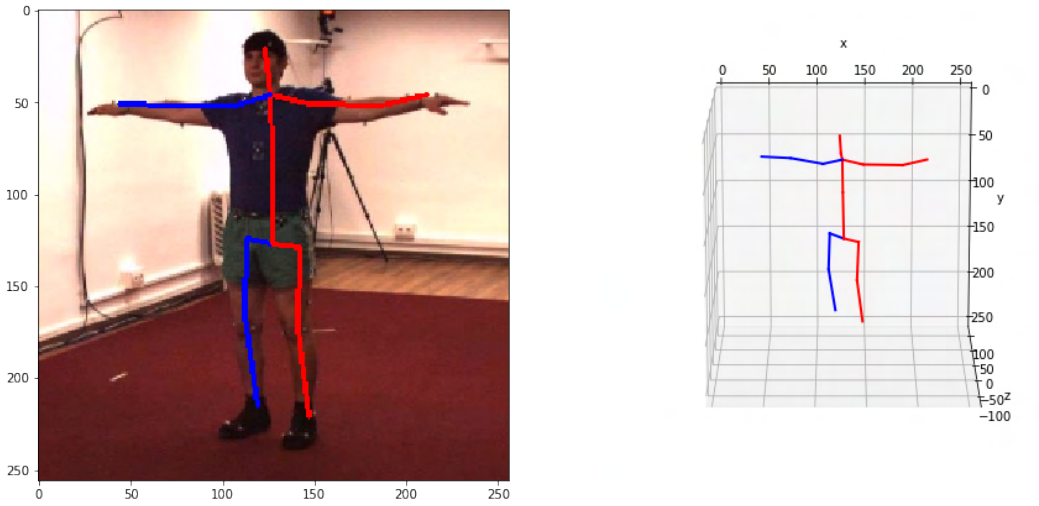


Figure 5.6: Example visualisation - standing pose

5.6 Conclusion

This chapter serves to evaluate the efficacy of the combination of epipolar geometric self-supervision and temporal methods (Temporal Dilated Convolutional Network (TDCN) and Kalman filter for 3D human pose estimation. The TDCN is a temporal convolutional neural network with a receptive field, and it is experimented both with fully annotated datasets and self-supervised pseudo ground truths.

The best performing self-supervised model (TDCN with Kalman filter) achieves a relatively well-performing 62.088mm MPJPE on unseen data, compared to 49.084mm on the equivalent fully-supervised TDCN. This is a 26% increase in average performance

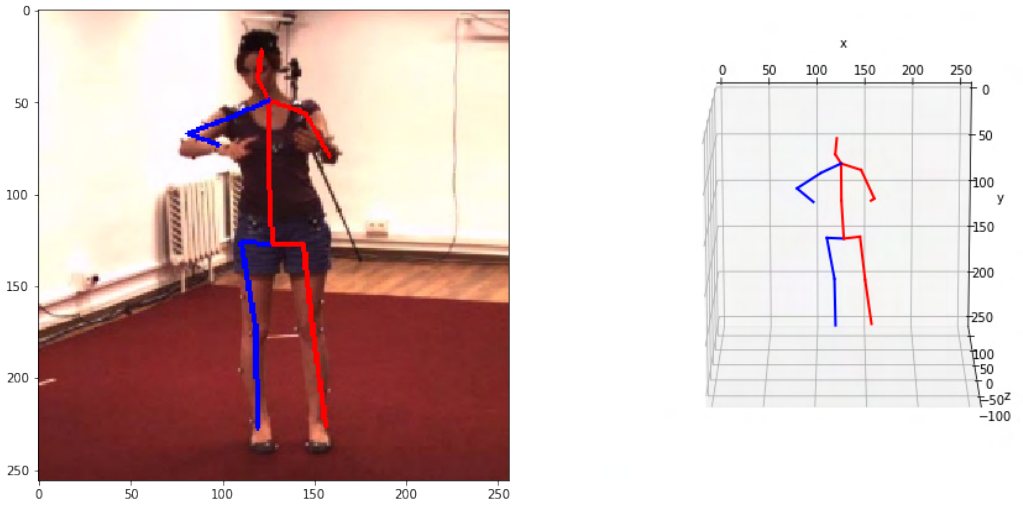


Figure 5.7: Example visualisations with some self-occlusion

error correlating to a 13.003mm MPJPE around 1cm of average increase in error per-joint over all frames. This is still a good result and highlights the potential for using self-supervised settings and leveraging the multi-view temporal nature of 3D HPE datasets.

This demonstrates the potential of self-supervised learning techniques in leveraging temporal information, even in situations where obtaining large datasets is difficult or not feasible. We determine the impact of supervision mode on the accuracy and reliability of the model's output, provide insights into the effectiveness of our proposed techniques.

Chapter 6

Future Work and Improvements

In this short chapter we will briefly express some of the potential improvements and future works of our research. We have split this up for easier comprehension into two sections, improvements (section 6.1) and future works (section 6.2). Improvements will focus on some of the small research-specific and experiment-specific details that would be nice to implement. Future works will discuss branching out of the research to related experiments and some newer methods that may keep our research contemporary.

6.1 Improvements

Several areas of improvement can be done to improve the clarity and reproducibility of our research:

6.1.1 Ablation Studies

To gain a deeper understanding of our models and pipelines, it would be beneficial to conduct full ablation studies. By systematically removing or modifying individual components, we can clearly assess their impact on the overall performance. This will provide valuable insights into the importance of each component and help optimise the models.

6.1.2 Increased Reproducibility

To improve the reproducibility of our work, we plan to run a larger number of model runs, more than 10 runs per model instead of the current 3. This will allow us to obtain more robust statistical evidence for our findings. Additionally, we will analyse and store the random seeds of non-deterministic elements to ensure consistent results across different runs.

6.1.3 Statistical Analysis

To validate the significance of our results, we intend to supplement our findings with statistical tests. Specifically by employing t-tests and analysis of variances (ANOVA) to assess the differences observed between different methods. These tests will provide a more rigorous evaluation of the performance of each method and establish their statistical significance [St *et al.* 1989].

By implementing these improvements, we aim to strengthen the comprehensibility and reliability of our research, ultimately contributing to its overall impact.

6.2 Future work

In this section, potential avenues for future research and expansion of our work are outlined:

6.2.1 Survey and comparison of geometric methods

One promising direction for future work is to conduct a survey and comparison of different geometric methods for obtaining pseudo ground-truth pose. By using the same experimental setups, we can compare and evaluate various geometric methods. This comparison can include an assessment of the accuracy, robustness, and computational efficiency of each method. Additionally, exploring other forms of self-supervision could provide valuable insights into alternative approaches.

This comparison may include evaluating the performance of methods such as epipolar geometry and simple triangulation [Iskakov *et al.* 2019]. Simple triangulation leverages multi-view geometry with simple trigonometric calculations instead of camera-specific geometric relations in epipolar geometry. Such a comparison may be able to illuminate the strengths and limitations of each method and provide guidance for future developments.

6.2.2 Exploration of additional geometric constraints

In addition to the geometric methods used for generating pseudo-ground truth annotations, there is potential for incorporating additional geometric constraints for pose consistency. For example, the incorporation of mirror-symmetry [Pavlakos *et al.* 2019] and self-consistency constraints [Iqbal *et al.* 2020] can be investigated. These additional constraints may offer new insights and further improve the accuracy and reliability of pose estimation.

6.2.3 Attention mechanisms

As outlined in our Related Works (chapter 3), attention mechanisms have gained significant recent attention in various fields of research, including natural language processing, computer vision, and machine learning. These mechanisms allow models to focus on specific parts of the input data that are deemed more relevant or informative for the

task at hand. In the context of 3D human pose estimation, attention mechanisms can prove to be a good model backbone to improve the accuracy and robustness of pose estimation models. By selectively attending to certain regions or keypoints in the input data, attention mechanisms can help the model better understand spatio-temporal relationships and dependencies between different body parts.

By pursuing these future research directions, we can expand the scope of our work, gain a deeper understanding of geometric methods, and potentially contribute to the development of novel techniques in the field of pose estimation.

Chapter 7

Conclusion

The introduction of deep learning in computer vision has led to many advancements in the pose estimation domain. With increasingly efficient estimation methods for 2D HPE, attention is turned to creating similarly efficient methods in 3D HPE. 3D human pose estimation is limited by its more constricted datasets, where image data is collected in a lab environment and rarely in the wild. Therefore, it is important to obtain efficient unsupervised and self-supervised methods that can train on unlabelled datasets and generalise better to real-life samples of pose images.

Geometric self-supervised methods have gained traction in the last few years in 3D pose estimation. This research hypothesises that by using a variety of temporal methods with geometric self-supervision, the accuracy of the existing self-supervised models may be improved. By combining modules that leverage temporal information, the goal is to improve overall prediction accuracies and that of difficult pose scenarios such as occlusion.

We demonstrate that the data error between using epipolar geometric self-supervision is fair to work with, showing just 5.29mm MPJPE when compared to ground-truth data. When used with an off-the-shelf 2D pose estimator on the H36M and MPII dataset, we obtain errors of 30.64mm and 47.38mm MPJPE, respectively.

By creating a self-supervised TDCN model using Kalman filter post-processing, we achieve our best self-supervised results achieving a MPJPE of 62.088mm on unseen data. An equivalent fully-supervised setting performs on average 49.084mm. We show 26% increase in average performance error correlating to a 13.003mm MPJPE. This is functionally around 1cm of average increase in error per-joint over all frames. Our results highlight the potential for using self-supervised settings and leveraging the multi-view temporal nature of 3D HPE datasets.

This study demonstrates the potential of self-supervised learning techniques, particularly their capacity to harness temporal information in the absence of extensive labelled datasets. Future research should explore the refinement of these methods, perhaps by incorporating more sophisticated temporal models or by extending the approach to more diverse and challenging datasets. The ultimate goal remains: push the boundaries of accuracy in 3D human pose estimation using minimal supervision, broadening the applicability and accessibility of this technology.

References

[Alldieck *et al.* 2017] Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conference on Pattern Recognition*, pages 347–360. Springer, 2017.

[Andriluka *et al.* 2014a] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[Andriluka *et al.* 2014b] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[Andriluka *et al.* 2018] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018.

[Anguelov *et al.* 2005] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416, 2005.

[Arnab *et al.* 2019] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019.

[Bogo *et al.* 2016] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016.

[Bouazizi *et al.* 2021] Arij Bouazizi, Julian Wiederer, Ulrich Kressel, and Vasileios Belagiannis. Self-supervised 3d human pose estimation with multiple-view geometry. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[Bourdev and Malik 2009] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th international conference on computer vision*, pages 1365–1372. IEEE, 2009.

[Cai *et al.* 2019] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2272–2281, 2019.

[Cao *et al.* 2017] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[Carreira *et al.* 2016] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[Chen and Yuille 2014] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. *Advances in neural information processing systems*, 27, 2014.

[Chen *et al.* 2016] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016.

[Chen *et al.* 2017] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 1212–1221, 2017.

[Chen *et al.* 2019a] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019.

[Chen *et al.* 2019b] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10895–10904, 2019.

[Chen *et al.* 2020] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.

[Cheng *et al.* 2019] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 723–732, 2019.

[Chu *et al.* 2017] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1831–1840, 2017.

[Ci *et al.* 2019] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2271, 2019.

[Collobert *et al.* 2016] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

[Cootes *et al.* 1995] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[Coskun *et al.* 2017] Huseyin Coskun, Felix Achilles, Robert DiPietro, Nassir Navab, and Federico Tombari. Long short-term memory kalman filters: Recurrent neural estimators for pose regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5524–5532, 2017.

[Dabral *et al.* 2018] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European conference on computer vision (ECCV)*, pages 668–683, 2018.

[Dabral *et al.* 2019] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *2019 international conference on 3D vision (3DV)*, pages 405–414. IEEE, 2019.

[Dalal and Triggs 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[Dang *et al.* 2019] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019.

[Dantone *et al.* 2013] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3041–3048, 2013.

[Dauphin *et al.* 2017] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[Doering *et al.* 2022] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20963–20972, 2022.

[Drover *et al.* 2018] Dylan Drover, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[Dsouza *et al.* 2020] Grandel Dsouza, Deepak Maurya, and Anoop Patel. Smart gym trainer using human pose estimation. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–4, 2020.

[Du *et al.* 2016] Yu Du, Yongkang Wong, Yonghao Liu, Feilin Han, Yilin Gui, Zhen Wang, Mohan Kankanhalli, and Weidong Geng. Marker-less 3d human motion capture with monocular image sequence and height-maps. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 20–36. Springer, 2016.

[Eichner and Ferrari 2010] Marcin Eichner and Vittorio Ferrari. We are family: Joint pose estimation of multiple persons. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 228–242. Springer, 2010.

[Fang *et al.* 2017] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[Felzenszwalb and Huttenlocher 2005] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61:55–79, 2005.

[Fish Tung *et al.* 2017] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4354–4362, 2017.

[Gehring *et al.* 2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017.

[Gong *et al.* 2016] Wenjuan Gong, Xuena Zhang, Jordi Gonzàlez, Andrews Sobral, Thierry Bouwmans, Changhe Tu, and El-hadi Zahzah. Human pose estimation from monocular images: A comprehensive survey. *Sensors*, 16(12):1966, 2016.

[Graves and Graves 2012] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[Güler *et al.* 2018] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[Guo *et al.* 2020] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

[Guo *et al.* 2022] Yongping Guo, Ying Chen, Jianzhi Deng, Shuiwang Li, and Hui Zhou. Identity-preserved human posture detection in infrared thermal images: A benchmark. *Sensors*, 23(1):92, 2022.

[Habibie *et al.* 2019] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10905–10914, 2019.

[Hafiz *et al.* 2021] Abdul Mueed Hafiz, Shabir Ahmad Parah, and Rouf Ul Alam Bhat. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. *arXiv preprint arXiv:2106.07550*, 2021.

[He *et al.* 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.* 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Hochreiter and Schmidhuber 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Huang *et al.* 2022] Zhangjin Huang, Rashid Khan, et al. A review of 3d human body pose estimation and mesh recovery. *Digital Signal Processing*, 128:103628, 2022.

[Hussein *et al.* 2013] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-third international joint conference on artificial intelligence*, 2013.

[Insafutdinov *et al.* 2016] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 34–50. Springer, 2016.

[Intel 2023] Intel. *Intel® RealSenseTM*. Intel® RealSenseTM Depth and Tracking Cameras, August 17 2023. Accessed on: March 17, 2023.

[Ionescu *et al.* 2014] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[Iqbal *et al.* 2020] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5243–5252, 2020.

[Iskakov *et al.* 2019] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019.

[Jack *et al.* 2019] Dominic Jack, Frederic Maire, Sareh Shirazi, and Anders Eriksson. Ige-net: Inverse graphics energy networks for human pose estimation and single-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7075–7084, 2019.

[Jain *et al.* 2011] Himanshu Prakash Jain, Anbumani Subramanian, Sukhendu Das, and Anurag Mittal. Real-time upper-body human pose estimation using a depth camera. In *Computer Vision/Computer Graphics Collaboration Techniques: 5th International Conference, MIRAGE 2011, Rocquencourt, France, October 10-11, 2011. Proceedings 5*, pages 227–238. Springer, 2011.

[Ji *et al.* 2020] Xiaopeng Ji, Qi Fang, Junting Dong, Qing Shuai, Wen Jiang, and Xiaowei Zhou. A survey on monocular 3d human pose estimation. *Virtual Reality & Intelligent Hardware*, 2(6):471–500, 2020.

[Jiang *et al.* 2023] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

[Johnson and Everingham 2010] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2–4, page 5. Citeseer, 2010.

[Joo *et al.* 2015] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[Joo *et al.* 2018] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018.

[Ju *et al.* 1996] Shanon X Ju, Michael J Black, and Yaser Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Proceedings of the second international conference on automatic face and gesture recognition*, pages 38–44. IEEE, 1996.

[Kanazawa *et al.* 2018] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[Karl-Bridge-Microsoft 2012] Karl-Bridge-Microsoft. *Kinect for Windows*. Microsoft Learn, February 1 2012. Accessed on: March 17, 2023.

[Kendall *et al.* 2015] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015.

[Kocabas *et al.* 2018] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018.

[Kocabas *et al.* 2019] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019.

[Kolotouros *et al.* 2019] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019.

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[Kundu *et al.* 2020] Jogendra Nath Kundu, Siddharth Seth, MV Rahul, Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11312–11319, 2020.

[Lassner *et al.* 2017] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017.

[Lee *et al.* 2018] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018.

[Li and Chan 2015] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12*, pages 332–347. Springer, 2015.

[Li *et al.* 2014] Sijin Li, Zhi-Qiang Liu, and Antoni B Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 482–489, 2014.

[Li *et al.* 2019] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2192–2201, 2019.

[Li *et al.* 2020a] Yang Li, Kan Li, Shuai Jiang, Ziyue Zhang, Congzhentao Huang, and Richard Yi Da Xu. Geometry-driven self-supervised method for 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34–07, pages 11442–11449, 2020.

[Li *et al.* 2020b] Yang Li, Kan Li, Xinxin Wang, and Richard Yi Da Xu. Exploring temporal consistency for human pose estimation in videos. *Pattern Recognition*, 103:107258, 2020.

[Liang and Lin 2019] Junbang Liang and Ming C Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4352–4362, 2019.

[Liao *et al.* 2023] Xianhua Liao, Jiayan Dong, Kangkang Song, and Jiangjian Xiao. Three-dimensional human pose estimation from sparse imus through temporal encoder and regression decoder. *Sensors*, 23(7):3547, 2023.

[Lin *et al.* 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Lin *et al.* 2017] Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, and Hui Cheng. Recurrent 3d pose sequence machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 810–819, 2017.

[Liu *et al.* 2015a] Jigang Liu, Dongquan Liu, Justin Dauwels, and Hock Soon Seah. 3d human motion tracking by exemplar-based conditional particle filter. *Signal Processing*, 110:164–177, 2015.

[Liu *et al.* 2015b] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: the body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015.

[Loper *et al.* 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866, 2015.

[Lowe 1999] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[Lowe 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[Luvizon *et al.* 2018] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018.

[Luvizon *et al.* 2019] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019.

[Martinez *et al.* 2017] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

[Mehta *et al.* 2017a] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[Mehta *et al.* 2017b] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[Mehta *et al.* 2017c] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017.

[Mehta *et al.* 2018] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[Moon *et al.* 2019] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10133–10142, 2019.

[Moreno-Noguer 2017] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2823–2832, 2017.

[Newell *et al.* 2016] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

[Newell *et al.* 2017] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017.

[Nibali *et al.* 2018] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

[Nibali *et al.* 2019] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 3d human pose estimation with 2d marginal heatmaps. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1477–1485. IEEE, 2019.

[Nie *et al.* 2017] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3467–3475. IEEE, 2017.

[Nie *et al.* 2019] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6951–6960, 2019.

[Novotny *et al.* 2019] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7688–7697, 2019.

[Omran *et al.* 2018] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

[Park *et al.* 2016] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 156–169. Springer, 2016.

[Pavlakos *et al.* 2017a] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

[Pavlakos *et al.* 2017b] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017.

[Pavlakos *et al.* 2018] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018.

[Pavlakos *et al.* 2019] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 803–812, 2019.

[Pavllo *et al.* 2019] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[Pishchulin *et al.* 2012] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012.

[Pishchulin *et al.* 2016] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[Qiu *et al.* 2019] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4342–4351, 2019.

[Radosavovic *et al.* 2018] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018.

[Ramanan *et al.* 2006] Deva Ramanan, David A Forsyth, and Andrew Zisserman. Tracking people by learning their appearance. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):65–81, 2006.

[Ren *et al.* 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Rhodin *et al.* 2015] Helge Rhodin, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 765–773, 2015.

[Rhodin *et al.* 2018a] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 750–767, 2018.

[Rhodin *et al.* 2018b] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018.

[Rogez *et al.* 2017] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017.

[Rogez *et al.* 2019] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1146–1161, 2019.

[Rostamzadeh *et al.* 2018] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

[Sapp and Taskar 2013] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3681, 2013.

[Sarafianos *et al.* 2016] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016.

[Schwarcz and Pollard 2018] Steven Schwarcz and Thomas Pollard. 3d human pose estimation from deep multi-view 2d pose. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2326–2331. IEEE, 2018.

[Shin *et al.* 2020] Michael Shin, Eduardo Castillo, Irene Font Peradejordi, and Shobhna Jayaraman. 3d human pose estimation with adaptive receptive fields and dilated temporal convolutions. *arXiv preprint arXiv:2005.13797*, 2020.

[Shotton *et al.* 2011] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

[Shuai *et al.* 2022] Hui Shuai, Lele Wu, and Qingshan Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4122–4135, 2022.

[Sidenbladh *et al.* 2000] Hedvig Sidenbladh, Michael J Black, and David J Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Computer Vision—ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II 6*, pages 702–718. Springer, 2000.

[Sigal *et al.* 2010] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[Simonyan and Zisserman 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sorkine 2006] Olga Sorkine. Differential representations for mesh processing. In *Computer Graphics Forum*, volume 25–4, pages 789–807. Wiley Online Library, 2006.

[St *et al.* 1989] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

[Stoll *et al.* 2011] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *2011 International Conference on Computer Vision*, pages 951–958. IEEE, 2011.

[Su *et al.* 2019] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5674–5682, 2019.

[Sun *et al.* 2017] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2602–2611, 2017.

[Sun *et al.* 2018] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018.

[Sutskever *et al.* 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[Suwajanakorn *et al.* 2018] Supasorn Suwajanakorn, Noah Snavely, Jonathan Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *arXiv preprint arXiv:1807.03146*, 2018.

[Tan *et al.* 2017] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. *British Machine Vision Association*, 2017.

[Tang and Wu 2019] Wei Tang and Ying Wu. Does learning specific features for related parts help human pose estimation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1107–1116, 2019.

[Tang *et al.* 2023] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023.

[Tekin *et al.* 2016] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–1000, 2016.

[Tekin *et al.* 2017] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3941–3950, 2017.

[Tome *et al.* 2017] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2500–2509, 2017.

[Tome *et al.* 2018] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *2018 international conference on 3D vision (3DV)*, pages 474–483. IEEE, 2018.

[Toshev and Szegedy 2014] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[Trumble *et al.* 2017] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017.

[Trumble *et al.* 2018] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018.

[Véges *et al.* 2019] Márton Véges, Viktor Varga, and András Lőrincz. 3d human pose estimation with siamese equivariant embedding. *Neurocomputing*, 339:194–201, 2019.

[Wandt and Rosenhahn 2019] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7782–7791, 2019.

[Wandt *et al.* 2021] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13294–13304, 2021.

[Wang and Li 2013] Fang Wang and Yi Li. Beyond physical connections: Tree models in human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 596–603, 2013.

[Wang *et al.* 2019] Jue Wang, Shaoli Huang, Xinchao Wang, and Dacheng Tao. Not all parts are created equal: 3d pose estimation by modeling bi-directional dependencies of body parts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7771–7780, 2019.

[Wang *et al.* 2021] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021.

[Wang *et al.* 2022] Zitian Wang, Xuecheng Nie, Xiaochao Qu, Yunpeng Chen, and Si Liu. Distribution-aware single-stage models for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13096–13105, 2022.

[Wei *et al.* 2016] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[Wu *et al.* 2017] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

[Xiao *et al.* 2018] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[Yang and Ramanan 2012] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

[Yang *et al.* 2018] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5255–5264, 2018.

[Yang *et al.* 2023] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. *arXiv preprint arXiv:2307.15880*, 2023.

[Zhang *et al.* 2018] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28(4):1837–1850, 2018.

[Zhang *et al.* 2021] Dejun Zhang, Yiqi Wu, Mingyue Guo, and Yilin Chen. Deep learning methods for 3d human pose estimation under different supervision paradigms: a survey. *Electronics*, 10(18):2267, 2021.

[Zhao *et al.* 2019] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019.

[Zhou *et al.* 2016a] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016.

[Zhou *et al.* 2016b] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 186–201. Springer, 2016.

[Zhou *et al.* 2017] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE international conference on computer vision*, pages 398–407, 2017.

[Zhou *et al.* 2019] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2344–2353, 2019.

[Zhu *et al.* 2017] Xiangyu Zhu, Yingying Jiang, and Zhenbo Luo. Multi-person pose estimation for posetrack with enhanced part affinity fields. In *ICCV PoseTrack Workshop*, volume 7, page 4321, 2017.