Robust and Efficient Methods in Visual 3D Human Pose Estimation

Der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University vorgelegte Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften von

István Sárándi, Master of Science
aus Budapest, Ungarn

Abstract

Many important machine perception tasks involve people in some ways: self-driving cars, service robots and industrial collaborative robots all need the ability to understand our actions and to anticipate what we might do next. This thesis makes contributions to a key task in visual human analysis: 3D human pose estimation. This active area of research is concerned with estimating the locations of various body joints and other anatomical landmarks in three-dimensional space, based on given sensor modalities, in our case a single color image. The estimated 3D poses are useful in several downstream tasks for autonomous systems, such as pose tracking over time, action recognition and motion forecasting.

Our main contributions address challenges that typically arise in real-world (mobile) robotics applications. This includes aspects of robustness, estimating pose in the robot's own coordinate frame, and building models that can generalize "in the wild," that is, beyond a specific motion capture studio setup.

We start by addressing robustness to occlusions, and present the first detailed analysis of how occlusions deteriorate 3D human pose estimation quality. Typical methods of the time did not address or evaluate this issue, as the standard benchmark, Human3.6M, contains almost no occlusions (other than self-occlusions). We show that augmenting training images with synthetic occlusions is an effective mitigation strategy. This augmentation also turns out to have benefits beyond robustness: our analysis shows that results improve on general benchmarks without occlusions as well. This is further demonstrated through winning a pose estimation competition at ECCV 2018.

In the next part, we turn to improving robustness to truncation, as well as reconstructing pose in the absolute camera space. Truncation means that parts of the body are not visible, being outside the input image boundaries. This is common for robots that move in crowds, as people close to the camera do not fit within its field of view. In the same work, we also address estimating the human pose at the metric

scale, performing learned scale recovery. Both capabilities are a consequence of our novel formulation of volumetric body joint heatmap estimation. While prior works use so-called 2.5D heatmaps, where two dimensions correspond to image space, we decouple the prediction volume from the input image space, allowing metric-scale truncation-robust estimation of poses. By combining this with 2D heatmap estimation, we construct an end-to-end trainable architecture for absolute 3D pose estimation in camera coordinates. Absolute pose estimation stands in contrast to the root-relative problem formulation, which forms the bulk of the literature and only estimates joint locations relative to the root (central) joint of the body. Our MeTRAbs approach surpasses the state of the art of the time on multiple benchmarks (Human3.6M, MPI-INF-3DHP and MuPoTS-3D) and forms the basis of our winning entry in an ECCV 2020 competition, trained on a combination of five datasets.

Given the success of joint training on five datasets, we set out to take multi-dataset training to the next level. Our goal is to explore how far data scale can improve 3D human pose estimation quality, and therefore we assemble the largest meta-dataset reported in the literature so far, combining 28 individual pre-existing datasets. For this, we have to reconcile the different annotation formats of the datasets, as they do not label the same set of body landmarks. We propose to learn the relations between these skeleton formats by discovering latent 3D landmarks that explain the full set of keypoints, using a novel affine-combining autoencoder formulation for dimensionality reduction. We show that data scale is important and that our allows enhanced information sharing among datasets.

Finally, we explore a novel use case for volumetric prediction, analogous to the volumetric heatmaps used in the rest of the chapters. We perform rich volumetric appearance feature prediction for the human reposing task, *i.e.*, transforming an image of a person to another pose. While prior works typically used 2D features, we show that this generative task also benefits from a 3D, volumetric representation that has been successful in pose estimation.

Zusammenfassung

Viele wichtige maschinelle Wahrnehmungsaufgaben beziehen den Menschen in irgendeiner Weise mit ein: Selbstfahrende Autos, Serviceroboter und kollaborierende Industrieroboter benötigen alle die Fähigkeit, unsere Handlungen zu verstehen und zu antizipieren, was wir als nächstes tun könnten. Diese Arbeit leistet einen Beitrag zu einer Schlüsselaufgabe in der visuellen Analyse von Menschen: der 3D-Körperposenschätzung. Dabei handelt es sich um ein aktives Forschungsgebiet, das sich mit der Schätzung der Positionen verschiedener Körpergelenke und anderer anatomischer Punkte im dreidimensionalen Raum befasst, basierend auf gegebenen Sensormodalitäten, in unserem Fall einem einzelnen Farbbild. Die geschätzten 3D-Posen sind in verschiedenen weiteren Komponenten autonomer Systeme anwendbar, wie z.B. Pose-Tracking über die Zeit, Handlungserkennung und Bewegungsvorhersage.

Unsere Hauptbeiträge befassen sich mit Herausforderungen, die typischerweise in realen (mobilen) Robotikanwendungen auftreten. Dazu gehören Aspekte der Robustheit, der Schätzung der Pose im eigenen Koordinatensystem des Roboters und der Erstellung von Modellen, die sich "in the wild" verallgemeinern lassen, d.h. jenseits eines spezifischen Studios für Bewegungserfassung.

Wir beginnen mit der Robustheit gegenüber Verdeckungen und präsentieren die erste detaillierte Analyse, wie Verdeckungen die Qualität der menschlichen 3D-Körperposenschätzung verschlechtern. Typische Methoden der damaligen Zeit haben dieses Problem nicht behandelt oder ausgewertet, da der Standard-Benchmark Human3.6M fast keine Verdeckungen (außer Selbstverdeckungen) enthält. Wir zeigen, dass die Anreicherung von Trainingsbildern mit synthetischen Verdeckungen eine wirksame Abhilfestrategie darstellt. Diese Erweiterung hat auch Vorteile, die über die Robustheit hinausgehen: Unsere Analyse zeigt, dass sich die Ergebnisse auch bei allgemeinen Benchmarks ohne Verdeckungen verbessern. Dies wird auch durch den Gewinn eines Posenschätzungswettbewerbs auf der ECCV 2018 demonstriert.

Im nächsten Teil wenden wir uns der Verbesserung der Robustheit gegenüber Trunkierung sowie der Rekonstruktion der Pose im absoluten Kameraraum zu. Trunkierung bedeutet, dass Teile des Körpers nicht sichtbar sind, da sie außerhalb der Grenzen des Eingabebildes liegen. Dies ist bei Robotern, die sich in Menschenmengen bewegen, häufig der Fall, da Personen in der Nähe der Kamera nicht in deren Sichtfeld passen. In derselben Arbeit befassen wir uns auch mit der Schätzung der menschlichen Pose auf der metrischen Skala und führen eine gelernte Skalenwiederherstellung durch. Beide Fähigkeiten sind eine Folge unserer neuartigen Formulierung der volumetrischen Körpergelenk-Heatmapschätzung. Während frühere Arbeiten so genannte 2,5D-Heatmaps verwenden, bei denen zwei Dimensionen dem Bildraum entsprechen, entkoppeln wir das Vorhersagevolumen vom Eingangsbildraum und ermöglichen so eine metrische, trunkierungsrobuste Schätzung der Posen. Indem wir dies mit der 2D-Heatmap-Schätzung kombinieren, konstruieren wir eine durchgängig trainierbare Architektur für die absolute 3D-Körperposenschätzung in Kamerakoordinaten.

Die absolute Posenschätzung steht im Gegensatz zur wurzelrelativen Problemformulierung, die den Großteil der Literatur ausmacht und die Gelenkpositionen nur relativ zum Wurzelgelenk (Zentralgelenk) des Körpers schätzt. Unser Ansatz übertrifft den damaligen Stand der Technik bei mehreren Benchmarks (Human3.6M, MPI-INF-3DHP und MuPoTS-3D) und bildet die Grundlage für unseren siegreichen Beitrag im ECCV 2020-Wettbewerb, trainiert auf einer Kombination von fünf Datensätzen.

Angesichts des Erfolgs unseres MeTRAbs-Modells, das auf fünf Datensätzen trainiert wurde, haben wir uns vorgenommen, das Multi-Datensatz-Training auf die nächste Stufe zu heben. Unser Ziel ist es, zu erforschen, inwieweit die Vergrößerung der Daten die Qualität der 3D-Körperposenschätzung verbessern kann. Daher stellen wir den größten Metadatensatz zusammen, der bisher in der Literatur beschrieben wurde, indem wir 28 einzelne, bereits existierende Datensätze kombinieren. Zu diesem Zweck müssen wir die unterschiedlichen Annotationsformate der Datensätze miteinander in Einklang bringen, da sie nicht dieselben Körpermerkmalen bezeichnen. Wir schlagen vor, die Beziehungen zwischen diesen Skelettformaten zu erlernen, indem wir latente 3D-Punkte entdecken, die die vollständige Menge von Keypoints erklären, indem wir eine neuartige, affin-kombinierende Autoencoder-Formulierung zur Dimensionalitätsreduktion verwenden. Wir zeigen, dass die Skalierung der Daten wichtig ist und dass unser Verfahren einen verbesserten Informationsaustausch zwischen Datensätzen ermöglicht.

Schließlich untersuchen wir einen neuen Anwendungsfall für volumetrische Vorhersagen, analog zu den volumetrischen Heatmaps, die in den übrigen Kapiteln verwendet werden. Wir führen volumetrische Merkmalschätzung für das menschliche Reposing durch, indem wir ein Bild einer Person in eine andere Pose transformieren. Während frühere Arbeiten typischerweise 2D-Merkmale verwendeten, zeigen wir, dass diese generative Aufgabe auch von einer volumetrischen 3D-Darstellung profitiert, die sich schon bei der Posenschätzung bewährt hat.

Contents

1 Introduction

1.1 Motivation

We titled this thesis *Robust and Efficient Methods in Visual 3D Human Pose Estimation*. Let us start by unpacking this, in order to clarify our motivations and aims.

Why Humans? Perceiving *people* holds special importance among the various tasks in artificial intelligence research. To be useful and safe in everyday life, intelligent systems need to understand us, our intentions and actions. Self-driving cars must stop if we are crossing the road ahead of them, indoor mobile robots need to find their way around us in the corridor, and manufacturing robots should be able to perform tasks based on our demonstrations—not to mention applications in medicine, sports, or digital media curation. The object category of humans has therefore received special attention in AI research. But despite decades of research and astonishing progress especially in the last decade, many technical challenges remain unsolved. The fact that the brain also dedicates plenty of precious resources for processing body images (Peelen and Downing, 2007; Vogels, 2022) indicates that we are indeed dealing with a difficult problem, and that computer vision systems should probably also contain specialized modules for human understanding.

Why Pose? When interpreting human images, it is important to choose the right level of abstraction for each particular use case. If a person is far away, modeling them as a point-like entity may be sufficient, but for a virtual try-on application we would want to infer a detailed, high-fidelity reconstruction of the body as well as the clothing. For many robotics applications, an intermediate level of abstraction is arguably the most relevant and actionable information source: the articulated body pose. Human poses are used in communication through gestures, for demonstrations of tasks, and they contain cues about current actions and future intents, such as walking direction or the

work phase in manufacturing. Human poses are informative in medical evaluations and athletic performance analysis, and can be used to drive the animation of virtual characters in entertainment.

Why 3D? To be most useful for further processing, the pose representation should not be confined to the image plane—it should be three-dimensional instead. While a 2D representation is inherently tied to one specific viewing angle, a 3D representation can be freely rotated to yield a view-invariant understanding of the scene (*cf*. mental rotation in humans, Shepard and Metzler, 1971). Furthermore, as robots need to act in 3D space, they also need to plan using 3D representations.

Why Visual? Cameras, images and videos are ubiquitous nowadays. While more specialized modalities (such as depth sensors) can be of use, requiring them would strongly restrict the pool of available training data and the possible application scenarios. Only requiring visual (RGB) input therefore makes methods more versatile and usable with many devices and data sources.

Why Robustness? When we deploy computer vision systems into the real world (outside curated datasets, "in the wild"), training-time assumptions can be violated, making robustness an important concern. In other words, it is not enough for algorithms to be accurate under the artificially controlled conditions of laboratory benchmarks, they need handle disturbances gracefully as well. In this thesis, we specifically investigate robustness w.r.t. partial occlusions (blockage of the line of sight) and image truncations (when only a part of the person is within the image boundaries). Occlusions arise whenever people handle objects or walk behind obstacles, and image truncation often happens when a person stands close to the camera, which is a common occurrence for robots navigating in crowds. It is important to prepare systems for such disturbances, already during their training.

Why Efficiency? Robots and other mobile devices have tight requirements for efficiency in terms of energy and costs. This makes it important to accomplish tasks with low amounts of compute power, while maintaining real-time performance.

Aims of This Thesis. In this thesis, we set out to advance the state of the art in visual 3D human pose estimation. Our guiding principles are to keep the models simple, modular and efficient, to tailor the problem formulations to the requirements of robotics applications, and to focus on learning strong representations from data instead of hand-crafted heuristics.

In recent years, we have witnessed enormous progress in computer vision and with it also in 3D human pose estimation. This is in large part thanks to harnessing the

representational power of deep convolutional neural networks (CNN). An especially powerful and intuitive way to encode 3D human pose within CNNs is via volumetric heatmaps, discrete probability distributions defined on a volumetric grid, expressing where each body landmark is likely to be located. Several of our works are within this paradigm, taking it further to achieve more robustness and better generalization, as well as to tackle broader tasks and encode richer representations, while maintaining simplicity and efficiency.

In detail, we first address the problem of *occlusion sensitivity* in 3D human pose estimation by performing the first comprehensive study of how various occlusion shapes degrade 3D pose estimation accuracy. We then propose to improve robustness through *synthetic occlusion augmentation*. Here we find that it is important to use realistic occluder objects as opposed to simple geometric shapes, and that this augmentation also improves results on non-occluded test images.

We then address the issue of learning to predict *metric-scale poses* within the volumetric heatmap paradigm. As opposed to typical "2.5D" heatmaps, our representation tackles scale recovery, an important aspect for robotics. We further show that our approach allows for *truncation-robust* recovery of human poses as well.

We harness this novel representation to construct an end-to-end trained *absolute* 3D human pose estimation architecture, where the 3D location of each person is also estimated, as opposed to the typical task definition of root-relative pose estimation. We achieve state-of-the-art results, all the while using low-resolution (8×8×8) heatmaps that are efficient to predict within real-time applications even on embedded hardware.

Next, to ensure the best in-the-wild accuracy for practical applications, we take inspiration from recent works that show the importance of large-scale data in training, and turn to *multi-dataset learning*. Since naive merging of datasets is made difficult through their different representations of the human skeleton (different sets of landmarks are annotated in each), we develop a novel approach for learning from heterogeneously labeled 3D keypoint data using our proposed *affine-combining autoencoder* formulation. This allows us to scale 3D human pose training to new heights, spanning 28 datasets and 13 million examples.

Finally, somewhat separately from the above line of works, we set out to better understand the characteristics of volumetric heatmap prediction, and propose to adapt it to more generic *volumetric feature maps*, for a finer-grained reconstruction task. Specifically, we tackle the image-generative task of human reposing (changing the human pose in a given to another desired pose), and achieve state-of-the-art scores, showing that this type of 3D processing is also useful beyond the pose estimation task.

We believe that these contributions are beneficial for advancing the science of visual 3D human understanding.

Social Impact and Ethical Considerations. As computer vision technology is maturing, more and more applications are being put into production that impact our

lives, prompting deeper considerations of ethical aspects in recent years. This holds especially for technology that can sense and perceive humans. How we record, analyze and store such data is an important emerging topic at the intersection of engineering, science, ethics, law and politics. Finding the correct solutions will require informed debate with broad participation from experts of various fields as well as the general public. It will be ultimately up to our wisdom to foster scientific and technological progress that can benefit us as individuals and society as a whole.

1.2 Contributions

We make the following contributions in this thesis:

- We perform the first systematic study on the effects of occlusions in deep learning–based 3D human pose estimation, using various different shapes (rectangles, circles, bars), as well as more realistic occluder objects added to test images. Since typical benchmarks do not model these effects, the aspect of occlusions had often been neglected in prior literature. Our analysis reveals that good benchmark performance on non-occluded images does not necessarily translate to robustness.

- We propose *synthetic occlusion augmentation* for addressing the issue of occlusion robustness. We perform a detailed analysis of what type of training-time occlusions perform best, and find it important to use realistic occluder objects for augmentation. We further discover that synthetic occlusion augmentation not only improves the robustness on occluded test images, but it also acts as an effective regularizer, and improves accuracy even on non-occluded test images. Our resulting model achieves first place in the *PoseTrack* challenge on 3D human pose estimation at the European Conference on Computer Vision in 2018.

- We introduce a novel *metric-scale truncation-robust heatmap* representation (*MeTRo*) to address two challenges simultaneously. First, this representation allows learning scale recovery from data, resulting in predictions that are made directly on a millimeter scale, instead of only in pixel scale. With this, we set a new state of the art on two commonly used 3D pose benchmarks. Second, our representation is naturally truncation-robust, since it decouples the input pixels space from the output volumetric heatmap space. Therefore, even when the person is partially outside the input image, we recover a complete skeleton estimate. This truncated scenario was previously underexplored despite its relevance to robotics, and our results significantly outperform the state of the art of the time.

- We further harness our MeTRo heatmap representation to construct an *end-to-end absolute 3D human pose estimator*, which we call *MeTRAbs*. We combine

MeTRo heatmaps with vanilla 2D heatmap estimation, and propose to use a differentiable geometric module to combine the 2D estimate with the root-relative 3D prediction to yield the absolute pose in 3D camera space. This allows us to backpropagate the gradients of an absolute loss, all the way to the backbone parameters. Through this method, we achieve state-of-the-art benchmark results, as well as first place in the *3D Poses in the Wild* (3DPW) challenge at the European Conference on Computer Vision in 2020. We furthermore show that the method is real-time-capable on embedded GPU hardware as well.

- We propose a novel method to scale fully-supervised 3D human pose estimation to a previously unexplored extreme multi-dataset regime. We tackle the so far rarely considered problem of supervising one pose estimator with heterogeneous annotations. Specifically, we propose a geometric autoencoding method for discovering the relations between multiple discrepant skeleton definitions that are used in the annotations of different publicly available datasets. Our *affine-combining autoencoder* is equivariant to relevant transformations (translation, rotation, chirality) and enables the transfer of learned relations among skeletons from the 2D image plane to the depth dimension. With this, we improve the consistency of multi-skeleton predictions and create the strongest publicly available 3D pose estimator models as measured on the 3DPW benchmark.
- We develop a method to extract rich 3D appearance features from human images, inspired by volumetric heatmap prediction. Specifically, we address the task of human reposing, which had previously been typically addressed with two-dimensional methods. We show that predicting volumetric appearance features—in analogy to the volumetric heatmaps we use in the other chapters—is effective in this task, and present state-of-the-art quantitative results on two benchmarks. Specifically, we warp the volumetric features according to the desired pose change and decode the result into an image in a generative adversarial network (GAN) framework. We show that our 3D volumetric appearance representation, as well as conditioning on the 3D target pose are important for good performance, outperforming directly comparable 2D baselines.

1.3 Structure of the Thesis

The thesis is structured as follows.

Chapter 1 introduces the overall motivation behind our research, summarizes our key contributions and outlines the structure of the thesis.

In **Chapter 2** – "Related Work," we present an overview of previous work on the task of visual 3D human analysis. We begin with a historic overview, followed by

a systematic review of the current state-of-the-art research landscape according to a broad range of aspects.

In **Chapter 3** – "Preliminaries," we cover fundamental topics that form the basis for later chapters. We start with a brief overview of deep learning, including key network architectures such as ResNet and EfficientNet that are used later in the thesis. We then discuss the fundamentals of image formation, camera models and the role of 3D geometry in these topics. We finish this chapter with an overview of the main evaluation metrics and datasets used in 3D human pose estimation.

In **Chapter 4** – "Occlusion-Robustness in 3D Human Pose Estimation," we describe our experiments on occlusion robustness in the context of 3D human pose estimation. We show that models trained on the most common academic dataset (Human3.6M) can suffer from severe occlusion sensitivity, even if they achieve excellent benchmark performance on non-occluded images. We also present an effective method to combat this problem through synthetic occlusion augmentations applied at training time. Our extensive experiments show that the shape of these synthetic occluders matter and realistic occluders are necessary to achieve the best performance. We show that such augmentations improve the prediction results even on non-occluded test images. This chapter is based on research originally presented in Sárándi *et al.* (2018a).

In **Chapter 5** – "Synthetic Occlusion Augmentation: A Case Study," we present a case study of applying synthetic occlusion augmentation within the context of the PoseTrack Challenge on 3D human pose estimation at ECCV 2018. This chapter describes our winning approach originally presented in Sárándi *et al.* (2018b).

In **Chapter 6** – "MeTRo: A Metric-Scale Truncation Robust Heatmap Representation," we describe a novel volumetric heatmap estimation method for truncation-robust metric-scale pose recovery. While prior work on volumetric heatmaps has tied the prediction volume's X and Y axes to the image space, we loosen this relationship and require the output to appear at a fixed metric scale regardless of the input zoom level. We furthermore do not fix the placement of the predicted skeleton within the prediction volume, allowing the recovery of complete skeletons even in truncated cases. We also propose a centered-striding method, adjusting the strided layers of the backbone network to perform better under the dense prediction scenario, *i.e.* when the test-time striding is decreased compared to the training time. With this method, we achieve state-of-the-art results on two 3D human pose estimation benchmarks. This chapter is based on research originally presented in Sárándi *et al.* (2020).

In **Chapter 7** – "MeTRAbs: An End-to-End Learned Absolute 3D Pose Estimator," we discuss our followup work to the previous chapter. This new method is capable of learning absolute 3D human pose estimation in an end-to-end manner, by combining

2D heatmap estimation on the one hand, and root-relative metric scale poses derived from the MeTRo representation on the other. We combine these two predictions using a differentiable, geometric reconstruction module based on either a full-perspective or a weak-perspective camera model. We evaluate our method on a multi-person absolute 3D pose estimation benchmark and achieve state-of-the-art results. We furthermore present performance experiments on desktop and embedded hardware, as well as several lightweight backbone networks, and find that our method can run efficiently on low-powered robot hardware as well. This chapter is based on research originally presented in Sárándi *et al.* (2021).

In **Chapter 8** – "Bridging Skeleton Formats via Geometric Autoencoding for Multi-Dataset Learning," we turn to multi-dataset learning. In 3D pose estimation competitions, best performing entries have used multiple datasets in combination, which raises the question how far this scaling can be pushed. Our goal in this chapter is to extend 3D pose estimation training to all currently available large-scale labeled datasets and supervise one model with them. The problem we face in this endeavor, however, is that the different datasets do not use the same skeleton format in their labels, *i.e.*, they label different sets of anatomical landmarks. Our novelty lies in devising an autoencoder-based method to discover how the different skeletons relate to each other without using any external reference such as a parametric body model. Specifically, we propose the affine-combining autoencoder equipped with constraints suitable for this task. It takes as input the union of all body joints defined across all datasets and compresses them to a lower-cardinality latent keypoint set. We show that this allows more consistent multi-skeleton prediction and better information sharing between the differently labeled data sources. By applying this method to the previously unexplored scale of 28 datasets and 13 million examples, we produce models with significantly higher prediction quality than other publicly available models, showing that dataset combinations are an effective way to introduce larger variability into the training process. This chapter is based on research originally presented in Sárándi *et al.* (2023).

In **Chapter 9** – "Reposing Humans by Warping 3D Features," we switch tasks and tackle human reposing, *i.e.*, the task of transforming an image of a person to depict them in a different pose. While prior works typically work in 2D, we take inspiration from the success of volumetric joint heatmap prediction (as used in previous chapters) and use the same paradigm to produce richer feature volumes. We then warp these feature volumes according to the desired pose change, and decode the warped features to the result image, conditioned on a 3D target pose. Our method achieves state-of-the-art quantitative results on two benchmarks and the ablations show the value in using 3D feature warping and conditioning on the 3D target pose, as opposed to doing these things in 2D. This chapter is based on research originally presented in Knoche *et al.* (2020). That paper is in turn based on Markus Knoche's master's thesis, supervised by myself and Prof. Leibe.

Finally, **Chapter 10** concludes this thesis with a recapitulation of the technical chapters and an outlook on exciting open challenges for the field.

Note: *This thesis is based on the technical contributions of my above mentioned publications (Sárándi* et al.*, 2018a,b, 2020, 2021, 2023; Knoche* et al.*, 2020). Several text passages and figures are reproduced from these articles, but new content has also been added to provide deeper insight into our approaches, including more detailed explanations, analyses and ablations.*

2 Related Work

In this chapter, we review the history and the current landscape of 3D human pose estimation research and closely related fields.

We proceed in two parts: In Section 2.1, we follow a chronological order describing major historic developments, allowing us to draw connections between different subfields that built on top of each other's progress in tandem. Then, in Section 2.2, we give an overview of the research landscape from a design-space perspective, categorizing currently important and promising categories of approaches along several axes. For where "history" ends and the present day begins, we will draw the cutoff line around the start of our work on this topic (*i.e.*, 2017-2018).

2.1 A Historical Overview

Computer vision for 3D human analysis is not a new research topic. While recent years have produced explosive progress, the field's history reaches back far, towards the very inception of computer vision as a discipline. Nested under the field of artificial intelligence, a major goal of computer vision has always been to equip machines with visual perception capacity so that they can naturally interact with and accomplish useful tasks for humans (Nilsson, 2009). This provided obvious motivation to develop algorithms to interpret images of humans. However, even if we were to disregard direct applicability, the visually complex and highly articulated nature of the human body has made it an interesting research object for benchmarking and pushing the limits of vision techniques in each era.

In the following, we review the major developments and milestones of the field. As the later chapters of this thesis are mainly concerned with single-image monocular 3D human pose estimation, we will keep the history of this task in the foreground.

2.1.1 Precursors

Even before the invention of the computer, people showed an interest in documenting and systematizing human pose for purposes such as arts, sciences, medicine or athletics. Highly abstract human figures expressing various poses already appeared in cave paintings of the Paleolithic. Klette and Tee (2008) present an overview of major accomplishments starting from the Antiquity. Of special note are the chronophotographic recordings of human and animal motion by Marey and Muybridge in the late 19th century. A formalized human motion representation was then developed by Laban (1928) for notating dance moves. In a seminal psychophysical study, Johansson (1973) attached bright markers to moving actors' body joints, and showed that even such sparse information allows observers to understand the pose semantics—an early inspiration for today's optical motion capture systems.

2.1.2 Early Algorithmic Attempts

Very early computer vision research was confined to simple worlds consisting of geometric primitives (Roberts, 1963). However, Badler (1975) already envisioned an algorithmic framework for interpreting motion (including human motion) from video. Marr and Nishihara (1976, 1978) fitted a cylinder-based hierarchical 3D human model to skeleton-like *primal sketches* (Marr, 1976) extracted from images. Badler and Smoliar (1979) described the outlines of a method to turn human motion videos into a Laban-inspired representation using stick figure, surface and volume-based human models. Rashid (1980) developed algorithms to process human keypoint motions (*moving light displays*, in the style of Johansson, 1973). O'Rourke and Badler (1980) designed a system to track 3D human pose from video, although it only worked on simple synthetic renderings, not real camera input.

Hogg's (1983) *WALKER* method could already track 3D pose in a real, albeit very simple, image sequence that showed a walking man from the side. This method fitted a cylinder-based human model to image edges obtained with the Sobel filter (Sobel and Feldman, 1968). Around the same time, Akita (1984) estimated 3D human motion from video by extracting binary contours and finding each body part with rule-based processing of the contours. Lee and Chen (1985) assumed known 2D joint positions and inferred plausible 3D poses by assuming fixed bone lengths.

2.1.3 Focus on Motion and Kinematics

Throughout the 1990s, research on 3D human analysis was focused on aspects of motion tracking, geometry and the 3D kinematics of articulation (Terzopoulos and Metaxas, 1991; Yamamoto and Koshikawa, 1991; Okawa and Hanatani, 1992; Rohr, 1994; Gavrila and Davis, 1996; Huber, 1996; Wren *et al.*, 1997; Bregler and Malik, 1998; Brand, 1999). These methods typically assumed a known first-frame pose initialization.

Single-image pose estimation was hard, since extracting rich visual features directly from images remained too difficult up until the mid-2000s. Hence, the features remained rather simple, with typical image processing pipelines starting out with some kind of binarization and continuing with operations on silhouettes or edges. For a detailed overview of this period, see the surveys by Aggarwal and Cai (1999) and Gavrila (1999).

2.1.4 Early Single-Image Pose Estimation

An early single-image, vision-based pose estimator (as opposed to video processing) is due to Rosales and Sclaroff (2000) but they predicted only 2D poses. They extracted Hu moment features (Hu, 1962) from the human silhouette, and employed multilayer perceptrons (MLP; Rosenblatt, 1962) to regress the joint locations. Barron and Kakadiaris (2000) presented a novel method to lift 2D poses into 3D, but did not tackle the image analysis part of the task.

Mori and Malik (2002) put the two stages together, resulting in an approach that could estimate 3D human pose from a single image. They estimated 2D pose using silhouette shape context matching and a lifted the result to 3D with a purely geometry-based step following Taylor (2000). Instead of a two-stage decomposition, Shakhnarovich *et al.* (2003) regressed body joint angles directly from edge direction histograms with locality-sensitive hashing and locally-weighted regression. Machine learning developments, especially kernel methods (Schölkopf and Smola, 2002) such as support vector machines (SVM; Cortes and Vapnik, 1995), fueled further progress. For example, Agarwal and Triggs (2006) used SVMs and relevance vector machines to regress directly from silhouette shape descriptors to joint angles. Such direct approaches are called *discriminative methods*, in contrast to *generative methods*, which optimize over the configuration space of a known human model (*e.g.*, "cylinder man") to match observed image evidence.

Generative (model-based) 3D approaches continued to improve as well. In a probabilistic graphical model framework, Sigal and Black (2006b) proceeded in two steps, first inferring 2D model state from silhouette and color features then learning to infer the state of a 3D tapered-cylinder human model from this 2D distribution. Going beyond such cylinder or sphere-based human models, Balan *et al.* (2007) used the higher-fidelity *SCAPE mesh model* (Anguelov *et al.*, 2005) to obtain more accurate shape and pose estimation from images.

For 2D pose estimation, most single-image methods were based on the idea of *pictorial structures*, which model the body in a *bottom-up* fashion, as individually detectable parts loosely connected by "springs." Initially proposed by Fischler and Elschlager (1973), this class of models gained renewed interest with Felzenszwalb and Huttenlocher's (2000) influential work (see also Felzenszwalb and Huttenlocher, 2005).

2D pose estimation approaches of this type include those by Ramanan and Forsyth (2003); Sigal *et al.* (2003, 2004); Hua *et al.* (2005).

2.1.5 Need for More Data

Around the mid-2000s, it became clear that a large obstacle to the further development of 3D human pose estimation was the lack of large benchmarks to consistently train, evaluate and compare the growing catalog of methods. While the spotlight often shines on methods and techniques, high-quality and large-scale data can be just as important for scientific progress (Sambasivan *et al.*, 2021).

The *HumanEva* dataset (Sigal and Black, 2006a; Sigal *et al.*, 2010) filled the void by providing synchronized multi-view videos along with about 40 000 reference 3D poses obtained through marker-based motion capture. The difficulty of the dataset was calibrated to the state-of-the-art pose estimation capabilities of the time. Correspondingly, HumanEva only depicts simple, upright poses, *e.g.*, walking, jogging or shadowboxing, making it less relevant for benchmarking today's stronger algorithms. The first large-scale dataset with more complex poses (*e.g.*, sitting, squatting, kicking or lying on the ground) appeared only in the mid-2010s (Human3.6M; Ionescu *et al.*, 2014). We give a detailed overview of these and later datasets in Section 3.5.

2.1.6 Rise of Strong Image Descriptors

The 2000s brought a wave of major developments in computer vision, which also impacted 3D human analysis. Gradient histogram–based image descriptors such as *SIFT* (Lowe, 2004) and *HOG* (Dalal and Triggs, 2005) offered robust tools to process detailed image patterns (such as texture), instead of being confined to working on silhouettes and contours. By clustering such descriptors, Csurka *et al.*'s (2004) *bag-of-visual-words* model enabled the extraction of high-level semantic content from images. An example for such visual word–based methods in 3D pose estimation is Ning *et al.* (2008).

The new techniques and the availability of HumanEva allowed 3D human pose estimation research to advance fast, with much more methods proposed than could be reviewed here, but Sarafianos *et al.*'s (2016) survey gives a good overview for the period between 2008 and 2015. Several of these works followed the pipeline of localizing 2D parts based on strong descriptors such as HOG or SIFT, and inferring the 3D configuration, for example though exemplar matching or probabilistic inference (Simo-Serra *et al.*, 2012, 2013; Wang *et al.*, 2014a).

While our review focuses on RGB methods, we have to mention the depth-based Kinect gaming sensor. Released in 2010, the Kinect made markerless 3D human pose technology known and affordable among a general audience for the first time. The Kinect (and its successors the Kinect v2 and Azure Kinect) also enabled the collection of

several large-scale datasets (Ni *et al.*, 2011; Ofli *et al.*, 2013; Göransson *et al.*, 2014; Wang *et al.*, 2014b; Shahroudy *et al.*, 2016; Alexiadis *et al.*, 2017; Liu *et al.*, 2017; Zimmermann *et al.*, 2018; Joo *et al.*, 2019), and was used extensively in many research projects (Han *et al.*, 2013). The original Kinect's underlying pose estimation algorithm (Shotton *et al.*, 2013) used decision forest learning from simple depth comparison features.

2.1.7 Deep Learning Era

The "ImageNet Moment." The year 2012 marks a watershed moment in computer vision: Krizhevsky *et al.*'s (2012) *AlexNet* convolutional neural network (CNN) decisively won the ImageNet Large Scale Visual Recognition Challenge (Russakovsky *et al.*, 2015), leading to a surge of renewed interest in artificial neural networks among the vision community. Within a few years, virtually every branch of computer vision, including human pose estimation, underwent a paradigm shift to using neural approaches.

Although neural networks have a long history, which we discuss in Section 3.1.2, several factors made this new incarnation more successful. The availability of large-scale datasets (*e.g.*, ImageNet, Deng *et al.*, 2009) and more compute power (esp. GPUs) were no doubt necessary. However, various seemingly small design choices (parameter initialization techniques, normalization layers, activation functions, optimizers, *etc.*) along with new automatic differentiation software frameworks added up to a qualitatively new paradigm, which became widely known as *deep learning* (LeCun *et al.*, 2015). This term is now also often applied retroactively to any layered or hierarchical representation learning (Chollet, 2021).

Heatmaps for 2D Pose. Toshev and Szegedy's (2014) *DeepPose* was the first deep-learning approach for human pose estimation (in 2D), using direct coordinate regression with fully connected layers at the end of a CNN. However, Jain *et al.* (2014) and Tompson *et al.* (2014) soon found that a classification-based, sliding-window formulation outperforms coordinate regression. These approaches produce belief maps (or *heatmaps*), *i.e.*, spatial output arrays, which contain high values at the likely positions for each body joint. CNNs are very effective in producing heatmaps in a sliding-window fashion, since computation can be shared between neighboring windows in the same forward pass. While the idea of generating such spatial classification output existed already in some early CNNs (Matan *et al.*, 1991; Wolf and Platt, 1993), it got popularized again in the influential *OverFeat* (Sermanet *et al.*, 2013) and *fully convolutional network* (FCN) by Long *et al.* (2015), for localization and segmentation, respectively.

We note that the methods by Jain *et al.* (2014) and Tompson *et al.* (2014) were not yet end-to-end deep learning–based. To impose plausible structure onto the joint positions (*e.g.*, plausible bones), they used probabilistic graphical models in a postprocessing step.

Early Deep 3D Pose Estimation. For 3D pose estimation, the best representation was less clear, and a major research question for several years was how to reconcile, combine or choose between heatmap prediction and coordinate regression. Li and Chan (2014) proposed a multi-head architecture with a 2D detection and a 3D regression branch. However, the direct 3D regression did not perform well, therefore early deep learning approaches such as Li *et al.*'s (2015) still used classical tools such as structured support vector machines (Tsochantaridis *et al.*, 2005) in combination with CNNs.

New Datasets. Given the transformative impact of ImageNet on object recognition, new large-scale datasets were also introduced for human pose estimation, such as MPII (Andriluka *et al.*, 2014) and COCO (Lin *et al.*, 2014) in 2D, as well as Human3.6M (Ionescu *et al.*, 2014) and later MPI-INF-3DHP (Mehta *et al.*, 2017a) in 3D.

Body Shape. The release of the *Skinned Multi-Person Linear* (SMPL; Loper *et al.*, 2015) parametric 3D mesh model reinvigorated research into visual body shape estimation alongside pose estimation and remains widely used to this day.

Multi-Stage Iterative Refinement Architectures. Carreira *et al.* (2016) proposed an *iterative error feedback* mechanism for 2D pose estimation, and the idea of stepwise refining pose predictions remained a major theme in later research. Wei *et al.*'s (2016) *convolutional pose machine* also used an iterative, multi-stage approach but represented the predictions after each stage as refined heatmaps instead of numerical offsets. Newell *et al.* (2016) devised a similar multi-stage architecture, the *stacked hourglass network*, but also incorporated ideas from encoder–decoder segmentation architectures (Noh *et al.*, 2015; Ronneberger *et al.*, 2015; Badrinarayanan *et al.*, 2017). These methods relied fully on end-to-end deep learning and no longer needed additional graphical models or similar postprocessing techniques.

Multi-Person Strategies. Given the successes in deep single-person pose estimation, many researchers turned to the topic of multi-person estimation, where multiple poses need to be estimated in one image. *DeepCut* (Pishchulin *et al.*, 2016) and *DeeperCut* (Insafutdinov *et al.*, 2016) formulated 2D multi-person pose estimation in two steps: a part candidate detection step, and a grouping step via integer linear programming to enforce geometrically consistent results.

Cao *et al.*'s (2017) *OpenPose* introduced *part affinity fields* (PAF) for multi-person 2D pose and became a de facto standard off-the-shelf tool for 2D pose estimation in downstream research. PAFs are per-bone vector fields that point towards neighboring joints of the same person.

Newell *et al.* (2017) proposed to predict *associative embedding* tags at each joint position and to cluster these to find which ones belong to the same person.

He *et al.* (2017) extended the *Faster R-CNN* (Ren *et al.*, 2015) object detector to *Mask R-CNN*, which could predict segmentation masks and 2D poses for multiple people.

In multi-person 3D pose estimation, *LCR-Net* (Rogez *et al.*, 2017, 2019) was also inspired by Faster R-CNN, and extended the anchor box refinement idea with anchor pose refinement.

Progress in 3D Pose. Tekin *et al.* (2017) fused a 2D heatmap prediction stream's output into the 3D pose regression stream to exploit the uncertainty information within the heatmaps and reap some of the benefits of heatmap representations for 3D estimation.

Zhou *et al.* (2017) introduced the idea of mixed-batch 2D/3D weakly-supervised training, where diverse, in-the-wild 2D data from MPII could be seamlessly blended with studio-based 3D data from Human3.6M.

Sun *et al.* (2017) argued for regression-based methods and showed that regression can achieve competitive results for both 2D and 3D estimation. For this, they use bone vectors as regression targets instead of joint positions, as well as a "compositional" loss, which supervises displacement vectors between joint pairs.

Martinez *et al.* (2017b) showed that lifting 2D poses to 3D through an MLP can be a surprisingly strong baseline, inspiring a line of further lifting-based methods.

Volumetric Heatmaps. Pavlakos *et al.* (2017) took the opposite approach and introduced *volumetric heatmap* prediction for 3D pose, based on the success of heatmaps in 2D. They used a stacked hourglass architecture with a coarse-to-fine strategy, each hourglass stage predicting finer-grained heatmaps along the depth dimension. In 2D, the benefit of heatmaps is clearer, as it exploits the convolutional sliding-window structure, performing a set of localized binary classifications of whether the joint is at the center of the receptive field or not. This automatically establishes a direct correspondence to the location, while such a correspondence must be learned in regression methods. However, it is less well understood why the heatmap construct also works well for the depth axis. Recent research by Stewart *et al.* (2022) on why classification tends to works better than regression (the "binning phenomenon") may hold clues to this.

Mehta *et al.* (2017b) proposed *VNect* (a video-based analog of Kinect), which uses 2D heatmaps along with novel *location maps* that contain X, Y, Z values for 3D joint coordinates, which need to be read out at 2D heatmap peaks to assemble the 3D pose. Mehta *et al.* (2018) extended this approach to *occlusion-robust pose maps*, where the 3D coordinates can also be read out from different 2D joint locations.

Sun *et al.* (2018a) (concurrently with Nibali *et al.*, 2018 and Luvizon *et al.*, 2018) introduced integral regression, also known as *soft-argmax*, for 2D and 3D pose estimation. Previously, soft-argmax was used in robotic policy learning (Levine *et al.*, 2016) and information retrieval (Chapelle and Wu, 2010), its main advantage being that it is

differentiable and reduces the quantization errors inherent in the hard-argmax decoding of traditional heatmaps. The method predicted volumetric heatmaps like Pavlakos *et al.* (2017), but dispensed with the iterative refinement stages, and rather adopted a simple ResNet backbone with a transposed convolutional head for increased output resolution.

From around this point on, the pace of research has become so fast that trying to weave it all into a single chronological narrative would be futile. Instead, we will continue based on topical categories.

2.2 Current Research Landscape

After the chronological overview of the field's history, let us now briefly survey the current state of the art and important areas of active research.

We can classify the methods along multiple axes based on the problem formulation (input and output definitions), the level of supervision, the stages within a method, the architecture, *etc*. We will proceed by mapping the current design space of human analysis, describing prominent example methods along the way. However, it is important to note that no such classification is complete or final, and interesting papers often break or transcend previously common categorizations.

2.2.1 Input Modalities

3D human pose estimation has been tackled using a wide range of sensors. RGB cameras are ubiquitous, simplifying training data collection and allowing wide deployability without needing special sensors at inference time. However, inferring 3D from RGB is ambiguous, and therefore depth cameras are also often used (Zimmermann *et al.*, 2018; Bashirov *et al.*, 2021), especially in robotics. Images are typically taken from a third-person perspective, but there is increasing research interest in using wearable, egocentric cameras mounted on the head (Jiang and Grauman, 2017; Xu *et al.*, 2019; Tome *et al.*, 2020) or the wrist (Lim *et al.*, 2022), since these do not require fitting a room with cameras in advance. Inertial measurement units (IMU) attached to the body can also be useful when the line of sight is blocked or cameras are difficult to calibrate or synchronize (Pons-Moll *et al.*, 2010; von Marcard *et al.*, 2018). 3D human pose estimation has also been addressed with LiDAR point clouds (Cong *et al.*, 2022; Li *et al.*, 2022b; Wu *et al.*, 2022), typically for automotive use cases. Other types of sensors used include tactile carpet signals (Luo *et al.*, 2021), polarization images (Zou *et al.*, 2020a), or even WiFi radio (Zhao *et al.*, 2018; Geng *et al.*, 2022) for through-wall human sensing.

2.2.2 Target Level of Abstraction

Depending on application requirements and available compute, we may aim at inferring various levels of details about humans. This can range from only recognizing the mere presence of a human, through bounding box detection and skeletal joint localization all the way to high-fidelity, clothed, textured mesh reconstruction with soft-tissue deformations, *etc*.

Body Joints. The term *3D human pose estimation* typically refers to estimating a list of 3D coordinates for major body joints such as shoulders, knees, *etc*. This representation abstracts away appearance details of the person such as body shape and clothes, and this can be useful for applications such as pedestrian motion forecasting, where appearance details are immaterial to the task.

In this representation, orientations are typically not reconstructed around the main axes of the bones. *E.g*., from the elbow and wrist joint positions it is not possible to tell the twist of the wrist (supination *vs*. pronation). Fisch and Clark (2021) tackle this ambiguity by estimating virtual *orientation keypoint* positions in addition to the body joint positions.

Geometric Parts. Representing the human body as a connected set of cylinder-like geometric primitives was a common approach historically (Marr and Nishihara, 1976; Sigal *et al.*, 2003; Sigal and Black, 2006b). Today, these have little relevance, with discriminative methods rather focusing on keypoint estimation and generative methods on more complicated parametric shape models.

Parametric Mesh Models. SCAPE (Anguelov *et al.*, 2005) is an early example of a detailed parametric body mesh model, and was already used for image-based shape and pose estimation in Balan *et al.* (2007). As opposed to the single mesh in SCAPE, a more sophisticated incarnation of part-based models is the *stitched puppet* of Zuffi and Black (2015), where the parts are made of triangle meshes. SMPL (*Skinned Linear Multi-Person Model*; Loper *et al.*, 2015) is perhaps the most influential body model in the literature and is widely used today. It consists of an artist-defined base template mesh with 6890 vertices. To model individual shape variation between people, the vertex positions can be adjusted within a high-dimensional PCA-derived shape space. It also models pose-dependent shape deformations (blend shapes) and can be posed by specifying joint rotations. Adam (Joo *et al.*, 2018) and SMPL-X (Pavlakos *et al.*, 2019) are extensions of SMPL with more details on hands and faces. STAR (Osman *et al.*, 2020) improves on SMPL by focusing on spatial locality of the blend shapes. SUPR (Osman *et al.*, 2022) models the body both as one unit and as body parts and also includes detailed feet. GHUM (Xu *et al.*, 2020a) dispenses with the linear aspect of

the SMPL family and learns a nonlinear shape space using variational autoencoders and normalizing flows.

An interesting recent direction is the estimation of mesh models in a nonparametric fashion, by modeling the vertices individually (Moon and Lee, 2020; Lin *et al.*, 2021a; Corona *et al.*, 2022). This essentially reduces body mesh estimation to the problem of keypoint estimation as in classic skeleton-based 3D human pose estimation, only with a much larger number of keypoints.

More generally, there is a converging trend between keypoint-based and parametric model–based estimation, *e.g.*, Zanfir *et al.* (2021); Wang *et al.* (2022). One problem with directly estimating body model–based pose and shape in the parameter space (*e.g.*, Kanazawa *et al.*, 2018) is the complicated nonlinear mapping between that space and where the body parts actually end up in the image. Using the more location-precise 3D keypoint estimation paradigm has been shown to help mesh estimation (Iqbal *et al.*, 2021). To ensure better matching to image evidence, many mesh estimation methods perform iterative optimization at test time, which was very expensive in earlier methods, *e.g.*, 1 minute per image in Bogo *et al.* (2016). As opposed to such classical analytic and numerical methods, recent research has also explored learned optimization and learned inverse kinematics solvers to accelerate body model fitting (Song *et al.*, 2020; Li *et al.*, 2021b; Shetty *et al.*, 2022).

DensePose. The DensePose (Güler *et al.*, 2018) representation specifies for every human pixel which surface point of the body is depicted there in reference to a 3D body template. It can be understood as a more detailed, continuous generalization of body part segmentation with discrete labels, and can be used as an intermediate representation for fitting body models.

Volume Models. Some methods model the human occupancy with voxel-based (Trumble *et al.*, 2018; Varol *et al.*, 2018) representation or using implicit distance fields (Saito *et al.*, 2019, 2020; Mihajlovic *et al.*, 2022).

NeRF (*neural radiance fields*; Mildenhall *et al.*, 2021) representations have also been successfully adopted for detailed human representation (Bergman *et al.*, 2022; Weng *et al.*, 2022).

2.2.3 Lifting *vs.* Direct Estimation

3D pose estimation approaches can be categorized by the steps or stages involved. Historically, the most common approach was two-stage lifting, where a model first estimates keypoint locations in 2D and uses a second stage to "lift" the 2D pose to 3D (Lee and Chen, 1985; Barron and Kakadiaris, 2000; Taylor, 2000) and several early deep-learning methods took this route as well.

The lifting itself can be performed, among others, by a fully connected feed-forward net (Martinez *et al.*, 2017b; Zhao *et al.*, 2017), distance matrix regression (Moreno-Noguer, 2017), exemplar matching (Chen and Ramanan, 2017) or probabilistic principal component analysis (Tome *et al.*, 2017).

The main benefit is that the two stages can be separately trained: the first only needs 2D annotations, the second stage only needs poses but no image data. The downside is that the lifting module has no access to detailed image information, subtle shading cues, *etc*. This can make the task very ambiguous, but using an entire temporal sequence of 2D poses as input can help in resolving ambiguities (Pavllo *et al.*, 2019).

2.2.4 Regression *vs*. Heatmaps

As already discussed in Section 2.1.7, the relative merits of predicting heatmaps or coordinates directly has been an important research question, with soft-argmax (Luvizon *et al.*, 2018; Nibali *et al.*, 2018; Sun *et al.*, 2018a) providing an effective compromise. Gu *et al.* (2021) show that soft-argmax can be biased towards the center of the image, and offer a method to compensate. Li *et al.* (2021c) introduce sampling-argmax to improve the probabilistic calibration of soft-argmax heatmaps, using the Gumbel-softmax reparametrization trick (Jang *et al.*, 2017). Gu *et al.* (2022) provide detailed comparison between argmax-based heatmap estimation and soft-argmax. Yu and Tao (2021) propose to reduce quantization effects inherent in heatmap discretization via randomized rounding. With the recent popularity of Transformers in vision (see Section 2.2.14), regression-based methods have seen some success again, *e.g.*, Lin *et al.* (2021a).

2.2.5 Absolute *vs*. Root-Relative Pose

Historically, the root-relative pose representation has been most common, without estimating the root joint location in the camera space. In bottom-up volumetric heatmap methods, estimating full heatmaps for the entire space would be too expensive, so Fabbri *et al.* (2020) proposed compressed volumetric heatmaps to alleviate this issue.

More recently, the absolute pose estimation problem has been tackled with explicit depth-order reasoning (Jiang *et al.*, 2020), depth map estimation (Véges and Lőrincz, 2019, 2020a) or bounding box size (Moon *et al.*, 2019).

Although it is called "absolute" pose estimation, the poses are still relative to the camera. In case of a moving camera, this may not be ideal, since camera-space motion intertwines human motion and camera motion. Henning *et al.* (2022) propose *BodySLAM* to disentangle human and camera motion and provide pose estimates in the world coordinate system, while *GLAMR* (Yuan *et al.*, 2022a) performs a global optimization over a full video sequence to get world-space predictions.

2.2.6 Multi-Person Pose Estimation

In many practical applications, it is common to encounter multiple people in the same image, and it is important to estimate everyone's pose. There are two main strategies to do this: *top-down* and *bottom-up* processing. These terms are used in analogy to types of parsing algorithms that output a tree-structured part–whole hierarchy (Aho *et al.*, 1986), and it is also common terminology in cognitive science (Kinchla and Wolfe, 1979). In the general sense, when building representations, the top-down order means starting with some prior knowledge about the overall scene and progressively filling in details about the parts. In contrast, bottom-up means finding parts or primitive components first, and gradually merging them into larger, nested structures of a part–whole hierarchy.

Note: In some human pose estimation texts, top-down *vs*. bottom-up refers to the distinction between generative methods (optimizing over the parameters of a human shape model to match image evidence) and discriminative methods (learned mapping from image features to poses directly, *e.g.*, Sminchisescu *et al.*, 2006). That sense is unrelated to the multi-person strategy employed.

In the **top-down** multi-person strategy, the system first localizes each person using a detector (Zou *et al.*, 2019; Cao *et al.*, 2021), then performs single-person pose estimation on resized image crops around each detection separately. The main advantages are conceptual simplicity and high accuracy. Indeed, it is simple, as good detectors are readily available, and the pose estimation part only needs to tackle the simpler single-person task. Further, this strategy handles different scales (people with different apparent size in the image) very well, as the pose estimator always runs on resized crops that depict the person at a consistent size. However, when bounding boxes overlap, the task of the pose estimator can be ambiguous as it is unclear which person's pose it should estimate. One remedy can be to also feed the target person's segmentation mask to the pose estimator along with the RGB crop (Rajasegaran *et al.*, 2022), at the cost of having to run a segmentation model which is typically more compute intensive than bounding box detection. Another disadvantage is that low-level feature computation happens multiple times: once in the person detector and once for every person crop. However, it is possible to avoid such redundant processing, as done in the influential Mask R-CNN[1] (He *et al.*, 2017) approach. The main idea is to compute backbone features just once, predict object proposals (approximate detections) and then to send *cropped feature maps* (instead of cropped RGB images) to a pose prediction module. Since the backbone features already have stronger semantics, the pose-specific prediction head can be much more lightweight than a typical single-person pose estimator network that starts from RGB input.

The **bottom-up** multi-person strategy first localizes all body parts in the image without regard to who they belong to, and subsequently groups the body parts into

[1] Despite the name, Mask R-CNN can also estimate 2D human poses, not only segmentation masks.

person instances. This grouping can be based on clustering associative embedding vectors (Newell *et al.*, 2017), or by predicting and following vector fields pointing towards other parts of a person (Cao *et al.*, 2017; Papandreou *et al.*, 2018; Kreiss *et al.*, 2019). While most of this research has been on 2D pose, there are extensions to 3D, as well (Liu *et al.*, 2019a). The main advantage of the bottom-up strategy is high-speed inference even in crowded scenarios, since the computational cost is nearly independent of the number of people present in the image. It can also use broader image context around people, which is cropped away in top-down methods. However, typically bottom-up approaches suffer from lower accuracy, mainly because there can be big differences in the scale of different people. Multi-scale estimation strategies can alleviate these issues to some degree.

Single-stage (or single-shot) regression (Mehta *et al.*, 2018; Sun *et al.*, 2021; Jin *et al.*, 2022) has been proposed to alleviate a shortcoming of both top-down and bottom-up strategies, namely that they consist of two separate stages, which are typically not trained together end-to-end. In single-stage regression, whole poses can be read out from numerical coordinate maps at the person center locations. Person centers are first found at center-heatmap maxima, so in some sense two steps are still required, but efficiency is improved. The disadvantage is that there is some spatial distance between the readout location and the actual body joints, and convolutional networks have difficulties in modeling such longer-range dependencies, leading to worse accuracy. (Note that the "single-stage" terminology is sometimes also used in a different sense in 3D human pose estimation, to mean direct 3D prediction as opposed to first predicting a 2D pose and then lifting it to 3D.)

Handling overlapping person instances has been a focus of several recent methods (Guo *et al.*, 2021; Khirodkar *et al.*, 2021; Wang and Zhang, 2022).

2.2.7 Supervision Level

Requiring full 3D-labeled supervision limits the set of applicable training datasets, and therefore many weakly supervised, self-supervised and unsupervised approaches have been proposed to tap into more data sources.

Mixing 2D data with 3D is a common form of weak supervision, already introduced in Zhou *et al.* (2017). Since annotating exact depth is hard for humans, Pavlakos *et al.* (2018) perform supervision based on depth order relations between joints (such ordinal labels were also used in Pons-Moll *et al.*, 2014).

Some methods can learn 3D pose only from 2D annotations through adversarially training a lifting network to yield plausible poses (Drover *et al.*, 2018; Chen *et al.*, 2019a; Wandt *et al.*, 2022). Mirrors appearing in in-the-wild scenes can provide an opportunity for multi-view triangulation as well (Fang *et al.*, 2021; Liu *et al.*, 2021).

Kocabas *et al.* (2019) use 2D pose predictions from multi-view images, and triangulation to self-supervise a 3D pose estimator. Iqbal *et al.* (2020) use the *MannequinChallenge*

dataset (Li *et al.*, 2019c) showing people "frozen" in place while the camera moves around the scene, as a form of multi-view self-supervision using a consistency loss. Rhodin *et al.* (2018b) use novel view synthesis as an auxiliary pretraining task to obtain geometric features that can be used to regress 3D pose from little labeled data.

Joo *et al.* (2021) perform test-time exemplar fine-tuning of the network weights, to lift 2D labels to 3D pseudolabels.

Further similar works include Rhodin *et al.* (2018a); Wang *et al.* (2019a); Li *et al.* (2020); Bouazizi *et al.* (2021); Roy *et al.* (2022).

2.2.8 Kinematic Constraints and Priors

Methods often employ some form of prior knowledge about the structure of humans to resolve ambiguities and to avoid implausible predictions. Explicit priors have been especially required for weaker image representations. Furthermore, too strong assumptions can be detrimental if the extracted image-based evidence is strong already.

Akhter and Black (2015) measured pose-conditioned joint angle limits to model plausibility. Bone lengths offer another proxy for pose plausibility, and making predictions as bones can therefore exploit such priors better (Sun *et al.*, 2017). The *kinematic chain space* (KCS; Wandt *et al.*, 2018) representation takes this further and also considers the angles between bones. Pavlakos *et al.* (2019) use a variational autoencoder (VPoser) to model pose plausibility, while Tiwari *et al.* (2022) train an implicit neural distance field in pose space.

Another useful constraint is to avoid the interpenetration of different body parts of a person (Bogo *et al.*, 2016; Pavlakos *et al.*, 2019), humans and scene objects (Hassan *et al.*, 2019) as well as collisions of multiple humans (Jiang *et al.*, 2020; Fieraru *et al.*, 2021c).

Reconstructing objects and humans together can also help with obtaining a more consistent spatial arrangement of them (Zhang *et al.*, 2020a; Dabral *et al.*, 2021; Bhatnagar *et al.*, 2022).

2.2.9 Occlusion Handling

Occlusions are a major problem in 3D pose estimation. Pose estimation is usually defined in an *amodal* manner, *i.e.*, we are interested in inferring even the occluded body parts. This stands in contrast to *modal* tasks, such as typical formulations of instance segmentation, where occluded object parts are not inferred. Mehta *et al.* (2018) proposed *occlusion-robust pose maps* (ORPM), an improvement over location maps (Mehta *et al.*, 2017b), where information about occluded joints are predicted at non-occluded parts of the body. *PARE* (Kocabas *et al.*, 2021a) improves the occlusion robustness of parametric mesh estimation by replacing global average pooling with attention-weighted pooling.

In the lifting paradigm, occlusion can be modeled as missing 2D joint detections. Carissimi *et al.* (2018) fill in those missing positions based on the non-occluded ones using a denoising autoencoder. Cheng *et al.* (2019, 2020) model (self-)occlusions in video, with in a temporal convolutional lifting network.

Further approaches tackling occlusions include Wang *et al.* (2019b); Zhang *et al.* (2020b); Qammaz and Argyros (2021); Khirodkar *et al.* (2022); Liu *et al.* (2022a).

2.2.10 Motion, Tracking, Temporal Modeling

There are several aspects to modeling human motion, and several different tasks.

Overall, as we mentioned in Section 2.1, the dominant approach in 3D pose estimation before deep learning used to be some form of tracking, *i.e.*, estimating poses over time, in image sequences. With the advent of deep learning–based pose estimators, the accuracy of single-frame estimation has increased substantially, such that temporal estimation is no longer indispensable. Nevertheless, temporal reasoning is still beneficial, especially in more difficult, noisier sequences, as well as for analyzing derivatives of position over time (*e.g.*, velocity, acceleration).

Image-Based Pose Tracking. Tekin *et al.* (2016) align a temporal stack of images via motion compensation and predict their poses jointly. Sun *et al.* (2019) estimate temporal joint offsets between two neighboring frames (similar to optical flow). Girdhar *et al.* (2018) propose a 3D Mask R-CNN based on tube proposals for 2D pose tracking. *TesseTrack* (Reddy *et al.*, 2021) is an end-to-end trained method that builds a 4D spatiotemporal feature tensor to estimate 3D pose tracks. *VoxelTrack* (Zhang *et al.*, 2022) performs tracking based on volumetric feature grids. *GLAMR* (Yuan *et al.*, 2022a) performs motion infilling and global multi-person optimization for prediction in world-coordinates.

Forecasting, Inpainting, Denoising, Smoothing. Performing temporal reasoning on the level of poses (instead of images) can be a way to reduce the dimensionality of the problem, to use more lightweight architectures and speed up the learning. The individual poses may be extracted from images using a single-frame pose estimator. Pose forecasting means predicting poses after time t, based on the poses up to time t. Martinez *et al.* (2017a) and Hossain and Little (2018) use recurrent networks for forecasting (GRU and LSTM, respectively). Cao *et al.* (2020) also take into account the scene layout, while Kundu *et al.* (2020) and Wen *et al.* (2022) jointly model the motions of two people.

Pose inpainting, infilling or interpolation is the task of predicting missing poses in a sequence (forecasting can be understood as outpainting, or extrapolation). Denoising, smoothing or refinement also modifies poses that were given in the input, to make the motion more plausible. *VIBE* (Kocabas *et al.*, 2020) applies a GRU network on the

pose sequence and model plausibility through an adversarial discriminator. Other inpainting or sequence refining methods include McLaughlin and Martinez del Rincon (2018); Hernandez *et al.* (2019); Kaufmann *et al.* (2020); Véges and Lőrincz (2020b); Yuan *et al.* (2022a); Zeng *et al.* (2022). Rajasegaran *et al.* (2022) showed that tracking human pose in 3D can also help with temporally-consistent identity association (a classic focus in multi-person tracking). Pavllo *et al.* (2019) perform 2D-to-3D temporal pose lifting with atrous temporal convolutions.

2.2.11 Physical Dynamics Modeling

Using physics-based models to obtain plausible motions is a long-researched direction (Brubaker *et al.*, 2009). This means modeling body parts as having mass, and considering the forces acting on them. Physics can constrain the predicted pose sequences to exclude impossible forces and accelerations, foot skating or other artifacts.

Physics-based approaches in 3D human pose estimation from recent years include Rempe *et al.* (2020); Shimada *et al.* (2020); Xie *et al.* (2021); Yuan *et al.* (2021, 2022b). Li *et al.* (2019b) also model human–object interactions using a physics-based approach. Furthermore, some methods use gravity as a scale reference (Bieler *et al.*, 2019; Dabral *et al.*, 2021), which is another type of physics-based information.

2.2.12 Multi-View Reconstruction

To reduce the inherent ambiguity in monocular 3D inference, several methods tackle multi-view human pose estimation. The straightforward approach is to perform 2D estimation from each view and triangulate the resulting poses using basic multi-view geometry (Hartley and Zisserman, 2004).

However, such direct triangulation of points does not take into account image details, the full multi-modality and uncertainty of the 2D keypoint estimates, nor does it model a human pose prior. Iskakov *et al.* (2019) introduced learnable triangulation, where 2D feature maps are backprojected into a volumetric, voxel-based feature grid, which is further processed to yield volumetric joint heatmaps.

Pirinen *et al.* (2019) use deep reinforcement learning to select the best camera viewpoints for triangulation.

Other learned multi-view methods include Remelli *et al.* (2020); Dong *et al.* (2021); Bartol *et al.* (2022); Ye *et al.* (2022).

2.2.13 Uncertainty, Probabilistic Output and Multiple Hypotheses

A single point estimate for the location of each joint is not always sufficient as output. For example, if we want to perform temporal smoothing or tracking, it is useful to

know how much we should trust each predicted joint location. In heatmap methods typical proxies for uncertainty can be the maximum value or the heatmap entropy.

However, a single position and a scalar uncertainty may not be informative enough if the pose is ambiguous and the predictive distribution is multi-modal. Estimating multiple hypotheses can be a way to model this multi-modality (Jahangiri and Yuille, 2017; Li and Lee, 2019; Li *et al.*, 2022b; Zheng *et al.*, 2022b). However, a finite set of hypotheses does not necessarily represent the full distribution well.

Recently, there has been progress in modeling arbitrary probability distributions through normalizing flows (NF; Kobyzev *et al.*, 2020), which can also be used to model the rich and complicated multi-modal predictive distribution in pose estimation. Wehrbein *et al.* (2021) condition an NF with the 2D pose, to perform probabilistic pose lifting to 3D. Kolotouros *et al.* (2021) perform image-based 3D human mesh estimation with NFs. Other methods using NFs for 3D pose include Zanfir *et al.* (2020); Li *et al.* (2021a); Hirschorn and Avidan (2022).

Even more recently, as diffusion probabilistic models (Sohl-Dickstein *et al.*, 2015) have become the state-of-the-art in image generation (Ho *et al.*, 2020), researchers have been looking to harness diffusion models for other tasks, too. *DiffPose* (Holmquist and Wandt, 2022) and *DiffuPose* (Choi *et al.*, 2022) are pioneering these methods in pose estimation.

2.2.14 Transformers

Originally introduced in natural language processing, the Transformer architecture (Vaswani *et al.*, 2017) has seen success in computer vision tasks recently (Khan *et al.*, 2022), including in human pose estimation.

Most of these approaches still use CNNs as backbone feature extractors. Yang *et al.* (2021) perform heatmap-based 2D pose estimation with a Transformer encoder on top of a CNN. Others perform direct coordinate regression. *TokenPose* (Li *et al.*, 2021f), *PRTR* (Li *et al.*, 2021d) and *Poseur* (Mao *et al.*, 2022) use keypoint queries for 2D pose estimation, similar to how *DETR* (Carion *et al.*, 2020) uses object queries. *METRO* (Lin *et al.*, 2021a) and *Mesh Graphormer* (Lin *et al.*, 2021b) perform full 3D pose and mesh recovery, with joint and vertex queries. Wang *et al.* (2021b) perform learnable multi-view triangulation in their multi-view pose Transformer.

In contrast to the above methods that operate on CNN backbone features, Dosovitskiy *et al.* (2021) introduced the Vision Transformer (ViT), which consists exclusively of attention layers, operating on image patches. ViT has been also adapted to 2D pose estimation (Xu *et al.*, 2022) and spatiotemporal 3D human pose estimation (Zheng *et al.*, 2021).

2.2.15 Applications

3D human analysis has uses in many other, downstream research tasks such as action recognition (Wang *et al.*, 2014b; Iqbal *et al.*, 2017; Luvizon *et al.*, 2018; Liu *et al.*, 2019b; Luvizon *et al.*, 2020), as well as practical applications.

The applications include fitness (Fieraru *et al.*, 2021a), sport (Colyer *et al.*, 2018), medicine (Belagiannis *et al.*, 2016; Chen *et al.*, 2018; Bigalke *et al.*, 2021; Cornman *et al.*, 2021), surveillance (Cormier *et al.*, 2022; Hirschorn and Avidan, 2022), art analysis (Zhao *et al.*, 2022), autonomous driving (Cong *et al.*, 2022; Zanfir *et al.*, 2022; Zheng *et al.*, 2022a), human–robot interaction (Villani *et al.*, 2018; Hentout *et al.*, 2019; Robinson *et al.*, 2022; Sampieri *et al.*, 2022) and biomechanical movement science (Seethapathi *et al.*, 2019).

Approaches developed for human pose estimation have also proved useful in animal pose estimation, for pigeons (Waldmann *et al.*, 2022), dogs (Kearney *et al.*, 2020), rodents (Dunn *et al.*, 2021), large mammals (Zuffi *et al.*, 2019) and primates (Sanakoyeu *et al.*, 2020).

Further Reading

As the literature on 3D human analysis is vast, the preceding overview was necessarily only a small selection. Good resources for further reading include survey papers and textbooks.

Survey Papers. For a comprehensive review of pre–deep learning developments, we recommend the survey articles Aggarwal and Cai (1999), Gavrila (1999), Bray (2001), Moeslund and Granum (2001), Moeslund *et al.* (2006), Forsyth *et al.* (2006), Wang *et al.* (2003) and Poppe (2007).

Sarafianos *et al.* (2016) already cover early deep learning era methods. Among more recent surveys, we especially recommend the CVIU papers Chen *et al.* (2020), Desmarais *et al.* (2021) and Wang *et al.* (2021a).

Further relevant surveys are Toshpulatov *et al.* (2022), Dubey and Dixit (2022), Kumar *et al.* (2022), Lan *et al.* (2022), Liu and Mei (2022), Dang *et al.* (2019), Zhang *et al.* (2021), Ji *et al.* (2020), Manesco and Marana (2022), Munea *et al.* (2020), Perez-Sala *et al.* (2014) and Bartol *et al.* (2020). Additonally, Tian *et al.* (2022) specifically survey mesh estimation methods.

Textbooks. Several textbooks cover pre–deep learning techniques in human analysis. We specifically recommend *Visual Analysis of Humans: Looking at People* (Moeslund *et al.*, 2011, esp. Part II), *Human Motion: Understanding, Modelling, Capture, and Animation* (Rosenhahn *et al.*, 2008, esp. Chapter 8), and the *Handbook of Virtual Humans* (Magnenat-Thalmann and Thalmann, 2004, esp. Chapter 3). The book *People Watching: Social,*

Perceptual, and Neurophysiological Studies of Body Perception (Johnson and Shiffrar, 2012) discusses the neuroscientific basis for body perception, also touching upon aspects of machine perception.

3 Preliminaries

3.1 Deep Learning

Modern computer vision cannot be imagined without machine learning (ML), and particularly deep learning (DL). Progress over the last decade has especially underlined the importance of learning representations from large-scale data in a fundamentally statistical and probabilistic paradigm (Sutton, 2019).

In the following, we give a bird's eye view on what machine learning is, and how deep neural networks were developed over time. Afterwards, we discuss specific architectures and components of the networks we use throughout this thesis.

3.1.1 Machine Learning Fundamentals

Paraphrasing Mitchell (1997), machine learning is the study of algorithms whose performance improves at some given task through experience. For this, it is important to define the task (*e.g.*, object categorization, 3D human pose estimation), to concretize what the experience is (*e.g.*, image observations along with human-annotated labels) and to design quantitative measures of performance (*e.g.*, classification accuracy, percentage of correct keypoints).

The essence of machine learning is the requirement that the algorithm's performance should improve on the *overall task*, not only on the parts it has experienced already. This is called *generalization*, which requires capturing stable patterns in the data while ignoring coincidences and noise. If a method fails to capture a real pattern, it is *underfitting*, while if it captures noise or coincidence, it is *overfitting*.

Typically, machine learning algorithms are applied in two distinct phases (though the above definition does not require this). During the *training* phase, the system is adjusted to reduce its error on the training data, which constitutes the "experience" in Mitchell's terms. During the *inference* phase, the trained model is either evaluated on a held-out test set or placed into production to accomplish something useful.

What sets apart machine learning from other uses of mathematical optimization and statistics is its open ended, inductive nature, and the willingness and focus on trading off performance on the known training set in exchange for test-time performance. Machine learning can therefore be seen as the search for computational solutions to the problem of induction (Henderson, 2022): deriving general rules from a finite set of observations. This requires *inductive biases*, *i.e.*, fundamental assumptions about the nature of patterns we are seeking, without which no learning is possible (Wolpert, 1996). A core difference between machine learning model classes then lies in what inductive biases they encode. Such inductive biases include the built-in translation equivariance of convolutional neural networks and various smoothness assumptions expressed in regularization techniques to combat overfitting. A core inductive bias expressed by today's deep learning neural network architectures is that the learned concepts should build a compositional hierarchy (Goyal and Bengio, 2022).

3.1.2 A Brief Neural Network History

Deep learning can be considered synonymous with machine learning using multilayer neural networks—except with fewer biological connotations. Neural network research has a long and tumultuous history, alternating between popularity and obscurity (Anderson and Rosenfeld, 2000; Schmidhuber, 2015; Goodfellow *et al.*, 2016; Baldi, 2021).

Perceptron. Mathematical modeling of neurons started with McCulloch and Pitts (1943), and gained broader interest with Rosenblatt's (1958) *perceptron*. He also proposed multilayer perceptron variants (MLP; Rosenblatt, 1962), though effective training mechanisms were not yet known for those. Rosenblatt considered the perceptron "first and foremost a brain model, not an invention for pattern recognition"; he wanted to understand natural instead of artificial intelligence. The neuroscientific and the engineering approaches have coexisted in the field of neural networks since then as well, with some strands of research only being concerned with one or the other, and others attempting to cross-pollinate ideas between them.

Following Minsky and Papert's (1969) in-depth analysis and critique of the perceptron's capabilities, interest in neural nets declined, especially in AI research (Olazaran, 1996).

Neocognitron. Inspired by Hubel and Wiesel's (1962) discovery of "simple" and "complex" cells in the cat visual system, Fukushima (1980) developed the *neocognitron*, a shift-robust image classifier neural net, trained layerwise. The alternating simple and complex layers performed local feature extraction and pooling, respectively. The neocognitron was intended both as neuroscientific attempt at modeling the

brain (Fukushima, 1980), as well as an engineering approach for building artificial pattern recognition systems (Fukushima and Miyake, 1982).

Backpropagation. A major milestone towards effective neural net training was the development of the *error backpropagation* algorithm (Werbos, 1982; Rumelhart *et al.*, 1986). This enabled computing error gradients w.r.t. the parameters, for use in gradient descent optimization, to train the entire network at once instead of layerwise heuristics. In handwritten digit recognition, LeCun *et al.* (1989) successfully used backpropagation in training a neocognitron-like network (*i.e.*, also a "multi-stage Hubel–Wiesel architecture;" LeCun *et al.*, 2010), consisting of convolutional layers followed by an MLP. This work was motivated purely from an engineering perspective. Although not yet called so in this work, such architectures are known today as *convolutional neural networks* (CNN).

Shallow ML. During the 1990s and 2000s, mainstream computer vision used "shallow" ML methods, such as support vector machines (SVM; Boser *et al.*, 1992; Cortes and Vapnik, 1995), on top of hand-engineered feature extractors. One reason for this was the better theoretical grasp on these models (Vapnik, 1995; Schölkopf and Smola, 2002), while neural nets were often regarded as tricky to train and too heuristic (Goodfellow *et al.*, 2016).

Renewed Interest. LeCun *et al.*'s (1998) improved *LeNet-5* architecture showed that CNNs can be competitive with (and in some respects superior to) state-of-the-art SVM-based approaches in handwritten digit recognition. LeNet used convolutional, average-pooling and fully-connected layers in a *conv-pool-conv-pool-fc-fc-fc* sequence, with hyperbolic tangent (tanh) nonlinearities in between. CNN improvements continued, however, it was not until *AlexNet*'s (Krizhevsky *et al.*, 2012) decisive victory in the 2012 ImageNet Large Scale Visual Recognition Challenge (Russakovsky *et al.*, 2015) that they became a staple of computer vision. AlexNet was larger than LeNet, with 5 convolutional and 2 fully-connected (MLP) layers. Furthermore, it used the *rectified linear unit* (ReLU) activation function (Fukushima, 1969; Nair and Hinton, 2010; $\mathrm{ReLU}(x) = \max(0, x)$) instead of tanh. ReLU's main advantage over tanh is that it does not saturate (on positive input), allowing better gradient flow and helping with the *vanishing gradient problem*.

Training AlexNet was, in large part, made possible through harnessing the massive parallel computational power of *graphical processing units* (GPU), originally designed for an entirely unrelated purpose—rasterizing video game graphics. GPUs have continued to be indispensable in deep learning to this day.

Since then, numerous architectures and techniques have been designed for deep neural networks, a selection of which we discuss in the following.

3.1.3 Architectures

Over the last decade, deep learning has converged to a modular paradigm, where deep networks for particular applications use off-the-shelf, so-called *backbone networks*, designed and pretrained by organizations with access to large-scale compute for hyperparameter optimization. The backbone extracts rich features for use in subsequent, application-specific layers or modules, often called *heads*. Such reuse of building blocks originally trained for different objectives than the target application is a form of *transfer learning*. In the following, we review some important backbone architectures that are still influential today and are relevant to the rest of the thesis.

VGGNet (Simonyan and Zisserman, 2015) has 13–16 conv layers and a more uniform design, with exclusively 3×3 filter kernels, inspired by Cireșan *et al.* (2011). Interestingly, despite having been proposed more than 8 years ago, one application where VGGNet still endures is in state-of-the-art image generation models (including the influential Stable Diffusion, Rombach *et al.*, 2022), for computing perceptual losses (Johnson *et al.*, 2016; Zhang *et al.*, 2018). We also use it for this purpose in Chapter 9.

Inception (Szegedy *et al.*, 2015), a.k.a. *GoogLeNet*, breaks with the strict sequential structure of earlier networks, and performs multiple convolutions on parallel branches with different kernel sizes and fuses their results back together. With 21 levels of convolutions, training it is more difficult, requiring intermediate supervision throughout the network, attaching auxiliary classication heads branching out from different layers, giving more direct gradient flow to early layers. One application of the Inception network today is in computing the Inception Score (Salimans *et al.*, 2016), an evaluation metric for image generation methods.

ResNet (He *et al.*, 2016a) adds shortcut or *skip connections* to very deep networks, which allows for more direct gradient flow. However, He *et al.* consider it "unlikely" that the ResNet's benefit lies in avoiding the *vanishing gradient problem*, since that problem was largely avoided already with ReLU activations, better initialization (Glorot and Bengio, 2010) and batch normalization (Ioffe and Szegedy, 2015, see below). Instead, to explain the ResNet's success, Veit *et al.* (2016) propose to understand them as ensemble models.

The original ResNet does not allow direct gradient flow across the individual residual blocks. The skip connections are used only within the blocks and do not build an unimpeded end-to-end pathway. *ResNetV2* (He *et al.*, 2016b), however, reorders the activation and normalization layers, enabling direct gradient flow throughout the network.

Besides V1 and V2, a so-called "ResNetV1.5" is also often used. This is a slight modification of ResNetV1. In V1, the striding in the bottleneck blocks happens on the 1×1 conv before the 3×3, while in V1.5 the stride is on the 3×3 conv. The PyTorch framework, for example uses this V1.5 variant for their main ResNet implementation, while TensorFlow uses the original V1 formulation.

For ImageNet training, three main concrete ResNet architectures are used, both in V1 and V2: ResNet-50, ResNet-101 and ResNet-152, with the numbers indicating the number of layers in each configuration. In the technical chapers, we make extensive use of these ResNet architectures, especially ResNet-50.

MobileNets (Howard *et al.*, 2017) are efficient convolutional architectures designed for use in mobile phones and other embedded devices. They employ *depthwise separable* convolutional modules (Sifre, 2014), which consist of a sequence of two convolutions: one depthwise and one pointwise. Depthwise convolution does not allow inter-channel interaction and filters each channel separately, while pointwise (a.k.a. 1×1) does the opposite: it allows the intermixing of channels but no spatial interactions. According to the experiments by Howard *et al.* (2017), this factorization provides a drastic reduction in computational cost (8–9x) and parameter count (6–7x), while empirically only reducing accuracy by 1% on ImageNet.

MobileNetV2 (Sandler *et al.*, 2018) introduces *inverted residual blocks*. While ResNet's residual blocks perform 3×3 convolutions on feature tensors whose channel dimension is first reduced via 1×1 convolution then expanded back, MobileNetV2 inverts this. Now the number of channels is *increased* before the 3×3 convolution, and the result is compressed to lower dimensionality afterwards. This remains efficient because the 3×3 convolution is done depthwise, as in MobileNetV1.

MobileNetV3 (Howard *et al.*, 2019) is the result of neural architecture search (Zoph and Le, 2017; Elsken *et al.*, 2019; Ren *et al.*, 2022) (NAS). NAS is a part of the automated machine learning (AutoML) paradigm (Hutter *et al.*, 2019) and can be seen as a next step towards learning more from data: as deep learning replaced hand-engineered features, NAS is supposed to obviate the need for hand-engineering neural architectures as well. However, most NAS approaches only learn how to best arrange a set of known modules, for example MobileNetV3 employs squeeze-and-excitation modules (Hu *et al.*, 2018), a module that had to be hand-designed first.

EfficientNet. Many neural net architectures are actually *families* of architectures, such as ResNet-50, 101 and 152 with increasing computational cost and accuracy. Tan and Le (2019) introduce a better way to design these architecture families, through *compound scaling*, where depth, width and spatial resolution are scaled jointly. Besides producing better-scaled ResNet and MobileNet families, they also introduce a new family via neural architecture search, the *EfficientNet*, optimized for number of floating-point operation (FLOPS).

EfficientNetV2 (Tan and Le, 2021) is again a NAS-based architecture, but optimized for training speed instead of FLOPS. FLOPS often do not correspond well with actual speed, since some operations have better dedicated hardware support than others. Another change in V2 is that some of the depthwise separable convolutional blocks are replaced by vanilla 3×3 convolutions, as these operations have better built-in hardware

support in tensor processing units (TPUs), and TensorCores of GPUs. The uniform scaling rule of V1 is also changed, and V2 scales later stages more strongly.

When comparing EfficientNets to ResNets, one important difference is that EfficientNets contain *squeeze-and-excitation* modules (Hu *et al.*, 2018), which enable a weak form of information exchange between spatially distant locations in the feature maps.

In this thesis, we apply EfficientNetV2 in Chapter 8 and found it to work well in 3D human pose estimation.

3.1.4 Normalization Layers

BatchNorm. Ioffe and Szegedy (2015) introduced *batch normalization* (BatchNorm) layers, first for the Inception architecture. While input data normalization is an old and established technique, the novelty of BatchNorm lies in normalizing the *internal* activations of a neural network. During training, this is done across all examples of a minibatch but separately for each feature map. At inference time, the normalization is done with stored statistics.

Why BatchNorm works so well is not fully understood yet. The original motivation was to reduce the *internal covariate shift* during training. This refers to the phenomenon that, as each layer is trained, the statistical distribution of the activations may drift, forcing the next layer to constantly "play catch-up" and adapt to it. Santurkar *et al.* (2018) challenge this reasoning, and rather argue that the improvements are due to a smoother loss landscape induced by the BatchNorm layers. Luo *et al.* (2019) understand it in terms of an implicit regularization effect.

The main disadvantages of BatchNorm are the discrepancy between the training-time and the inference-time behavior of the layer, as well as the complex cross-example interactions that may require more careful minibatch sampling strategies. BatchNorm inspired further normalization techniques that address these shortcomings, *e.g.*, batch renormalization (Ioffe, 2017), which addresses the train-test discrepancy.

While BatchNorm still remains a very effective and popular technique today, there are efforts to replace it with better-behaved methods such as LayerNorm (Ba *et al.*, 2016), *e.g.*, in the Vision Transformer (ViT; Dosovitskiy *et al.*, 2021) and the ConvNeXt architecture (Liu *et al.*, 2022b). There are also works that seek to eliminate the need for any normalization layers at all (Shao *et al.*, 2020; Brock *et al.*, 2021).

GroupNorm. To avoid the cross-example interactions of BatchNorm, group normalization (GroupNorm; Wu and He, 2018) performs the normalization only within one example's features. The feature channels are split up into multiple groups and normalization happens across different feature channels of the same group, as well as spatially. This has been shown to perform better in the small-batch regime, and we also apply it in Chapter 9 for this reason. We have also successfully applied GroupNorm for person-centric multi-task learning (Pfeiffer *et al.*, 2019).

Ghost BatchNorm. Similar to the grouping of channels in GroupNorm, the technique of *ghost batch normalization* (Hoffer *et al.*, 2017) splits up a minibatch and performs BatchNorm only within each chunk. While originally developed for large-batch training, Summers and Dinneen (2020) also effectively apply it in the small and medium-batch regime as a regularizer. In Chapter 8, we use it in the context of multi-dataset learning of 3D human pose estimation.

3.1.5 Optimizers

The typical way to optimize deep neural networks is by gradient descent (Cauchy, 1847). That is, the network parameters θ are updated by first computing the gradient of the loss function $\mathcal{L}(\theta)$ w.r.t. θ (using backpropagation) and taking a step in the negative gradient direction (direction of steepest descent). The size of the step is called the learning rate η.

When the loss function is evaluated by taking into account the whole training dataset, the algorithm is called *batch* gradient descent. If it only takes into account one random example from the dataset, then it is referred to as *stochastic* gradient descent (SGD). However, the typical way to train deep nets is *minibatch* stochastic gradient descent, where the loss is computed on a smaller but non singleton subset of the examples. Since using minibatches (often also just called "batches") is so common, SGD without further qualifiers typically means minibatch SGD.

Vanilla SGD chooses its update direction purely based on the current minibatch, which can lead to undesired oscillations. The *momentum* technique (Polyak, 1964) improves on this by using an exponential moving average of the gradient history, such that oscillating gradients are smoothed out and steps in such directions become smaller.

Weight decay (Hanson and Pratt, 1988) furthermore subtracts a certain fraction of the previous weight value. In SGD, this is equivalent to ℓ_2 regularization, *i.e.*, adding a squared penalty on the weights to the loss function.

AlexNet (Krizhevsky *et al.*, 2012) was trained using SGD with momentum and weight decay. Since then, more sophisticated adaptive optimizers have been invented. A detailed overview is provided in Ruder (2016).

RMSprop. One disadvantage of SGD is that it handles every parameter the same way, even though they may have very different natural scales. *RMSprop* (Tieleman *et al.*, 2012) addresses this by rescaling the gradients by their exponentially decaying root mean square (RMS) values. This way, the optimization process becomes invariant to a rescaling of the parameters.

Adam. Bringing together the benefits of momentum and RMSprop, Kingma and Ba (2015) propose *Adam* (adaptive moment estimation). Adam keeps track of exponentially

decaying averages of the gradients (like momentum) as well as the squared gradients (like RMSprop). Additionally, these averages are bias-corrected to take into account the fact that they are initialized at zero, and therefore are biased towards zero in the initial phase of training.

AdamW. Researchers often combine multiple techniques together, but this can lead to unexpected interactions. As Loshchilov and Hutter (2019) show, L_2 regularization and weight decay are no longer equivalent when using Adam. They find that weight decay achieves better results than ℓ_2 regularization, and it allows the weight decay factor to be selected independently of the learning rate. This weight-decayed version of Adam is called AdamW. In our earlier work presented in this thesis, we use the Adam optimizer, and later on we switch to AdamW.

Choosing a "best" among these optimizers is difficult, since the choice of hyperparameters can greatly affect their performance. Sivaprasad *et al.* (2020) perform a systematic study using the same hyperparameter optimization budget for each, and find Adam to perform best in their experimental setup. This, however, does not necessarily generalize to every task and dataset.

3.1.6 Further Reading

Machine Learning Books. For details on machine learning we refer the reader to the textbooks by Murphy (2022), Alpaydin (2021), Hastie *et al.* (2009), Bishop (2006), Mitchell (1997) and Shalev-Shwartz and Ben-David (2014).

Deep Learning Books. Specifically for deep learning, the books by Goodfellow *et al.* (2016), Baldi (2021), Chollet (2021) and Prince (2022) can be recommended.

3.2 Geometry of Image Formation

There is a school of thought that understands computer vision as inverse graphics (Grenander, 1978), *i.e.*, figuring out what are the objects, lights, materials, *etc*. in the scene that gave rise to the image. Whether this view truly encompasses all of vision—including higher-level semantics, saliency, *etc*.—is debatable. Nevertheless, understanding how the 3D world gets mapped to our input image (*i.e.*, the forward process) is certainly important for the kind of vision that aims to infer 3D structure (an inverse process), *e.g.*, 3D human pose estimation.

Here, we summarize the most important pieces of background knowledge about the 3D geometry of image formation, which underlie the algorithms presented in the technical chapters of this thesis.

More detailed treatment of this topic can be found in Ma *et al.* (2004) and Hartley and Zisserman (2004).

3.2.1 Pinhole Cameras

The simplest perspective camera model is the *pinhole camera model*. Its *extrinsic* parameters describe the camera's location and orientation in the world coordinate system. In contrast, the *intrinsic* parameters specify where each incoming ray of light ends up on the final pixel grid.

We use the the *homogeneous coordinate* representation of points, since this allows both rotations and translations to be expressed as matrix multiplications. Let us denote conversion to and from the homogeneous representation (in both 2D and 3D) with Π^{-1} and Π, respectively:

$$\Pi^{-1}\left(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}\right) = \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \quad \Pi\left(\begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix}\right) = \begin{bmatrix} X/W \\ Y/W \\ Z/W \end{bmatrix}, \tag{3.1}$$

$$\Pi^{-1}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad \Pi\left(\begin{bmatrix} x \\ y \\ w \end{bmatrix}\right) = \begin{bmatrix} x/w \\ y/w \end{bmatrix}. \tag{3.2}$$

$$\tag{3.3}$$

The projection operator Π divides by the last coordinate and drops it—this is perspective projection with the last coordinate understood as the depth. The unprojection operator Π^{-1} increases the dimensionality by one and places the point at unit depth along the new dimension.

The world point $\mathbf{P} = (X, Y, Z)^T$ gets transformed to the image point $\mathbf{p} = (x, y)^T$ as

$$\mathbf{p} = \Pi\left(K\left[\begin{array}{c|c} R & \mathbf{t} \end{array}\right]\Pi^{-1}(\mathbf{P})\right), \tag{3.4}$$

where $R \in \mathrm{SO}(3)$ is a rotation matrix expressing the camera orientation, $\mathbf{t} \in \mathbb{R}^3$ is a translation vector, and $K \in \mathbb{R}^{3\times3}$ is the intrinsic matrix.

The intrinsic matrix K is built from the horizontal and vertical focal lengths f_x, f_y (sometimes decomposed into optical focal length, sensor size and resolution), the skew s and the principal point coordinates c_x and c_y as

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.5}$$

3.2.2 Lens Distortion

Real cameras with lenses introduce additional distortions (*e.g.*, radial and tangential) compared to the ideal pinhole camera model. This is modeled as a nonlinear transformation φ, before applying the intrinsic matrix K.

$$\mathbf{p} = \Pi\left(K\,\Pi^{-1}\left(\varphi\left(\Pi\left(\left[\begin{array}{c|c} R & \mathbf{t} \end{array}\right]\Pi^{-1}\left(\mathbf{P}\right)\right)\right)\right)\right), \tag{3.6}$$

These intermediate two-dimensional coordinates before applying K are also called *normalized image coordinates*, and are importantly independent of the focal length.

In the 12-parameter formulation used by the OpenCV library (Bradski, 2000), the distortion function $\varphi(\cdot)$ uses the following coefficients: six radial coefficients $k_1, k_2, k_3, k_4, k_5, k_6$, two tangential coefficients p_1, p_2, as well as four thin prism coefficients s_1, s_2, s_3, s_4. The distorted point is calculated as follows:

$$r = \left\|\mathbf{p}\right\|, \tag{3.7}$$

$$a = \frac{1 + k_1 r^2 + k_2 r^4 + k_3 r^6}{1 + k_4 r^2 + k_5 r^4 + k_6 r^6}, \tag{3.8}$$

$$b = 2\left[p_2, p_1\right]\mathbf{p}, \tag{3.9}$$

$$\mathbf{c} = r^4\left[s_2, s_4\right] + r^2\left[p_2 + s_1, p_1 + s_3\right], \tag{3.10}$$

$$\varphi\left(\mathbf{p}\right) = (a + b)\mathbf{p} + \mathbf{c} \tag{3.11}$$

There is no closed-form formula for the inverse operation φ^{-1}, but it can be approximated with numerical, iterative algorithms. For example, if $\tilde{\mathbf{p}}$ is the distorted point, we can approximate $\varphi^{-1}\left(\tilde{\mathbf{p}}\right)$ by an iterative sequence of estimates starting at $\mathbf{p}_0 = \tilde{\mathbf{p}}$, then applying a rearranged version of (3.11):

$$\mathbf{p}_{n+1} = \frac{\tilde{\mathbf{p}} - \mathbf{c}_n - b_n\mathbf{p}_n}{a_n} \tag{3.12}$$

where the values $a_n, b_n, \mathbf{c}_n$ are computed from $\mathbf{p}_n$ as in (3.8)-(3.10). We found $\mathbf{p}_5 \approx \varphi^{-1}\left(\tilde{\mathbf{p}}\right)$ to already be a good approximation. This inverse computation is needed, for example, if we estimate 2D keypoints on a distorted image and need to know where those points would appear on a non-distorted image. We would also need the same computation for warping an image to simulate the effect of distortion.

While this 12-parameter formulation is usually sufficient for many practical applications, we note that more detailed distortion models methods are also available (Schops *et al.*, 2020).

3.2.3 Perspective Undistortion

Objects projected far from the principal point can appear severely distorted. In our 3D human pose estimation approaches we counter this problem by reprojecting the image

onto a virtual camera, whose orientation is changed (R') such that its principal point is at the person's bounding box center. The intrinsics are adjusted (K') to have the person fill the desired crop resolution, to remove the skew s and have equal focal length in horizontal and vertical direction. Two cameras with a shared optical center but different orientations and intrinsics yield images that map onto each other according to a *homography* $H \in \mathbb{R}^{3\times3}$:

$$\mathbf{p}' = \Pi(H\,\Pi^{-1}\mathbf{p}), \tag{3.13}$$

$$H = K'R'R^TK^{-1}, \tag{3.14}$$

where $\mathbf{p}'$ is the image point in the new camera's coordinate frame. This homography can be used to warp the input image to achieve perspective undistortion.

In the presence of lens distortions, the warping is more complicated, and the transformations needs to be computed for the coordinates of every individual pixel to get a warp field, and this can be expensive. (Undistortion as a preprocessing step can therefore make sense in case of offline processing or training.)

The interpolation method and the issue of antialiasing need to be considered and evaluated when performing this kind of warping, as they can have considerable impact on the resulting image quality and hence the pose estimation accuracy as well.

If we make pose predictions based on the reprojected image of this virtually rotated camera, we need to transform the result back, using the rotation RR'^{-1}.

We have used this reprojection method starting from (Sárándi *et al.*, 2018a) but did not quantitatively evaluate its impact. Since then, Yu *et al.* (2020a) have shown that naive cropping performs significantly worse than the perspective-correct cropping.

However, all this relies on knowing the camera intrinsics. For the case that they are unknown, Kocabas *et al.* (2021b) have proposed a method to estimate them.

3.2.4 Weak *vs*. Full Perspective

The so-called weak-perspective camera model is often adopted in 3D human pose estimation. In the weak-perspective model, the points within the same object (*e.g.*, the joints of one 3D pose) are projected as if they had the same depth (Z) coordinate. In other words, there is some perspective effect because *different* objects are scaled differently, but *within* one object the perspective effects are ignored. This can be a convenient simplifying assumption because it means that, *e.g.*, moving a person twice as far away from the camera would result in exactly scaling the projected coordinates by one half.

The assumption is a reasonable approximation only if the distance of the person from the camera is much larger than the depth differences between individual body joints. Kissos *et al.* (2020) found that this often does not hold in practice, *e.g.*, on the 3DPW dataset.

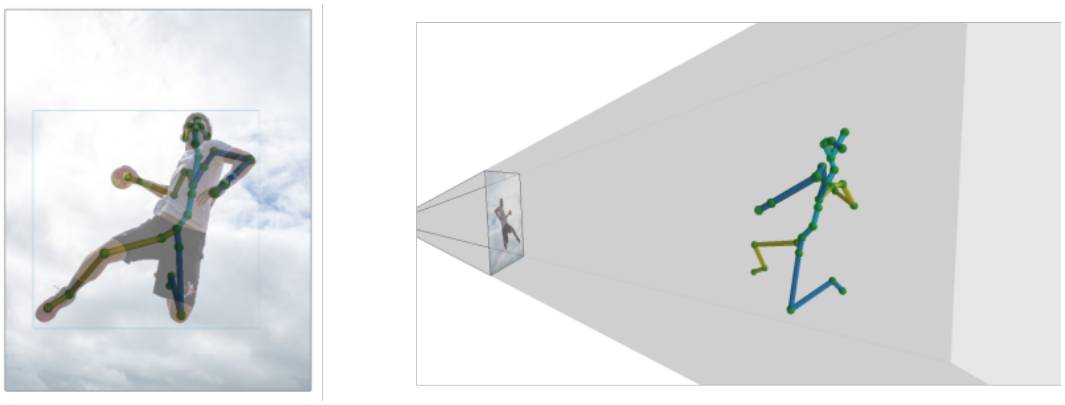


Figure 3.1: Single-image 3D human pose estimation is the computer vision task of estimating the 3D location for a set of anatomical landmarks of the target person.

3.3 Problem Formulation and Terminology

The task of 3D human pose estimation has been formalized in various ways (in terms of positions, angles, *etc*.). In this thesis, we understand it as a 3D keypoint localization task from a color image, as illustrated in Figure 3.1.

The goal is to find a good function (hypothesis) h, mapping from the space of RGB images to the space of poses, *i.e.*, $h : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^{J \times 3}$, where J is the number of body joints in the particular skeleton format. We will interchangeably refer to these as keypoints, joints or landmarks.

The joints are usually arranged into a kinematic tree, where the edges are referred to as bones or limbs. *E.g.*, we have the left lower arm bone connecting the left wrist joint to the left elbow joint.

3.4 Evaluation Metrics

To compare the quality of different human pose estimation models, we need quantitative evaluation metrics. Computing a metric at the end of an experiment helps decide whether, *e.g.*, a particular model change has led to an improvement. However, a model's behavior is much richer than a single number could represent. It is important to keep the wider application context in mind, as we rarely estimate a pose purely for its own sake and some kinds of errors may cause larger problems in some applications than others. If we need to remain application-agnostic, it is good practice to report multiple different metrics that highlight different types of errors. Nevertheless, no evaluation metric can be a substitute for careful visual inspection and testing under a wide range of real-world, application-dependent conditions.

Let J denote the number of body joints in the given skeleton format. We need to compare an estimated 3D pose $\hat{P} \in \mathbb{R}^{J \times 3}$ predicted by a model to a reference

pose $P \in \mathbb{R}^{J\times 3}$ given in the dataset. We denote the individual joint positions as $\hat{\mathbf{p}}_i$, $\mathbf{p}_i \in \mathbb{R}^3$. Many metrics are computed root-relatively. This can be understood either as subtracting the respective root joint position from both the prediction and the reference pose, or as translating the prediction towards the reference to aligning the root joints, before computing evaluation metrics. We take the first approach here, and use superscript "rr" to denote root-relative coordinates, *i.e.*, $\mathbf{p}_i^{\text{rr}} = \mathbf{p}_i - \mathbf{p}_{\text{root}}$ and $\hat{\mathbf{p}}_i^{\text{rr}} = \hat{\mathbf{p}}_i - \hat{\mathbf{p}}_{\text{root}}$. Let us further denote the Euclidean distance between prediction and reference as $d_i = \left\|\hat{\mathbf{p}}_i - \mathbf{p}_i\right\|$ and $d_i^{\text{rr}} = \left\|\hat{\mathbf{p}}_i^{\text{rr}} - \mathbf{p}_i^{\text{rr}}\right\|$, since many metrics are based on this distance.

Typically, metrics are averaged over all poses of the test set, hence we will define the following metrics with regard to one instance.

3.4.1 MPJPE

The *mean per joint position error* (MPJPE) is the average root-relative Euclidean error:

$$\text{MPJPE} = \frac{1}{J}\sum_{i=1}^{J} d_i^{\text{rr}} = \frac{1}{J}\sum_{i=1}^{J} \left\|\hat{\mathbf{p}}_i^{\text{rr}} - \mathbf{p}_i^{\text{rr}}\right\|. \tag{3.15}$$

Typically, the root joint itself (whose root-relative error is zero by definition) is *included* in the averaging. Advantages of the MPJPE include its straightforward interpretation and that it has no parameters. However, its main drawback is its sensitivity to outliers.

3.4.2 P-MPJPE

The *Procrustes-aligned mean per joint position error* (PMPJPE or PA-MPJPE) is similar to MPJPE, but it first aligns the prediction to the reference pose using the least-squares optimal Helmert transformation, *i.e.*, translation, rotation and uniform scaling (Schönemann, 1966), with

$$s^*, R^*, \mathbf{t}^* = \underset{s\in\mathbb{R},\, R\in\text{SO}(3),\, \mathbf{t}\in\mathbb{R}^3}{\arg\min} \sum_{i=1}^{J} \left\|(sR\hat{\mathbf{p}}_i + \mathbf{t}) - \mathbf{p}_i\right\|^2, \tag{3.16}$$

after which the average Euclidean error is calculated as

$$\text{P-MPJPE} = \frac{1}{J}\sum_{i=1}^{J} \left\|(s^*R^*\hat{\mathbf{p}}_i + \mathbf{t}^*) - \mathbf{p}_i\right\|. \tag{3.17}$$

P-MPJPE is tolerant of various misalignments (*e.g.*, the incorrectly "leaning" poses), and also does not evaluate the quality of scale estimation.

3.4.3 PCK

The percentage of correct keypoints (PCK) measures the fraction of joints, for which the root-relative Euclidean error is at most a certain threshold $\tau \geq 0$:

$$\mathrm{PCK}(\tau) = \frac{1}{J} \sum_{i=1}^{J} \left[d_i^{\mathrm{rr}} \leq \tau \right] = \frac{1}{J} \sum_{i=1}^{J} \left[\left\| \hat{\mathbf{p}}_i^{\mathrm{rr}} - \mathbf{p}_i^{\mathrm{rr}} \right\| \leq \tau \right], \tag{3.18}$$

using the Iverson bracket notation $[x]$ that yields 1 if x is true and 0 otherwise. It is often denoted as *e.g.*, PCK@150mm for $\tau = 150\,\mathrm{mm}$.

The main advantage is robustness to outlier errors. However, it requires the choice of a specific threshold τ, and hence it can be more informative to plot PCK as a curve over a range of thresholds. Furthermore, it does not reward more accurate predictions than τ, which may be an advantage or a disadvantage in different settings. When the reference poses are themselves not very precise, trying to match them beyond a certain error is futile, therefore the evaluation metric need not be sensitive below τ error. This is the original motivation for introducing it in Mehta *et al.* (2017a). Earlier, Ionescu *et al.* (2014) defined the opposite of this metric, called mean per joint localization error (MPJLE), which counts the fraction of joints with *higher* error than the threshold. However, over time, the PCK has seen more adoption in the literature.

3.4.4 AUC

The *area under the curve* (AUC) is the average PCK as the threshold ranges from 0 to τ. It can be defined using the following definite integral:

$$\mathrm{AUC}(\tau) = \frac{1}{\tau} \int_0^{\tau} \mathrm{PCK}(t)\, dt. \tag{3.19}$$

AUC shares the outlier robustness property of PCK, while still rewarding error reduction beyond τ. Typically, the integral in the AUC formula is evaluated through a discrete approximation, *e.g.*, by averaging over K uniformly spaced thresholds from 0 to τ as

$$\mathrm{AUC}(\tau) \approx \frac{1}{K} \sum_{k=0}^{K-1} \mathrm{PCK}\left(\frac{k\tau}{K-1} \right), \tag{3.20}$$

whose computation requires $O(NK)$ operations. However, as we now show, there is in fact a compact formula to compute the AUC in $O(N)$ time, *without approximation or discretization of the threshold range*, as

$$\mathrm{AUC}(\tau) = \frac{1}{J} \sum_{i=1}^{J} \left[1 - \frac{d_i^{\mathrm{rr}}}{\tau} \right]_+, \tag{3.21}$$

Discretization (K)	8	16	32	64	128	∞
AUC result	57.18	57.87	58.21	58.38	58.47	58.56
Discretization error	1.38	0.68	0.35	0.18	0.09	0

Table 3.1: Comparison of AUC@150mm (%) values as computed with the approximate (3.20) and exact (3.21) formulas, in the context of experiments we will discuss in detail in Section 6.8.

where $[x]_+ = \max\{0, x\}$.

Comparing AUC values reported in different papers is difficult if authors use different (or unreported) discretization. We argue that computing the AUC exactly, without discretization, is the best way to arrive at a consistently defined metric and it is also faster to compute than the discretized version. We found no prior mention of this derivation in the pose estimation literature, despite the wide use of the AUC metric.

Derivation. Substituting (3.18) into (3.19) and then switching the order of integration and summation we get

$$\mathrm{AUC}(\tau) = \frac{1}{\tau}\int_0^{\tau} \frac{1}{N}\sum_{i=1}^{N} \left[d_i^{\mathrm{rr}} \leq t\right] dt = \tag{3.22}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\tau}\int_0^{\tau} \left[d_i^{\mathrm{rr}} \leq t\right] dt = \tag{3.23}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\tau}\left(\int_0^{\min\{d_i^{\mathrm{rr}},\tau\}} 0\, dt + \int_{\min\{d_i^{\mathrm{rr}},\tau\}}^{\tau} 1\, dt\right) = \tag{3.24}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\tau}\Big(0 + (\tau - \min\{d_i, \tau\})\Big) = \tag{3.25}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left(1 - \min\left\{\frac{d_i^{\mathrm{rr}}}{\tau}, 1\right\}\right) = \tag{3.26}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \max\left\{0, 1 - \frac{d_i^{\mathrm{rr}}}{\tau}\right\} = \frac{1}{N}\sum_{i=1}^{N} \left[1 - \frac{d_i^{\mathrm{rr}}}{\tau}\right]_+ . \tag{3.27}$$

Experimental Check. Model details will be introduced later on in the thesis, however, we find it apt to already present a quantitative evaluation of the impact of exact AUC computation compared with various levels of discretization, in Table 3.1. Overall, a difference on the order of 0.1% can be expected for different discretizations.

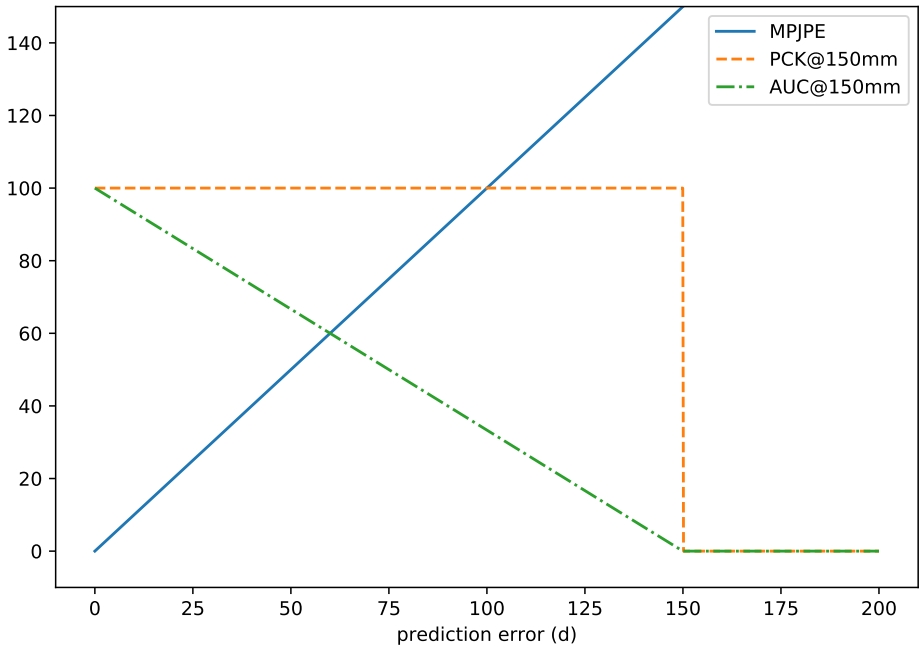


Figure 3.2: The MPJPE, PCK and AUC metrics can be all understood as averaging some function f of the jointwise errors, which we plot here.

3.4.5 Absolute Metrics

MPJPE, PCK and AUC can also be computed without root joint alignment, to evaluate absolute pose estimation quality. These are denoted as A-MPJPE, A-PCK and A-AUC, respectively:

$$\text{A-MPJPE} = \frac{1}{J}\sum_{i=1}^{J} d_i, \tag{3.28}$$

$$\text{A-PCK}(\tau) = \frac{1}{J}\sum_{i=1}^{J} [d_i \leq \tau], \tag{3.29}$$

$$\text{A-AUC}(\tau) = \frac{1}{J}\sum_{i=1}^{J} \left[1 - \frac{d_i}{\tau}\right]_+. \tag{3.30}$$

3.4.6 A Unified View

Based on the above, the MPJPE, PCK and AUC metrics all have the form $(1/J)\sum_{i=1}^{J} f(d_i)$, that is, the individual Euclidean joint errors are transformed with some function f and the results are averaged: $f_{\text{MPJPE}}(d) = d$, $f_{\text{PCK}}(d) = [d \leq \tau]$, $f_{\text{AUC}}(d) = [1 - d/\tau]_+$. Figure 3.2 shows a plot of these functions.

3.4.7 Dataset-Specific Protocols

Reproducible evaluation of a specific model is important, even if the training process has a component of randomness. Therefore, throughout the thesis, we make the effort to use the same protocol (including AUC discretization) as the official evaluation script if there is one available. There are various details that we will point out in the technical chapters, such as which joints to evaluate, how to average the results (per sequence, or per pose, *etc.*), what kind of alignment to the reference pose is used (root-alignment, rigid-only Procrustes, rigid+scale Procrustes, bone vector rescaling, *etc.*).

Specifically for AUC, discretization can be emulated with minor adjustments to the exact formula we derived above. For example, the official 3DPW evaluation script computes PCK at 1 mm intervals from 0 to 200 mm (start inclusive, end exclusive), then integrates using the `scipy.integrate.quad` function, which by default uses only 50 samples. On MPI-INF-3DHP, the official evaluation script computes PCK at 5 mm intervals from 0 to 150 mm (inclusive, 31 samples) then averages these sampled values for the AUC result.

These can be computed faster—with verified exact correspondence—as

$$f_{\mathrm{AUC}}^{\mathrm{3DPW}}(d) = \left[1 - \frac{\left\lfloor \frac{d}{200-1} \cdot 50 \right\rfloor + 0.5}{50}\right]_+ \qquad f_{\mathrm{AUC}}^{\mathrm{3DHP}}(d) = \left[1 - \frac{\left\lfloor \frac{d}{150} \cdot 30 \right\rfloor + 1}{30+1}\right]_+ . \tag{3.31}$$

3.5 Datasets

Since around the late 2000s, datasets have played a key role in the improvements we have witnessed in computer vision. This has been enabled by advances in recording technologies (*e.g.*, widely available digital cameras, better motion capture products), as well as more advanced machine learning models that can "absorb" large amounts of training data without saturating accuracy.

As research interest in human analysis has been on the rise, there are also numerous new datasets released every year, such that it becomes a challenge to even list them all. In the following we summarize some of the most important and largest ones. See Figure 3.3 for a selection of representative frames from some datasets.

3.5.1 Real 3D Datasets

Human3.6M (Ionescu *et al.*, 2014) was for long the largest publicly available 3D human pose estimation dataset and thus the main benchmark for comparing methods. The initial baseline method was derived from Ionescu *et al.* (2011).

The dataset is captured with 4 cameras in a motion capture studio and provides hardware-synchronized, calibrated RGB videos with 1000×1000 px resolution at 50 fps,

Figure 3.3: Datasets. Recent years have spurred the creation of many 3D human pose estimation datasets, here we present a representative frame from a selection of these datasets.

time-of-flight depth data, segmentation masks and person bounding boxes. Reference poses are recorded with a ten-camera Vicon motion capture setup based on optical marker tracking. Hence, the subjects wear tight-fitting but realistic clothes, as well as visible markers. Raw marker positions are not available, instead Vicon's proprietary software performs tracking in an angle-based representation and the authors calculate 3D joint positions by applying forward kinematics on the Vicon skeleton. In the dataset, a total of 11 subjects (one at a time) imitate 15 different activities such as asking for directions, talking on the phone, walking a dog or sitting down. The inclusion of such activities was a major improvement over previously available datasets such as HumanEva. In some scenes a chair is present for sitting, otherwise the capture area only includes the target person.

Two evaluation protocols are in wide use in the literature. In Protocol 1, the training subjects are S1, S5, S6, S7, S8, while S9 and S11 are used for testing according to the MPJPE metric. In Protocol 2, subjects S1, S5, S6, S7, S8, S9 are used in training and S11 in evaluation according to Procrustes-aligned MPJPE. Every 64th frame is evaluated.

According to the original protocol, the test subjects are S2, S3, S4 and S10. For these subjects, the reference poses are not publicly available (additionally, RGB data for S10 is withheld for privacy reasons). Evaluation according to this protocol can be performed through submission to an evaluation server, but most publications follow the aforementioned Protocols 1 or 2.

The main Human3.6M skeleton used for evaluation consists of 17 joints but the dataset also provides some further joint positions.

MPI-INF-3DHP (Mehta *et al.*, 2017a) is another important dataset, based on a commercial markerless motion capture system ("The Captury"). The training set consists of 8 subjects performing a diverse set of actions in a green-screen studio. Synchronized RGB videos are provided from 14 camera views with resolution 2048×2048 px and 25 fps, of which five are positioned similarly to Human3.6M (*i.e.*, at a chest height) and are often the only ones used for training. Chroma key sequences are available based on the green screen, enabling masking out and replacing the background as augmentation. Since the mocap system is markerless, the subjects have no visible markers on them unlike in Human3.6M, and they wear more general and looser-fitting clothes.

Test frames come from 3 different types of scenes: the same green-screen studio as in training, the studio with the green-screen drapes removed, as well as outdoor sequences. Each of these setups has two sequences with one subject each.

Especially the outdoor sequences make this benchmark more challenging than Human3.6M. The dataset provides two ground truth variants: unnormalized metric-space poses and "universal" (height-normalized) ones. Since its publication, the official ground truth has been updated twice, making not all published results comparable. In our experience, the first update changes scores by 1–3%, while the second one only by 0.1%, which is within experimental fluctuation, making the two latest versions comparable.

The full skeleton contains 28 joints, with 17 of them being similar to the Human3.6M joints, though not exactly equally positioned.

HumanEva (Sigal *et al.*, 2010) was the pioneer of 3D pose datasets, as we understand them today. It contains 40k poses captured using markers, along with synchronized multi-view video. While large for its time, today's datasets are orders of magnitude larger and contain more challenging poses and appearances, and hence HumanEva is mostly of historical relevance today.

CMU-Panoptic (Joo *et al.*, 2019) is a large-scale multi-person 3D pose dataset recorded in a specially constructed sphere-shaped studio. Calibrated RGB data is available from 31 high-definition cameras (1920×1080 px, 29.97 fps), 480 VGA cameras (640×480 px, 25 fps) and 10 Kinect v2 sensors (RGB and depth). There are 66 sequences available, showing dozens of people interacting in various ways, playing social games like Mafia, haggling with each other, playing musical instruments, *etc*. The reference poses were derived from candidate joints obtained from 2D predictions of a *convolutional*

pose machine (Wei *et al.*, 2016) from every camera, and a complicated association, triangulation, filtering and tracking procedure on these candidates.

3DPW (von Marcard *et al.*, 2018) is the first large-scale, *in-the-wild* 3D human pose estimation dataset, *i.e.*, recorded from free-hand moving cameras in everyday urban, indoor and nature environments instead of studios. In-the-wild 3D pose capture is difficult due to frequent occlusions and the difficulties in camera calibration and synchronization. Von Marcard *et al*. avoid this issue by using *inertial measurement units* (IMU) attached to the subjects for recovering the poses. The dataset contains multi-person sequences, with SMPL annotations given for one or two of the depicted people. The original train/validation/test split has been superseded and the dataset is now mostly used only for evaluation purposes.

MuCo-3DHP (Mehta *et al.*, 2018) is a synthetically composited multi-person dataset, based on MPI-INF-3DHP. The authors sample four random frames from 3DHP and use the green screen–based segmentation masks to paste the four people's foreground segment onto one image, according to their depth order. *MuPoTS-3D* was also proposed in Mehta *et al.* (2018), and is a mixed indoor and outdoor multi-person test set, compatible with MuCo-3DHP, consisting of 20 sequences showing people performing various actions and interactions.

AIST-Dance++ (Li *et al.*, 2021e) is a large-scale studio dataset of various dances, based on AIST-Dance (Tsuchida *et al.*, 2019), captured from 10 cameras arranged in a circle. The background is uniform white and the poses are often very challenging. The reference poses were obtained in a markerless way, via triangulation of 2D estimates and are therefore somewhat noisy.

HUMBI (Yu *et al.*, 2020b) is another large-scale markerless dataset, showing subjects perform various fairly simple dance moves and game-playing actions. The distinguishing feature of HUMBI is the large number of subjects (772) with various appearances.

HuMMan (Cai *et al.*, 2022) is one of the most recent human sensing datasets, and includes recordings of 1000 subjects, 400k sequences and 60M frames with point clouds, meshes, textures, SMPL body parameters and keypoint annotations. Besides 10 RGB-D cameras, the dataset also provides recordings from smartphone cameras.

Some Further Labeled Datasets. *Berkeley-MHAD* (Ofli *et al.*, 2013), *UMPM* (van der Aa *et al.*, 2011), *TotalCapture* (Trumble *et al.*, 2017), *BML-MoVi* (Ghorbani *et al.*, 2021), *Human4D* (Chatzitofis *et al.*, 2020) and *MADS* (Zhang *et al.*, 2017) are further marker-based 3D human pose datasets with RGB videos. *GPA* (Wang *et al.*, 2019b) and *3DOH50K* (Zhang *et al.*, 2020b) are datasets specifically focused on occlusions. *IKEA ASM* (Ben-Shabat *et al.*, 2021) shows people assembing furniture and provides markerless, triangulated reference poses. *Fit3D* (Fieraru *et al.*, 2021a), *HumanSC3D* (Fieraru *et al.*, 2021b) and *CHI3D* (Fieraru *et al.*, 2020) are three datasets

recorded in the same studio as Human3.6M, focusing on fitness actions, self-contact and close interactions, respectively. *BEHAVE* (Bhatnagar *et al.*, 2022) focuses on human–object interactions and *RICH* (Huang *et al.*, 2022) on human–scene contact. *ASPset* (Nibali *et al.*, 2021) is a large, sports-focused dataset recorded on a football pitch, and is released under a public-domain license. *3D-Yoga* (Li *et al.*, 2022a) contains 117 categories of yoga poses by 22 subjects and has multi-view RGB and 3D skeleton annotations. *CHICO* is focused on human–robot collaborative scenarios and pose forecasting.

Pseudo-Labeled Datasets. Since recording 3D-labeled datasets is difficult in the wild, pseduo-labeling 2D datasets with 3D labels can be a fruitful approach. UP-3D (Lassner *et al.*, 2017) and EFT (Joo *et al.*, 2021) are examples of such datasets.

Multi-View Non-Pose Human Datasets. As single-image (2D) pose estimation is becoming stronger, further multi-view datasets of people may also be triangulated with small effort, especially with the rise of learnable triangulation methods (Iskakov *et al.*, 2019). Multi-view datasets without (high-quality) pose labels include *NTU-RGB+D* (Liu *et al.*, 2019b), *PKU-MMD* (Liu *et al.*, 2017), *MPI08* (Baak *et al.*, 2010; Pons-Moll *et al.*, 2010), *NW-UCLA* (Wang *et al.*, 2014b), *PHPS* (Zou *et al.*, 2020b), as well as Alexiadis *et al.* (2017).

3.5.2 Synthetic 3D Datasets

Real 3D pose datasets are difficult to record, since they need complex, synchronized, multi-camera setups that are challenging to use outdoors and "in the wild." Inertial measurement units and other sensors can help with this to an extent, but challenges remain for larger scale collection.

In contrast, synthetic data generation through computer graphics techniques is a useful way to obtain virtually unlimited, diverse data, with rare appearances, and environments that would be difficult to capture in reality. Another benefit of synthetic data is the possibility of building large-scale datasets without privacy issues. Synthetic data has a long history in pose estimation, going back all the way to O'Rourke and Badler (1980), but of course graphics quality has increased immensely over time.

SURREAL (Varol *et al.*, 2017) is a synthetic dataset consisting of millions of frames of SMPL (Loper *et al.*, 2015) body meshes rendered in the Blender software, according to pose sequences from the CMU MoCap database (CMU, 2003), overlaid on random background images. The dataset provides ground-truth pose, mesh parameters, body part segmentations and depth maps. While the rendering quality is not photorealistic ("sim-to-real gap"), the dataset depicts many extreme poses such as cartwheels, many different textures, *etc.*, making it valuable as an addition to the training process.

GTA5. Several datasets have been produced using the *Grand Theft Auto V* (GTA5) video game published by Rockstar Games. GTA5 is one of highest-budget video games ever produced (McGinty, 2013), and correspondingly has a vast array of animations, characters, object models, indoor and outdoor environments, weather conditions, *etc*. that can be combined in endless ways to yield varied datasets for computer vision. The *JTA* dataset (Fabbri *et al.*, 2018) focuses on large-scale crowds of walking pedestrians. *SAIL-VOS* (Hu *et al.*, 2019) is primarily an amodal video object segmentation dataset, but also includes 3D human pose labels, from many cinematic cutscenes of the game, thereby showing more complicated poses. *GTA-IM* (Cao *et al.*, 2020) is focused on 3D human pose forecasting and everyday motions such as walking or sitting down. *GTA-Human* (Cai *et al.*, 2021) contains a large number of short sequences or randomly sampled actions, weather, characters, *etc*., with SMPL-based annotations.

3DPeople (Pumarola *et al.*, 2019) is another large-scale synthetic dataset, with 80 characters in more realistic clothing than in the case of SURREAL. The dataset provides ground-truth meshes, poses, depth, normals, body part segmentation and cloth segmentation.

AGORA (Patel *et al.*, 2021) is a much more realistic-looking synthetic dataset in terms of graphics quality, environmental conditions and clothing. The characters stem from the high-fidelity RenderPeople model repository. *SPEC* (Kocabas *et al.*, 2021b) follows a similar pipeline for data generation, focusing on camera setups with more extreme perspective distortions. 3D pose and mesh annotations are available according to the SMPL (Loper *et al.*, 2015) and SMPL-X (Pavlakos *et al.*, 2019) body models.

HSPACE (Bazavan *et al.*, 2021) is similar to AGORA in terms of realism. It is somewhat larger and uses the Google's GHUM body model (Xu *et al.*, 2020a) instead of SMPL.

PeopleSansPeople. (Ebadi *et al.*, 2021) and *PSP-HDRI+* (Ebadi *et al.*, 2022) are synthetic human datasets generated with the Unity rendering engine. In contrast to HSPACE and AGORA, the backgrounds are random images, and not realistic scenes corresponding with the humans.

3.5.3 2D Datasets

LSP (Leeds Sports Poses; Johnson and Everingham, 2010) is an early full-body 2D human pose dataset, with images collected from Flickr, depicting athletes in complex sports and acrobatic poses. Originally consisting of 1000 training and 1000 test images, LSP was later extended with 10k further images with noisier annotations (Johnson and Everingham, 2011).

Other early 2D pose datasets, such as *Buffy* (Ferrari *et al.*, 2009) and *FLIC* (*Frames Labeled In Cinema*; Sapp and Taskar, 2013), were based on TV shows and movies, and only labeled upper bodies.

MPII (Andriluka *et al.*, 2014) is a dataset with 25k training images collected from YouTube, annotated with 16 keypoints (of which typically 14 are evaluated).

MS-COCO (*Common Objects in Context*; Lin *et al.*, 2014) is a large-scale dataset for object detection, segmentation, captioning and 2D human pose estimation. With 250k person instances with keypoints, it is one of the largest available pose datasets.

PoseTrack (Andriluka *et al.*, 2018) is a large 2D human pose dataset, containing videos annotated with poses, associated with consistent IDs over time.

JackRabbot, or JRDB (Martin-Martin *et al.*, 2021) is a dataset of 54 sequences collected from a mobile robot platform moving around the Stanford University campus, recording with multiple RGB cameras and LiDAR. Originally the dataset was annotated box-level 3D object tracking, and was later extended with activity labels, and more recently with 2D pose annotations for 600k person instances (Vendrow *et al.*, 2022).

4 Occlusion-Robustness in 3D Human Pose Estimation

Occlusion is commonplace in realistic human-robot shared environments, yet its effects are not considered in standard 3D human pose estimation benchmarks. This leaves the question open: how robust are state-of-the-art 3D pose estimation methods against partial occlusions?

In this chapter, we study the effect of superimposing several types of synthetic occlusions on the Human3.6M dataset and find that a state-of-the-art method of the time, as measured on the Human3.6M benchmark, can still be sensitive even to low amounts of occlusion. Addressing this issue is key to progress in applications such as collaborative and service robotics. We take a first step in this direction by improving occlusion-robustness through training data augmentation with synthetic occlusions. This also turns out to be an effective regularizer that is beneficial even for non-occluded test cases.

This chapter is based on our publication (Sárándi *et al.*, 2018a), presented at the 2018 IROS Workshop on Robotic Co-Workers 4.0.

4.1 Overview

To collaborate with humans and to understand their actions, collaborative and service robots need the ability to reason about human pose in 3D space. An important challenge in realistic environments is that humans are often only seen partially, *e.g.*, standing behind machine parts or carrying objects in front of the body (see Figure 4.1). Robust robotics solutions need to handle such disturbances gracefully and make use of the visual cues still present in the scene to reason around the occlusion.

Although recent years have brought significant advances in 3D human pose estimation, as measured on standard computer vision benchmarks such as Hu-

Figure 4.1: Example of partial occlusions in the context of shared human-robot workspaces. Note how easily we humans can guess the rough pose of the person behind the occlusion. Can current 3D human pose estimation methods do that as well?

man3.6M (Ionescu *et al.*, 2014), the behavior of models under occlusion remains largely unexplored, as the benchmarks do not systematically model occlusion effects.

To our knowledge, we present the first systematic study of various types of test-time (synthetic) occlusions in 3D human pose estimation from a single RGB image. As we will see, ignoring the aspect of occlusions may cause model accuracy to rapidly deteriorate, even under mild occlusion levels, despite the good benchmark performance. Such sudden and unexpected failures in the robot's perception would prevent smooth and comfortable human–robot interaction and may lead to safety hazards. Furthermore, we demonstrate that simple occlusion data augmentation during training increases model robustness. This augmentation also improves performance even for non-occluded test images. Our approach is efficient and suitable for high frame-rate applications.

4.2 Related Work

4.2.1 3D Human Pose Estimation

3D human pose estimation has seen rapid progress in recent years. For a thorough overview of prior work, see Chapter 2. Current state-of-the-art methods use deep neural networks, either directly on the input image or on the output of a 2D pose estimator. Based on the sweeping success of heatmap-based representations in 2D human pose estimation (*e.g.*, Newell *et al.*, 2016), heatmaps have recently been also adopted in 3D methods, including volumetric (Pavlakos *et al.*, 2017; Sun *et al.*, 2018a) and marginal heatmaps (Nibali *et al.*, 2019).

4.2.2 Occlusions, Erasing and Copy-Pasting

In a pre-deep learning study based on silhouettes and HOG features, Huang and Yang (2009) tackled occlusions in 3D pose estimation from RGB, but their analysis was limited to walking actions and occlusions with two rectangles. Occlusion effects have also been studied in 3D pose estimation from depth input (Rafi *et al.*, 2015), where exploiting semantic information from the occluder itself was found to improve predictions.

Data augmentation by erasing a rectangular block from the input has recently been concurrently investigated under the names *Random Erasing* (Zhong *et al.*, 2020) and *Cutout* (DeVries and Taylor, 2017), for image classification, object detection, and person re-identification. Similarly, synthetically placing objects into a scene by image-level *copy-pasting* has been shown to help object detection (Dwibedi *et al.*, 2017; Georgakis *et al.*, 2017; Dvornik *et al.*, 2018). However, those methods are trained to detect these pasted objects, while in our case the task is to infer what lies behind them. Ke *et al.* (2018) augment training images for 2D human pose estimation by copying background patches over some of the body joints. Research on facial landmark localization has investigated and modeled occlusions for a long time (Burgos-Artizzu *et al.*, 2013; Ghiasi and Fowlkes, 2014), including augmenting training images with randomly pasted occluding objects (Yuen and Trivedi, 2017).

4.3 Method

In this chapter, we study the effect of occlusion on the accuracy of 3D human pose estimation. To this end, we have devised a 3D pose estimation approach that reaches state-of-the-art benchmark performance, leading us to expect that the observations drawn from our experiments also transfer to other models.

4.3.1 Architecture

We use a fully convolutional net to predict volumetric body joint heatmaps from the input RGB image, based on a ResNet-50 (He *et al.*, 2016a) backbone architecture. After discarding the global average pooling layer, we adjust the number of output channels of the ResNet to be the product of the number of joints and the number of heatmap-voxels along the depth axis. Reshaping the resulting tensor yields the volumetric heatmaps. Nominal stride and depth discretization are configured to yield heatmaps of size $16 \times 16 \times 16$ for an image of size 256×256 px. Given the volumetric heatmap, coordinate predictions are obtained using soft-argmax (Levine *et al.*, 2016; Nibali *et al.*, 2018; Sun *et al.*, 2018a). As in Pavlakos *et al.* (2017), the x and y coordinates are interpreted as image space coordinates, while z is the depth of the particular joint relative to the root (pelvis) joint depth, with the 16 voxels covering 2 meters. In order

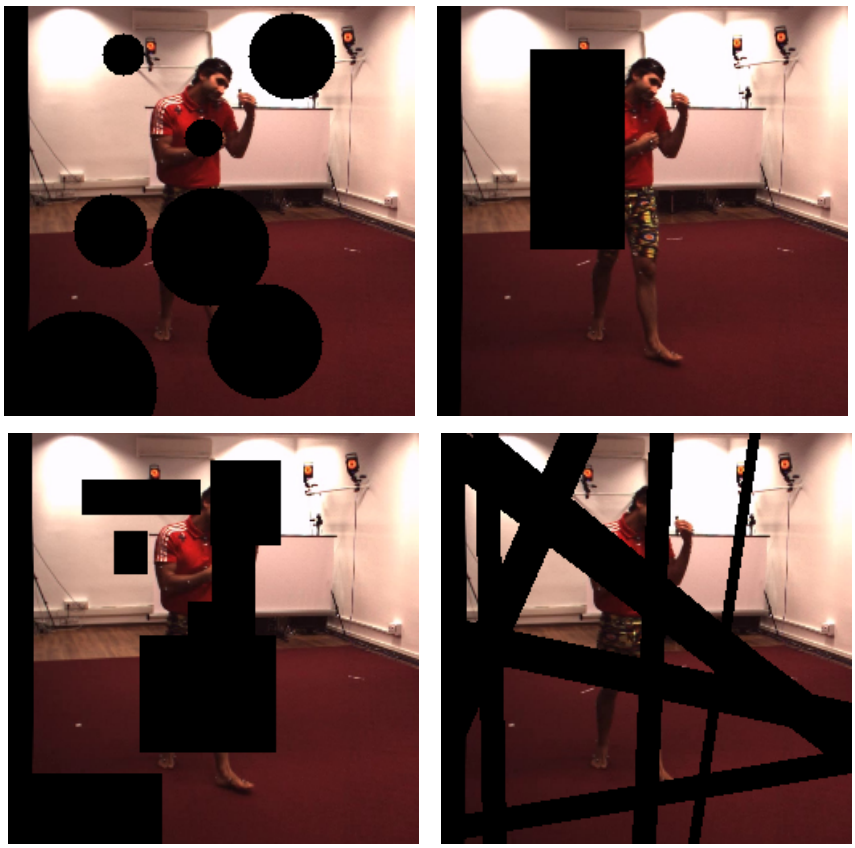


Figure 4.2: Examples of the investigated geometric occlusions: circles, a single rectangle (Zhong *et al.*, 2020), rectangles, oriented bars. See Figure 4.3 for an example with Pascal VOC objects.

to concentrate on the aspect of articulated pose, as opposed to person localization, we assume that the true root joint depth is given by an oracle at test time. The coordinates are back-projected to camera space using the known camera intrinsics. Finally, the ℓ_1 loss is computed on the predicted and ground truth 3D coordinates in camera space. Since all of the preceding operations are differentiable, the network can be trained end-to-end.

4.4 Experimental Setup

4.4.1 Dataset

Human3.6M (Ionescu *et al.*, 2014) is the largest public 3D pose estimation dataset. It contains 11 subjects imitating 15 actions in a controlled indoor environment while being recorded with 4 cameras and a motion capture system. Following the most common experimental protocol in the literature, we use five subjects (S1, S5, S6, S7, S8)

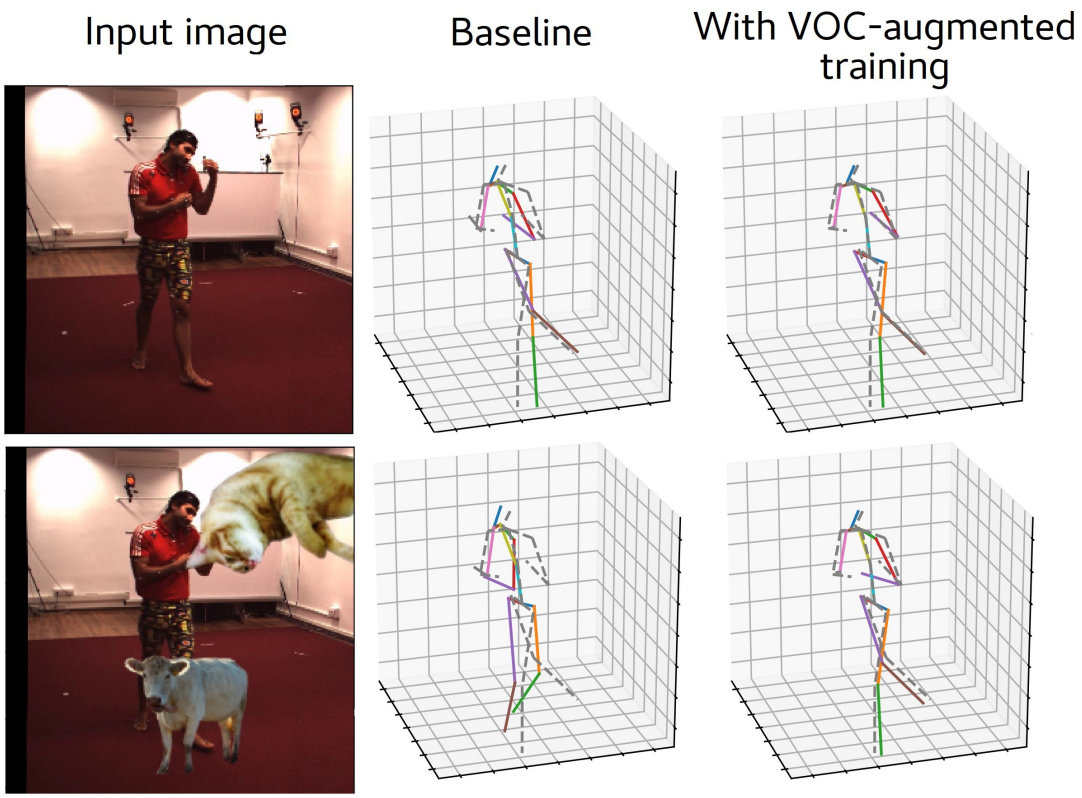


Figure 4.3: Prediction change in the presence of synthetic test-time occlusion. Ground truth is shown with grey dashed lines, predictions with colorful ones. The baseline model fails to predict the pose of the occluded limbs, while the model trained with occlusion augmentation behaves more robustly.

for training and two (S9, S11) for testing. We train action-agnostic models, as opposed to action-specific ones.

4.4.2 Data Sampling

To reduce the redundancy in training poses, we adaptively subsample the frames similarly to Mehta *et al.* (2017b), only keeping a frame when at least one body joint moves by at least 30 mm compared to the last kept frame. For the test set we follow prior work and use every 64th frame.

4.4.3 Image Preprocessing

Before feeding an image to the network, we center and zoom it on the person, at a resolution of 256×256 px. To ensure correct perspective (with the principal point at the image center), we reproject the image onto a virtual camera pointed at the center of the

person's bounding box, as provided in the dataset. Scaling is applied so that the larger side of the person's bounding box covers about 80% of the image side length. Common data augmentation techniques are used in training, including random rotation, scaling, translation, horizontal flipping, as well as image filtering such as color distortions and blurs.

4.4.4 Evaluation Metrics

Following standard practice on Human3.6M, we evaluate prediction accuracy by the so-called mean per joint position error (MPJPE), which is the mean Euclidean error of all joints after skeleton alignment at the root (pelvis) joint. Procrustes alignment is not used.

4.4.5 Synthetic Occlusions for Robustness Analysis

We consider solid black shapes and some more realistic object segments from the Pascal VOC 2012 dataset (Everingham *et al.*, 2012) as occluders in this study (see Figures 4.2 and 4.3). The number, position and size of the objects are generated at random. We define the *degree of occlusion* as the percentage of occluded pixels inside the person's bounding box and vary this quantity between 0% and 70%.

4.4.6 Occlusion-Augmented Training

We hypothesize that synthetic occlusion data augmentation during training can improve test-time occlusion robustness. To verify this, we use the same kinds of occlusions as described in the previous section, with an additional *mixture* variant, which uses one of the other types at random for each frame. We make sure to strictly separate the VOC objects used for training and testing. Furthermore, we try the RE-0 variant of *random erasing* by Zhong *et al.* (2020), generating a single occluding black rectangle of random size according to their pseudocode. We refer to this mode as *single rectangle* in this chapter.

To make these strategies comparable, we parameterize them such that the distribution of the number of occluded pixels is similar. Notably, we only apply these augmentations with 50% probability for each frame. This was found important in prior work on occlusion augmentation (DeVries and Taylor, 2017).

4.4.7 Implementation Details

We use the implementation of ResNet50V1 and the corresponding ImageNet-pretrained initial weights from the TensorFlow-Slim library (Silberman and Guadarrama, 2016).

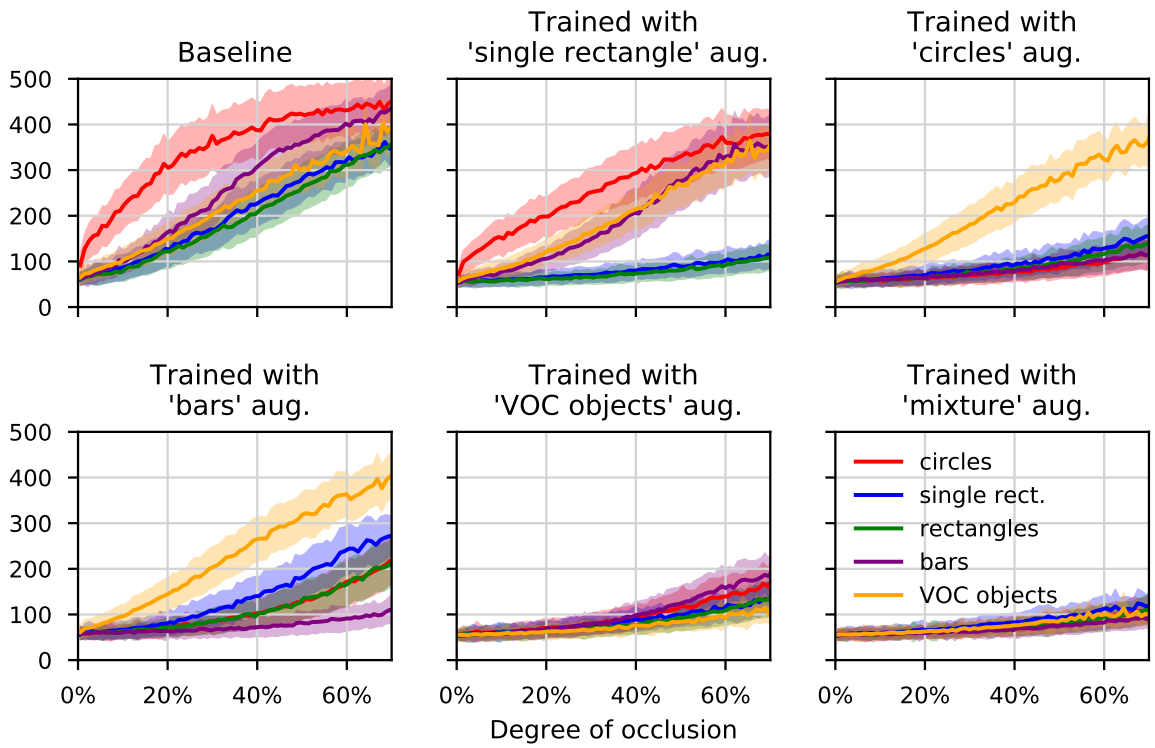


Figure 4.4: Assessing occlusion robustness on Human3.6M. Each subplot shows the performance when training with a particular augmentation method. Within a subplot, each line shows the mean and standard deviation of MPJPE under increasing degrees of occlusion of a particular type.

Training is done with the Adam optimizer and a minibatch size of 64, for 40 epochs, taking approximately 24 hours on an NVIDIA GeForce Titan X (Pascal) GPU.

4.5 Results

We start presenting our results by showing that our baseline model has state-of-the-art performance. We then show how performance deteriorates with test-time occlusions and that this can be mitigated using occlusion data augmentation. The augmentations are then shown to help even when the test images do not contain synthetic occlusions.

4.5.1 Baseline Performance

The current state of the art among published methods which use no extra 2D pose datasets for training is by Pavlakos *et al.* (2017), as shown in Table 4.1. Since our evaluation assumes knowledge of the root joint depth at test time, we compare with

Pavlakos *et al.*'s performance under the same conditions, for which the results can be found in their supplementary material. Our baseline's MPJPE of 63.3 mm is already better than Pavlakos *et al.*'s 64.8.

The method by Sun *et al.* (2018a) achieves an MPJPE of 64.1 mm, but it is unclear whether this approach uses the known root joint depth or resolves scale ambiguity by other means.

4.5.2 Robustness Analysis Under Occlusion

We evaluate the robustness of our baseline model using different degrees and types of occlusions (see the top left plot of Figure 4.4). We observe that circular occlusions cause by far the largest increase in error, the reason for which needs further investigation. Occlusions with oriented bars, VOC objects and rectangles lead to comparable performance loss. We note that rectangles are the least problematic type of occlusion, despite being a widely used test case in the literature.

Figure 4.3 visualizes an example. The baseline network gives good predictions for the non-occluded case, but when we paste two Pascal VOC objects onto the image, prediction visibly fails for the affected limbs.

4.5.3 Augmentation Improves Occlusion-Robustness

We now turn to the evaluation of occlusion augmentation at training time for increased test-time occlusion-robustness. Figures 4.4 and 4.5 show the results. Erasing a single rectangle (as in Zhong *et al.*, 2020) results in robustness against multiple rectangles at test time, but is much less effective for the other types of occlusions, being most sensitive to circles. Using several rectangles during training works slightly better than single-rectangle random erasing, but it, too, has difficulty in generalizing to other types of occlusion structures. Circular occlusion augmentation generalizes to all other simple geometric occlusion shapes, but barely helps when more realistic VOC objects are used as occluders at test time. VOC-augmentation, however, does generalize to both simple geometric shapes and other VOC objects (the objects used in training and testing are strictly separated). The qualitative difference in robustness when using this augmentation type is illustrated in Figure 4.3. The network learned to use context cues and gives good prediction even for the almost fully occluded lower left leg. Finally, the combination of all these strategies proves to be effective against all of the analyzed occlusion types together.

4.5.4 Regularization via Occlusion Augmentation

In the previous section, we have seen that training-time occlusion augmentation is helpful when evaluating on occluded test examples. Let us now look at the effect

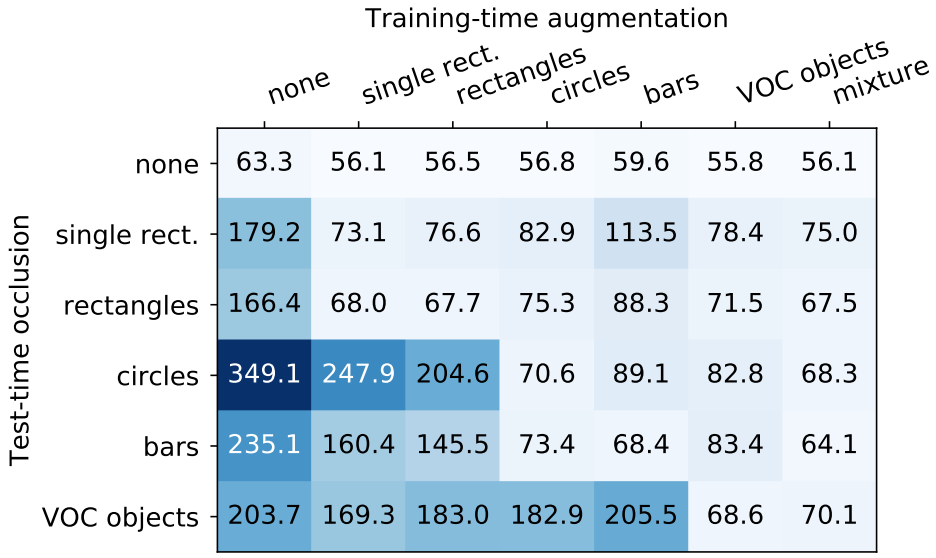


Figure 4.5: Exploring how much each type of training-time data augmentation protects against each type of test occlusions. The numbers are the MPJPE averaged for degrees of occlusion between 10% and 50%.

of these augmentation schemes when evaluating on the original test data without synthetic occlusions (see Table 4.1). All occlusion augmentation strategies are found to improve upon the baseline result, with the *VOC objects* performing the best and *bars* the worst. Although this was not our original aim in conducting this study, it is a valuable finding, which we are going to exploit in Chapter 5.

4.5.5 Runtime

Inference of the whole pipeline runs at 64, 165, and 204 fps for batch sizes of 1, 8, and 64 images, respectively, on a single NVIDIA GeForce Titan X (Pascal) GPU. This makes the method suitable for high frame rate applications.

4.6 Conclusion

We have presented a systematic study of occlusion effects on 3D human pose estimation from a single RGB image, using an efficient ResNet-based test model.

We found that despite producing state-of-the-art benchmark results, the network's performance quickly drops when synthetic occlusions are added. Circular structures turned out to be particularly problematic, the reason of which needs further study. We then showed that training-time occlusion data augmentation is effective in reducing

occlusion-induced errors, while also improving the performance without test-time occlusions.

Future experiments should also target other datasets besides Human3.6M and it remains to be seen how well our findings about synthetic occlusions generalize to real ones.

	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg
Zhou *et al.* (2015)	87.4	109.3	87.1	103.2	116.2	139.5	106.9	99.8	124.5	199.2	107.4	118.1	79.4	114.2	97.7	113.0
Tekin *et al.* (2016)	102.4	147.7	88.8	125.4	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	55.1	126.3	65.8	125.0
Zhou *et al.* (2016)	91.8	102.4	97.0	98.8	113.4	125.2	90.0	93.8	132.2	159.0	106.9	94.4	79.0	126.0	99.0	107.3
Sun *et al.* (2017)	90.2	95.5	82.3	85.0	87.1	94.5	87.9	93.4	100.3	135.4	91.4	87.3	78.0	90.4	86.5	92.4
Sun *et al.* (2018a)	63.8	64.0	56.9	64.8	62.1	70.4	59.8	60.1	71.6	91.7	60.9	65.1	51.3	63.2	55.4	64.1
Pavlakos *et al.* (2017)	67.4	72.0	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	59.1	74.9	63.2	71.9
Pavlakos *et al.* (2017)*	59.3	64.9	59.4	61.3	65.1	69.0	57.1	60.1	75.1	91.9	64.5	59.6	66.8	53.7	56.8	64.8
Ours (no occlusion aug.)	60.2	64.1	55.9	58.3	63.8	69.5	58.8	64.4	67.7	90.8	61.9	59.2	66.0	56.9	50.8	63.3
w/ circles aug.	52.9	58.0	51.8	54.8	56.9	62.6	51.4	55.0	64.7	79.2	56.3	52.5	58.8	47.9	43.0	56.8
w/ single rectangle aug.	52.0	58.6	51.0	53.5	56.1	62.6	51.5	54.2	65.7	71.2	56.1	52.9	58.2	47.8	42.9	56.1
w/ rectangles aug.	51.9	57.9	52.5	54.2	57.3	61.9	51.7	55.2	63.4	76.7	56.5	51.7	58.8	47.8	43.4	56.5
w/ bars aug.	55.0	60.1	54.1	56.4	59.9	64.9	52.4	59.5	67.7	88.7	58.5	54.2	62.4	50.0	45.4	59.6
w/ VOC objects aug.	51.2	58.7	51.7	53.4	56.8	59.3	50.7	52.6	65.5	73.2	56.8	51.4	56.6	47.0	42.4	55.8
w/ mixture aug.	51.3	57.8	52.5	53.8	55.9	58.7	50.9	52.8	66.7	77.1	56.6	51.7	56.6	47.6	42.8	56.1

Table 4.1: Mean per joint position error on Human3.6M for methods using no extra pose datasets in training. Methods below the line have access to the ground-truth root joint depth at test time. (No synthetic occlusions are used on the test inputs.) (*we show the results with known root depth)

5

Synthetic Occlusion Augmentation: A Case Study

In Chapter 4, we have seen how occlusion data augmentation can lead both to a substantial improvement of occlusion robustness in 3D human pose estimation, and provide a regularizing effect, reducing test error even on non-occluded images.

In this chapter, we describe a case study of adapting and applying our method to the 2018 ECCV PoseTrack Challenge on 3D human pose estimation, where our submission achieved first place.

Since this challenge uses held-out test data, it allows for a more unbiased evaluation of competing methods than the typical Human3.6M evaluation protocol, where gradual overfitting over the years could be a concern. In addition to reaching first place in the challenge, our method also surpasses the state of the art on the usual Human3.6M benchmark, when measured against other methods that use no additional pose datasets in training. We have released the code for applying synthetic occlusions publicly.[1]

This chapter is based on our paper (Sárándi *et al.*, 2018b), which is, in turn, a longer version of our extended abstract presented at the 2018 ECCV PoseTrack Workshop.

5.1 Task and Dataset

The 3D human pose estimation part of the 2018 ECCV PoseTrack Challenge invited participants to tackle the following task. Given an uncropped, static RGB image containing a single person, estimate the position of 17 body joints in 3D camera space, relative to the root (pelvis) joint position.

The dataset in this challenge is a subset of Human3.6M (Ionescu *et al.*, 2014), with 35 832 training, 19 312 validation and 24 416 test examples. There are a few key differences in the challenge protocol compared to the full benchmark. First, the challenge version does not provide person bounding boxes and camera intrinsics.

[1] `https://github.com/isarandi/synthetic-occlusion`

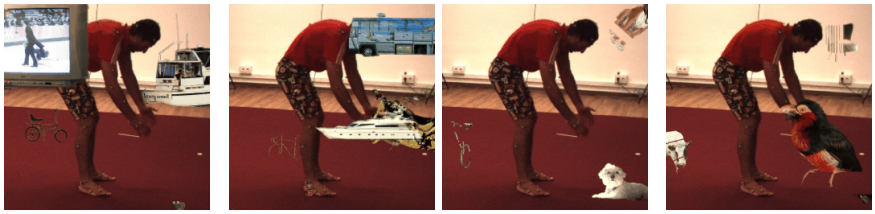


Figure 5.1: Examples of synthetic occlusions with Pascal VOC objects (geometric and color augmentations not depicted).

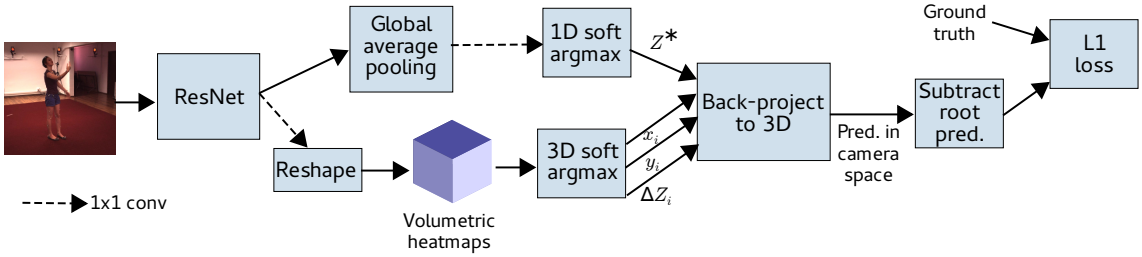


Figure 5.2: Overview of our architecture.

Second, the ground truth labels are more restricted for the training set, consisting only of root-relative 3D joint coordinates. Notably, image-space 2D joint coordinates are not available either.

5.2 Method

We present a modified version of our method described in Chapter 4, which we used for studying occlusion robustness in 3D pose estimation. We extend the method to handle the above-mentioned differences in the experimental protocol.

5.2.1 Image Preprocessing

Since bounding boxes are not given, we obtain them using the YOLOv3 detector (Redmon and Farhadi, 2018). The camera's intrinsic parameters are not specified in the data, however the images all originate from the Human3.6M dataset. This motivates us to treat the focal length f as a global hyperparameter. Using this focal length, we reproject the image to be centered on the person box through a homography, at a scale where the larger side of the box fills 90% of the resulting image.

5.2.2 Backbone Network

We feed the transformed image (with resolution 256×256 px) into a fully convolutional backbone network (ResNet50V2; He *et al.*, 2016b). We directly obtain volumetric

	Dir.	Dis.	Eat	Greet	Phn.	Pose	Pur.	Sit	SitD	Sm.	Pht.	Wait	Walk	WD	WT	Avg
Zhu	58	59	64	62	65	60	68	77	92	65	68	62	60	70	59	66
Rhodin	51	53	58	52	64	53	67	94	132	65	64	57	53	67	53	66
Zhou	52	56	55	51	57	53	64	73	81	61	60	57	49	61	53	59
Park	53	52	52	53	55	55	54	71	84	56	60	58	51	64	57	58
Shen	53	54	54	52	56	55	58	70	78	60	59	57	48	61	56	58
Pavlakos	44	46	50	47	56	47	52	63	70	54	54	48	46	58	46	52
Sun	**38**	43	46	41	46	**40**	49	65	73	48	49	43	38	52	**38**	47
Ours	**38**	**40**	**43**	**40**	**43**	**40**	**47**	**58**	**64**	**43**	**48**	**42**	**36**	**50**	**38**	**45**

Table 5.1: Mean per joint position errors achieved by participants of the 2018 ECCV PoseTrack Challenge on 3D human pose estimation (Ionescu *et al.*, 2018), on a subset of the Human3.6M dataset. In contrast to our method, some participants used extra 2D pose datasets in training (in accordance with the challenge rules).

heatmaps from the backbone net by adding a 1×1 convolutional layer on the last spatial feature map of the backbone, producing $J \cdot D$ output channels. The resulting tensor is reshaped to yield J volumes, one per body joint, each with depth D.

5.2.3 Volumetric Heatmaps

Following Pavlakos *et al.* (2017), we use 2.5D volumetric heatmaps: the X and Y axes correspond to image space and the depth axis to camera space, relative to the person center. Root-relative depth predictions, however, are not sufficient. In order to back-project the image-space coordinates to camera space, we would need to know the absolute depths. Pavlakos *et al.* optimize the root joint depth in postprocessing, based on a fixed bone-length assumption. By contrast, we predict this distance using a second prediction head attached to the backbone network (see Figure 5.2). This head outputs a 1D heatmap, discretized to 32 units, representing a 10 meter range in front of the camera. We will discuss a more general approach to tackle absolute 3D human pose estimation in Chapter 7.

5.2.4 Soft-Argmax

We extract coordinate predictions from both heatmaps using soft-argmax (Levine *et al.*, 2016; Nibali *et al.*, 2018). Since this operation is differentiable, there is no need to provide ground-truth heatmaps at training time (Sun *et al.*, 2018a). Instead, the loss can be computed deeper in the network and backpropagated through the soft-argmax operation. Soft-argmax also reduces the quantization errors inherent in hard argmax and gives fine-grained, continuous results without requiring memory-expensive, high-resolution heatmaps (Sun *et al.*, 2018a). Indeed, we use a heatmap resolution as low as

16^3 for the results presented in this chapter, and will further reduce this to 8^3 in the later ones.

5.2.5 Camera Intrinsics

Having predicted image coordinates x_i, y_i, depth coordinates ΔZ_i relative to the person center and the absolute depth Z^* of the person center by soft-argmax, we now need camera intrinsics to transform the coordinates from image space to 3D camera space. As mentioned earlier, the original camera's focal length f is treated as a hyperparameter, and we must also take into account the zooming factor s applied in preprocessing.

To avoid the need for precise hyperparameter tuning of f, we learn an additional, input-independent corrective factor c for the focal length during training, to achieve better alignment of image and heatmap locations. Denoting the image height and width as H and W, back-projection is performed as

$$\begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} = (Z^* + \Delta Z_i) \begin{bmatrix} fsc & 0 & W/2 \\ 0 & fsc & H/2 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}. \tag{5.1}$$

5.2.6 Loss

After subtracting the root joint coordinates, we compute the ℓ_1 loss in the original camera space w.r.t. the provided root-relative ground truth. No explicit heatmap loss is used. Since all above operations are differentiable the whole network can be trained end-to-end.

5.2.7 Data Augmentation

In Chapter 4, we found that augmenting training images with synthetic occlusions acts as an effective regularizer. Starting with the objects in the Pascal VOC dataset (Everingham *et al.*, 2012), we filter out persons, segments labeled as *difficult* or *truncated* and segments with area below 500 px, leaving 2638 objects. With probability p_{occ}, we paste a random number (between 1 and 8) of these objects at random locations in each frame. We also apply standard geometric augmentations (scaling, rotation, translation, horizontal flip) and appearance distortions (blurs and color manipulations). At test time only horizontal flipping augmentation is used.

5.2.8 Training Details

The backbone network is initialized with ImageNet-pretrained weights from Silberman and Guadarrama (2016). We train the final method for 410 epochs on the union of the training and validation set using the Adam optimizer (Kingma and Ba, 2015) and

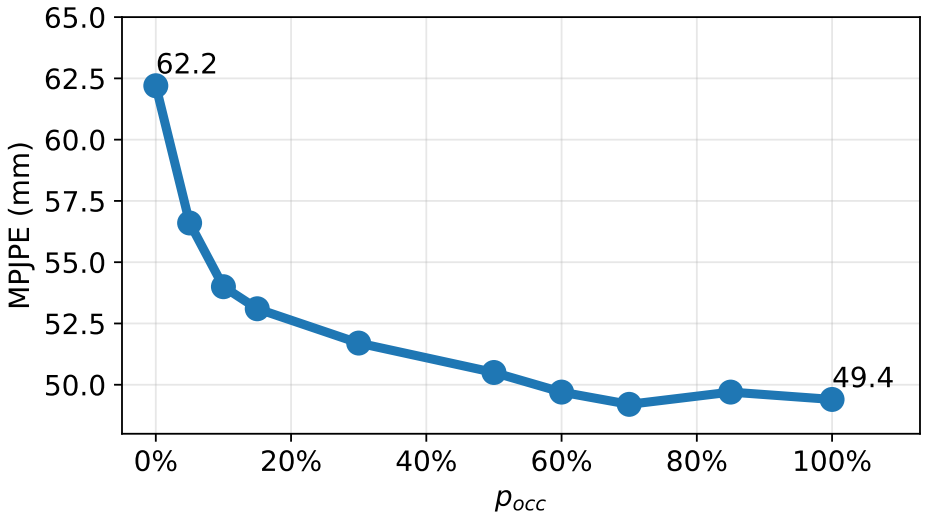


Figure 5.3: Effect of the per-frame probability of occlusion augmentation (p_{occ}) evaluated on the Challenge validation set.

cyclical (triangular) learning rates (Smith, 2017). Our final challenge predictions were produced using a snapshot ensemble (Huang *et al.*, 2017), averaging the predictions of snapshots taken at the last three learning rate minima of the cyclical schedule. We used $f = 1500$ and $p_{\text{occ}} = 0.5$ for the submission.

5.3 Results

The evaluation metric is the mean per joint position error (MPJPE) over all joints after subtraction of the root joint position. Our method achieves best results for all actions, even ahead of methods using extra 2D pose datasets in training (see Table 5.1). The margin is largest for the actions *Sitting* and *Sitting Down*, showing that our method is more robust to the presence of a chair, which is the only occluding object in the Human3.6M dataset.

5.3.1 Effect of Occlusion Augmentation

Figure 5.3 shows how synthetic occlusion augmentation improves results on the Challenge validation set as we vary the probability p_{occ} of applying occlusion augmentation to each frame. Augmenting just 10% of the images already improves MPJPE by 8.2 mm and improvements continue to about $p_{\text{occ}} = 70\%$, after which performance is only influenced slightly.

5.3.2 Full Human3.6M Benchmark

For comparison with prior work, we train and evaluate our method on the full Human3.6M benchmark as well. Here we use the bounding boxes and camera intrinsics provided with the dataset and minimize the ℓ_1 loss computed on the absolute (*i.e.*, non-root-relative) coordinates in camera space for 40 epochs. The person center depth Z^* is estimated as described in Section 5.2. We follow the common protocol of training on five subjects (S1, S5, S6 S7, S8) and evaluating on two (S9, S11), without Procrustes alignment. We use no snapshot ensembling here, for better comparability. The occlusion probability p_{occ} is set to 1. As seen in Table 5.2, our method outperforms all prior work on Human3.6M in the setting where no additional pose datasets are used for training.

5.4 Conclusion

We have presented an architecture and data augmentation method for 3D human pose estimation and have shown that it outperforms other methods both by achieving first place in the 2018 ECCV PoseTrack Challenge and by surpassing the state of the art on the full benchmark among methods using no additional pose datasets in training.

	Dir.	Dis.	Eat	Greet	Phon.	Pose	Pur.	Sit	SitD	Smo.	Phot.	Wait	Walk	WalkD	WalkT	Avg
* Zhou *et al.* (2017)	54.8	60.7	58.2	71.4	62.0	65.5	53.8	**55.6**	75.2	111.6	64.2	66.0	51.4	63.2	55.3	64.9
* Martinez *et al.* (2017b)	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0	94.6	62.3	78.4	59.1	65.1	49.5	52.4	62.9
* Sun *et al.* (2017)	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
* Pavlakos *et al.* (2018)	48.5	54.4	54.4	52.0	59.4	49.9	52.9	65.8	71.1	56.6	65.3	52.9	60.9	**44.7**	47.8	56.2
* Luvizon *et al.* (2018), single-crop	51.5	53.4	49.0	52.5	53.9	50.3	54.4	63.6	73.5	55.3	61.9	50.1	46.0	60.2	51.0	55.1
* Luvizon *et al.* (2018), multi-crop	49.2	51.6	**47.6**	50.5	51.8	48.5	51.7	61.5	70.9	53.7	60.3	48.9	44.4	57.9	48.9	53.2
* Sun *et al.* (2018a)	**47.5**	**47.7**	49.5	**50.2**	**51.4**	**43.8**	**46.4**	58.9	**65.7**	**49.4**	**55.8**	**47.8**	**38.9**	49.0	**43.8**	**49.6**
Tekin *et al.* (2016)	102.4	147.7	88.8	125.4	118.0	112.4	129.2	138.9	224.9	118.4	182.7	138.8	55.1	126.3	65.8	125.0
Zhou *et al.* (2016)	91.8	102.4	97.0	98.8	113.4	90.0	93.8	132.2	159.0	106.9	125.2	94.4	79.0	126.0	99.0	107.3
Zhou *et al.* (2015)	87.4	109.3	87.1	103.2	116.2	106.9	99.8	124.5	199.2	107.4	139.5	118.1	79.4	114.2	97.7	113.0
Sun *et al.* (2017)	90.2	95.5	82.3	85.0	87.1	87.9	93.4	100.3	135.4	91.4	94.5	87.3	78.0	90.4	86.5	92.4
Pavlakos *et al.* (2017)	67.4	72.0	66.7	69.1	72.0	65.0	68.3	83.7	96.5	71.7	77.0	65.8	59.1	74.9	63.2	71.9
Sun *et al.* (2018a)	63.8	64.0	56.9	64.8	62.1	59.8	60.1	71.6	91.7	60.9	70.4	65.1	51.3	63.2	55.4	64.1
Ours (no occlusion augm.)	63.3	65.5	56.0	62.1	64.0	60.7	64.8	76.7	93.0	63.3	69.7	62.0	54.1	68.8	61.3	65.7
Ours (full)	*49.1*	*54.6*	*50.4*	*50.7*	*54.8*	*47.4*	*50.1*	*67.5*	*78.4*	*53.1*	*57.4*	*50.7*	*40.1*	*54.0*	*46.1*	*54.2*

Table 5.2: Mean per joint position error on the full Human3.6M dataset. Results marked with an asterisk (*) were achieved using extra 2D pose dataset(s) in training. Boldface indicates the overall best results, while italic indicates the best when using no extra 2D pose datasets.

6

MeTRo: A Metric-Scale Truncation Robust Heatmap Representation

In this chapter, we introduce a simple and effective heatmap-based method to predict metric-scale, root-relative 3D human pose from monocular RGB images, even under image truncation.

Heatmap representations have formed the basis of 2D human pose estimation systems for many years, but their generalizations for 3D pose have only recently been considered. This includes 2.5D volumetric heatmaps, familiar from the previous chapters, whose X and Y axes correspond to image space and the Z axis to metric depth around the subject. To obtain metric-scale predictions, these methods must include a separate, explicit postprocessing step to resolve scale ambiguity. Further, they cannot encode body joint positions outside of the image boundaries, leading to incomplete pose estimates in case of image truncation.

We address these limitations by proposing metric-scale truncation-robust (*MeTRo*[1]) volumetric heatmaps, whose dimensions are defined in metric 3D space near the subject, instead of being aligned with image space. We train a fully convolutional network to estimate such heatmaps from monocular RGB in an end-to-end manner. This reinterpretation of the heatmap dimensions allows us to estimate complete metric-scale poses without test-time knowledge of the focal length or person distance and without relying on anthropometric heuristics in postprocessing. Furthermore, as the image space is decoupled from the heatmap space, the network can learn to reason about joints beyond the image boundary. Using ResNet50V2 without any additional learned layers, we obtain state-of-the-art results on the Human3.6M and MPI-INF-3DHP benchmarks. As our method is simple and fast, it can become a useful

[1]Not to be confused with *METRO*, the *MEsh TRansfOrmer* (Lin *et al.*, 2021a), published shortly after our work.

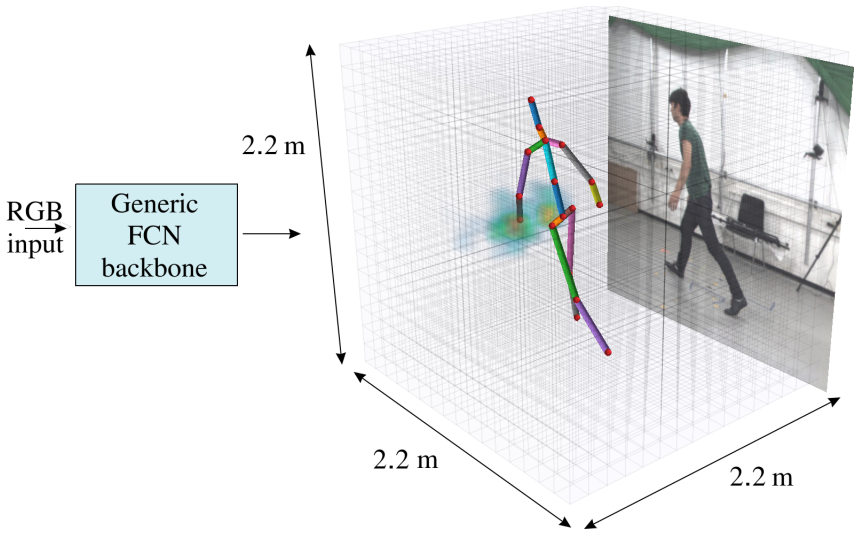


Figure 6.1: We propose to use a generic fully convolutional network (ResNet50V2, in this chapter) to directly predict volumetric heatmaps in 3D metric space around the subject. This visualization shows a $16 \times 16 \times 16$ heatmap for the left wrist and the full skeleton predicted by our model.

component for real-time top-down multi-person pose estimation systems. We make our code publicly available to facilitate further research.[2]

This chapter is based on our publication (Sárándi *et al.*, 2020), presented at the 2020 IEEE Conference on Automatic Face and Gesture Recognition. Additional materials beyond the contents of that paper provide more detailed ablation experiments, as well as a detailed study of how the occlusion augmentation remains effective with our novel representation. We also analyze whether the shape of the texture content of the occluders matters more.

6.1 Overview

Human pose estimation from camera input is a long-standing problem in computer vision with a wide range of applications including human–robot interaction (Zimmermann *et al.*, 2018), virtual reality (Alldieck *et al.*, 2018), medicine (Belagiannis *et al.*, 2016; Srivastav *et al.*, 2018) and commerce (Neverova *et al.*, 2018). Since the adoption of deep convolutional neural networks (CNN), and especially heatmap representations, we have witnessed rapid progress in pose estimation research (Newell *et al.*, 2016; Yang *et al.*, 2017; Ke *et al.*, 2018).

Recently, deep CNNs have been successfully applied to the monocular 3D human pose estimation task as well (Martinez *et al.*, 2017b; Mehta *et al.*, 2017b; Zhou *et al.*,

[2] `https://vision.rwth-aachen.de/metro-pose3d`

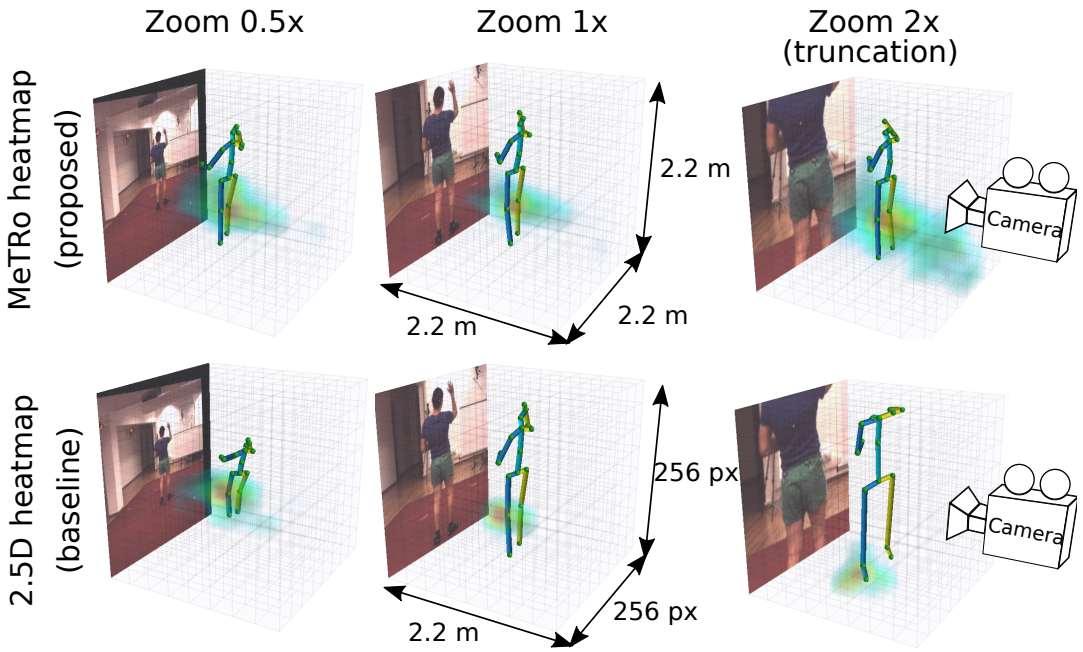


Figure 6.2: By defining heatmaps in the 3D metric space around the person (*bottom row*) we can directly predict scale-correct and complete poses. This is in contrast to prior work (*top row*) that defines the X and Y heatmap axes in image space and requires further postprocessing to obtain a metric-scale skeleton. The three columns show how zooming affects the heatmap representation (a knee heatmap is shown along with the soft-argmax decoded skeleton). Notice that our heatmap-space representation is largely invariant to image scaling and estimates a complete pose even under body truncation at the image boundaries.

2017; Luo *et al.*, 2018a; Nibali *et al.*, 2019). Here a person's anatomical landmarks are sought in 3D space, *i.e.*, in millimeters, instead of pixels. These advances tie into one of the major themes of computer vision research, reconstructing 3D structure from images. Such tasks are especially challenging due to inherent geometric ambiguities. One class of ambiguities arise because different 3D articulations may share the same 2D projection. Another ambiguity is between the size of an object and its distance, since small objects near the camera look the same as large ones far away.

There is no clear consensus yet about the most effective way to represent and tackle these problems. One promising line of approaches extend 2D joint heatmaps with a depth axis, resulting in a 2.5D volumetric representation (Pavlakos *et al.*, 2017; Iqbal *et al.*, 2018; Luvizon *et al.*, 2018; Sun *et al.*, 2018a). Finding heatmap maxima gives the estimated pixel coordinates and root-relative depths per joint (a 2.5D pose). While these estimates can be highly accurate, the 2.5D representation does not address the challenging ambiguity between scale (person size) and distance. Indeed, to bridge the gap between a 2.5D and a 3D pose, one needs to perform scale recovery as a separate

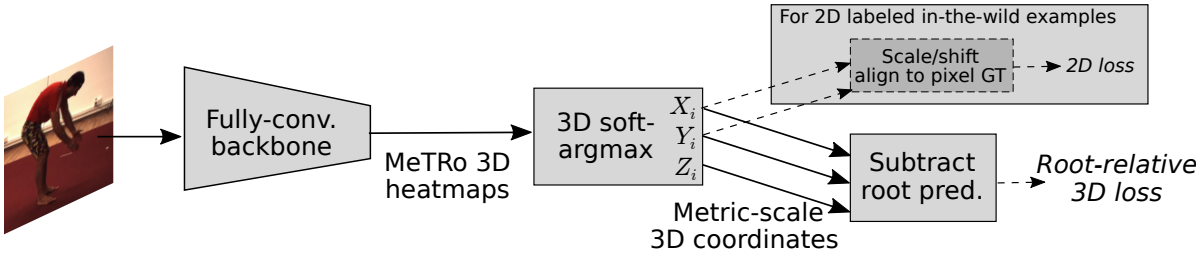


Figure 6.3: Overview of MeTRo. We first generate volumetric heatmaps using an off-the-shelf fully convolutional backbone. Applying soft-argmax on these heatmaps and scaling by an image-independent constant factor yields joint coordinates in metric space up to translation. We minimize the root-relative ℓ_1 loss. Focusing on simplicity, no learnable parameters are introduced outside the standard backbone. Note that reasoning about truncated body parts, scale recovery and back-projection also happen implicitly within the backbone. Weak supervision from in-the-wild 2D-labeled data is incorporated by aligning the metric prediction to the 2D ground truth by scaling and translation and computing the ℓ_1 loss (dashed arrows and boxes).

postprocessing step. Multiple explicit anthropometric heuristics have been proposed as scale cues, *e.g.*, bone length priors (Pavlakos *et al.*, 2017) or a skeleton length prior (Sun *et al.*, 2018c), computed by averaging over the training poses. However, these simple heuristics have difficulties when the experimental subjects have diverse heights. A further limitation is that 2.5D formulations are constrained to the estimation of joints that lie within the image boundaries. This can be problematic in practical applications with noisy bounding box detectors. While one could use an additional module to estimate missing joints, it is preferable to learn the complete skeleton estimation in a single unified stage.

Our goal in this chapter is to tackle the above limitations in a simple and efficient manner, while keeping the structural advantages of fully convolutional heatmap estimation, as opposed to numerical coordinate regression.

To this end, we propose training a fully convolutional network to output what we call *metric-scale truncation-robust (MeTRo) heatmaps* as illustrated in Figure 6.1. All dimensions of these heatmaps are defined to have a fixed metric extent in meters, a concept illustrated in Figure 6.2. This is an unconventional task definition for fully convolutional networks (FCN). FCNs are predominantly applied for pixelwise prediction tasks, such as semantic segmentation, where the input and output are pixel-to-pixel aligned, or at least are in the same coordinate frame. In our proposed approach, the input pixel positions and the output metric positions only satisfy a looser form of spatial correspondence. Nevertheless, we show that somewhat surprisingly, such a mapping can still be learned effectively by a standard modern FCN backbone.

While explicit prior knowledge of problem structure is known to be beneficial, it is still an open question how much geometric computation needs to be performed

explicitly and how much can be learned by deep networks from data. By skipping the 2.5D stage, we train the backbone FCN to implicitly reason about out-of-image joints, discover scale cues and learn the geometric perspective back-projection in an end-to-end manner. Our MeTRo heatmap representation can naturally encode body parts lying outside the image, since the prediction volume's bounds do not correspond to the image bounds. As there is no need to design an explicit scale recovery step, the pipeline becomes simpler, and the prediction of the root-relative pose requires neither the focal length nor the root joint distance to be known at test time.

Recent approaches have achieved good generalization performance to in-the-wild images by using abundant and diverse images with 2D pose labels in the training procedure besides 3D data (Zhou *et al.*, 2017; Luvizon *et al.*, 2018; Sun *et al.*, 2018a). Applying such weak supervision is challenging in our representation, since the network does not make any pixel-based predictions, its outputs are directly on a metric scale. We tackle this by proposing a scale and translation invariant loss computation method for 2D-annotated examples using an alignment layer. Combined with the recently introduced differentiable soft-argmax (Levine *et al.*, 2016; Nibali *et al.*, 2018; Sun *et al.*, 2018a; Luvizon *et al.*, 2019) layer, our method becomes end-to-end learned all the way from image to final 3D metric-scale prediction as shown in Figure 6.3. Soft-argmax also allows rapid training with low-resolution heatmaps and using dense prediction with smaller strides at test time for higher quality results, without the need for a decoder module. Here we find that the details of the striding mechanism are crucial and propose a "centered striding" method that distributes the output neuron receptive fields evenly over the image. Experimentally, our MeTRo heatmap estimation achieves state-of-the-art results on the two largest 3D pose benchmarks, Human3.6M and MPI-INF-3DHP. To isolate the effect of the representation, we perform direct comparisons with 2.5D heatmap learning using bone length–based scale recovery (Pavlakos *et al.*, 2017), under otherwise equal training conditions. We find that scale cues can indeed be learned implicitly in this fashion and MeTRo outperforms the baseline on most test sequences.

6.2 Related Work

3D human pose estimation has had a long research history starting with hand-crafted features and part-based models. Similar to other computer vision problems, the transition to deep convolutional networks has led to a dramatic performance increase in this task as well. For details, see Chapter 2, here we recapitulate only the most relevant related works for this chapter.

6.2.1 Deep 3D Human Pose Estimation

Much of the inspiration in recent 3D pose estimator design has come from lessons learned in 2D pose research. DeepPose (Toshev and Szegedy, 2014), the first neural method for 2D pose estimation, directly regressed 2D body joint coordinates on the RGB input via convolutional and fully connected layers. Later top-performing methods have transitioned to predicting body joint heatmaps by fully convolutional networks (*e.g.*, Newell *et al.*, 2016) as an intermediate representation. These heatmaps are spatially discretized arrays (one for each joint), in which higher values indicate higher confidence that the particular joint is located at the corresponding position.

One line of 3D pose research builds on top of 2D heatmaps and infers the 3D pose from them by exemplar matching (Chen and Ramanan, 2017), regression (Martinez *et al.*, 2017b) or probabilistic inference (Tome *et al.*, 2017). One inherent limitation of such approaches is that the image content only indirectly influences the 3D estimation, as it acts on the result of the 2D estimation stage. Furthermore, 2D-to-3D lifting is performed in a numerical coordinate representation, which does not benefit from the built-in convolutional structure of CNNs.

Nibali *et al.* (2019) predict three marginal heatmaps per body joint, for the XY, XZ and YZ planes, respectively. Pavlakos *et al.* (2017) have proposed extending 2D heatmaps with a root-relative metric depth axis. One can obtain the 2D pixel positions and root-relative depths of the joints by finding maxima in the heatmaps.

One downside of heatmap representations has been the requirement of a dense output, which can become especially costly in 3D. The recently proposed soft-argmax (Levine *et al.*, 2016; Nibali *et al.*, 2018; Luvizon *et al.*, 2019), a.k.a. integral regression (Sun *et al.*, 2018a), method greatly alleviates this problem. As opposed to hard-argmax, which simply finds the location of the highest heatmap activation, soft-argmax is computed as the weighted average of all voxel grid coordinates, using softmaxed heatmap activations as the weights. For example, a low resolution heatmap can encode a joint position lying halfway between two bin centers by outputting 0.5 for both bins. By virtue of being differentiable unlike hard-argmax, it also obviates the need for explicit heatmap-level supervision (*e.g.*, voxelwise cross-entropy). Instead, the loss can be computed (and its gradients back-propagated) from the coordinates yielded by soft-argmax.

Besides 2D heatmaps, Mehta *et al.* (2017b) estimate three further output channels per joint, the so-called *location maps*. These are read out at the position of the corresponding heatmap's peak to obtain the X, Y and Z coordinates on a metric scale. Note how in this approach the final 3D joint coordinates are generated in the form of activation *values* (of the location maps at the heatmap peaks), as opposed to high-activation *locations*. We can thus think of it a conceptual hybrid of direct numerical coordinate regression and heatmap estimation. A downside of this method is that it requires high-resolution location maps and cannot benefit from the soft-argmax approach.

6.2.2 Scale Ambiguity

It is well known that projecting a 3D world onto a 2D image plane results in ambiguity between size and distance (depth). However, the end goal for 3D scene understanding and 3D human pose estimation in particular is a metric-space output at the true scale. The ambiguity can only be resolved using semantic scale cues, *i.e.*, prior knowledge of the usual size of humans and other objects appearing in the scene. Unfortunately, not all papers include a description of how this step is performed. Some authors report their results assuming a known focal length and known ground-truth root joint distance (Sun *et al.*, 2018a,b; Chen *et al.*, 2019b; Nibali *et al.*, 2019) and leave their estimation as a separate task. A simple anthropometric approach is used by Pavlakos *et al*.Given 2D pixel positions and root relative depth estimates from volumetric heatmaps, they optimize the absolute person distance such that the back-projected skeleton's bone lengths match the average over the training set in a least squares sense (Pavlakos *et al.*, 2017). A detailed description of this convex optimization problem is given in Pavlakos *et al.* (2017, supp.). We use this scale recovery approach as our main baseline comparison throughout the chapter. Sun *et al.* (2018c) employ a similar idea, but use the overall skeleton length and a weak perspective model instead. Some recent works have shown that direct regression of person height from an image is a challenging task (Dantcheva *et al.*, 2018; Günel *et al.*, 2019). Véges and Lőrincz (2019) make use of a monocular depth prediction network pretrained on various indoor and outdoor datasets to help with absolute person distance estimation.

6.2.3 Truncated Pose Estimation

Single-person 3D pose estimation benchmarks, such as Human3.6M (Ionescu *et al.*, 2014), assume that the whole person is visible in the input image. In practical applications, however, bounding boxes are obtained using imperfect detectors, which can result in body truncation, especially in high-occlusion scenes. A possible remedy could be extending the detection crops by amodal completion (Kar *et al.*, 2015), but this would result in a loss of image resolution. Generally, pose estimation performance under truncation has not been studied extensively in the literature. Recent work by Park *et al.* (2020) uses cropping data augmentation to improve 2D pose estimation. Vosoughi and Amer (2018) create randomly truncated crops from Human3.6M images, and show that current methods perform poorly on truncated person images, even when only considering the present (within-boundary) joints. They tackle the problem using direct numerical coordinate regression, similar to early 2D pose estimation methods (Toshev and Szegedy, 2014). We show that our approach performs significantly better in the truncated setting. Other methods, such as LCR-Net (Rogez *et al.*, 2017), can also produce out-of-image predictions, but this aspect has not been explicitly evaluated by its authors.

6.3 Method

The input to our model is an RGB image crop $I \in \mathbb{R}^{w \times h \times 3}$ depicting a person. The desired output is a 3D skeleton, consisting of J joint coordinates $\left\{(\Delta X_j, \Delta Y_j, \Delta Z_j)^T\right\}_{j=1}^{J}$ in millimeters, up to arbitrary translation (hence the Δ symbols).

6.3.1 Metric-Space Volumetric Heatmap Representation

As is common in heatmap-based approaches, we apply a fully convolutional backbone network, with effective stride s to produce an array with $d \cdot J$ spatial output channels. Here d is the number of discretization bins along the depth axis of the prediction volume. We then split the array along the channel axis into J volumes, each of shape $(w/s) \times (h/s) \times d$. 3D spatial softmax is applied over each of them, resulting in volumetric heatmap activations $V^{(j)} \in \mathbb{R}^{(w/s) \times (h/s) \times d}$. Up to this point the process is similar to other volumetric heatmap approaches (Pavlakos *et al.*, 2017; Sun *et al.*, 2018a). The difference lies in how the heatmap axes are interpreted to yield metric-scale coordinates. In particular, the 3D joint coordinates are decoded using soft-argmax with *fixed* scaling factors:

$$\begin{bmatrix} \Delta X_j \\ \Delta Y_j \\ \Delta Z_j \end{bmatrix} = \sum_{p,q,r} V^{(j)}_{p,q,r} \cdot \begin{bmatrix} p \cdot s/w \cdot W \\ q \cdot s/h \cdot H \\ r \cdot 1/d \cdot D \end{bmatrix}, \tag{6.1}$$

where the p, q, r are zero-based integer indices into the volumetric heatmap array and W, H, D are the fixed metric width, height and depth extents of the full prediction volume. We set these extents as 2.2 meters in our work, which allows capturing people of usual height even when stretched out. In fact, the largest bounding cube needed to capture all joints of a person in the training sets of Human3.6M and MPI-INF-3DHP is 2.06 and 1.92 meters, respectively. Depending on striding logic (see Section 6.3.3), Equation 6.1 needs to be adjusted slightly, *e.g.*, the volume size may change with denser striding (Figure 6.4). The final root-relative prediction is obtained by subtracting the predicted root coordinates from all joint positions. Supervision is applied on these root-relative coordinates. This means that the position of the root joint prediction within the volume is not explicitly supervised and the network can place the skeleton anywhere within the prediction volume. The gradients are backpropagated through the root joint subtraction operation. No camera calibration–based back-projection, nor bone or skeleton size–based rescaling is needed for this root-relative prediction. The network is trained to perform these operations implicitly within the backbone.

6.3.2 Architecture

In contrast to prior work that employs decoders with upsampling layers and multiple refinement stages, we show that the task can be tackled in a significantly simpler

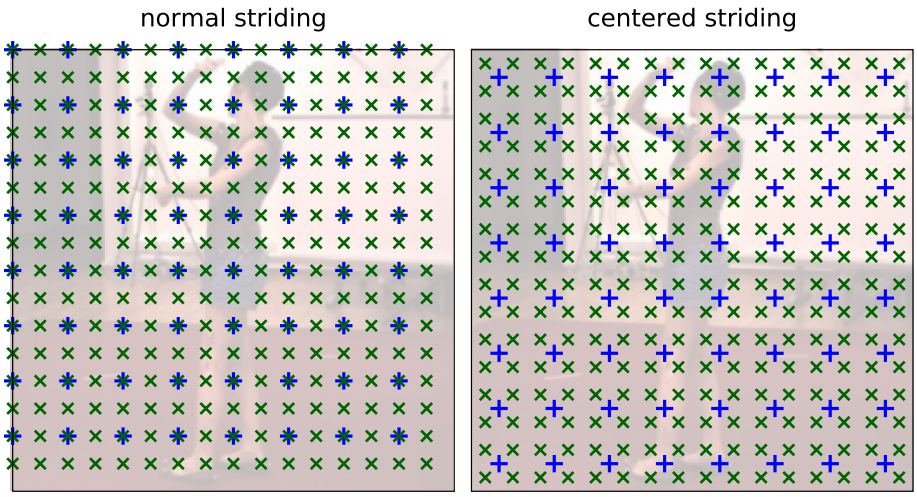


Figure 6.4: Receptive field centers of the output neurons in a strided FCN operating on a 256×256 image (+: stride 32, ×: stride 16). *left*: normal striding logic, where the top left result is kept per 2×2 block. Note that denser striding skews the sample density towards the bottom and right in the border areas. *right:* by reversing the stride logic in the last strided layer (*i.e.*, bottom right result taken, instead of top left), the samples are centered and the increased striding density is distributed evenly.

fashion. Indeed, we simply apply the ResNet50V2 (He *et al.*, 2016b) backbone to directly predict spatial heatmaps, without any additional learnable layers, such as transposed convolutions. By default, ResNet has an effective stride of 32, resulting in heatmaps of spatial size 8×8 from the input image of size 256×256 during training. The depth of the volumetric heatmap is set to 8. When testing on single-person datasets, we apply the trained network with an effective stride of 4, to obtain heatmaps with spatial size 64, which is the typical size used in prior work (Pavlakos *et al.*, 2017; Sun *et al.*, 2018a). This is called dense prediction and is commonly used in image segmentation (Chen *et al.*, 2017a). In this technique, striding is removed from a given number of convolutional layers and the dilation rate of subsequent convolutions is increased correspondingly. As we will see, dense prediction increases the compute requirements but also improves accuracy, while still allowing real-time execution.

6.3.3 Centered Striding

When changing striding density at test time compared to training time, it is important to consider how the distribution of heatmap receptive field centers is affected. The left side of Figure 6.4 shows a 256×256 image processed with training stride 32 (+) and test stride 16 (×). The coverage changes significantly between training and test, and is not symmetric over the image. While not an issue for pixel-labeling tasks, soft-argmax is a

weighted vote-averaging scheme and introducing new voting positions in an uneven manner skews the prediction result. To tackle this issue, we propose *centered striding*, where the striding logic in the last convolutional layer of the backbone is "reversed," such that it outputs the *bottom right* result per each 2×2 block. The result is a more evenly distributed coverage over the image, with each original sampling position replaced with four new ones equally spaced around it. This benefit is evaluated in Section 6.6.

6.3.4 Scale and Translation Agnostic 2D Loss

Similar to recent approaches (Zhou *et al.*, 2017; Luvizon *et al.*, 2018; Sun *et al.*, 2018a), we train simultaneously on 3D-labeled data from motion capture studios and 2D-labeled, in-the-wild data from the MPII dataset (Andriluka *et al.*, 2014), to incorporate more appearance variation in the training process. Half of each mini-batch is filled with examples of either kind. Supervision via 2D labels is straightforward when using 2.5D heatmaps, as the X and Y heatmap axes correspond to the space in which the 2D labels are defined. However, since our prediction volume is defined on a metric scale and is not aligned with image space, we propose a 2D loss computation method that is invariant to prediction scale and translation. To this end, we first orthographically project the predicted 3D skeleton onto the image plane by discarding the Z coordinate. Then we align the projected prediction to the 2D pixel-scale ground truth by translation and uniform scaling to the least-squares optimal fit before computing the loss. This alignment layer is differentiable and gradients can be backpropagated through it. We note that a similar scale-invariant loss has been used by Rhodin *et al.* (2018a) to enforce multi-view consistency of 3D poses.

6.3.5 Truncated Pose Estimation

Our metric-space heatmap representation decouples the image boundary from the heatmap boundary. This enables the prediction of joint locations outside the image frame without additional design effort, the network is simply trained to output complete poses at a metric scale, regardless of how the input image is scaled or cropped. To evaluate this aspect, we follow Vosoughi and Amer (2018) by randomly cropping Human3.6M inputs, keeping at least 1/4 of the area of the person bounding square. Examples of such crops are in the second row of Figure 6.13. We consider two scenarios. In the first one, the above described sampling of truncated crops is only performed at test time. In the second case, such crops are used for training as well.

6.3.6 Training Details

Loss. Prior work has shown that the ℓ_1 loss is preferable in soft-argmax-based pose estimation (Sun *et al.*, 2018a). To balance the losses computed on 3D and 2D-annotated examples, we use a fixed weighting factor tuned on a separate validation set of Human3.6M, yielding the overall loss as

$$\mathcal{L} = \mathcal{L}_{\text{ann3D}} + \lambda \mathcal{L}_{\text{ann2D}}. \tag{6.2}$$

Training Schedule. We initialize the network with ImageNet-pretrained weights and use the Adam optimizer with weight decay (Loshchilov and Hutter, 2019) and a batch size of 64. We decay the learning rate exponentially by an overall factor of 100, in two parts: from 10^{-4} to 3.33×10^{-5} over 25 epochs and from 3.33×10^{-6} to 10^{-6} in 2 final cooldown epochs.

Randomness. As usual in deep learning, several sources of randomness influence the exact results of an experiment: random weight initialization, data shuffling, data augmentation and hardware-level non-determinism of execution order. We control these (except the last) by consistently seeding the random number generators. To distinguish random fluctuations from algorithmic differences, we repeat our main experiments with 5 different seeds and report the mean and standard deviation of the evaluation metrics. In Section 6.8.1, we will analyze the repeatability of our experiments in more detail.

6.3.7 Intuition

As described above, our network is trained to output complete skeletons at a fixed metric scale, regardless of image zooming and truncation. However, at this point it is not clear how such predictions are produced by the network. To gain more intuition, we visualize projected heatmaps in Figure 6.5, allowing us to better understand how this fully convolutional model is able to achieve approximate invariance to image scale and truncation. In particular, we can see that the soft-argmax output is not necessarily in the middle of the heatmap's most prominent peak. As soft-argmax yields the heatmap's center of mass, even distant heatmap values have an influence. Intuitively, this allows the network to move the prediction result towards different heatmap locations by adding counter-balancing correction weights, for example at the image sides or at the person center. Regarding truncation, the last row shows that the model can infer that the arms must lie above waist level, since there is no visual evidence of them in the image. To understand how a fully convolutional network can "know" where the truncation happens, we refer to Islam *et al.* (2020), who show that even fully convolutional networks can encode positional information as a result of the zero paddings within convolutional layers. This means that the location of the top

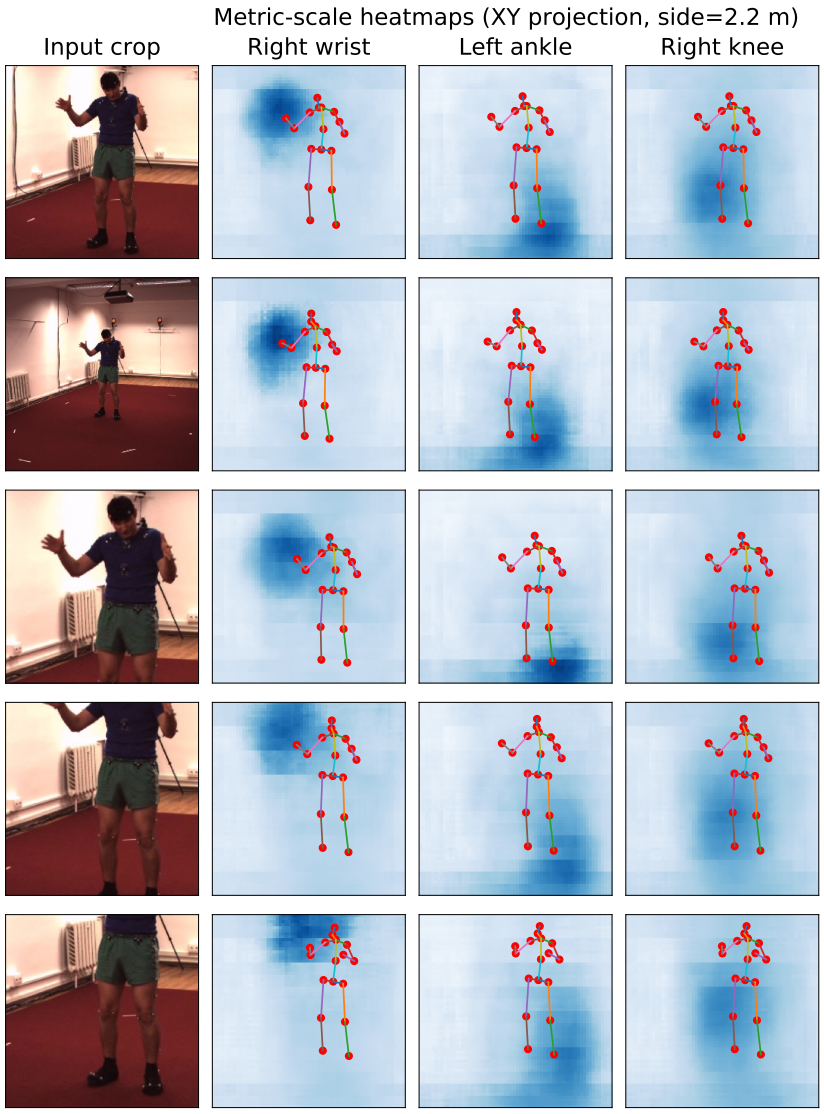


Figure 6.5: A closer look at how scale and truncation robustness is achieved in the heatmaps. We plot the projected *metric-scale* heatmaps for 3 joints with the full soft-argmax skeleton for reference. We observe that the predicted skeleton is approximately invariant to change in scale and truncation. Since the metric size of the person does not change with image scaling, the backbone learns to output heatmaps with a similar center of mass, regardless of image scale. Note that the heatmaps do not align with image space and this is intended by design. (The broad peaks are a result of training the model at low, 8×8 heatmap resolution.)

image border can be used as a cue for the network to shift the full skeleton downwards inside the heatmap volume, such that it fits. Note that the network is free to place the skeleton anywhere within the volume, since the root prediction is subtracted before

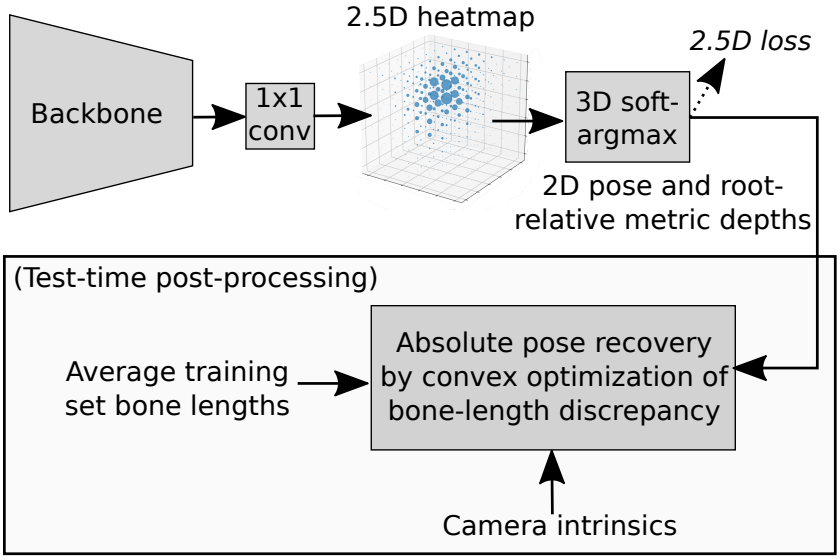


Figure 6.6: Baseline 2.5D Architecture. We use this alternative inspired by Pavlakos *et al.* (2017, supp.), for ablative comparison experiments.

computing the root-relative loss. This means that the exact position of the skeleton in this visualization has no effect on the actual model outputs. Instead, the network can place the skeleton such that it best fits inside the prediction volume.

6.4 Baseline using 2.5D Heatmaps

For comparison, we implement a 2.5D baseline derived from Pavlakos *et al.* (2017), who introduced volumetric heatmaps for 3D human pose estimation. Pavlakos *et al.* use a coarse-to-fine estimation scheme with a stacked hourglass architecture (Newell *et al.*, 2016) and no soft-argmax. To make the baseline directly comparable to our results, we instead use the architecture depicted in Figure 6.6. This baseline directly estimates 2.5D heatmaps through a 1×1 convolution at the end of the backbone. We then use soft-argmax, and compute the ℓ_1 loss on the resulting coordinates. This makes the baseline similar to the method by Sun *et al.* (2018a), except the latter uses additional learned layers and does not perform scale recovery. As a test-time postprocessing step, the baseline uses the bone length–based optimization method from Pavlakos *et al.* (2017, supp.) to recover the root joint depth, which we briefly reiterate here. Given an assumed value for the root joint depth Z_0 and known camera intrinsics, the 2.5D pose can be back-projected into metric space and each bone's resulting length $b_i(Z_0)$ can be

calculated. The optimal Z_0 is then the one that minimizes the squared bone length discrepancy, as compared to the average training bone lengths t_i:

$$Z_0^* = \arg\min_{Z_0} \sum_{i \in \text{bones}} \left(b_i(Z_0) - t_i\right)^2 , \tag{6.3}$$

where we only use bones, whose both ends are predicted to lie within the image (further from the border than 1 stride length). This is a nonlinear least-squares problem, and we solve it using the Levenberg–Marquardt algorithm initialized at $Z_0 = 2$ m. To reiterate, as in Pavlakos *et al.* (2017), the absolute pose is not supervised during the baseline's training and the optimization of Z_0 is not backpropagated through, for simplicity. We note, however, that the recent development of differentiable optimization layers (Amos and Kolter, 2017; Agrawal *et al.*, 2019) could, in principle, enable such a solution as well.

6.5 Datasets and Preprocessing

We conduct experiments on Human3.6M (Ionescu *et al.*, 2014) and MPI-INF-3DHP (3DHP; Mehta *et al.*, 2017a).

On **Human3.6M**, we evaluate according to both Protocols 1 and 2. We use the provided bounding boxes and downsample videos from 50 to 10 fps. To further reduce redundancy, training frames are only used if at least one body joint moves at least 100 mm since the previous kept frame.

We use the 2D-labeled **MPII** (Andriluka *et al.*, 2014) for weak supervision, following the idea by Zhou *et al.* (2017). Only arm and leg joints are used from MPII, as we found these to be the most consistently labeled across datasets. We only use instances explicitly marked as "well separated" from other people and take the provided person centers and sizes as the center and side length of the bounding box.

On **3DHP**, we follow Zhou *et al.* (2017) in moving the hips towards the neck by a fifth of the pelvis–neck vector before comparing with MPII-annotated skeletons for 2D loss computation. We evaluate w.r.t. both ground truth variants: unnormalized metric-space poses and "universal" (height-normalized) ones. We use only the chest-height cameras as Mehta *et al.* (2017a), and only examples where all joints are within the image. We generate 3DHP bounding boxes by combining the bounding box of labeled joints and the most confident person detection of YOLOv3. The same frame sampling strategy is used as described above for Human3.6M.

We crop the image to the person's bounding square and resize it to 256×256 px. Perspective effects must be taken into account when centering the image on the subject as this induces an implicit rotation of the camera (Mehta *et al.*, 2017a). We compensate for this effect by transforming image and the target joint positions to match the rotated camera frame. The green-screen 3DHP sequences are gamma-adjusted with an exponent of 0.67.

We apply geometric **augmentations** (scaling, rotation, translation, horizontal flip) and color distortion (brightness, contrast, hue, saturation). Synthetic occlusion is added with 70% probability, half of which are rectangles with uniform white noise as in Zhong *et al.* (2020), half are segmented non-person objects from the Pascal VOC dataset (Everingham *et al.*, 2012) as in Chapters 4 and 5. On the 3DHP dataset we also apply background augmentation with 70% probability following (Mehta *et al.*, 2017a), but no compositing for clothes and chair. The backgrounds are taken from the INRIA Holidays dataset (Jegou *et al.*, 2008) excluding person images. We do not use ensembling or test-time augmentation, all evaluation is done on a single crop.

We use the standard metrics from the literature. The main metric on 3DHP is the percentage of correct keypoints (PCK), *i.e.*, the fraction of joints predicted within a certain distance of the ground truth (150 mm by convention). The AUC metric is the area under the PCK curve as the threshold ranges from 0 to 150 mm. The metric on Human3.6M is the mean per joint position error (MPJPE). We follow the usual protocols, evaluating 14 joints on 3DHP, excluding the root, and 17 joints on Human3.6M, including the root.

6.6 Main Results

On **Human3.6M** without ground truth depth or scale information, we achieve 49.3 mm MPJPE, which is within the margin of error compared to the state of the art by Xu *et al.* (2020b) (49.2 mm), while using a considerably simpler approach (see Table 6.1). (In all tables, the number after "±" is the standard deviation of 5 repeated experiments with different random seeds, therefore the standard error of the mean is a fifth of this value.) This is only surpassed by Chen *et al.* (2019b) (48.4), however they do use the ground truth root joint depth for back-projection at test-time and do not perform scale recovery. Similarly, Sun *et al.* (2018a) obtain comparable results (49.6), however they also access the ground-truth root joint depth at test time, for image cropping (Sun *et al.*, 2018b).

Besides simplifying the prediction pipeline and allowing for truncation-robust prediction (see below), our metric heatmap representation also performs better than the 2.5D baseline with bone length–based scale recovery under the exact same experimental conditions.

Table 6.7 shows that training data augmentations improve performance by a large margin.

On Protocol 2 (Table 6.2), the benefit of our method is masked by the use of Procrustes alignment, which explicitly ignores the quality of scale recovery. It is therefore unsurprising that our method performs about equally well as the 2.5D variant (within the standard deviation of repeated experiments).

On **3DHP**, our method outperforms prior works by a large margin, including ones trained on more datasets as well (Table 6.3). Both with universal (height-normalized)

	Dir.	Dis.	Eat	Gre.	Phn.	Pose	Pur.	Sit	SitD	Sm.	Pht.	Wait	Walk	WD	WT	Avg ↓
Methods using ground-truth scale or depth information at test time																
Sun *et al.* (2017)	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Nibali *et al.* (2019)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	57.0
Luvizon *et al.* (2018)	51.5	53.4	49.0	52.5	53.9	50.3	54.4	63.6	73.5	55.3	61.9	50.1	**46.0**	60.2	51.0	55.1
Luvizon *et al.* (2020)	43.7	48.8	45.6	**46.2**	49.3	43.5	46.0	56.8	67.8	50.5	57.9	**43.4**	40.5	53.2	45.6	49.5
Sun *et al.* (2018a)	47.5	47.7	49.5	50.2	51.4	43.8	46.4	58.9	65.7	49.4	55.8	47.8	38.9	**49.0**	43.8	49.6
Chen *et al.* (2019b)	45.3	49.8	46.1	49.6	**48.2**	**41.7**	47.4	**53.1**	**55.2**	**48.0**	57.7	45.6	40.8	52.4	45.2	**48.4**
Methods using no ground truth scale or depth information at test time																
Pavlakos *et al.* (2017)	67.4	72.0	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Zhou *et al.* (2017)	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2	111.6	64.2	65.5	66.0	51.4	63.2	55.3	64.9
Martinez *et al.* (2017b)	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Fang* *et al.* (2018)	50.1	54.3	57.0	57.1	66.6	53.4	55.7	72.8	88.6	60.3	73.3	57.7	47.5	62.7	50.6	60.4
Yang *et al.* (2018)	51.5	58.9	50.4	57.0	62.1	49.8	52.7	69.2	85.2	57.4	65.4	58.4	43.6	60.1	47.7	58.6
Pavlakos *et al.* (2018)	48.5	54.4	54.4	52.0	59.4	49.9	52.9	65.8	71.1	56.6	65.3	52.9	44.7	60.9	47.8	56.2
Liu *et al.* (2019a)	47.0	53.1	50.3	48.8	56.0	48.1	47.6	65.9	72.6	52.3	61.4	49.1	39.3	54.2	**40.6**	52.4
Xu *et al.* (2020b)	**40.6**	**47.1**	45.7	**46.6**	50.7	45.0	47.7	56.3	**63.9**	49.4	63.1	**46.5**	**38.1**	51.9	42.3	**49.2**
Sharma *et al.* (2019)	48.6	54.5	54.2	55.7	62.6	50.5	54.3	70.0	78.3	58.1	72.0	55.4	45.2	61.4	49.7	58.0
Cai *et al.* (2019)	46.5	48.8	47.6	50.9	52.9	48.3	45.8	59.2	64.4	51.2	61.3	48.4	39.2	53.5	41.2	50.6
2.5D baseline	45.1	50.4	45.4	47.8	**50.0**	**44.6**	49.8	59.0	69.4	49.4	56.5	48.0	39.6	**49.4**	45.0	50.2±0.3
MeTRo (ours)	46.3	48.3	**43.3**	48.2	50.2	45.1	**46.1**	**56.2**	66.8	**49.3**	**54.5**	46.7	40.1	49.6	46.2	**49.3±0.7**

Table 6.1: Evaluation on Human3.6M Protocol 1 (subjects 9 and 11), using mean per joint position error (MPJPE) without Procrustes alignment. All methods use extra 2D-labeled pose data in training.

	PA-MPJPE
Nie *et al.* (2017)	79.5
Pavlakos *et al.* (2017)	51.9
Sun *et al.* (2017)	48.3
Martinez *et al.* (2017b)	47.7
Sun *et al.* (2018a)	40.6
Nibali *et al.* (2019)	40.4
Habibie *et al.* (2019)	49.2
Xu *et al.* (2020b)	38.9
Chen *et al.* (2019b)	**33.7**
2.5D baseline	34.5±0.4
MeTRo (ours)	34.7±0.5

Table 6.2: Comparison of Procrustes-aligned MPJPE with prior work on Human3.6M under Protocol 2 (test subject 11).

skeletons and true metric-scale ones, the MeTRo representation outperforms the baseline on green-screen studio images, however, the outdoor scenes were recorded on an empty field without scale cues and the explicit bone length–based scale recovery performs better there. Qualitative results are in Figure 6.13.

We analyze **scale recovery** in more detail in Table 6.4. As expected, the idealized method with test-time access to the ground truth root joint depth performs best on both Human3.6M and 3DHP. The proposed approach performs better than the 2.5D baseline using average bone lengths on Human3.6M and comparably on 3DHP. On Human3.6M, MeTRo closes most of this scale recovery gap between the 2.5D average bone length baseline and the idealized variant using the true root. Interestingly, our approach outperforms even the 2.5D variant using ground truth bone lengths for each test frame. On 3DHP, MeTRo's scale recovery performance is similar to the 2.5D baseline (equal PCK, better AUC, slightly worse MPJPE). Further, on this dataset, access to ground truth scale information provides a larger improvement than on Human3.6M, highlighting the importance of testing on many subjects.

When tested on **truncated crops**, our method by far outperforms prior approaches (Table 6.5). This is true even for our default training configuration, but performance improves substantially when training on truncated images as well. The method is robust to truncation of up to 7 or 8 joints (of the 17) before overall performance substantially degrades (Figure 6.7). Given the obvious ambiguity introduced by truncation, it is noteworthy that even truncated joints can be estimated with as little as about 100 mm average error. Qualitative examples are in the second row of Figure 6.13, showing that our method can handle strongly truncated cases as well.

	Stand/ walk	Exer- cise	Sit on chair	Cro./ reach	On floor	Sport	Misc.	Green screen	No gr.sc.	Out- door	Total		
	PCK↑										PCK↑	AUC↑	MPJPE↓
Universal, height-normalized skeletons (simplified scale recovery task													
Rogez *et al.* (2017)*	70.5	56.3	58.5	69.4	39.6	57.7	57.6	–	–	–	59.7	27.6	158.4
Zhou *et al.* (2017)*H+M	85.4	71.0	60.7	71.4	37.8	70.9	74.4	71.7	64.7	72.7	69.2	32.5	137.1
Zhou *et al.* (2019)H+M	–	–	–	–	–	–	–	75.6	71.3	80.3	75.3	38.0	–
Mehta *et al.* (2017b)*3+M+L+H	87.7	77.4	74.7	72.9	51.3	83.3	80.1	–	–	–	76.6	40.4	124.7
Mehta *et al.* (2017a)*3+M+L+H	86.6	75.3	74.8	73.7	52.2	82.1	77.5	84.6	72.4	69.7	75.7	39.3	117.6
Mehta *et al.* (2018)*3+M+L+C	83.8	75.0	77.8	77.5	55.1	80.4	72.5	–	–	–	75.2	37.8	122.2
Luo *et al.* (2018a,b)3+M+H	**95.5**	82.3	89.9	84.6	66.5	92.0	93.0	–	–	–	84.3	47.5	84.5
Nibali *et al.* (2019)3+M	–	–	–	–	–	–	–	–	–	–	87.6	48.8	87.6
2.5D baseline3+M	95.1	90.7	86.8	**92.4**	**74.2**	94.1	91.7	92.1	89.0	**87.7**	89.9±0.2	52.8±0.4	79.7±0.6
MeTRo (ours)3+M	95.0	**91.8**	**90.2**	92.1	73.4	**95.1**	**91.8**	**93.4**	**90.3**	86.5	**90.6**±0.4	**56.2**±0.5	**74.9**±1.4
Metric-scale skeletons (full scale recovery task													
2.5D baseline3+M	93.1	**89.3**	83.6	**93.1**	**73.7**	**93.2**	**91.1**	89.0	**87.9**	**89.4**	**88.7**±0.6	48.6±1.3	**87.1**±2.2
MeTRo (ours)3+M	**94.0**	89.2	**87.1**	89.1	68.9	92.6	90.3	**90.1**	87.8	85.7	88.2±0.5	**48.7**±0.7	88.4±1.3

Table 6.3: Comparison on MPI-INF-3DHP with prior methods. *Evaluated with the first version of the dataset, with some annotation difference. Dashes (–) reflect a lack of published information. Superscripts indicate the training data (first characters of 3DHP, Human3.6M, MPII, LSP and COCO).

	Human3.6M			MPI-INF-3DHP		
	PCK↑	AUC↑	MPJPE↓	PCK↑	AUC↑	MPJPE↓
2.5D GT root depth	96.6	*68.8*	*49.0*	*90.8*	*56.1*	*74.2*
2.5D GT bone length	96.4	67.0	51.9	90.3	*56.1*	74.6
2.5D avg train bones	96.6	68.1	50.2	**89.6**	52.1	**80.6**
MeTRo (ours)	**97.0**	**68.6**	**49.3**	**89.6**	**52.6**	81.1

Table 6.4: Comparison with baseline methods of scale recovery, with or without access to ground truth information. For both datasets, metric-scale skeletons are used with the same 17 joints for comparability. The first two comparison methods access the ground truth at test time.

	All joints	Present joints
Mehta *et al.* (2017b)*	396.4	338.0
Zhou *et al.* (2017)*	400.5	332.5
Vosoughi and Amer (2018)	185.0	173.6
MeTRo*	124.7	76.8
MeTRo	**77.8**	**59.8**

Table 6.5: MPJPE scores on Human3.6M under truncation, evaluating all or only the present (within-frame, non-truncated) joints. (*Training was not performed with truncated crops.) Results of other methods are taken from Vosoughi and Amer (2018).

6.6.1 Speed–Accuracy Tradeoff

Given a bounding box crop, inference only requires a single forward pass of a standard backbone. Table 6.6 shows that 511 crops can be processed per second on an RTX 2080 Ti desktop GPU when operating on batches of 8 crops at stride 32 (the time cost of performing the detection stage is not considered). Varying the heatmap resolution using dense prediction provides diminishing returns (Table 6.6), showing that soft-argmax can cope with heatmaps of very coarse resolution. This means our method is attractive for use in top-down multi-person pose estimation systems as well.

6.7 Detailed Occlusion Experiments

In this section, we provide a deeper analysis the MeTRo method regarding occlusion robustness and occlusion augmentation effects, going beyond the experimental results published in Sárándi *et al.* (2020), the main paper this chapter is based on. These experiments were carried out using an earlier version of the MeTRo method, without

	Striding variant	Test stride			
		32	16	8	4
MPJPE	normal strides	53.1	52.5	52.7	52.9
	center-aligned	50.9	50.2	50.0	**49.3**
Speed (crop per sec.)	no batching	160	150	105	38
	batch size 8	**511**	475	292	92

Table 6.6: Impact of Striding on Speed and Accuracy. We measure inference speed (crops per second) and error (Human3.6M MPJPE) tradeoff with the two striding variants from Figure 6.4.

Geometry	Color	Occlusion	MPJPE
✓	–	–	58.0
✓	✓	–	52.8
✓	✓	✓	**49.3**

Table 6.7: Augmentations. We perform an ablation study on data augmentations on Human3.6M, showing that strong augmentation is important.

weak supervision with 2D pose estimation (see Figure 6.8), which is why we dedicate a separate section for these results.

First, we explore what makes augmentation with textured real objects so effective. Does it result from the realistic low-level image statistics of the synthetic occluder (the content of the occluder), or from the irregular silhouette (or shape), resulting from the use of object masks?

Next, we explore if the benefits of occlusion augmentation can be fully attributed to the fact that 3D datasets show limited appearance variability, which was our original motivation in pursuing this line of work. To test this, we perform an experiment with occlusion augmentation on 2D datasets consisting of in-the-wild images, as opposed to the 3D-annotated studio data that we used so far in this chapter. Any improvements that occlusion augmentation brings on such in-the-wild data cannot simply be attributed to an obvious lack of appearance variation in the images.

6.7.1 Default Setup

Here we use a simplified model depicted in Figure 6.8, where the weak supervision with 2D examples is not applied, and instead we perform pretraining based on 2D pose estimation on MPII. For the 3D task, the prediction heatmap has dimensions 16×16×16 except when marked otherwise.

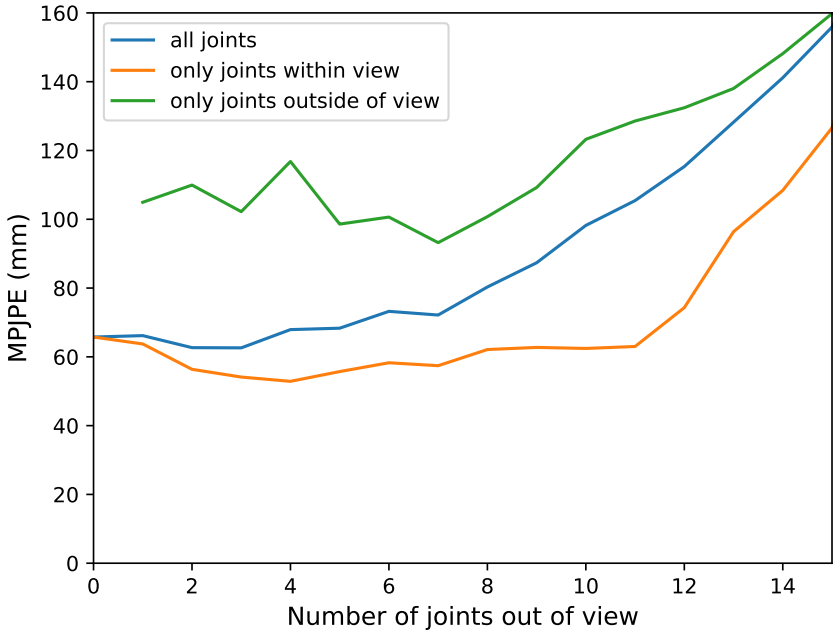


Figure 6.7: Analysis of robustness to truncation on Human3.6M. Average performance remains relatively stable up to 7 truncated joints.

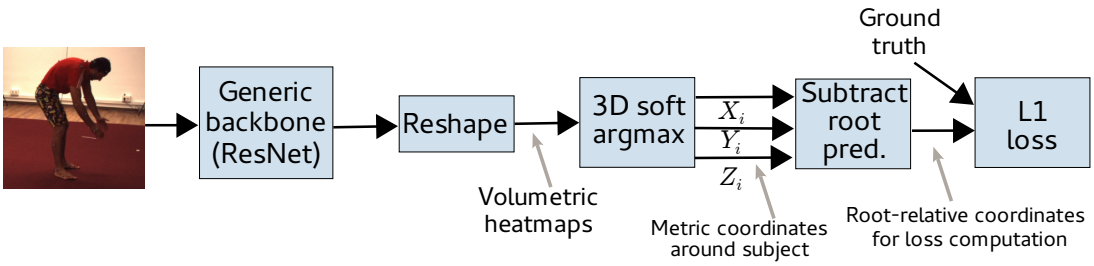


Figure 6.8: Overview of our approach used in Section 6.7. Here we do not use weak supervision with 2D data.

Adam optimizer is used with batch size 32 and exponentially decaying learning rate from 10^{-4} to 10^{-5}. We use 20 epochs on Human3.6M and 30 on 3DHP.

Training on MPII is performed using effective stride 32 in the backbone, resulting in 8x8 heatmaps for the 256×256 px inputs. The learning rate is decayed from 10^{-4} to 10^{-5} over 400 epochs, after which 10 epochs with 10^{-6} were performed. We trained on the 14 joints used in the official MPII evaluation protocol. Since the model trained with Obj-Texture augmentation performed best on the validation set, we use the weights obtained using this augmentation for all experiments designated as "MPII-pretrained." The pretraining was performed on the whole MPII train+val set.

Figure 6.9: Examples of random synthetic occlusion augmentations in conjunction with all other augmentations. Each column, in order, shows a Human3.6M and an MPI-INF-3DHP training example augmented with Rect-Noise, Rect-Texture, Obj-Noise and Obj-Texture augmentation, respectively. Background augmentation is also shown for MPI-INF-3DHP.

6.7.2 Augmentation: Shape *vs*. Content

To disentangle the effect of occluder shapes (silhouettes) and contents (the pixel values of the occluder), we experiment with two kinds of shapes and two kinds of contents. For shapes, we consider a rectangle (Rect) and object silhouettes (Obj; from Pascal VOC). For texture, we use uniform random white noise (Noise) and image content from Pascal VOC (Texture). There are thus four possible combinations of shape and content. In this terminology, Rect-Noise is the original random erasing augmentation, Obj-Texture was used in the previous chapter, Rect-Texture samples a rectangle as random erasing does but fills it with content from Pascal VOC images, and vice versa, Obj-Noise uses object silhouettes filled with random noise (Obj-Noise). See Figure 6.9 for samples of fully augmented training examples. Occlusions are applied with 70% probability for each image (independently from any background augmentation on 3DHP).

Table 6.8 shows how the synthetic occluder's shape and filling affect accuracy on the two 3D pose estimation datasets. We can see that on Human3.6M, textured objects yield best performance, while on 3DHP, which already uses background augmentation, rectangles filled with noise give best results.

6.7.3 Robustness Analysis

In this section, we report on occlusion-robustness experiments similar to those included in Chapter 4, but we now use our MeTRo representation, perform pretraining on MPII,

Figure 6.10: Occlusion robustness analysis on Human3.6M (top) and MPI-INF-3DHP (bottom). The four columns visualize robustness to three types of synthetic occlusion at test time: Pascal VOC objects (disjoint set from those used in augmentation), black circles, white oriented bars and a rectangle filled with a random solid color (see Figure 6.11). The degree of occlusion is the percentage of occluded pixels in the person bounding box. These experiments on MPI-INF-3DHP use background augmentation and universal skeletons.

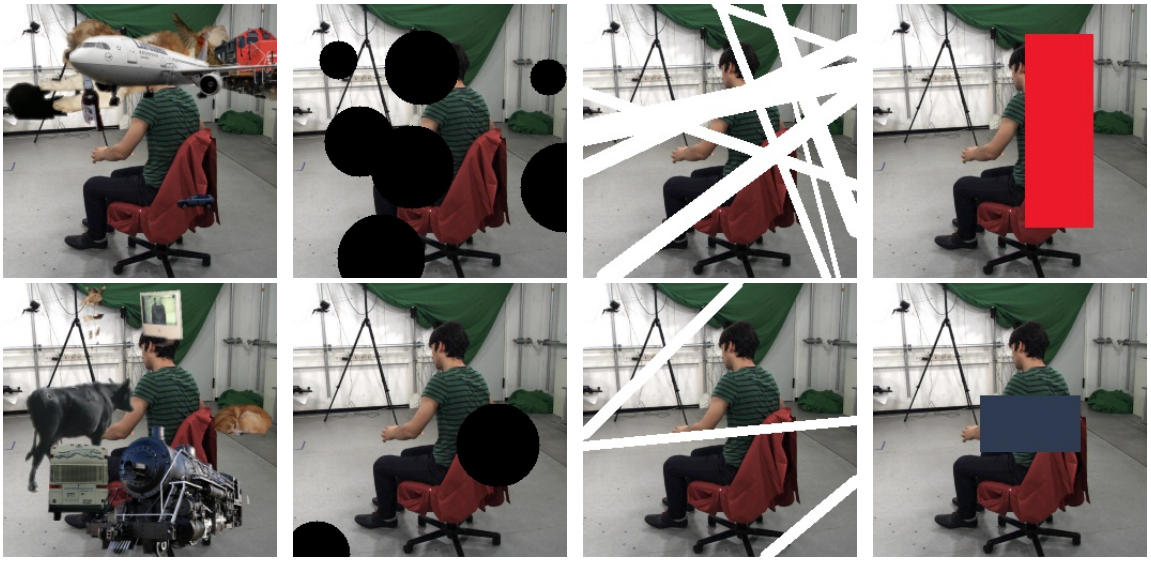


Figure 6.11: Examples of random synthetic occlusions over an MPI-INF-3DHP test example, used for robustness analysis. Each column shows two samples with objects from Pascal VOC, black circles, white oriented bars and a randomly-colored rectangle, respectively.

experiment with the MPI-INF-3DHP dataset besides Human3.6M and vary the colors of test-time occluders as well (as opposed to only using solid black as in Chapter 4). We apply three types of challenging synthetic occlusion patterns, not seen by any of the models during training: Pascal VOC objects (disjoint from those used for training augmentation), multiple black circles, multiple white oriented bars and a colored rectangle (see Figure 6.11).

As is shown in Figure 6.10, while augmenting with VOC objects (red line) generalizes to all test occlusion patterns, random erasing (blue line) is not as good for robustness.

Furthermore, the improvement in robustness is more modest on 3DHP than on Human3.6M. This can be attributed to the more difficult and diverse poses contained in 3DHP and that we already use background augmentation on 3DHP, which already injects additional appearance variation.

6.7.4 Distribution of Occluded Pixel Ratio

It is important to make sure that the two occlusion augmentation shapes (object and rectange) cover a similar number of pixels in the input image. This way we can isolate the effect of occluder shape, not size, when comparing augmentation results. We follow the pseudocode by Zhong *et al.* (2020) for generating the rectangle shaped occluder. We configure the Pascal VOC object occlusion generation such that the occluded pixel ratio within the image has similar distribution to the rectangle's: The number of objects

Occlusion augmentation	Human3.6M			MPI-INF-3DHP		
	PCK↑	AUC↑	MPJPE↓	PCK↑	AUC↑	MPJPE↓
None	89.5	54.3	76.0	85.4	54.7	85.8
Rect-Noise	93.4	59.2	65.8	**87.9**	**57.3**	**77.9**
Rect-Texture	94.1	59.8	64.2	85.6	55.3	84.6
Obj-Noise	93.2	57.7	68.2	85.4	55.1	86.4
Obj-Texture	**94.6**	**59.8**	**64.0**	85.4	55.2	87.6

Table 6.8: Occluder Shape *vs.* Filling. We analyze how the augmenting occluder's shape and filling affects accuracy on 3D pose estimation benchmarks. PCK and AUC are thresholded at standard 150 mm. 3DHP experiments here also use background augmentation. Training was performed only on the respective dataset.

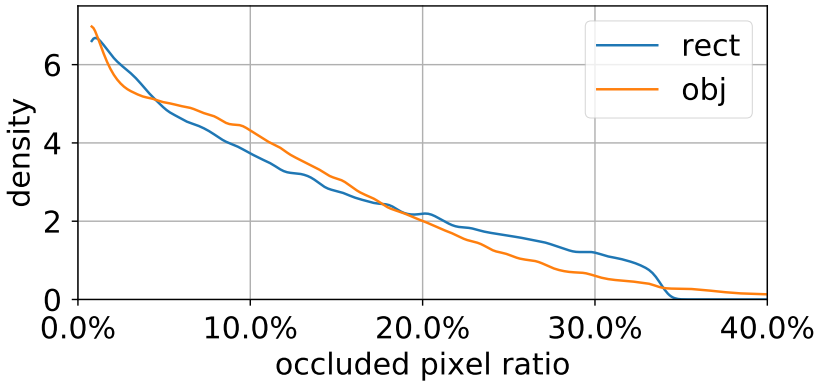


Figure 6.12: Probability density of the occluded pixel ratio for the two random shapes we use for occlusion augmentation. The curves were obtained by simulation and Gaussian smoothing. This verifies that the differences in results do not stem merely from different amount of pixels being occluded with the different augmentation methdos.

pasted is uniformly picked between 1 and 8, and each object is scaled by a uniformly distributed factor between 0.1 and 0.5 compared to their original size in the Pascal VOC dataset.

Random erasing's rectangle occludes 11.6% of pixels on average, with a standard deviation of 8.9 percentage points. Our Pascal VOC object augmentation has mean 11.1% and st. dev. 8.8 percentage points. Both skew towards small occlusion ratio. See Figure 6.12 to compare density plots.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Avg.
No occl. aug.	93.4	92.7	85.2	78.7	85.7	81.1	75.5	85.3
Rect-noise	94.2	**93.5**	86.2	78.9	86.9	81.2	75.4	85.8
Rect-texture	93.8	**93.5**	86.2	79.3	**87.1**	81.9	76.4	86.1
Obj-noise	93.7	93.3	85.7	78.9	86.6	81.0	76.3	85.7
Obj-texture	**94.3**	93.4	**86.3**	**79.6**	86.7	**82.8**	**77.1**	**86.3**

Table 6.9: PCKh@0.5(%) comparison of different occlusions augmentation schemes on the MPII validation set. For reference, performance with obj-texture augmentation on the official, held-out MPII test set is 87.6%.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
No occl. aug.	95.8	91.8	82.7	77.2	92.3	90.3	86.3	88.2
Rect-Noise	95.7	**92.1**	84.1	77.8	93.1	90.8	86.5	88.7
Obj-Texture	**97.1**	91.9	**84.5**	**78.6**	**93.3**	**91.0**	**87.0**	**89.1**

Table 6.10: PCK@0.2(%) comparison on LSP-test.

6.7.5 2D In-The-Wild Occlusion Experiment

Recall that our original motivation in detailed investigation of occlusion augmentations was the limited appearance variation in current 3D pose datasets that are recorded in motion capture studios. To test whether this type of augmentation holds broader relevance in pose estimation than simply making up for a lack of appearance variation, we set out to apply the same augmentations to the much more varied, *in-the-wild* MPII 2D human pose dataset and evaluate the change in 2D human pose estimation accuracy.

Our baseline network for this experiment is a simplified variant of Nibali *et al.* (2018): a ResNet50V2 backbone predicts 2D body joint heatmaps of size 8×8 which are decoded using soft-argmax and the ℓ_1 loss is minimized on the resulting coordinates. The results are shown in Table 6.9. We observe that even when applied to this in-the-wild dataset, all four combinations of occluder shape and filling bring improvements and Obj-Texture performs best. Similar results are achieved on the LSP (Table 6.10) and FLIC (Table 6.11) datasets, where we compare Rect-Noise with Obj-Texture.

6.8 Detailed Ablations

We report more ablations to evaluate trade-offs and justify design choices in our MeTRo model. We focus on the recently introduced and more challenging MPI-INF-3DHP dataset, as it has been less studied in the past. The starting point for all these experiments is the configuration with background and Rect-Noise (*i.e.*, random erasing;

	Wrist PCK
No occl. aug.	96.3
Rect-Noise	96.5
Obj-Texture	**97.2**

Table 6.11: PCK@0.2(%) comparison on FLIC-test. Today, FLIC counts as a relatively simple dataset, hence we focus on the most difficult joint, the wrist in this ablation.

	Same MPII pretraining			Randomized MPII pretraining		
	PCK	AUC	MPJPE	PCK	AUC	MPJPE
	88.9	58.6	73.9	88.9	58.6	73.9
	89.1	58.6	74.2	89.2	58.6	73.2
	89.4	58.9	73.0	89.4	58.6	73.1
	88.8	58.4	75.8	88.7	58.2	74.7
	89.3	58.7	73.0	89.3	58.3	74.0
Mean	89.1	58.6	74.0	89.1	58.5	73.8
St. deviation	0.25	0.16	1.05	0.27	0.15	0.60
Coeff. of var. (%)	0.28	0.27	1.42	0.30	0.27	0.81

Table 6.12: Repeatability study on MPI-INF-3DHP for our default configuration as described in Section 6.8.

Zhong *et al.*, 2020) augmentation and MPII-pretraining, but no weak 2D supervision. This is referred to as the *default configuration* or *default experiment* from here on. We indicate the default setting with italics in all tables.

6.8.1 Repeatability

As usual in deep learning, our training procedure is stochastic due to random augmentations, training set shuffling and initialization. So that we can better interpret the ablation results, in Table 6.12, we quantify the stochasticity of the training process by rerunning the default experiment four more times on MPI-INF-3DHP from the same MPII-pretrained weights and also using a new pretraining each time. The obtained mean is used for further comparisons.

6.8.2 Backbone Size

By using a more powerful backbone, one can trade off computational complexity for higher accuracy. For our main experiments, we use ResNet50V2 for practical reasons. In Table 6.13, we can see the results for two larger backbones that further improve

Backbone	PCK	AUC	MPJPE	Train fps	Test fps	#Params
ResNet50V2	89.1	58.6	74.0	**126***	**250**	**24M**
ResNet101V2	90.0	59.5	71.6	78	168	43M
ResNet152V2	**90.5**	**59.9**	**69.8**	60	124	58M

Table 6.13: Comparison of performance on MPI-INF-3DHP with different backbones from the ResNetV2 family. * indicates that CPU-based image preprocessing becomes the bottleneck.

Method	w/o Procrustes			w/ Procrustes		
	PCK	AUC	MPJPE	PCK	AUC	MPJPE
Single-crop						
Nibali *et al.* (2019)	87.6	48.8	87.6	**94.8**	61.4	61.6
Ours (*default*)	**89.1**	**58.6**	**74.0**	92.9	**63.2**	**61.5**
Ours (no occl. aug.)	87.2	57.1	79.7	91.5	61.4	66.1
Multi-crop						
Nibali *et al.* (2019)	88.3	49.6	85.2	**95.1**	62.2	60.1
Ours (*default*)	**89.9**	**59.7**	**70.8**	93.3	**64.6**	**58.8**
Ours (no occl. aug.)	87.8	58.0	76.7	91.9	62.5	63.8

Table 6.14: Evaluation with and without Procrustes alignment and multi-crop evaluation on MPI-INF-3DHP universal skeletons.

results. Speed is evaluated in terms of frames processed per second (fps), taking into account minibatching with size 32 during training and 8 during testing. The experiment was run on a single NVIDIA TitanX (Pascal) GPU.

6.8.3 Procrustes Analysis and Multi-Crop Evaluation

In Table 6.14, we compare our results with a top-performing concurrent work on MPI-INF-3DHP by Nibali *et al.* (2019) (MargiPose). Nibali *et al.* use background and clothing augmentation. We include results with background augmentation and either with or without random-erasing occlusion augmentation.

We also compare under the multi-crop test-time augmentation protocol of Luvizon *et al.* (2018) and Nibali *et al.* (2019) in Table 6.14, as this enables direct comparison. Despite the fact that Nibali *et al.* use a multi-stage, custom architecture with intermediate supervision, our method with just a ResNet50V2 backbone outperforms their approach on several metrics and is competitive on the others. Our method compares especially well when no Procrustes alignment is used, *i.e.*, when scale recovery is necessary.

Initialization / Pretraining	PCK	AUC	MPJPE
Random (He) initialization	76.9	46.7	112.8
ImageNet	87.9	57.3	77.9
MPII (without ImageNet)	85.5	54.3	85.4
ImageNet + MPII	89.1	58.6	74.0
ImageNet + MPII + Human3.6M	**89.6**	**59.3**	**73.0**

Table 6.15: Evaluation under different weight initialization and pretraining methods on MPI-INF-3DHP. Results get better as more data is seen in pretraining.

6.8.4 Initialization and Pretraining

Table 6.15 shows results obtained with different initializations. As expected, the more data we use to pretrain the network, the better the accuracy.

6.8.5 Multi-Dataset Training

We report multi-dataset experiments in Table 6.16. When jointly training, we follow Zhou *et al.* (2017) in adjusting the pelvis and hips to be more compatible across datasets: we move these joints of MPI-INF-3DHP skeletons towards the neck by a fifth of the neck-pelvis vector. In this case background augmentation is also used on Human3.6M to treat the datasets consistently.

We find that jointly training on both MPI-INF-3DHP and Human3.6M results in a single model that has superior performance on both test sets simultaneously, compared to training dataset–specific models or mere pretraining. This becomes an important motivation for further pursuing joint multi-dataset training in Section 7.6 and Chapter 8.

6.8.6 Skeleton Normalization

Our main results presented for MPI-INF-3DHP were obtained on the universal (*i.e.*, height-normalized) skeletons to remain comparable to prior work. However, our metric 3D heatmap formulation does not require such normalization and can also be trained on the actual (non-normalized, non-universal) skeletons by simply using those as the target labels during training.

An interesting question is whether training on actual skeletons helps the model learn scale recovery or whether it silently "ignores" the person scale information and simply learns what can already be learned from universal skeletons. Table 6.17 shows that training on actual skeletons does result in improved scale recovery (total PCK increases from 86.5% to 87.8% by 1.3 points). This works best (+1.9 PCK points) on the green-screen studio test examples (same scene as in training), since the network

Training	Test dataset					
	Human3.6M			MPI-INF-3DHP		
	PCK	AUC	MPJPE	PCK	AUC	MPJPE
Metric skeletons						
Dataset-specific	95.5	61.5	60.8	87.8	52.3	85.0
Jointly trained	**96.4**	**63.5**	**57.2**	**90.1**	**53.4**	**79.9**
Universal skeletons						
Dataset-specific	–	–	–	89.1	58.6	74.0
Pretrained on the other	–	–	–	89.6	**59.3**	73.0
Jointly trained	–	–	–	**91.3**	59.0	**70.2**

Table 6.16: Joint Two-Dataset Training. We obtain improved results by training a single model on both MPI-INF-3DHP and Human3.6M. While pretraining on another dataset also helps, joint training is more effective. (We perform the universal-skeleton experiments on 3DHP only, as common in the literature.)

Training skeletons	Test PCK on metric skeletons			
	Green Screen	No Gr.Sc.	Outdoor	Total
Universal	90.6	85.1	82.3	86.5
Metric	**92.5**	**86.4**	**82.4**	**87.8**

Table 6.17: Analysis of the method's ability to learn scale recovery when trained on actual, raw skeletons. Good performance on the actual skeletons requires implicit scale estimation by the backbone.

can relate the test person's size to cues in the known background. Scale recovery is less successful when the green screen is removed in the studio (+1.3 points), and the advantage essentially disappears when testing in the outdoor setting (+0.1 points). We hypothesize that learning scale recovery that generalizes to unknown scenes would require observing more people with various heights during training.

6.8.7 Training Length

To make sure that we have trained our models for long enough, we show that doubling the training length does not result in further significant improvement. Recall that our learning rate schedule is exponential decay from 10^{-4} to 10^{-5} over 30 epochs and 1 epoch with 10^{-6}. In Table 6.18 we show that extending the exponential decay phase from 30 to 60 epochs gives no further benefit and the numbers remain within one standard deviation of the repeatability study shown in Table 6.12.

Training epochs	PCK	AUC	MPJPE
30	**89.1**	58.6	**74.0**
60	89.0	**58.8**	74.7

Table 6.18: Training Length. We verify that 30 epochs are sufficient to train the model.

Heatmap res.	PCK	AUC	MPJPE	Train fps	Test fps
$8 \times 8 \times 8$	87.9	57.2	78.5	**126***	303
$16 \times 16 \times 16$	89.1	58.6	74.0	**126***	250
$32 \times 32 \times 32$	89.4	**59.3**	72.3	43	117
Reduced depth resolution					
$8 \times 8 \times 4$	89.0	57.8	75.0	**126***	**305**
$16 \times 16 \times 4$	89.0	58.2	74.8	**126***	250
$32 \times 32 \times 4$	**89.8**	**59.3**	**71.7**	48	120

Table 6.19: Impact of Heatmap Resolution. Higher spatial resolution in the heatmaps yields better performance but coarse depth is also sufficient. * indicates that our CPU-based image preprocessing becomes the bottleneck.

6.8.8 3D Heatmap Resolution

A computational cost *vs*. accuracy tradeoff exists in the choice of the heatmap resolution. As shown in Table 6.19, increasing spatial resolution in the X and Y axes improves results (slightly). However, the resolution along the depth axis can be decreased to as low as 4 with good performance.

6.8.9 Metric Size of the Volume

We chose 2200 mm as the metric side length of the volume represented by the 3D heatmap predicted by the backbone, so that all training skeletons fit inside with some margin. Table 6.20 shows that the results are not very sensitive to this hyperparameter.

6.9 Conclusion

We proposed metric-scale truncation-robust (MeTRo) volumetric heatmaps in the context of 3D human pose estimation. These heatmaps directly represent the metric space around the person instead of being tied to the image space and can be predicted with any standard fully convolutional network. With a modified weak supervision scheme for 2D labels, careful stride alignment considerations and strong data augmentation,

Volume side length	PCK	AUC	MPJPE
1800 mm	**89.3**	**58.7**	**73.6**
2000 mm	89.3	58.6	73.4
2200 mm	89.1	58.6	74.0
2400 mm	89.0	58.6	74.3

Table 6.20: Impact of Volume Size. Results as the metric volume size is varied show little sensitivity to this hyperparameter.

Figure 6.13: Qualitative Results. Predictions are shown in color, ground truth in gray (except for MPII, where it is unavailable). Green spheres mark predictions within 150 mm of the ground truth, red cubes beyond that threshold. Note that our method performs well on truncated (partial body) images as well (second row).

we achieved state-of-the-art results on two important benchmarks: Human3.6M and MPI-INF-3DHP. In carefully controlled experiments, we showed that our approach can implicitly discover scale cues from the data and outperforms a previously proposed explicit bone length based heuristic on all test scenarios except the two outdoor sequences of MPI-INF-3DHP. Future research should consider possibilities for learning similar scale cues from large-scale outdoor data as well. We also performed a detailed analysis of occlusion augmentation and robustness in conjunction with the MeTRo

method, as well as detailed ablations and hyperparameter studies. Another interesting future direction can be the evaluation on people with widely differing heights, if such data becomes available on a large scale. Beyond scale recovery, we demonstrated the second benefit of the MeTRo representation, the prediction ("hallucination") of complete skeletons even when only a part of the body is contained in the image. Given its speed and robustness to detection noise, we expect our approach to be useful in designing top-down multi-person pose estimation systems in the future.

7 MeTRAbs: An End-to-End Learned Absolute 3D Pose Estimator

In this chapter, we apply the MeTRo heatmap concept introduced in Chapter 6 to the task of absolute (*i.e.*, non-root-relative) 3D human pose estimation.

Absolute pose estimation is especially important in multi-person scenarios, in order to recover the spatial layout of the whole group. We therefore turn to the multi-person setting in this chapter, after having focused on single-person estimation in the previous chapters. For this, we use the top-down multi-person paradigm, first running detection then pose estimation for each person.

State-of-the-art results of the MuPoTS-3D benchmark, as well as achieving first place in the 3D Poses in the Wild competition at ECCV 2020 demonstrate the effectiveness of our approach.

This chapter is based on our journal article (Sárándi *et al.*, 2021), published in the IEEE Transactions on Biometrics, Behavior, and Identity Science. As additional content, we include performance measurements on embedded hardware using more efficient backbone networks, showing that our method is real-time capable on low-powered hardware as well.

7.1 Overview

Most of the 3D human pose estimation literature is concerned with the so-called *root-relative* version of this task. This means that each body joint's position is estimated relative to the person's center joint, *i.e.*, the pose is only recovered up to translation and the position of the root joint relative to the camera is not estimated. This may be sufficient for applications such as gesture and action recognition, but for a robot navigating in a crowd, it is important to know where each person is located in the absolute 3D space, *e.g.*, for pathfinding.

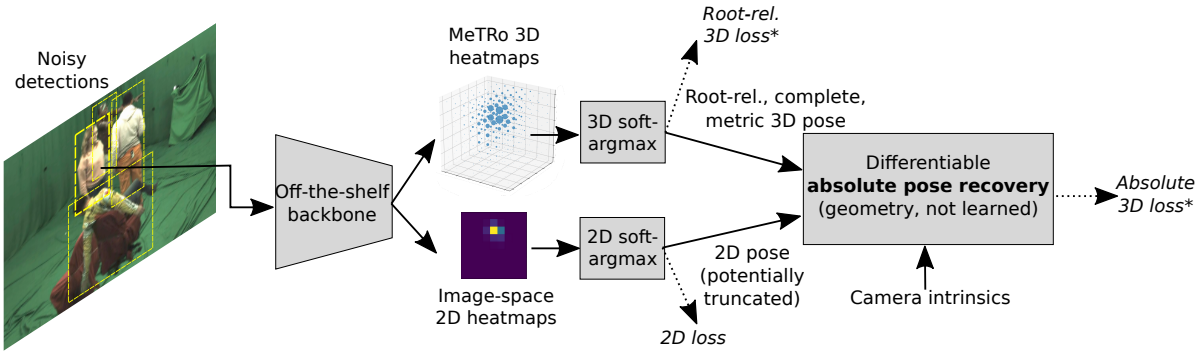


Figure 7.1: MeTRAbs Architecture. We predict two types of body joint heatmaps with a standard fully convolutional backbone: metric-scale truncation-robust 3D heatmaps (as defined in Chapter 6), as well as pixel-space 2D heatmaps. We decode the heatmaps to coordinates via soft-argmax. Intermediate ℓ_1 losses are applied on these decoded 2D and root-relative 3D poses. Finally, we reconstruct the absolute pose through a differentiable linear-least-squares optimization module derived from the pinhole camera model. Importantly, supervision is also applied on this final combined output, and its gradients are backpropagated all the way to the backbone, resulting in end-to-end optimized absolute pose prediction. (*For 2D-labeled examples, the root-relative loss is replaced by a scale and translation-invariant 2D loss and the absolute 3D loss is not used.)

To obtain absolute predictions, we combine 3D metric-scale root-relative heatmaps and 2D image-space heatmaps in a two-headed CNN architecture, and subsequently reconstruct the absolute 3D root position in a differentiable manner. We evaluate our network in a top-down fashion combined with an off-the-shelf person detector and refer to this combined approach as MeTRAbs. While prior approaches have tackled the root reconstruction problem, to our knowledge we are the first to backpropagate gradients through this type of reconstruction, allowing us to end-to-end supervise the absolute pose task. We show that this is crucial for good distance estimation, and extensively evaluate strong and weak perspective–based reconstruction variants.

Quantitatively, we achieve state-of-the-art results of the time on the popular multi-person dataset MuPoTS-3D. Further demonstrating the effectiveness and scalability of our approach, we achieve first place in the 2020 ECCV 3D Poses in the Wild (von Marcard *et al.*, 2018) Challenge using additional training data.

Similar to Chapter 6, we perform extensive comparisons with 2.5D heatmap learning using bone length–based scale recovery (Pavlakos *et al.*, 2017), under otherwise equal training conditions and find that our end-to-end formulation is more effective in this task.

We release our code publicly to enable further followup research.[1].

[1] `https://vision.rwth-aachen.de/metrabs`

7.2 Related Work

Most monocular 3D pose estimation methods have typically only been evaluated in a root-relative manner. However, some recent works have also explicitly tackled the absolute (non-root-relative) pose estimation task, where every joint position is predicted within the 3D camera coordinate frame. This is closely related to metric-scale prediction discussed at length in the previous chapter: if both the image-space pose and the metric-scale root-relative pose are known, one can reconstruct the absolute distance (assuming a calibrated camera).

Mehta *et al.* (2017a) and Dabral *et al.* (2019) reconstruct the root offset by assuming a weak perspective model. Mehta *et al.* (2019) assume the foot touches the known ground plane in the first frame. Moon *et al.* (2019) predict the metric area of the human bounding box as a numerical value via a separate deep network (RootNet), besides their root-relative 2.5D PoseNet. In contrast to Moon *et al.*, we estimate the scaled pose fully convolutionally and do not require multiple separate backbones. In our earlier work Sárándi *et al.* (2018b) (described in Chapter 5), we estimated the distance directly from the image crop, however that approach does not generalize well to novel environments. Dabral *et al.* (2019) propose to estimate the focal length jointly with the distance, implicitly relying on the perspective distortion of people far from the optical axis. As the authors note, this cannot work well when the camera is turned directly towards the target person. Véges and Lőrincz (2019) make use of a monocular depth prediction network pretrained on various indoor and outdoor datasets to help with absolute person distance estimation.

In video estimation with known frame rate, Bieler *et al.* (2019) used gravitational acceleration as a scale reference.

Finally, some recent works also consider the depth relations among people: Jiang *et al.* (2020) optimize the depth ordering by occlusion cues, while Fieraru *et al.* (2020) explicitly localize contact points between people to help with coherent reconstruction. In contrast, we perform our estimation for each person independently.

7.3 Method

In this section, we describe MeTRAbs, consisting in a combination of MeTRo 3D heatmap estimation presented in Chapter 6 with traditional 2D pose heatmaps in a single end-to-end trained network for absolute 3D pose estimation. The overall idea is that the MeTRo approach implicitly estimates the scale, which we can then use to infer the distance. By applying this method within a top-down paradigm (detection, cropping, pose estimation), we obtain a fast and accurate way to tackle multi-person absolute 3D pose estimation.

As discussed in Chapter 6, we first estimate a complete metric-scale pose $\left\{(\Delta X_j, \Delta Y_j, \Delta Z_j)^T\right\}_{j=1}^J$ up to translation (where J is the number of joints).

By additionally estimating the 2D, image-space pose $\left\{(x_j, y_j)^T\right\}_{j=1}^J$, we obtain all the necessary information to recover the absolute 3D pose in the (calibrated) camera coordinate system, as we will see in the following. We assume known camera intrinsics, since monocular focal length estimation (Kar *et al.*, 2015; Workman *et al.*, 2015) is a very challenging task (*cf*. the "dolly zoom" effect; Liang *et al.*, 2020).

The absolute pose can be expressed as

$$\left\{(X_0 + \Delta X_j,\ Y_0 + \Delta Y_j,\ Z_0 + \Delta Z_j)^T\right\}_{j=1}^J, \tag{7.1}$$

with (X_0, Y_0, Z_0) being the absolute pose offset, which we aim to recover. For this, we first calculate the normalized image coordinates as $(\tilde{x}_j, \tilde{y}_j)^T = K^{-1}(x_j, y_j)^T$, where K is the intrinsic matrix.

Mehta *et al.* (2017a) derive a formula to reconstruct the absolute root position using the weak perspective projection model. Véges and Lőrincz (2019), while still operating in the weak perspective model, note that an approximation step involved in Mehta *et al.*'s algorithm leads to worse performance. Motivated by this, we derive a reconstruction method under the full perspective pinhole camera model and extensively compare it with Mehta *et al.*'s weak perspective method. In a full perspective model, a perfect estimate would satisfy

$$\begin{bmatrix}\tilde{x}_j \\ \tilde{y}_j\end{bmatrix} = \begin{bmatrix}(X_0 + \Delta X_j)/(Z_0 + \Delta Z_j) \\ (Y_0 + \Delta Y_j)/(Z_0 + \Delta Z_j)\end{bmatrix}, \tag{7.2}$$

which can be rearranged to

$$\begin{bmatrix}X_0 - \tilde{x}_j Z_0 \\ Y_0 - \tilde{y}_j Z_0\end{bmatrix} = \begin{bmatrix}\tilde{x}_j \Delta Z_j - \Delta X_j \\ \tilde{y}_j \Delta Z_j - \Delta Y_j\end{bmatrix}. \tag{7.3}$$

Considering all joints, we obtain $2J$ linear equations in the three variables (X_0, Y_0, Z_0). Since $\tilde{x}, \tilde{y}, X, Y$ and Z are estimates, the equation system is noisy and over-determined. Hence we opt to solve it by linear least squares, with a differentiable solver based on Cholesky decomposition. This differentiability allows us to directly supervise the network with a loss $\mathcal{L}^{\text{abs3D}}$ computed on the final absolute 3D pose output, which turns out to be crucial for accurate distance estimation.

For truncated images, (7.2) only holds for body joints inside the image frame, since the 2D heatmap method cannot estimate out-of-image joint locations. We therefore exclude joints from the optimization, which are predicted to lie closer to the image border than one stride length. After reconstructing the root joint position, we can obtain the absolute pose in two ways. Either as $(\Delta X_j + X_0, \Delta Y_j + Y_0, \Delta Z_j + Z_0)^T$ (adding

the reconstructed offset to the 3D head's root-relative output), or as $(\tilde{x}_j, \tilde{y}_j, 1)^T \cdot (\Delta Z_j + Z_0)$ (back-projecting the 2D head's output). For joints that lie within the image, we use the latter option, while for truncated ones we use the former. Both the individual prediction heads and the final absolute output are supervised with the ℓ_1 loss. As in the root-relative MeTRo network, we apply weak supervision from 2D-labeled data for MeTRAbs as well, on both heads. Extending Equation 6.2, the loss becomes

$$\mathcal{L} = \mathcal{L}^{\text{abs3D}}_{\text{ann3D}} + \mathcal{L}^{\text{head3D}}_{\text{ann3D}} + \mathcal{L}^{\text{head2D}}_{\text{ann3D}} + \lambda \left(\mathcal{L}^{\text{head2D}}_{\text{ann2D}} + \mathcal{L}^{\text{head3D}}_{\text{ann2D}} \right), \tag{7.4}$$

where we again set $\lambda = 0.1$.

We found that the absolute loss can introduce numerical instabilities very early during training, since at this point the two prediction heads do not yet produce sufficiently compatible outputs, making the reconstruction problem ill-conditioned. Hence, we only turn on the absolute loss after 5000 update steps.

In a multi-person scenario, inference speed becomes a priority, since the model is evaluated on each person detection separately. To retain real-time performance, we do not apply dense prediction with MeTRAbs; the network is trained and tested with coarse, 8×8×8 heatmaps.

7.4 Datasets, Preprocessing, Evaluation

We evaluate our method in a multi-person context by training on MuCo-3DHP and testing on MuPoTS-3D.

MuCo-3DHP (Mehta *et al.*, 2018) is a synthetically composited multi-person dataset, derived from 3DHP by pasting persons over each other based on their root joint depth order. As Véges and Lőrincz (2019), we generate 150k training images, each with 4 people. We run YOLOv3 on these images to get realistic bounding boxes.

MuPoTS-3D (Mehta *et al.*, 2018) is a mixed indoor and outdoor multi-person test set, compatible with MuCo-3DHP, consisting of 20 sequences showing people performing various actions and interactions. MuPoTS-3D provides normalized and unnormalized skeletons. Indoor sequences are gamma-corrected with an exponent of 0.67

We use the same preprocessing and augmentation configuration as in Chapter 6, with the only difference that synthetic occlusion probability is reduced to 30% since some occlusion is already introduced from compositing person segments over each other.

Since we are studying the multi-person setting here, we use all person instances from the 2D MPII dataset and obtain realistic bounding boxes for them with YOLOv3 (Redmon and Farhadi, 2018).

We use the standard **evaluation measures**, which we defined in Section 3.4. The main one on MuPoTS-3D is the the percentage of correct keypoints (PCK) with threshold 150 mm. The AUC is the area under the PCK curve as the threshold ranges from

	A-MPJPE↓	MPJPE↓	A-PCK↑	PCK↑	Det.Rate↑
Rogez *et al.* (2017)	–	146‡	–	–	86
Mehta *et al.* (2018)	–	132‡	–	–	93
Baseline in Véges and Lőrincz (2019)	320†	122‡	–	–	91
Véges and Lőrincz (2019)	292†	120‡	–	–	91
Véges and Lőrincz (2020a)*	257.2 (255†)	119.4 (108‡)	38.1	75.4	93
2.5D baseline	317.6 (313.6†)	114.0 (110.0‡)	**40.0**±1.0	79.3±0.3	**94.2**±0.0
MeTRAbs	**248.2** (**246.9**†)	**108.2** (**104.3**‡)	**40.2**±1.9	**81.1**±0.4	94.1±0.1
w/o abs. loss	328.8 (327.8†)	108.4 (104.7‡)	36.7±3.2	80.9±0.4	94.1±0.1

Table 7.1: Results on MuPoTS-3D. Detected, unnormalized poses, no bone rescaling. (*Re-evaluated public results; joint count: †17, ‡16, else 14)

0 to 150 mm. The official MuPoTS-3D evaluation script rescales the bone lengths of the prediction to match the ground truth bone lengths before computing metrics, leading to some confusion and inconsistency between reported results. In Mehta *et al.* (2018) rescaling was only used for evaluating LCR-Net (Rogez *et al.*, 2017), but it has since been adopted by other authors as well. For consistency and simplicity, we train MeTRAbs only with unnormalized skeletons. When evaluating on universal (normalized) skeletons, we use bone rescaling. On unnormalized skeletons, we do not use bone rescaling, in order to directly evaluate the raw metric-space outputs of the methods. Note that bone rescaling to the ground truth can counter-intuitively lead to worse scores due to error accumulation along the kinematic chain. For example, if the estimated wrist position is correct but the elbow is wrong, bone rescaling can shift the wrist prediction away and make it appear worse.

Following Véges and Lőrincz (2020a), on MuPoTS-3D we also evaluate absolute (*i.e.*, non-root-relative) metrics, prefixed with "A-," *e.g.*, A-PCK. For absolute MPJPE, Véges and Lőrincz (2019, 2020a) evaluate all 17 joints, and for relative MPJPE only 16 (no pelvis), and use the 14 MPII joints for PCK and A-PCK. By default we use 14 joints on MuPoTS-3D, except when marked otherwise in the tables.

7.5 Results

On **MuPoTS-3D**, our approach yields state-of-the-art results. For height-normalized skeletons with bone rescaling (standard setting in prior work, Table 7.4 and Table 7.5), MeTRAbs outperforms the 2.5D baseline, and the baseline already reaches state-of-the-art results. Our method performs particularly well on test sequence 2, with heavy occlusions (*e.g.*, Figure 7.2, left). Removing the supervision with the absolute 3D loss worsens the absolute PCK of all poses from 38.4% to 35.0%. Surprisingly, the

Persp. assumption		All annotations		Matched annotations	
Training	Test	A-PCK↑	PCK↑	A-PCK↑	PCK↑
F	F	37.2±1.7	76.2±0.5	39.3±1.7	79.9±0.5
F	W	**39.4**±1.6	76.2±0.5	**41.6**±1.6	80.0±0.5
W	F	35.6±1.8	77.1±0.4	37.6±1.8	81.0±0.5
W	W	38.1±1.8	**77.2**±0.4	40.2±1.9	**81.1**±0.4
–	F	33.0±3.3	77.0±0.4	34.9±3.3	80.8±0.4
–	W	34.8±3.1	77.0±0.4	36.7±3.2	80.9±0.4

Table 7.2: Comparison of weak (W) and full (F) perspective-based absolute pose reconstruction. The two letters denote the training and the test time variant. (Unnormalized skeletons, without bone rescaling.)

	MPJPE↓	MPJPE-PA↓	PCK@50mm↑	AUC@200mm↑
Known association to GT identity				
Sun *et al.* (2020)	81.8	58.6	37.3	59.9
Kissos *et al.* (2020)	83.2	59.7	42.4	62.3
MeTRAbs (ours)	**68.8**	**49.7**	**48.8**	**66.8**
Unknown association to GT identity				
MeTRAbs (ours)	85.1	56.7	45.8	63.2

Table 7.3: Results on the 3DPW challenge dataset. (PA=Procrustes analysis)

root-relative accuracy seems to improve when turning off the absolute loss. This is, however, hard to interpret, as Table 7.4 shows an artificial evaluation setting with normalized-height skeletons and bone rescaling, thereby removing some of the scale recovery aspect from the evaluation. When evaluating on unnormalized skeletons without bone rescaling (Table 7.1), it becomes clear that the absolute loss helps: absolute MPJPE improves from 328.8 mm to 248.2 mm, absolute PCK from 36.7% to 40.2%, with the root-relative metrics slightly improving as well. These are state-of-the-art results and improve over methods that are pre- or jointly trained on ground truth pixel-wise depth prediction datasets (Véges and Lőrincz, 2019, 2020a). Further, we can see that the absolute PCK score has high variance and therefore small differences are not necessarily meaningful. The standard deviation across 5 repeat experiments is around 1.4–3.2%, and the absolute results for individual test sequences varies strongly as well across different configurations. This is because the test examples are strongly correlated since they come from video sequences. Lastly, we note that the detection rate is essentially the same for all of our configurations (Table 7.1), since we use the same detections, and the official evaluation script performs matching based on the 2D projection, which is very similar across these methods.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg↑
Root-relative PCK for all annotated poses																					
Rogez *et al.* (2017)	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
Mehta *et al.* (2018)	81.0	60.9	64.4	63.0	69.1	30.3	65.0	59.6	64.1	83.9	68.0	68.6	62.3	59.2	70.1	80.0	79.6	67.3	66.6	67.2	66.0
Rogez *et al.* (2019)	87.3	61.9	67.9	74.6	78.8	48.9	58.3	59.7	78.1	89.5	69.2	73.8	66.2	56.0	74.1	82.1	78.1	72.6	73.1	61.0	70.6
Moon *et al.* (2019)	**94.4**	77.5	79.0	81.9	85.3	72.8	81.9	**75.7**	**90.2**	90.4	79.2	79.9	75.1	**72.7**	81.1	89.9	89.6	81.8	**81.7**	**76.2**	81.8
Dabral *et al.* (2019)	85.1	67.9	73.5	76.2	74.9	52.5	65.7	63.6	56.3	77.8	76.4	70.1	65.3	51.7	69.5	87.0	82.1	80.3	78.5	70.7	71.3
Véges and Lőrincz (2020a)	89.5	75.9	85.2	83.9	85.0	73.4	83.6	58.7	65.1	90.4	76.8	81.9	67.0	55.9	80.8	90.6	**90.0**	81.1	81.1	68.6	78.2
Mehta *et al.* (2019)	89.7	65.4	67.8	73.3	77.4	47.8	67.4	63.1	78.1	85.1	75.6	73.1	65.0	59.2	74.1	84.6	87.8	73.0	78.1	71.2	72.1
Benzine *et al.* (2021)	78.1	62.5	55.5	63.8	70.2	50.8	73.8	65.3	55.1	79.3	70.4	72.3	65.4	55.3	65.2	81.3	77.2	75.9	74.2	71.6	67.5
2.5D baseline	93.0	76.4	88.6	85.2	86.3	75.7	84.3	67.9	84.3	**93.4**	81.6	89.8	77.3	67.7	83.8	91.0	86.1	84.8	77.1	71.2	82.3±0.1
MeTRAbs	93.8	80.8	**89.3**	**87.0**	86.6	74.5	83.7	66.2	85.0	92.9	80.4	89.6	77.1	68.7	**86.3**	92.0	86.6	84.4	77.3	71.4	82.7±0.3
w/o abs. loss	94.0	**82.6**	88.4	86.5	**87.3**	**76.2**	**85.9**	66.9	85.8	92.9	**81.8**	**89.9**	**77.6**	68.5	85.6	**92.3**	89.3	**85.1**	78.2	71.6	**83.3**±0.2
Root-relative PCK for detected poses																					
Rogez *et al.* (2017)	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Mehta *et al.* (2018)	81.0	65.3	64.6	63.9	75.0	30.3	65.1	61.1	64.1	83.9	72.4	69.9	71.0	72.9	71.3	83.6	79.6	73.5	78.9	90.9	70.8
Rogez *et al.* (2019)	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
Moon *et al.* (2019)	**94.4**	78.6	79.0	82.1	86.6	72.8	81.9	75.8	**90.2**	90.4	79.4	79.9	75.3	81.0	81.0	90.7	89.6	83.1	81.7	77.3	82.5
Dabral *et al.* (2019)	85.8	73.6	61.1	55.7	77.9	53.3	75.1	65.5	54.2	81.3	82.2	71.0	70.1	67.7	69.9	90.5	85.7	86.3	85.0	91.4	74.2
Véges and Lőrincz (2020a)	89.5	81.6	85.9	84.4	90.5	73.5	85.5	68.9	65.1	90.4	79.1	82.6	72.7	68.1	81.0	94.0	**90.4**	87.4	**90.4**	92.6	82.7
Mehta *et al.* (2019)	89.7	78.6	68.4	74.3	83.7	47.9	67.4	75.4	78.1	85.1	78.7	73.8	73.9	77.9	74.8	87.1	88.3	79.5	88.3	**97.5**	78.0
Benzine *et al.* (2021)	78.3	75.0	56.9	64.1	76.1	51.3	74.7	**79.1**	55.2	79.3	74.5	74.5	70.2	69.5	67.6	85.7	82.6	78.7	79.1	89.6	72.7
2.5D baseline	93.0	80.1	89.2	85.8	90.1	76.9	88.6	75.6	84.3	**93.4**	85.9	90.6	83.4	80.9	83.8	93.0	86.6	89.3	85.0	90.8	86.3±0.1
MeTRAbs	93.8	84.4	**90.0**	**87.6**	90.5	75.7	88.1	74.9	85.0	92.9	84.7	90.4	83.3	**82.2**	**86.3**	93.9	87.1	88.9	85.2	91.3	86.8±0.4
w/o abs. loss	94.0	**86.5**	89.0	87.1	**91.1**	**77.4**	**90.2**	75.7	85.8	92.9	**86.0**	**90.7**	**83.8**	82.0	85.6	**94.3**	89.8	**89.6**	86.5	91.7	**87.5**±0.2

Table 7.4: Root-relative pose comparison to prior work on the MuPoTS-3D benchmark for normalized skeletons with bone rescaling to the ground truth before computing the percentage of correct keypoints (PCK). (For the direct evaluation of the metric-space poses, see Table 7.1).

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg↑
Absolute PCK for all annotated poses																					
Moon *et al.* (2019)	59.5	44.7	51.4	46.0	52.2	27.4	23.7	26.4	**39.1**	23.6	18.3	14.9	38.2	26.5	36.8	23.4	14.4	19.7	18.8	25.1	31.5
Benzine *et al.* (2021)	22.2	18.1	16.1	18.5	20.4	14.7	21.2	18.9	16.0	22.9	20.3	20.9	18.9	16.0	18.9	23.5	22.3	21.8	21.5	20.8	19.8
Véges and Lőrincz (2020a)	50.4	33.4	52.8	27.5	53.7	31.4	22.6	33.5	38.3	56.5	24.4	35.5	45.5	34.9	49.3	23.2	**32.0**	30.7	26.3	**43.8**	37.3
2.5D baseline	**77.6**	**50.5**	**58.6**	40.3	**74.6**	21.9	7.3	27.0	22.4	38.6	**32.2**	37.6	25.2	43.9	50.4	**35.0**	25.5	**41.1**	**31.9**	27.8	**38.5**±0.9
MeTRAbs	21.2	21.1	45.5	**48.2**	40.9	**34.9**	**33.0**	**51.5**	34.9	**85.6**	18.0	36.7	**50.3**	**53.1**	**54.3**	28.1	28.8	26.8	20.0	35.1	**38.4**±1.9
w/o abs. loss	48.9	32.9	15.3	18.9	48.7	11.8	19.1	42.3	28.9	78.4	27.5	**60.6**	38.6	42.8	43.1	28.4	28.7	28.6	23.3	33.8	35.0±3.1
Absolute PCK for detected poses																					
Moon *et al.* (2019)	59.5	45.3	**51.4**	46.2	53.0	27.4	23.7	26.4	**39.1**	23.6	18.3	14.9	38.2	29.5	36.8	23.6	14.4	20.0	18.8	25.4	31.8
Benzine *et al.* (2021)	22.7	21.2	17.1	18.6	22.0	14.8	21.5	22.9	16.0	22.9	21.5	21.6	20.3	20.0	19.4	18.9	23.8	22.6	22.9	25.8	20.9
Véges and Lőrincz (2020a)	50.4	35.9	53.3	27.7	57.2	31.4	23.1	39.3	38.3	56.5	25.2	35.8	49.3	42.5	49.4	24.1	**32.1**	33.1	29.3	**59.2**	39.6
2.5D baseline	**77.6**	**53.0**	**59.1**	40.5	**77.9**	22.2	7.6	30.1	22.4	38.6	**33.9**	37.9	27.2	52.4	50.4	**35.8**	25.7	**43.3**	**35.2**	35.5	**40.3**±1.0
MeTRAbs	21.2	22.1	45.8	**48.5**	42.8	**35.4**	**34.8**	**58.3**	34.9	**85.6**	19.0	37.0	**54.3**	**63.6**	**54.3**	28.8	29.0	28.2	22.0	44.9	**40.5**±1.9
w/o abs. loss	48.9	34.5	15.5	19.0	50.8	12.0	20.1	48.0	28.9	78.4	29.0	**61.1**	41.7	51.2	43.1	29.0	28.8	30.1	25.8	43.3	36.9±3.1

Table 7.5: Absolute pose comparison to prior work on the MuPoTS-3D benchmark for normalized skeletons with bone rescaling to the ground truth before computing the percentage of correct keypoints (PCK). (For the direct evaluation of the metric-space poses, see Table 7.1).

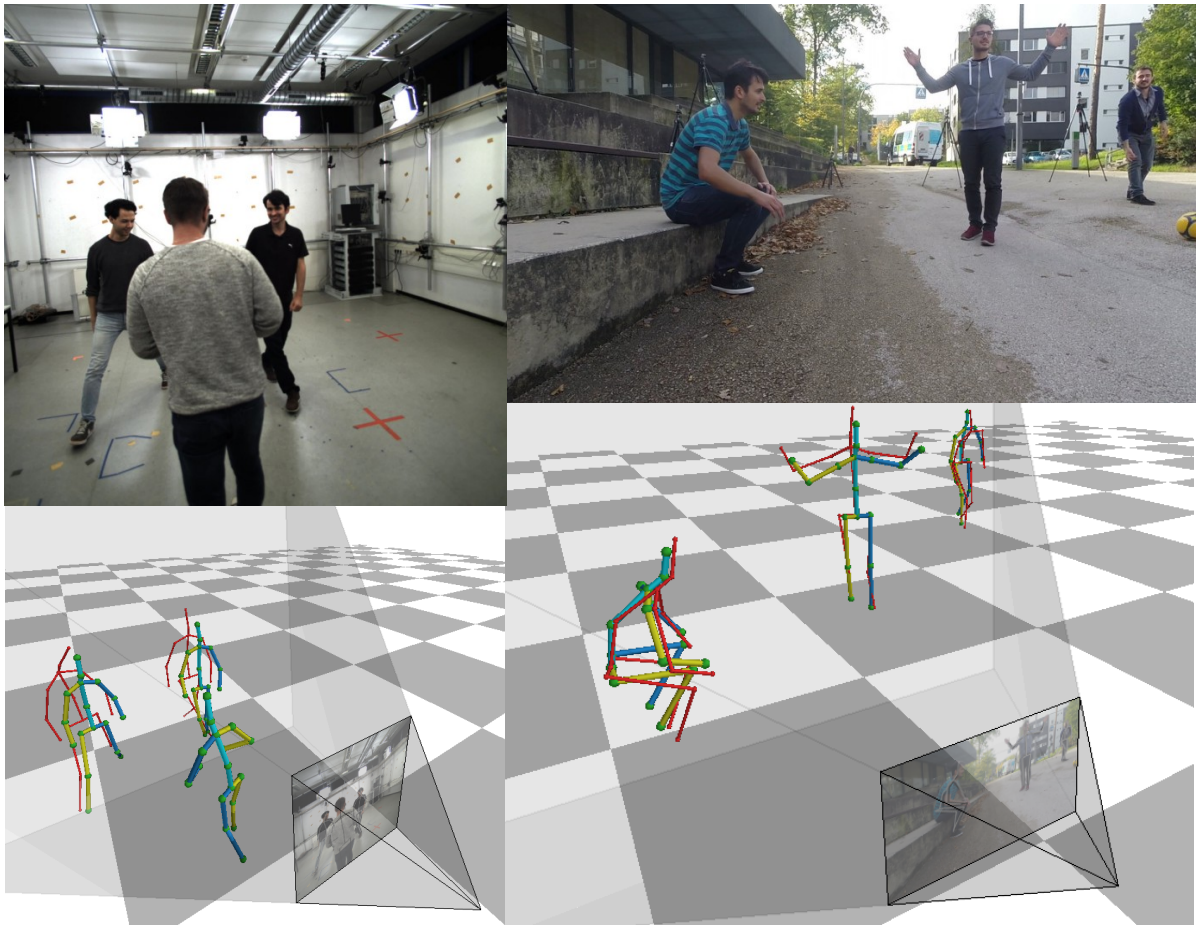


Figure 7.2: Qualitative results on MuPoTS-3D (prediction in blue-yellow, ground truth in red).

In Table 7.2, we evaluate whether using the full perspective pinhole camera model in the absolute pose reconstruction module brings benefits. In the last two rows, the absolute loss is not used at training time. In the other cases we backpropagate the absolute loss gradients either through the weak or full perspective reconstruction method. We find that training on MuCo-3DHP with the full perspective model improves the absolute results, but when testing on MuPoTS-3D, it is better to use the weak model. This may be explained by the fact that people in the MuCo-3DHP dataset are closer to the camera than in MuPoTS-3D, resulting in stronger perspective effects in MuCo-3DHP. To verify this, we computed the ratio of the farthest and closest joint's depth $\max_j Z_j / \min_j Z_j$ per pose. If this ratio is large, the weak perspective assumption is a bad approximation. The median and the 90th percentile of this ratio on MuCo-3DHP is 1.22 and 1.41, while on MuPoTS-3D it is only 1.16 and 1.26, respectively. This confirms that perspective effects are stronger in MuCo-3DHP.

Another possible reason is that the model may output slightly perspective distorted results in the metric 3D head, which are better handled by the weak-perspective model in the next step, as opposed to training time, when the network learns to output the correct metric, perspective-undistorted pose, for which the full perspective model works better afterwards. Nevertheless, as there is no clear overall winner between the weak and full perspective models, and changing the method across training and test is clearly not desirable, we use the more commonly applied weak perspective method for all other experiments.

7.5.1 Inference Speed

Our method is capable of real-time inference. By gathering all person instances of a frame in a batch, MeTRAbs can process 128, 118, 98, 67, 41 frames per second for 1, 2, 4, 8 and 16 people per frame, respectively. The above calculations assume the image crops are available instantly and the time cost of detection is excluded.

7.6 ECCV 2020 3DPW Challenge

Finally, we note that our MeTRAbs method has won the 3D Poses in the Wild (3DPW; von Marcard *et al.*, 2018) challenge, organized as a workshop event at the 2020 European Conference on Computer Vision. Table 7.3 compares results using the 3DPW protocol. Having seen the effectiveness of multi-dataset training in Section 6.8.5, we train our network on a combination of the Human3.6M, MuCo-3DHP, SURREAL (Varol *et al.*, 2017), SAIL-VOS (Hu *et al.*, 2019) and CMU-Panoptic (Joo *et al.*, 2019) datasets. We use ResNet101V2 as the backbone, and additionally apply upper body crop (truncation) augmentation at training time, as well as and 5-crop averaging at test time. When identity tracking is needed, we perform frame-to-frame matching based on absolute pose distance. The listed methods are not directly comparable due to different training data. Even with this caveat, our top results show that our approach can scale with further training data and performs well even in challenging in-the-wild scenarios. This motivates us to scale multi-dataset training of MeTRAbs even further, which we will discuss in the next chapter.

7.7 System Implementation and Embedded Evaluation

We perform detailed performance measurements on robot-compatible embedded hardware, in the context of the Horizons 2020 project “CROWDBOT” funded by the

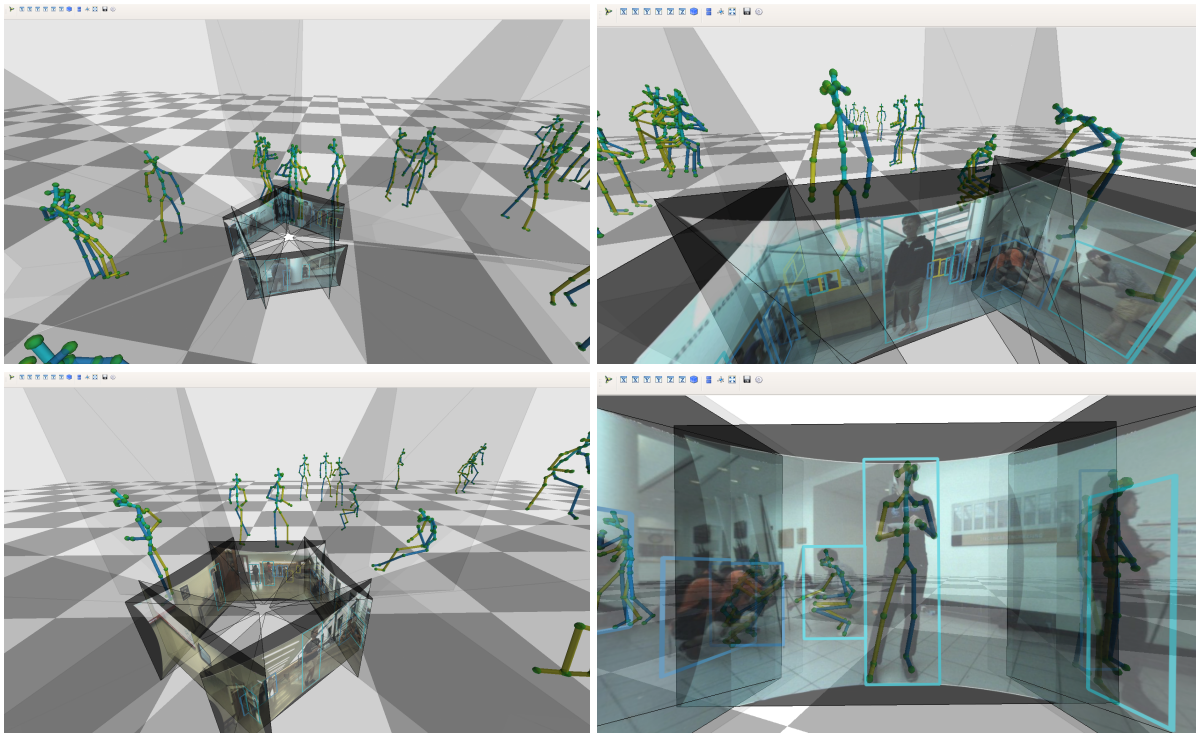


Figure 7.3: Qualitative results on the JackRabbot (JRDB) dataset.

European Union (H2020-ICT-2017-779942), aimed at enabling robots to navigation safely in dense crowds. [2]

Furthermore, a live demonstration of the system running on a laptop and a webcam was presented in the Demo Track of the 2022 European Conference on Computer Vision (ECCV).

In Section 7.5, we used the ResNet50V2 as the backbone and in Section 7.6 ResNet101V2. The ResNet family is an extensively studied architecture and is therefore ideal for good comparability with the related literature. However, newer backbones have been proposed in recent years as we described in Section 3.1.3, and these have different speed–accuracy tradeoff characteristics as well as some hardware-specific optimizations. Specifically, the recent EfficientNetV2 (Tan and Le, 2021) is preferable in the high-accuracy regime, while the lightweight MobileNetV3 (Howard *et al.*, 2019) is better in the low-compute, high-speed regime.

In the following, we present a detailed analysis of the speed–accuracy tradeoff with respect to the following aspects. We concentrate on the pose estimation part here and do not measure the time for person detection, cropping, *etc*.

[2] Some text passages that I wrote for the corresponding public technical report (CROWDBOT, 2021) are included in the following.

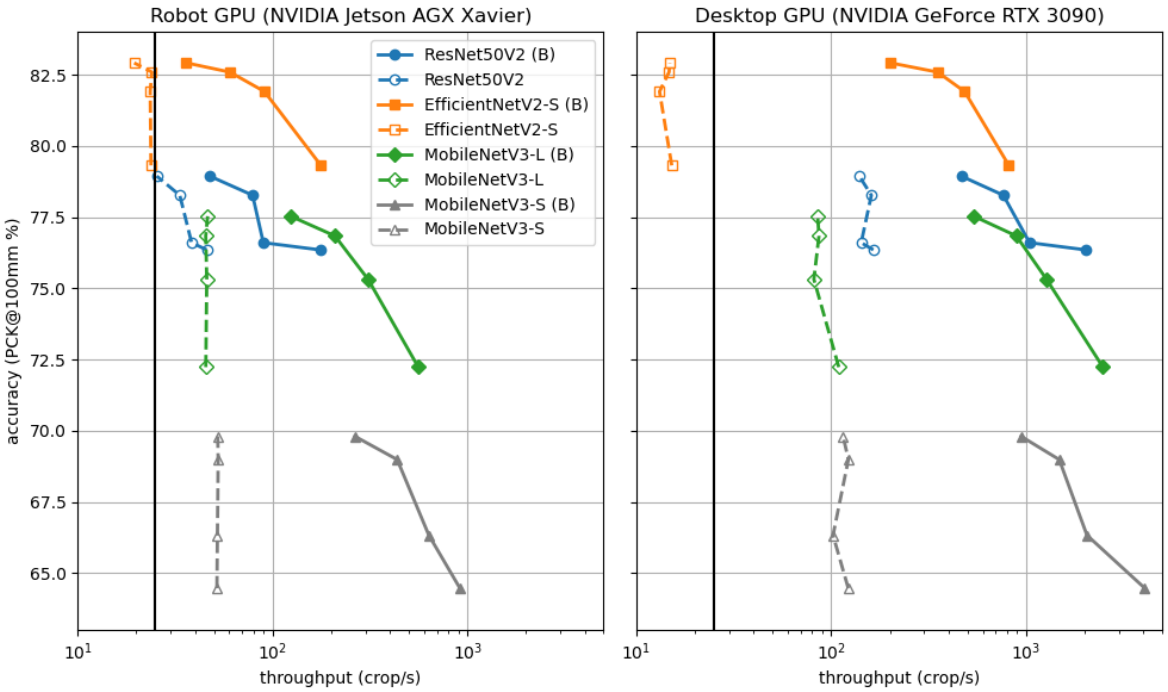


Figure 7.4: Throughput *vs.* accuracy. Performance evaluation on robot and desktop hardware. (B) denotes batching 64 crops together for increased parallelization.

Backbones. We use MobileNetV3-Small, MobileNetV3-Large, ResNet50V2, and EfficientNetV2-S (see Section 3.1.3 for details on these architectures).

Test-Time Augmentation. The prediction quality can be significantly improved by transforming the input image crop in multiple ways (rotation, mirroring, scaling, gamma adjustment) and averaging the resulting predictions (after transforming the individual predictions back to a common coordinate frame). Here we analyze such test-time augmentation (TTA) with 1, 2, 3 and 5 crops.

Batching. Performing inference on multiple image crops at once improves throughput due to the highly parallel computational architecture of GPUs. This can mean batching multiple persons detected in a single frame, or batching over multiple frames. The latter case introduces additional latency in the system. Here, we consider two extremes: no batching and batching 64 crops.

Hardware. We perform measurements both on a high-end desktop GPU (NVIDIA GeForce RTX 3090) and an embedded edge computing device (NVIDIA Jetson AGX Xavier).

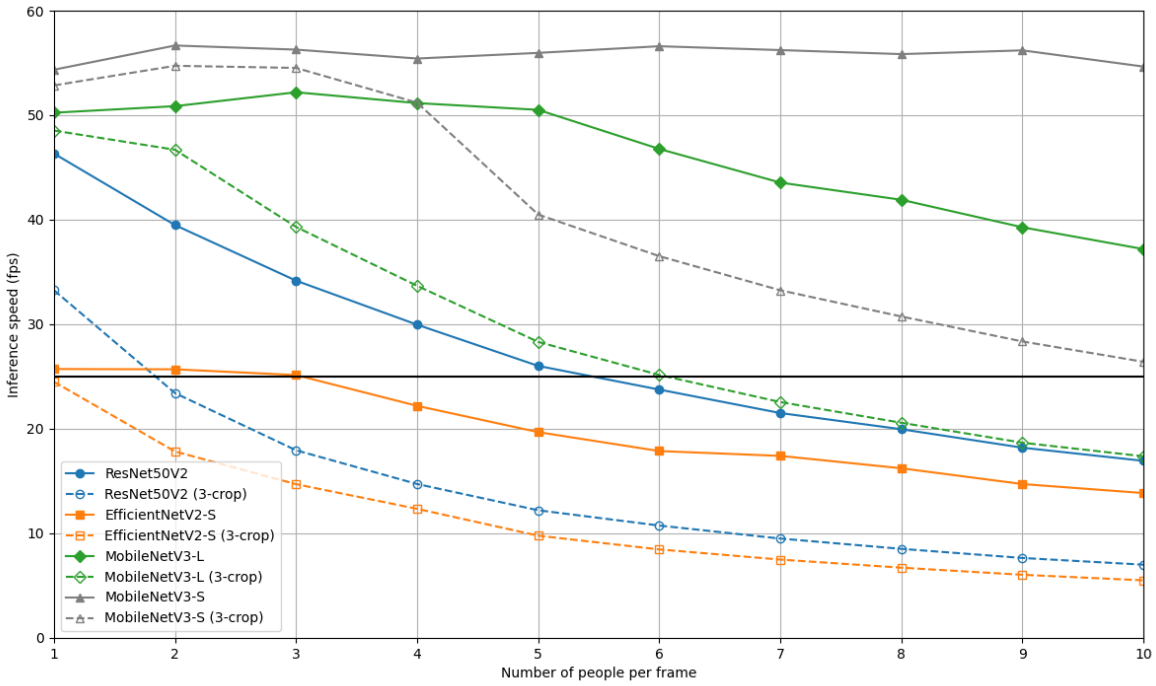


Figure 7.5: Multi-Person Performance. We analyze inference speed (as frames per second) on the Jetson AGX Xavier embedded device, as a function of the number of people in each frame.

Estimation quality is measured on the 3DPW benchmark dataset according to the PCK thresholded at 100 mm.

Results. Figure 7.4 shows the results. The different colors represent different backbone networks. The line style indicates whether the GPU is used with batching (batch size 64) or without batching (batch size 1). The points on each line represent different amounts of test-time augmentation, going from bottom to top: 1, 2, 3 and 5 augmentations per crop. The horizontal axis shows logarithmic throughput. The black vertical line shows 25 crops/s, the throughput needed for online processing of standard video input with a single person. We make the following observations: Test-time augmentation is highly effective (*e.g.*, MobileNetV3-Large improves from 72.2% to 77.5% PCK@100mm with 5-crop augmentation). Especially if we do not use cross-frame batching, the overhead of test-time augmentation is small (there is idle computational capacity otherwise).

Batching can increase throughput by more than an order of magnitude, hence the batch size needs to be chosen as the highest possible value, according to latency requirements. EfficientNetV2-S has significantly higher accuracy than ResNet50V2 with little additional computational cost on the Jetson hardware. Interestingly, EfficientNetV2-S is less efficient on the desktop: non-batched inference is faster on the Jetson than

on the RTX 3090. MobileNetV3-Small has too low accuracy to be a viable practical choice. MobileNetV3-Large is on the Pareto frontier in case of the Jetson, but is strictly outperformed by ResNet50V2 on the desktop.

Figure 7.5 shows Jetson timing measurements from a different perspective. Batching is only performed within a single frame, *i.e.*, the multiple people present in a single frame are processed in parallel, but not across different frames, in order to evaluate inference speed at minimum latency. Solid lines show inference without test-time augmentation, dashed lines use 3-crop augmentation. 25 frames per second (a common video frame rate) is highlighted for reference.

MobileNetV3-Large performs well in this low-latency scenario and can process 6 people per frame at 25 fps with 3-crop augmentation. The fastest inference is possible with MobileNetV3-Small without test-time augmentation, in which case processing speed remains above 50 fps even for crowds with more than 10 people, although at the cost of a large decrease in accuracy. Single-person inference is possible in real-time even with the highly accurate setting of using EfficientNetV2-S with 3-crop augmentation.

We show qualitative results obtained on the JRDB robotics dataset in Figure 7.3.

Overall, we conclude from these results that our method is real time–capable on an edge device and that the EfficientNetV2 backbone is highly promising for this task. In Chapter 8, we adopt it as the main backbone of choice, including its large, EfficientNetV2-L variant.

7.8 Conclusion

We have proposed a method for truncation-robust absolute 3D human pose estimation, building upon the metric-scale truncation-robust volumetric heatmap (MeTRo) concept in combination with 2D heatmap estimation. We have seen the importance of supervising the absolute pose prediction end-to-end by employing a differentiable combination of 2D and root-relative 3D poses. For this, we tested two alternatives, based on weak and full perspective geometry, but neither performed clearly better than the other in our experiments, likely due to the limited camera diversity in the training and test data. Applying MeTRAbs in the top-down multi-person paradigm, we have achieved state-of-the-art results on the challenging MuPoTS-3D dataset while keeping the method real-time capable. From these experiments, we can conclude that heatmap estimation is a versatile paradigm, and it is possible to tackle absolute 3D human pose estimation through exclusively estimating heatmaps and encoding all quantities such as coordinates or sizes as activation locations, instead of as activation values. Finally, we have discussed some system implementation details for high performance and efficiency. We have also seen that our model can be deployed to low-powered hardware with lightweight architectures, while maintaining its real-time capability.

8

Bridging Skeleton Formats via Geometric Autoencoding for Multi-Dataset Learning

Deep learning–based 3D human pose estimation performs best when trained on large amounts of labeled data, making combined learning from many datasets an important research direction. One obstacle to this endeavor are the different skeleton formats provided by different datasets, i.e., they do not label the same set of anatomical landmarks. There is little prior research on how to best supervise one model with such discrepant labels. We show that simply using separate output heads for different skeletons results in inconsistent depth estimates and insufficient information sharing across skeletons. As a remedy, we propose a novel affine-combining autoencoder (ACAE) method to perform dimensionality reduction on the number of landmarks. The discovered latent 3D points capture the redundancy among skeletons, enabling enhanced information sharing when used for consistency regularization.

Our approach scales to an extreme multi-dataset regime, where we use 28 3D human pose datasets to supervise one model, which outperforms prior work on a range of benchmarks, including the challenging 3D Poses in the Wild (3DPW) dataset. Our code and models are available for research purposes.[1]

This chapter is based on our publication (Sárándi *et al.*, 2023), presented at the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

8.1 Overview

Research on 3D human pose estimation has gone through enormous progress in recent years (Moon *et al.*, 2019; Fabbri *et al.*, 2020; Kocabas *et al.*, 2020; Liu *et al.*, 2020; Cheng *et al.*, 2021; Joo *et al.*, 2021; Lin *et al.*, 2021a; Sun *et al.*, 2021). While semi-supervised and self-supervised approaches are on the rise (Zhang *et al.*, 2021; Kundu *et al.*, 2022),

[1] `https://vision.rwth-aachen.de/wacv23sarandi`

Dataset name	#Examples	#Real Subj.	#Keypoints	Subset	Skeleton
Real images with markerless MoCap					
MuCo-3DHP (Mehta *et al.*, 2018)	677k	8	28	♢♣♡	3DHP
CMU-Panoptic (Joo *et al.*, 2019)	2.81M	>60	19	♣♡	COCO
AIST-Dance++ (Tsuchida *et al.*, 2019; Li *et al.*, 2021e)	1.86M	30	19	♣♡	COCO
HUMBI (Yu *et al.*, 2020b)	1.26M	772	19	♣♡	COCO
MPI-INF-3DHP (Mehta *et al.*, 2017a)	627k	8	28	♣♡	3DHP
RICH (Huang *et al.*, 2022)	96k	15	42	♣♡	SMPL-X (Pavlakos *et al.*, 2019)
BEHAVE (Bhatnagar *et al.*, 2022)	42k	7	43	♡	SMPL, COCO
ASPset (Nibali *et al.*, 2021)	124k	15	17	♡	*
3DOH50K (Zhang *et al.*, 2020b)	50k	<10	14	♡	LSP (Johnson and Everingham, 2010)
IKEA ASM (Ben-Shabat *et al.*, 2021)	23k	48	17	♡	*
Real images with marker-based MoCap					
Human3.6M (Ionescu *et al.*, 2014)	165k	5	25	♢♣♡	H3.6M
TotalCapture (Trumble *et al.*, 2017)	130k	5	21	♣♡	*
BML-MoVi (Ghorbani *et al.*, 2021)	553k	13	87	♡	*
Berkeley-MHAD (Ofli *et al.*, 2013)	526k	12	43	♡	*
UMPM (van der Aa *et al.*, 2011)	164k	30	15	♡	*
Fit3D (Fieraru *et al.*, 2021a)	147k	8	25	♡	H3.6M
GPA (Wang *et al.*, 2019b)	109k	13	34	♡	*
HumanSC3D (Fieraru *et al.*, 2021b)	72k	4	25	♡	H3.6M
CHI3D (Fieraru *et al.*, 2020)	46k	6	25	♡	H3.6M
Human4D (Chatzitofis *et al.*, 2020)	40k	4	32	♡	*
MADS (Zhang *et al.*, 2017)	33k	5	15	♡	*
Synthetic images					
SURREAL (Varol *et al.*, 2017)	1.9M		24	♢♣♡	SMPL (Loper *et al.*, 2015)
3DPeople (Pumarola *et al.*, 2019)	946k		29	♣♡	*
JTA (Fabbri *et al.*, 2018)	562k		22	♣♡	*
HSPACE (Bazavan *et al.*, 2021)	195k		35	♣♡	GHUM (Xu *et al.*, 2020a)
SAIL-VOS (Hu *et al.*, 2019)	101k		26	♣♡	*
AGORA (Patel *et al.*, 2021)	79k		66	♣♡	SMPL[-X]
SPEC (Kocabas *et al.*, 2021b)	59k		24	♡	SMPL
Real images with 2D annotations (weak supervision)					
COCO (Lin *et al.*, 2014)	47k		17		
MPII (Andriluka *et al.*, 2014)	27k		16		
PoseTrack (Andriluka *et al.*, 2018)	40k		15		
JRDB (Martin-Martin *et al.*, 2021)	59k		17		
Totals (for 3D-labeled data					
Small (3 datasets)	2.8M	13	77	♢	
Medium (14 datasets)	10.8M	>900	277	♣	
GRAND TOTAL (28 ds.)	13.4M	>1k	555	♡	

Table 8.1: We study the extreme multi-dataset setting of 3D human pose estimation, using all datasets below in one training process. We define three dataset combinations (indicated by ♢, ♣, and, ♡) to study the effect of training data amount. (* marks custom dataset-specific skeletons.)

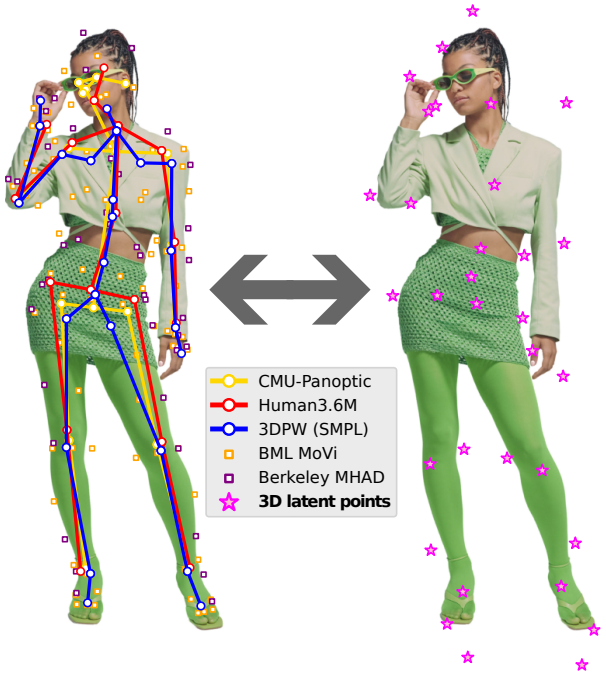


Figure 8.1: Different 3D human pose datasets (*e.g.*, CMU-Panoptic and Human3.6M) provide annotations for different sets of body landmarks (*left*). To best leverage such discrepant labels for multi-dataset 3D pose estimation, we discover a smaller set of latent 3D keypoints (*right*), from which the dataset-specific points can be reconstructed. This allows us to capture the redundancy among the different skeleton formats and enhance information sharing between datasets, ultimately leading to improved pose accuracy.

best results are still achieved when using as much labeled training data as possible. However, individual 3D pose datasets tend to be rather small and lacking in diversity, as they are often recorded in a single studio with few subjects. Therefore, to provide the best possible models for downstream applications (*e.g.*, action recognition, sports analysis, medical rehabilitation, collaborative robotics), it becomes important to use many datasets in the training process. Thanks to sustained efforts by the research community, numerous publicly released, labeled datasets exist.

However, as prior published works only train on at most a handful of them, it remains unknown what performance could be achieved by combining more than a decade of dataset collection efforts into a single model. Unfortunately, this is not a trivial undertaking, since different datasets do not use the same skeleton format for their labels (see Figure 8.1), *e.g.*, the hip keypoints are at different heights, some body parts are only labeled in some datasets, some provide surface markers while others provide keypoints inside the body, *etc.* Prior work has rectified such differences through a handful of individually defined rules (*e.g.*, shrink the hip–pelvis distance by a certain factor, Rapczyński *et al.*, 2021), but this does not scale to many keypoints and

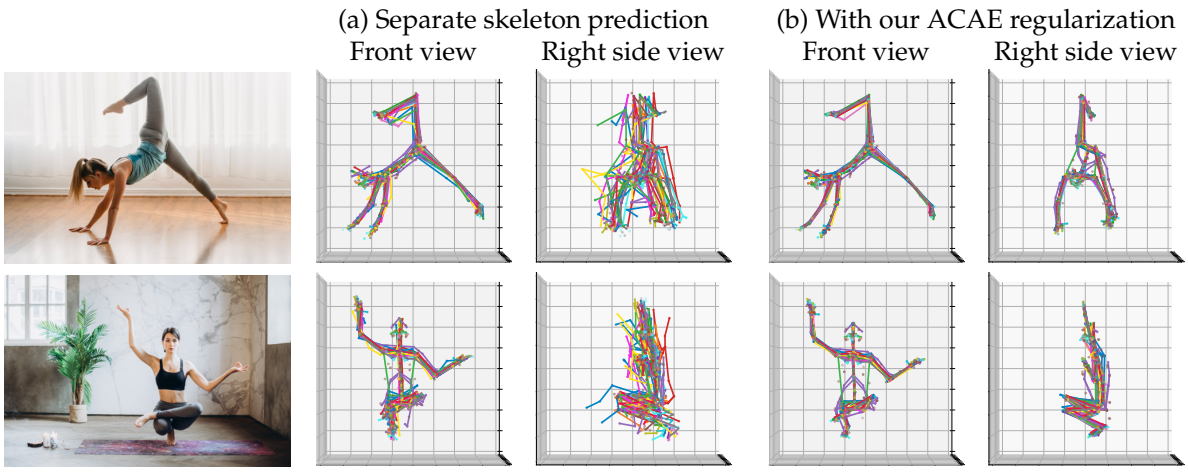


Figure 8.2: We train models to jointly estimate multiple 3D human pose skeletons, allowing us to train on many datasets at once. a) Simply using separate prediction heads on a shared backbone is an insufficient solution, as we obtain inconsistent outputs along the depth axis. b) We propose to capture and exploit the redundancy among the different skeletons using a novel affine-combining autoencoder–based (ACAE) regularization. This leads to a clear improvement in skeleton consistency.

datasets—we need a more systematic and automatic method. The question we tackle in this work is therefore: *How can we automatically merge dozens of 3D pose datasets into one training process, given the label discrepancies?* We refer to this task as *multi-skeleton 3D human pose estimation*.[2]

If we ignored the discrepancies altogether and proceeded as if keypoints with the same name represented the same body landmark, the model would be supervised with inconsistently labeled examples and would learn to output a skeleton format that is some kind of average of the true ones, leading to subpar benchmark performance. Alternatively, we may consider this as a multi-task learning problem, and predict the skeletons on separate output heads on a shared backbone, without assuming any skeleton correspondences. But as we will see, this is not ideal either, as there is insufficient information sharing between skeletons, which is most apparent in inconsistencies between the depth predictions of such a model, as shown in Figure 8.2a.

To strike the right balance between those two extremes, we aim to establish *some* connections between the skeleton formats without assuming them to be the same. To learn such geometric relations between skeletons, we introduce a novel autoencoder-based dimensionality reduction technique to compress a larger set of 3D keypoints (the joints from all datasets) into a lower-cardinality representation (a smaller latent keypoint set). The encoder and decoder compute affine combinations of their input

[2]For simplicity, we call any set of landmarks provided in a particular dataset a "skeleton," and use "landmark" and "joint" synonymously.

points, and are thus equivariant to rotation and translation. We further induce chirality equivariance (left–right symmetry) via weight sharing (Yeh *et al.*, 2019). We call this model an *affine-combining autoencoder* (ACAE). We employ the ACAE in pose estimation training as an output regularizer, to encourage consistent predictions. This improves prediction results both qualitatively and quantitatively. As an alternative to the regularization approach, we can also directly predict the latent keypoints of the ACAE with a 3D pose estimator. This latter variant avoids the need for the underlying pose estimator to estimate a large number of joints, which may be costly for some methods. In both cases, the final predictions become consistent, showing the value of our approach in tackling multi-dataset 3D pose estimation.

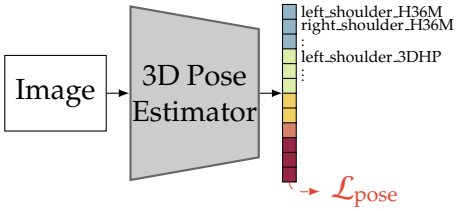
Through an extensive literature review, we have identified 28 datasets with high-quality 3D human pose labels. By systematically preprocessing these datasets and discarding redundant poses, we constructed a meta-dataset of 13 million examples, spanning more than a thousand people. This is almost two orders of magnitude more data than in typical research papers (*e.g.*, Human3.6M has 165k examples after redundancy filtering). We show that using more data indeed helps, and that our approach scales to 28 datasets providing a total of 555 joints in their skeleton formats, summarized in Table 8.1. Our final models show excellent in-the-wild performance, outperforming currently available models, making them highly useful for downstream research.

In summary, we make the following contributions in this chapter. (1) We assemble the largest scale meta-dataset for 3D human pose estimation to date, consisting of 28 individual datasets, and release scripts for reproducing the process. We call special attention to the problem of disparate skeleton annotation formats in these datasets, which has rarely been addressed in the literature so far. (2) We propose affine-combining autoencoders (ACAE), a novel linear dimensionality reduction technique applicable to keypoint-based representations such as poses. (3) We apply the ACAE to regularize model predictions to become more consistent, leading to qualitative and quantitative improvements, and we show that the latent points can be predicted directly as well. (4) We release high-quality 3D pose estimation models with excellent and consistent in-the-wild performance due to diverse supervision and our regularization tying together different skeleton formats.

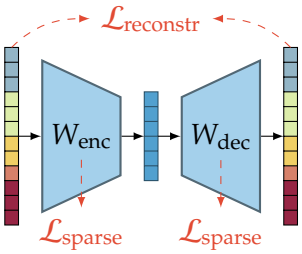
8.2 Related Work

3D Human Pose Estimation. For an overview on the history and current trends in 3D pose estimator design, see Chapter 2. We emphasize that our approach is independent of the internals of the pose estimation method.

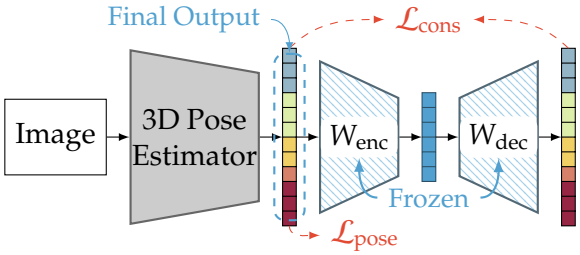
Handling Discrepancy in Skeleton Formats. In 2D-to-3D pose lifting, Rapczyński *et al.* (2021) combine pairs of datasets in training by concatenating the training data sets,



(a) Step 1: Train initial model.



(b) Step 2: Train the autoencoder.



(c) Step 3: Fine-tune the model for consistency.

Figure 8.3: Our complete training workflow. We train an initial model on all skeletons of multiple datasets without enforcing consistency. Using this model, we create pseudo-ground truth, needed to train an autoencoder that learns a latent keypoint space. In turn, we use this frozen autoencoder to regularize the initial model during fine-tuning, encouraging consistent predictions. Alternative strategies for the last step are shown in Figure 8.4.

and harmonize the joints through hand-crafted rules. In 2D pose tracking, Guo *et al.* (2018) train dataset-specific output heads and combine their results via hand-crafted rules. To unify pose representations, some prior works on (image-independent) MoCap data standardized the height and bone length of skeletons (Holden *et al.*, 2016; Mandery *et al.*, 2016). The AMASS dataset (Mahmood *et al.*, 2019) addresses the problem of discrepancy in MoCap data representations by mapping them to the SMPL (Loper *et al.*, 2015) representation, but the dataset does not provide corresponding images and cannot be used for image-based pose estimation. Furthermore, the underlying MoSH++ algorithm relies on a complex, multi-stage procedure requiring temporal sequence data and a pre-existing body mesh model. More generally, in the mesh model–based paradigm, researchers have handled different skeleton formats by learning joint regressors from the mesh vertices. However, learning such regressors requires high-quality mesh fits, and as Hedlin *et al.* (2022) show, these are difficult to obtain. In contrast, our method has different goals and is much simpler in comparison. We do not aim to generate a definitive, universal ground truth representation for all datasets, instead our latent keypoint set is only used as an intermediate representation for

the pose estimator, but the losses and evaluations are still computed in the original skeleton formats, after decoding the latent points into full skeletons.

Keypoint Discovery. Discovering a good set of landmarks to describe objects has been investigated in other contexts in computer vision. 2D keypoint discovery has been used to disentangle pose and shape in 2D human pose estimation (Jakab *et al.*, 2018, 2020). In 3D, Jakab *et al.* (2021) discover control points for deforming 3D shapes. Rhodin *et al.* (2018b) learn a 3D human representation that consists of a set of 3D points, which encode both pose and appearance, optimizing for the unsupervised auxiliary task of novel view synthesis. Loper *et al.* (2014) optimize the placement of sparse markers on the body to best capture both human shape and pose.

Linear Subspace Learning. Linear dimensionality reduction has a long history, with principal component analysis being the best known representative (Pearson, 1901). Its relation to autoencoders was discovered by Bourlard and Kamp (1988), and a recent paper by the same first author reviews the developments since (Bourlard and Kabil, 2022). Linear autoencoders have been employed in robust and sparse (Guerra-Urzola *et al.*, 2021) variants, a detailed overview is presented by Cunningham and Ghahramani (2015). Our proposed affine-combining autoencoders are related, but have different constraints, tailored to our use case, *i.e.*, that the weights sum to unity, and there is no requirement of orthogonality, unlike in PCA.

8.3 Method

Our goal is to obtain a strong, monocular RGB-based 3D human pose estimation model by integrating numerous datasets into one mixed training process, even when the different datasets provide annotations according to different skeleton formats.

Suppose we have D skeleton formats, with $\{J_d\}_{d=1}^{D}$ joints in each, for a total of $J = \sum_{d=1}^{D} J_d$ joints overall. Further, we have a merged dataset with N training examples, each consisting of an image of a person and annotations for a subset of the J body joints in 3D.

Our proposed workflow consists of three main steps. First, we train an initial model that predicts the different skeletons on separate prediction heads, branching out from a common backbone network (Figure 8.3a). With the resulting model, we can run inference and produce a pseudo-ground-truth "parallel corpus" of many poses given in every skeleton format. From this, the geometric relations between skeleton formats can be captured. We accomplish this in the second step, by training an undercomplete geometry-aware autoencoder, which discovers a latent 3D body landmark set that best captures human pose variations in the pseudo-GT data (Figure 8.3b). Finally, equipped with the trained autoencoder, we rely on its learned latent space to make the model

output consistent across skeleton formats through output regularization (Figure 8.3c). We also experiment with direct latent point prediction, and a hybrid variant for the last step.

8.3.1 Initial Model Training

The first step of our workflow is to train an initial pose estimator to predict all J joints separately (Figure 8.3a). This means that no correspondences or relations across different skeletons are assumed, *i.e.*, without specifying or enforcing that the left shoulder joint of one skeleton should be predicted near the left shoulder of another skeleton. This is akin to multi-task architectures that use different task-specific heads on one backbone. The pose loss we minimize is $\mathcal{L}_{\text{pose}} = \mathcal{L}_{\text{meanrel}} + \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{abs}}\mathcal{L}_{\text{abs}}$, where $\mathcal{L}_{\text{meanrel}}$ is an ℓ_1 loss computed after aligning the prediction and ground truth at the mean, $\mathcal{L}_{\text{proj}}$ is an ℓ_1 loss on the 2D coordinates after projection onto the image, and $\mathcal{L}_{\text{abs}}$ is an ℓ_1 loss on the absolute pose (in camera coordinates). Since each training example is annotated only with a subset of the J joints, we ignore any unlabeled joints when averaging the loss.

When visualizing the different skeleton outputs of this trained model, we see inconsistencies among them along the challenging depth axis (see Figure 8.2). This is understandable, since we have not employed any training mechanism that would ensure any relations between the output skeletons (except that they are predicted from shared backbone features). On the other hand, when projected onto the image plane, the predictions appear sufficiently consistent.

8.3.2 Pseudo–Ground Truth Generation

To characterize how the joints of the different skeletons relate to one another, we need pose labels according to all skeleton formats for the same examples, to function as a "Rosetta Stone." Since no such ground truth is available (datasets only provide one type of skeletons, rarely two), we generate pseudo–ground truth using the initial separate-head model. It is important to use images that the model can handle well in this step, hence we choose a relatively clean, clutter-free subset of the *training* data for this purpose (Human3.6M and MoVi). This yields a set of K pseudo-ground-truth poses, with all J joints: $\{P_k \in \mathbb{R}^{J\times 3}\}_{k=1}^{K}$.

8.3.3 Affine-Combining Autoencoder

To capture the redundancy among the full set of J joints, and ultimately to improve the consistency in estimating them, we introduce a simple but effective dimensionality reduction technique. Since the pseudo-GT is more reliable in 2D (the X and Y axes) than in the depth dimension, the transformation to and from the latent representation should

be viewpoint-independent, in other words the representation should be equivariant to rotation and translation. This equivariance in turn requires the latent representation to be geometric, *i.e.*, to consist of a list of L latent 3D points $Q_k \in \mathbb{R}^{L\times 3}$ ($L < J$).

This makes intuitive sense: the way the different skeletons relate to each other is only dependent on how joints are defined on the human body, not on the camera angle. The latent points are then responsible for spanning the overall structure of a pose. Specific skeleton formats can then be computed in relation to these latents. Further, the latent points should only have sparse influence on the joints, *e.g.*, some latent points should be responsible for the positioning of the left arm and these should have no influence on the right leg's pose.

We find that these requirements can be fulfilled effectively by adopting a novel constrained undercomplete linear autoencoder structure, which we call *affine-combining autoencoder* (ACAE). Instead of operating on general n-dimensional vectors, an ACAE's **encoder** takes as input a list of J points $\mathbf{p}_j \in \mathbb{R}^3$ and encodes them into L latent points $\mathbf{q}_l \in \mathbb{R}^3$ by computing affine combinations according to

$$\mathbf{q}_l = \sum_{j=1}^{J} w_{l,j}^{\text{enc}} \mathbf{p}_j, \quad \sum_{j=1}^{J} w_{l,j}^{\text{enc}} = 1, \quad \forall l = 1, \ldots, L. \tag{8.1}$$

Similarly, the **decoder**'s goal is to reproduce the original points from the latents, again through affine combinations:

$$\hat{\mathbf{p}}_j = \sum_{l=1}^{L} w_{j,l}^{\text{dec}} \mathbf{q}_l, \quad \sum_{l=1}^{L} w_{j,l}^{\text{dec}} = 1, \quad \forall j = 1, \ldots, J. \tag{8.2}$$

Since affine combinations are equivariant to any affine transformation, our encoder and decoder are guaranteed to be rotation and translation equivariant. (Note that the same weighting is used for the X, Y and Z coordinates.)

The learnable parameters of the ACAE are the affine combination weights $w_{l,j}^{\text{enc}}$ and $w_{j,l}^{\text{dec}}$, which can also be understood as (potentially negative) generalized barycentric coordinates (Hormann and Sukumar, 2017) for the latents w.r.t. the full joint set and vice versa. Allowing negative coordinates is necessary, as this allows the latents to spread outwards from the body, similar to a cage used in graphics (Nieto and Susín, 2013). Restricting the encoder and decoder to convex combinations would severely limit its expressiveness. We adopt the ℓ_1 reconstruction loss, as it is robust to outliers which may be present due to noise in the pseudo-GT.

To achieve sparsity in the weights (*i.e.*, spatially localized influence), we use ℓ_1 regularization.

Problem Statement. We can now formally state our proposed ACAE problem in matrix notation for the weights. Given K training poses with J joints $\{P_k \in \mathbb{R}^{J\times 3}\}_{k=1}^{K}$,

$$\begin{aligned} \underset{W_{\text{enc}}\in\mathbb{R}^{L\times J},\, W_{\text{dec}}\in\mathbb{R}^{J\times L}}{\text{minimize}} \quad & \mathcal{L}_{\text{reconstr}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}} \\ \mathcal{L}_{\text{reconstr}} &= \frac{1}{K}\sum_{k=1}^{K} \|P_k - W_{\text{dec}}W_{\text{enc}}P_k\|_1 \\ \mathcal{L}_{\text{sparse}} &= \|W_{\text{enc}}\|_1 + \|W_{\text{dec}}\|_1 \\ \text{s. t.} \quad & W_{\text{enc}}\mathbf{1}_J = \mathbf{1}_L, \quad W_{\text{dec}}\mathbf{1}_L = \mathbf{1}_J, \end{aligned} \tag{8.3}$$

where $\mathbf{1}_a$ is a vector of dimension a filled with ones and λ_{sparse} controls the strength of the sparsity regularization. The sum-to-one (partitioning of unity) constraints ensure that the weights express affine combinations, which is necessary for translation equivariance.

L1 Regularization Discourages Negative Weights. We note here that in the context of the ACAE, the ℓ_1 regularization plays another role as well, besides inducing sparsity: it reduces the amount of negative weights, thus preferring nearly convex combinations. To see that this is the case, we can partition the weights to negative and non-negative ones and sum them up separately as

$$w_{l,+}^{\text{enc}} = \sum_{j\,:\,w_{l,j}^{\text{enc}}\geq 0} w_{l,j}^{\text{enc}} \tag{8.4}$$

$$w_{l,-}^{\text{enc}} = \sum_{j\,:\,w_{l,j}^{\text{enc}}< 0} w_{l,j}^{\text{enc}} \tag{8.5}$$

$$w_{l,+}^{\text{enc}} + w_{l,-}^{\text{enc}} = 1, \tag{8.6}$$

and analogously for the decoder weights. Now, the ℓ_1 penalty (sum of absolute values) can be written as

$$\ell_1\left(w_{l,\cdot}^{\text{enc}}\right) = \sum_{j=1}^{J}\left|w_{l,j}^{\text{enc}}\right| = w_{l,+}^{\text{enc}} - w_{l,-}^{\text{enc}} = \left(1 - w_{l,-}^{\text{enc}}\right) - w_{l,-}^{\text{enc}} = 1 - 2\cdot w_{l,-}^{\text{enc}} = 1 + 2\cdot\left|w_{l,-}^{\text{enc}}\right|. \tag{8.7}$$

This means that the ℓ_1 penalty is equivalent to penalizing the absolute sum of the negative weights. When all weights are non-negative, we get convex combinations. In other words, the ℓ_1 regularization in the ACAE encourages constructing close-to-convex combinations besides sparsity.

Reconstruction Loss on 2D Projection. As discussed above, the pseudo-GT is more reliable in its 2D projection than along the depth axis. We therefore adapt the above general problem formulation to take this into account by defining the reconstruction loss on 2D projections:

$$\mathcal{L}_{\text{reconstr}}^{\text{proj}} = \frac{1}{K} \sum_{k=1}^{K} \left\| \Pi(P_k) - \Pi\left(W_{\text{dec}} W_{\text{enc}} P_k\right) \right\|_1 , \tag{8.8}$$

where $\Pi(\cdot)$ denotes camera projection.

Our key insight here is that it is sufficient to observe the high-quality 2D image-plane projections of this model's outputs to characterize how the joints of different skeleton formats geometrically interrelate, because these relations are viewpoint-independent. As a simplified example, if we observe on many poses, that a certain joint tends to be halfway in between two other joints in 2D, then this will also have to hold along the depth axis.

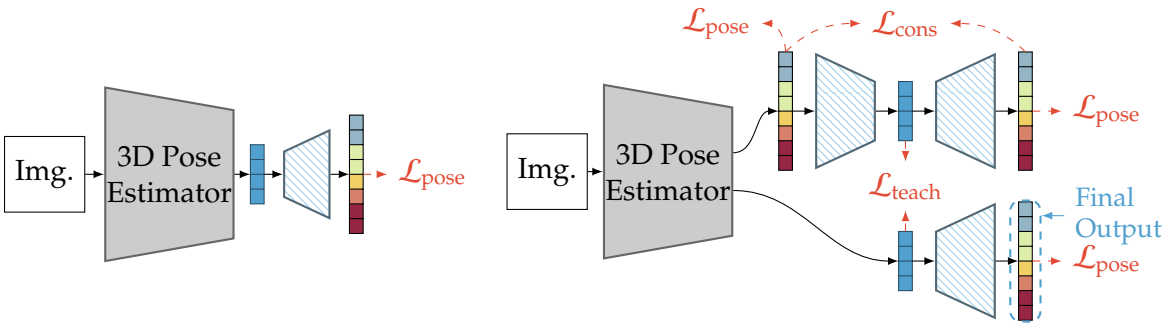
Chirality Equivariance. As humans have bilateral symmetry, it is natural to expect the autoencoder to be chirality-equivariant, *i.e.*, to process the left and right sides the same way (Yeh *et al.*, 2019). To this end, we partition the latent keypoints into three disjoint sets: left, right and central latents, following the same proportions as in the full joint set. Assume, without loss of generality, that the points are sorted and grouped into left-side, right-side and center blocks. We then impose the following weight-sharing block structure on both the encoder and decoder weight matrices:

$$W = \left[\begin{array}{c|c|c} W_1 & W_2 & W_3 \\ \hline W_2 & W_1 & W_3 \\ \hline W_4 & W_4 & W_5 \end{array}\right]. \tag{8.9}$$

This structure indeed ensures chirality equivariance, since the matrix remains the same if we permute both its rows and columns by swapping the first two sections, *i.e.*, swapping the left and right points in the inputs and the outputs.

Head Keypoint Weighting. Based on the intuition that smaller motions of head and facial keypoints can be more semantically relevant, we weight these joints higher (by a factor of 10) in the loss, ensuring that the latents sufficiently cover the head as well. (We later found that this is not strictly necessary and the method also works without this as well.)

Training. We train the autoencoder using the Adam optimizer (Kingma and Ba, 2015) with batch size 32. To enforce the sum-to-one constraints, we normalize the weight matrices within the computational graph.



(a) Direct prediction of latent keypoints.

(b) Hybrid of the models from Fig. 8.3c and 8.4a with a further student–teacher loss.

Figure 8.4: Alternative model structures for the fine-tuning phase, to be used instead of Figure 8.3c in our training workflow.

8.3.4 Consistency Fine-Tuning

Once our affine-combining autoencoder is trained on pseudo-GT, we freeze its weights and use it to enhance the consistency of 3D pose estimation outputs, with one of three alternative methods.

Output Regularization. In this case (Figure 8.3c), we estimate all J joints $\hat{P} \in \mathbb{R}^{J\times 3}$ with the underlying pose estimator, but we feed this output through the autoencoder, and apply an additional loss term that measures the consistency of the prediction with the latent space, through an ℓ_1 loss, as

$$\mathcal{L}_{\text{cons}} = \left\| \hat{P} - W_{\text{dec}} W_{\text{enc}} \hat{P} \right\|_1 . \tag{8.10}$$

This encourages that the separately predicted skeletons can be projected to latent keypoints and back without information loss, thereby discouraging inconsistencies between them. The pose loss $\mathcal{L}_{\text{pose}}$ (from Section 8.3.1) is applied on $\hat{P}$.

Direct Latent Prediction. To avoid having to predict a large number of J joints in the base pose estimator, we define an alternative approach where the latents $\hat{Q} \in \mathbb{R}^{L\times 3}$ are directly predicted and then fed to the frozen decoder (Figure 8.4a). The last layer is reinitialized from scratch, as the number of predicted joints changes from J to L. The pose loss $\mathcal{L}_{\text{pose}}$ is applied on $W_{\text{dec}}\hat{Q}$.

Hybrid Student–Teacher. In a hybrid of the above two variants, we keep the full prediction head and add a newly initialized one to predict the latents $\hat{Q}$ directly

(Figure 8.4b). To distill the knowledge of the full prediction head to the latent head, we add a student–teacher-like ℓ_1 loss

$$\mathcal{L}_{\text{teach}} = \left\| \hat{Q} - \texttt{stop\_gradient}(W_{\text{enc}}\hat{P}) \right\|_1, \tag{8.11}$$

where the `stop_gradient` operation ensures that gradients from this loss are only backpropagated to the latent predictor (the student), as typical in student–teacher setups. During inference, we use $W_{\text{dec}}\hat{Q}$ as the output, to be as lightweight as direct latent prediction.

8.4 Experimental Setup

We now describe our experimental setup, consisting of the base 3D pose estimator model, the details of the training procedure, the used datasets and the evaluation metrics.

8.4.1 Base Model

We adopt the recent state-of-the-art MeTRAbs 3D human pose estimator (see Chapter 7) as the platform for our experiments, but we note that our method is agnostic to the specifics of the underlying pose estimator. Unless mentioned otherwise, the backbone is EfficientNetV2-S (Tan and Le, 2021).

8.4.2 Training Details

We crop a 256×256 px square around the person, apply perspective undistortion with camera intrinsics and perform augmentation as in Chapter 7.

Optimizer Settings. We perform 400k training steps with AdamW (Loshchilov and Hutter, 2019) by default. Our learning rate schedule is shown in Figure 8.5. The learning rate starts at 2.12e-4 and exponentially decays by a factor of 3 over 92% of training, then drops by a factor of 10 and then further decays exponentially by a factor of 3 until the end of the initial training.

The final fine-tuning phase has 40k iterations with smaller learning rate on the backbone than the heads. We perform a warm restart on the last layer (the prediction head) in order to ensure that the regularization loss can take effect, without disrupting the already mostly converged weights of the backbone. For the head, we follow a similar recipe as in the initial training, but we perform the large learning rate drop at 50% of the fine-tuning phase. For the backbone, we repeat the last, decaying segment of the initial schedule.

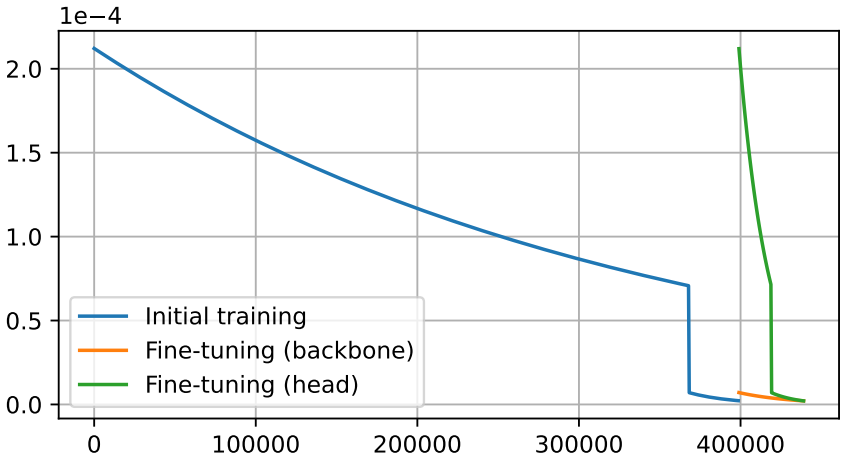


Figure 8.5: Learning rate schedule.

Loss Details. We make minor adjustments to the MeTRAbs model described in Chapter 7 which we use as the basis of our experiments. In Chapter 7, we performed internal supervision on the output of a 2D heatmap head and a 3D heatmap head, besides supervision on the final output. Instead, we simplify this and only use the absolute pose output for supervision, *i.e.*, the 2D projection loss and the mean-relative loss are computed on this single, absolute output. This makes the implementation cleaner when more losses are added for consistency regularization or student–teacher latent matching, *etc.*, since the model can be treated as a single-output black box.

The weak supervision loss for 2D-annotated examples only consists of the 2D projection loss. For this, we do not specifically predict skeletons according to the skeleton formats of the 2D datasets. Instead, the prediction is derived by averaging the corresponding 3D joint predictions for every output skeleton format. In other words, for the calculation of the 2D weak loss, we consider our prediction for the left shoulder to be the average of all the left shoulder joints in every skeleton format that we use.

We use $\lambda_{\text{proj}} = 1$ and $\lambda_{\text{abs}} = 0.1$ and scale the weak-supervision-loss by a factor of 0.2.

The absolute loss $\mathcal{L}_{\text{abs}}$ is only turned on after 5000 steps (also in fine-tuning, for consistency), similarly with the teacher loss. In some datasets the absolute distance to the person can be very large (*e.g.*, JTA, SAIL-VOS, ASPset). Here the absolute loss would overwhelm the total loss, so we scale down the absolute Z component to a maximum effective distance of 10 m for loss computation.

Batch Composition. We use a total batch size of 128, with every dataset represented with a fixed number of examples per batch. In Table 8.2, we specify the number of examples from each dataset per batch. This is based on the total number of examples in each dataset but not linearly, as we oversample smaller datasets compared to their size, in order to provide more diverse supervision to the model. For batch generation, we set up one queue per dataset that iterates over epochs of that dataset, then we interleave

the streams and chunk it into batches (as opposed to independently sampling each batch).

Initialization. We initialize with ImageNet-pretrained weights. For the RN50 experiment in Table 4 (SOTA), we use ResNet50V1.5 as implemented in PyTorch, ported to TensorFlow, along with the ImageNet weights, which we found to be superior to the ones provided with TensorFlow.

We precisely control the random seeds, which guarantees that bitwise equal batches are fed to each training run, improving comparability.

BatchNorm Aspects. We use Ghost BatchNorm (Hoffer *et al.*, 2017; Summers and Dinneen, 2020) with size 16, as this improved convergence in multi-dataset training, together with switching the BatchNorm layers to inference mode for the last 1000 updates.

In this *inference-mode fine-tuning* the BatchNorm layers use the stored, fixed statistics for normalization instead of the usual training mode of using the statistics of the current minibatch. In Ghost BN, the stored statistics may be suboptimal, since they are updated based on parts of the batch, instead of the overall batch statistics. A final fine-tuning in "inference mode" allows the network to fine-tune its weights to the setting that it will be used in during inference (*i.e.*, to adapt the weights to work well with the stored statistics).

Max-Norm Constraint to Combat Weight Explosion. Further, with long trainings, we found that training in 16-bit floating point (FP16) precision is unstable, as the activations tend to grow out of the representable range. This happens because the convolutional kernels of the backbone grow in scale during gradient descent. Indeed, since they are always followed by BatchNorm layers in EfficientNetV2, the weight scale has no impact on the network output (though it has an effect on the effective learning rate, van Laarhoven, 2017). Therefore, the gradient vector is always orthogonal to the weight vector for each kernel (as scaling the weights has no effect on the loss), leading to gradual growth in scale. Roburin *et al.* (2022) provide a detailed analysis of this spherical view of training with normalization layers and their first figure illustrates how this growth happens.

We mitigate the problem by applying a max-norm constraint on the convolutional kernels, to keep them from growing without bound. Some numerical instability remains in case of very long trainings, the cause of which would require more detailed investigation.

Implementation Details. We use TensorFlow version 2.9 with Keras, CUDA 11.4 and CuDNN 8.2.4 for the implementation. Training takes about 2 days with the EffV2-S

Dataset name	Small	Medium	Full
Real images with markerless MoCap			
MuCo-3DHP	32	9	6
CMU-Panoptic	–	9	7
AIST-Dance++	–	9	6
HUMBI	–	7	5
MPI-INF-3DHP	–	5	3
RICH	–	7	4
BEHAVE	–	–	3
ASPset	–	–	4
3DOH50K	–	–	3
IKEA ASM	–	–	2
Real images with marker-based MoCap			
Human3.6M	32	9	4
TotalCapture	–	5	3
BML-MoVi	–	–	5
Berkeley-MHAD	–	–	3
UMPM	–	–	2
Fit3D	–	–	2
GPA	–	–	4
HumanSC3D	–	–	1
CHI3D	–	–	1
Human4D	–	–	1
MADS	–	–	2
Synthetic images			
SURREAL	32	8	5
3DPeople	–	6	4
JTA	–	5	3
HSPACE	–	5	3
SAIL-VOS	–	7	5
AGORA	–	5	3
SPEC	–	–	2
Real images with 2D annotations (weak supervision)			
COCO	8	8	8
MPII	8	8	8
PoseTrack	8	8	8
JRDB	8	8	8

Table 8.2: Batch composition for the experiments with the three different levels of dataset combinations. Each minibatch consists of 96 examples with 3D labels and 32 with 2D labels.

backbone and about 6 days with EffNetV2-L on a single NVIDIA A40 GPU (48 GB) in mixed FP16/FP32 precision.

The autoencoder weights are trained on pseudo-GT obtained with EffNetV2-L.

8.4.3 Datasets

See Table 8.1 for an overview of all used datasets, which employ a variety of skeleton formats. In some cases, *e.g.*, when annotations are derived through triangulating COCO-like predictions (of *e.g.*, OpenPose), or through fitting a body model (*e.g.*, SMPL), we can assume that multiple datasets use the same convention (indicated in the last column). For other datasets, we assume the skeleton is a custom one, yielding 555 distinct keypoints in total. As most 3D human datasets contain videos, rather than isolated images, the number of sufficiently different poses is smaller than the total number of annotated frames. We hence discard examples where all joints remain within 100 mm of the last stored example. Our overall processing ensures that each training example has a person-centered image crop, camera intrinsics, 3D coordinates for some subset of the joints, a bounding box and a segmentation mask.

Where missing, we obtain person bounding boxes with YOLOv4 (Bochkovskiy *et al.*, 2020) and person segmentation with DeepLabv3 (Chen *et al.*, 2017b). Examples with implausible bone lengths are removed to avoid training on erroneous annotations. We use all cameras of 3DHP, and all HD cameras of CMU-Panoptic (and all sequences with labels). We further calibrated all cameras of BML-MoVi that did not have calibration provided in the dataset, and use all of them in training (based on pose predictions from an earlier version of our model). We use 200k composited images for MuCo-3DHP, generated with the official Matlab script.

8.4.4 Evaluation Metrics

We evaluate on four datasets: MuPoTS (Mehta *et al.*, 2018), 3DPW (von Marcard *et al.*, 2018), 3DHP (Mehta *et al.*, 2017a) and Human3.6M (Ionescu *et al.*, 2014). Over the years, different evaluation metrics and protocols have become customary on different datasets, whose details can be very arcane. Especially in a multi-dataset setting, we find it important to use consistent metrics. For our main experiments, we therefore adopt the following four metrics everywhere: MPJPE: mean Euclidean distance between predicted and ground truth joints after alignment at the root joint. PMPJPE: mean Euclidean distance after Procrustes alignment. PCK@100mm: percentage of joints predicted within 100 mm of the ground truth after root alignment. CPS@200mm: percentage of poses where *all* joints are within 200 mm distance of the ground truth after root alignment (Wandt *et al.*, 2021).

We evaluate all 24 SMPL joints for 3DPW, and 17 joints for 3DHP and MuPoTS. In case of 3DPW, the entire dataset is used for testing, none of it for training. For 3DHP

	MuPoTS-3D				3DPW				MPI-INF-3DHP				Human3.6M			
	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑
Single dataset	91.3	62.9	65.3	53.5	–	–	–	–	66.5	46.7	83.0	78.4	48.3	**33.2**	92.1	89.8
Small (◊)	88.4	61.3	67.3	59.9	81.0	54.5	72.9	35.8	64.8	45.9	83.7	78.9	**42.1**	33.8	**94.6**	**90.3**
Medium (♣)	86.1	59.4	69.0	**67.9**	64.3	45.6	82.5	70.0	61.7	44.6	85.6	80.2	43.2	34.5	94.3	90.3
Full (♡)	**84.6**	**59.0**	**70.1**	66.0	**61.8**	**43.4**	**83.8**	**71.1**	**59.6**	**44.1**	**86.6**	**81.8**	44.7	34.3	94.3	90.1

Table 8.3: Results using different amounts of datasets when training a separate-head model. Table 8.1 defines which datasets belong in which combination size. Using more datasets improves results on the 3DPW, 3DHP and MuPoTS benchmarks. On Human3.6M the small dataset combination gives better results, but this studio benchmark it less suited for studying real-world generalization capacity, as opposed to in-the-wild and outdoor benchmarks such as 3DPW and MuPoTS.

we use the official split, for Human3.6M the most common split from the literature, *i.e.*, subjects S9 and S11 are used for testing.

For MuPoTS, we evaluate the matched poses. We use the same YOLOv4 detector in all our experiments (with a high recall of 94.6%), hence the matched-pose results are directly comparable between our different configurations. For our main evaluations, in each benchmark, we simply calculate the average metrics over all metric-scale poses.

In Table 8.5, for SOTA comparison, we use the more complex standard evaluation metrics. That is, for MuPoTS, here we use bone rescaling, normalized skeletons, and averaging is performed first per sequence and the final value is the average of per-sequence averages. In this, and also other details, we follow the same protocols as in Chapter 7 (*e.g.*, which joints to evaluate).

8.5 Results

We present results showing that using many datasets makes the model more accurate, our consistency regularization brings qualitative and quantitative improvements, the final models are significantly above the performance reported in state-of-the-art papers, as well as further results and ablations.

8.5.1 Benefit of Training Data Scale

Since one contribution of this chapter is the study of the large-scale multi-dataset training regime, an important question is whether this brings improvements or whether performance saturates with just a few large-scale datasets. As a simple baseline, we train models on individual datasets and evaluate on the corresponding test splits.

(With MuPoTS, we use MuCo-3DHP for training). We then train on three dataset combinations, as shown in Table 8.3. There is a clear trend showing performance improvement when training with more datasets, and the small dataset combination also outperforms single-dataset baselines.

We note that Human3.6M scores sometimes suffer from additional data. Human3.6M uses the same studio environment in the training and evaluation split, therefore the model works better when a large part of the training batches are filled with Human3.6M examples, allowing it to specialize on images from this room, but this does not reflect true generalization ability. The model trained on the large dataset combination achieves very strong scores across the board, confirming that using many datasets makes a difference.

Despite the good benchmark scores, we qualitatively observe (Figure 8.2) that the different skeleton outputs can still be inconsistent among themselves.

8.5.2 Consistent Multi-Skeleton Prediction

A first naive baseline for achieving consistent predictions is to merge joints from different skeletons (*e.g.*, we predict only one "left shoulder" joint), reducing the joint count from 555 to 163. This leads to weaker results than predicting all joints separately (see Table 8.4), since joints with similar names may represent somewhat different keypoints.

Human3.6M is again an outlier, as the prediction with merged joints works well for it. Since the model can easily recognize that a test image comes from the Human3.6M studio, it can adapt its prediction to match the Human3.6M skeleton format. This is not possible on *e.g.*, 3DPW, since the model cannot know in advance what skeleton format will be used for the reference poses of these images, since they come from diverse in-the-wild scenes.

When using our proposed ACAE-based regularization (*cf*. Figure 8.3c), we can see consistent improvements for almost all metrics. However, the improvement in the qualitative performance of the model is even more striking. As seen in Figure 8.2, the regularized model creates significantly more consistent skeleton predictions. Especially the depth-consistency is improved, but some errors in the frontal view are also corrected. Figures 8.7–8.9 show further predictions for a variety of images, showing that these observations hold broadly. Furthermore, we see excellent in-the-wild performance, even on challenging poses, or in suboptimal lighting conditions.

Overall, the model that estimates latent keypoints (Figure 8.4a) has slightly lower performance than the separate-head baseline, likely because latent keypoints may be placed at less characteristic locations on the body and can thus be harder to localize. Further, the latent keypoint head's weights are initialized from scratch, whereas the regularization-based method fine-tunes a pre-trained head. The hybrid combination from Figure 8.4a performs slightly worse than the model that is only regularized, but

		MuPoTS-3D			3DPW			MPI-INF-3DHP			Human3.6M		
		MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑
EffNetV2-S	Merged joints	91.9	67.3	63.2	69.9	72.5	48.3	79.5	69.7	69.8	51.6	80.7	79.4	**44.6**	34.2	93.9	89.8
	Separate joints (F. 8.3a)	84.6	59.0	70.1	66.0	61.8	43.4	83.8	71.1	59.6	44.1	**86.6**	81.8	44.7	34.3	94.3	**90.1**
	Consistency regul. (F. 8.3c)	**81.8**	**57.8**	**72.5**	**72.9**	**61.5**	**43.0**	**84.0**	**71.9**	**59.2**	**43.6**	**86.6**	**82.7**	45.2	**33.3**	**94.4**	**90.1**
	Latent pred. (F. 8.4a)	83.0	58.9	71.4	71.2	62.0	43.6	**84.0**	71.7	60.2	44.7	86.1	80.2	46.5	34.4	93.9	89.5
	Hybrid (F. 8.4b)	82.7	58.5	71.6	72.1	61.8	43.3	**84.0**	71.8	60.4	44.8	85.9	80.9	46.1	34.2	94.1	89.4
EffNetV2-L	Separate joints	82.9	57.7	71.0	70.9	60.9	42.1	84.4	73.4	59.1	42.2	88.0	**85.3**	41.6	32.0	95.1	92.1
	Consistency regul.	**81.0**	**57.4**	**72.8**	**74.8**	**60.6**	**41.7**	**84.7**	**74.3**	**57.9**	**41.8**	**88.2**	84.7	**40.6**	**30.7**	**95.7**	**92.6**
	Hybrid	81.3	57.9	72.4	73.9	61.1	42.0	84.6	**74.3**	59.2	42.8	87.2	84.3	41.8	31.4	95.6	**92.6**

Table 8.4: Main results. We evaluate different strategies for handling different skeleton annotation formats during training.

	MuPoTS-3D	3DPW			MPI-INF-3DHP		Human3.6M
	PCK_{150}↑	MPJPE↓	PMPJPE↓	PCK_{50}↑	MPJPE↓	PCK_{150}↑	MPJPE↓
Sun *et al.* (2021)	–	80.1	56.8	36.5	–	–	–
Lin *et al.* (2021b)	–	74.7	45.6	–	–	–	51.2
Gong *et al.* (2021)	–	–	–	–	71.1	89.2	50.2
Cheng *et al.* (2023)	89.6	–	–	–	–	–	49.3
Ours with crop resolution 256x256 and 400k steps							
ResNet-50	92.2	65.5	47.2	49.0	64.2	93.3	45.8
EffNetV2-S	93.7	61.5	43.0	51.8	60.0	95.3	45.2
EffNetV2-L	94.1	60.6	41.7	52.1	59.2	95.8	40.6
Ours with crop resolution 384x384 and 800k steps							
EffNetV2-S	94.9	59.5	41.0	53.1	58.7	96.2	41.4
EffNetV2-S 5-crop TTA	95.2	58.9	39.9	53.6	57.5	96.7	40.1
EffNetV2-L	95.4	58.9	39.5	53.9	55.4	97.1	36.5
EffNetV2-L 5-crop TTA	95.7	57.0	38.1	55.4	53.6	97.6	35.5

Table 8.5: Comparison to recent state-of-the-art works. TTA means test-time augmentation.

in many cases still outperforms the baseline. This shows that a direct estimation of the discovered latent keypoints is also a viable option. By design, this approach also produces consistent results, since we compute a single latent set of keypoints from which we decode all skeletons.

We also train the regularization and hybrid variants with EffNetV2-Large (lower part of Table 8.4). Overall, the results follow the same order, and they are better across the board. Regularization improves results and also leads to consistent predictions, and the hybrid approach is somewhat better than the initial model trained to predict separate joints.

This means that our autoencoder-based regularization is effective at improving results both quantitatively and qualitatively, and the discovered latent keypoints can be predicted directly. This opens up interesting future research directions, as the latent keypoints can be seen as a model agnostic interface, potentially allowing us to incorporate new skeleton formats by expanding the decoder, without a need for model specific fine-tuning or probing.

8.5.3 Comparison to Prior Work

In Table 8.5, we compare our final results to recent state-of-the-art published works (using standard protocols) and observe much better accuracy than SOTA models. We emphasize that this comparison is not "fair" w.r.t. the amount of training data. However, our goal in this chapter is to show the value in large-scale multi-dataset training, and to investigate how to best supervise models in that setting.

	Chirality	MuPoTS-3D			3DPW			MPI-INF-3DHP		
		MPJPE↓	PCK_{100}	CPS_{200}	MPJPE↓	PCK_{100}	CPS_{200}	MPJPE↓	PCK_{100}	CPS_{200}
Cons. regul.		**81.8**	72.4	**73.1**	61.6	83.9	71.9	**59.2**	**86.6**	82.1
Cons. regul.	✓	**81.8**	**72.5**	72.9	**61.5**	**84.0**	71.9	**59.2**	**86.6**	**82.7**
Hybrid		83.2	71.2	71.7	61.7	**84.0**	**72.0**	60.3	85.9	80.7
Hybrid	✓	82.7	71.6	72.1	61.8	**84.0**	71.8	60.4	85.9	80.9

Table 8.6: Effect of enforcing chirality equivariance constraints on the autoencoder's weight matrices.

	#latents	MuPoTS-3D			3DPW			MPI-INF-3DHP		
		MPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PCK_{100}↑	CPS_{200}↑
Cons. reg.	24	**81.6**	72.4	**73.1**	62.0	83.9	71.7	**58.9**	86.5	81.7
	32	82.2	72.1	72.5	61.8	83.9	71.8	59.2	86.5	81.9
	48	81.8	**72.5**	72.9	**61.5**	**84.0**	**71.9**	59.2	**86.6**	**82.7**
	64	82.3	72.0	73.0	61.8	83.8	71.8	59.2	**86.6**	82.1
Hybrid	24	86.0	69.4	65.0	67.2	81.2	64.6	70.2	80.0	62.2
	32	**82.7**	71.4	**72.1**	62.3	83.9	**71.8**	**60.2**	**86.2**	**81.4**
	48	**82.7**	**71.6**	**72.1**	**61.8**	**84.0**	**71.8**	60.4	85.9	80.9
	64	84.1	70.6	67.0	62.2	83.7	71.3	62.0	84.5	80.2

Table 8.7: Effect of the number of latent points on final performance. We use 48 as our default setting.

8.5.4 Ablations

Chirality Equivariance Constraints. In Table 8.6, we analyze the effect of enforcing chirality equivariance on the ACAE. In the quantitative metrics, we see approximately no change or a slight positive effect on both evaluated models. Given that symmetry makes sense as an inductive bias, we use chirality equivariance in our default setting.

Latent Keypoint Count. As Figure 8.6 shows, once a minimum number of latent points is reached, the reconstruction error only decreases slowly (evaluated on a held-out pseudo-GT validation set). We evaluate several latent sizes for fine-tuning in Table 8.7. 48 points work well in practice, and our regularization method is robust w.r.t. this hyperparameter. When directly predicting latent keypoints, using too few or too many latent keypoints has a negative effect, but the differences are small beyond 32.

Training Length. In Table 8.8, we study the effect of the length of training on the final model performance. Clearly, longer training can further improve the results and especially the correct pose score improves. Do note that every new line doubles the number of training steps, so this is expensive.

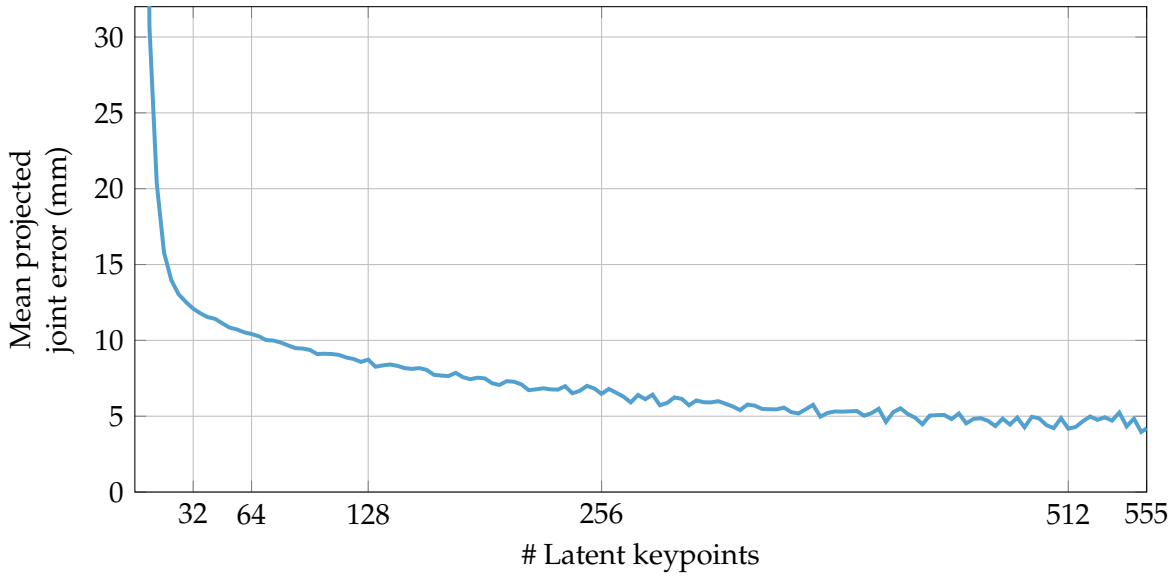


Figure 8.6: Intrinsic dimension analysis of our pseudo–ground truth with 555 joints. The residual error curve shows a characteristic elbow shape.

	MuPoTS-3D				3DPW				MPI-INF-3DHP				Human3.6M			
	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑
Initial, separate-skeleton model																
100k	88.6	62.4	67.8	60.0	65.9	47.0	81.5	65.3	66.3	51.2	82.2	71.7	47.7	37.6	91.9	86.6
200k	87.3	60.3	68.3	65.1	63.1	44.6	82.6	69.5	63.1	47.4	84.4	77.1	46.8	36.2	93.1	88.7
400k	84.6	59.0	70.1	66.0	61.8	43.4	**83.8**	71.1	59.6	44.1	86.6	81.8	44.7	34.3	94.3	90.1
800k	**82.9**	**57.8**	**70.5**	**69.8**	**61.7**	**42.6**	83.7	**73.1**	**58.8**	**42.9**	**87.3**	**83.0**	**43.0**	**33.2**	**94.8**	**91.4**
Fine-tuned with consistency regularization for 40k steps																
100k	85.5	60.6	70.3	66.6	65.0	46.0	81.8	66.9	63.6	48.7	83.9	74.1	46.7	36.4	92.5	87.6
200k	84.2	59.2	70.8	70.4	63.4	44.3	82.7	69.9	61.4	46.3	85.4	79.1	46.7	35.1	93.3	88.6
400k	81.8	57.8	72.5	72.9	61.5	43.0	84.0	71.9	59.2	43.6	86.6	82.7	45.2	33.3	94.4	90.1
800k	**80.5**	**56.8**	**72.7**	**74.4**	**61.3**	**42.1**	**84.5**	**73.3**	**57.7**	**42.2**	**87.7**	**84.3**	**42.0**	**31.9**	**95.3**	**91.4**

Table 8.8: Ablation for the length of training.

Table 8.9 shows that further extending the fine-tuning phase can bring minor performance benefits. For reasons of practicality, we chose 400k training steps and 40k fine-tuning step as the default setting for the main experiments, albeit one could achieve slightly better results with longer schedules.

	MuPoTS-3D				3DPW				MPI-INF-3DHP				Human3.6M			
	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑
20k	82.0	58.0	72.2	72.1	61.7	43.0	83.9	71.3	59.3	43.6	86.5	82.1	44.9	33.5	94.3	89.8
40k	81.8	**57.8**	**72.5**	72.9	61.5	43.0	84.0	71.9	59.2	43.6	86.6	82.7	45.2	**33.3**	94.4	**90.1**
80k	**81.6**	**57.8**	72.4	**73.2**	**61.4**	**42.9**	**84.1**	**72.1**	**58.4**	**43.2**	**87.2**	**82.9**	**44.5**	**33.3**	**94.6**	**90.1**

Table 8.9: Ablation for the length of consistency-regularized fine-tuning with an initial training length of 400k steps.

Ghost BatchNorm. In Table 8.10, we show an ablation on using Ghost BatchNorm. We compare three options: normal BatchNorm, Ghost BatchNorm where the 96 3D annotated examples are normalized as one group and the 32 2D-labeled ones as another, and Ghost BatchNorm with ghost batch size 16. While the differences are not very large, the Ghost BatchNorm options tend to perform better. This is probably due to the discrepancies in BatchNorm statistics among datasets.

Furthermore, Table 8.10 also demonstrates that it is important to fine tune the network at the end in inference mode when using Ghost BatchNorm.

8.6 Conclusion

We have proposed a principled, automatic approach to the problem of large-scale multi-skeleton training of 3D human pose estimation. Despite its practical relevance in exploiting a large number of 3D pose datasets in one training, this problem has been largely overlooked in the literature.

Our approach relies on a novel formulation of dimensionality reduction of sets of keypoints, via an affine-combining autoencoder with guaranteed built-in equivariances to common transformations. By regularizing a 3D human pose estimator's output to stay close to the learned latent space discovered by the autoencoder, we can more effectively share information between the different datasets, resulting in an overall more accurate and consistent pose estimator. We release code for data processing and training, as well as trained models to serve as high-quality off-the-shelf methods for downstream research.

	MuPoTS-3D				3DPW				MPI-INF-3DHP				Human3.6M			
	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑	MPJPE↓	PMPJPE↓	PCK_{100}↑	CPS_{200}↑
*Initial, separate-skeleton model, **with** fine-tuning at the end with BN in inference mode*																
Normal BN	84.5	59.2	70.0	65.9	62.6	43.6	83.6	**71.9**	61.3	**43.7**	85.7	**82.0**	46.3	34.6	94.2	89.9
Ghost BN (3D/2D)	**83.6**	**58.7**	**70.4**	**70.0**	62.8	43.5	83.2	71.2	61.8	45.1	85.7	80.3	46.6	34.7	94.1	**90.5**
Ghost BN 16	84.6	59.0	70.1	66.0	**61.8**	**43.4**	**83.8**	71.1	**59.6**	44.1	**86.6**	81.8	**44.7**	**34.3**	**94.3**	90.1
*Initial, separate-skeleton model, **without** fine-tuning at the end with BN in inference mode*																
Normal BN	84.2	59.1	70.4	66.4	62.6	43.6	83.5	71.9	60.4	43.4	86.1	82.3	45.9	34.4	94.2	89.8
Ghost BN (3D/2D)	88.2	63.5	66.7	62.0	67.3	47.5	80.9	68.0	66.3	49.0	81.7	77.4	50.8	40.4	90.7	87.7
Ghost BN 16	85.8	60.4	69.0	63.5	63.4	44.7	83.2	70.8	59.8	44.3	86.3	81.8	45.4	35.7	93.6	89.6
*Fine-tuned with consistency regularization for 40k steps, **with** fine-tuning at the end with BN in inference mode*																
Normal BN	83.3	58.5	70.9	73.2	63.1	43.7	83.4	71.8	60.5	**43.4**	85.9	**82.7**	46.0	33.5	**94.4**	89.9
Ghost BN (3D/2D)	**81.2**	**57.7**	**72.5**	**74.0**	62.2	**43.0**	83.7	**72.3**	60.5	44.4	86.1	80.9	46.2	33.7	94.1	**90.6**
Ghost BN 16	81.8	57.8	**72.5**	72.9	**61.5**	**43.0**	**84.0**	71.9	**59.2**	43.6	**86.6**	**82.7**	**45.2**	**33.3**	**94.4**	90.1
*Fine-tuned with consistency regularization for 40k steps, **without** fine-tuning at the end with BN in inference mode*																
Normal BN	83.1	58.7	71.3	72.9	62.8	43.5	83.5	71.7	59.6	43.2	86.4	82.9	45.5	33.4	94.5	89.9
Ghost BN (3D/2D)	89.1	64.6	66.4	62.1	68.5	49.4	79.6	65.2	73.4	52.9	76.4	70.9	58.4	43.4	85.7	82.7
Ghost BN 16	84.0	60.0	70.7	69.7	63.3	45.3	82.7	69.4	63.2	45.7	83.8	80.1	49.0	36.6	92.2	88.3

Table 8.10: Ablation for Ghost Batch Normalization and inference-mode fine-tuning for 1000 steps.

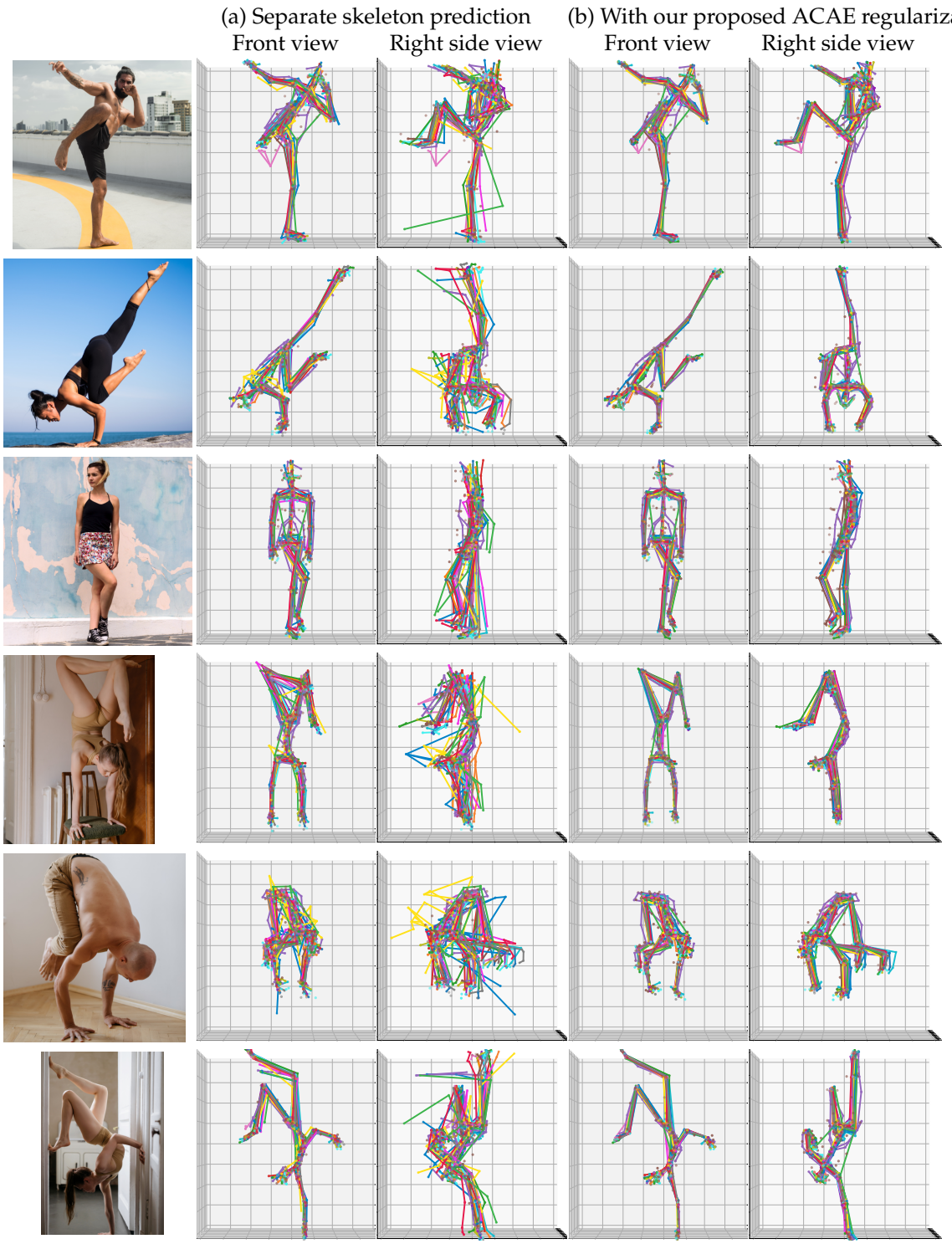


Figure 8.7: A qualitative result comparison between a model trained without (a) and with our ACAE regularization (b). It can clearly be seen that our regularization leads to improved skeleton consistency.

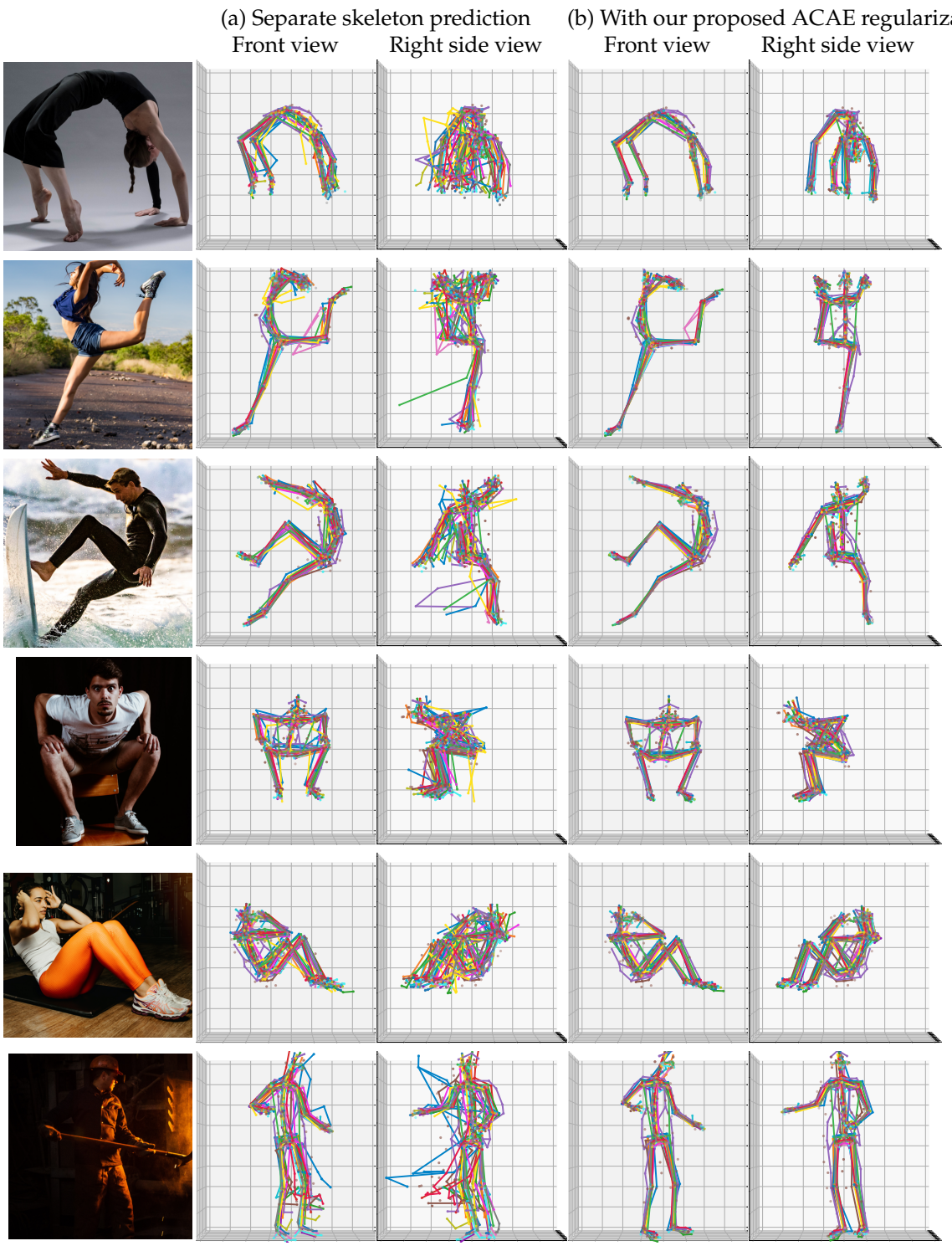


Figure 8.8: A qualitative result comparison between a model trained without (a) and with our ACAE regularization (b). It can clearly be seen that our regularization leads to improved skeleton consistency.

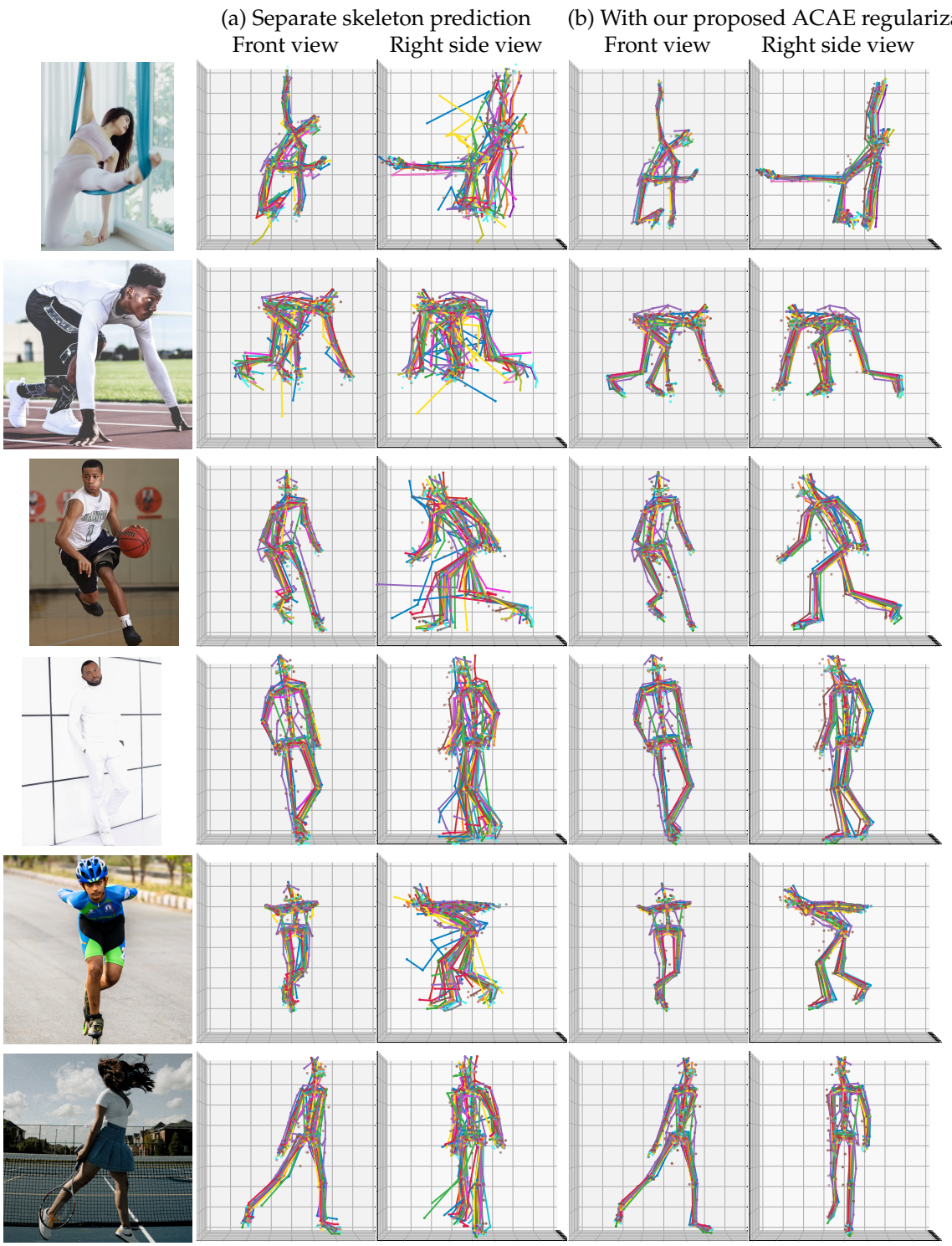


Figure 8.9: A qualitative result comparison between a model trained without (a) and with our ACAE regularization (b). It can clearly be seen that our regularization leads to improved skeleton consistency.

9 Reposing Humans by Warping 3D Features

In the previous chapters, we have presented a series of methods using volumetric body joint heatmap representations, predicted directly through fully convolutional networks. Motivated by the state-of-the-art performance achieved through this representation, we pose the research question: Could we predict richer information than body joints in the same way?

To explore this possibility, we consider the task of pose-conditioned human image synthesis, also known as person reposing. In this task, the inputs are an image of a person and a desired target pose, while the output needs to be a synthesized image of the same person in the target pose. This requires representing not only the human pose but also the appearance, as well as reasoning about occlusions and unseen body parts. Our aim in this chapter is to demonstrate the versatility of direct volumetric prediction, allowing us to create a three-dimensional *volumetric feature map*, which can then be easily manipulated by geometric warping operations to change the person's pose, as illustrated in Figure 9.1. Crucially, the warping will be performed in 3D, which, as we demonstrate, improves the resulting image quality, over a 2D baseline.

This chapter is based on our publication (Knoche *et al.*, 2020), presented at the 2020 CVPR Workshop Towards Human-Centric Image/Video Synthesis, in turn based on Markus Knoche's master's thesis project, supervised by Prof. Bastian Leibe and myself.[1]

9.1 Overview

The ability to freely change a human's pose has a variety of real-world applications as well, from generating large crowds or performing stunts in film making to data

[1] While I devised the motivation and overall idea behind this work, most of the implementation was performed by Markus Knoche as part of his master's thesis. The detailed experimental and architectural design choices were made in collaboration between us.

augmentation for human-centric computer vision tasks. Several prior works employ fully convolutional neural networks for this task. However, unlike typical image-to-image translation tasks (*e.g.*, colorization), reposing requires moving information over large spatial distances in the image, since the same body part may appear at a very different image position in the input and the output. The only mechanism available to typical fully convolutional networks for moving information over image space is to gradually pass it on to neighboring pixel locations in the convolutional layers. This, however, requires a large number of layers to reach the required distance. To make "information shuttling" quicker, many recent approaches apply some form of explicit geometric transformations. Some warp 2D features such that they become aligned with the target pose, which is also specified in 2D (Balakrishnan *et al.*, 2018; Dong *et al.*, 2018; Neverova *et al.*, 2018; Siarohin *et al.*, 2018; Grigorev *et al.*, 2019; Horiuchi *et al.*, 2019). We argue that such a 2D approach is insufficient to capture complex, three dimensional changes in articulated human pose.

Mesh-based approaches fit a 3D body model to the input, infer the texture and render the mesh in the target pose (Zanfir *et al.*, 2018; Liu *et al.*, 2019c). While capturing the 3D aspect, these approaches have the downside that a specific human might not be captured well by a general model, for example due to uncommon hairstyles and spacious clothing.

Inspired by recent volumetric approaches for related tasks (Pavlakos *et al.*, 2017; Nguyen-Phuoc *et al.*, 2019), we propose a novel reposing method, illustrated in Figure 9.1, which warps 3D volumetric CNN features without requiring an explicit mesh model. Using only a 2D image as input, our model implicitly learns a latent volumetric representation of the input person. This representation is then warped using 3D transformations based on input and target pose to align it to the target pose. We process the warped features along with 3D target pose heatmaps with a decoder, to synthesize the reposed image.

By ablation, we show the benefits of the two 3D aspects of our work: first the 3D warping, and second, representing the target pose in 3D. Overall, our method achieves state-of-the-art scores on the commonly used DeepFashion and the newer iPER benchmarks. Our code is publicly available to enable further research.[2]

9.2 Related Work

Initial Methods. Image generation methods have come a long way since the introduction of generative adversarial networks (GAN; Goodfellow *et al.*, 2014). Building upon Isola *et al.*'s (2017) image-conditioned GAN, Ma *et al.* (2017) were the first to tackle pose-conditioned person image generation. Their method feeds the image and 2D target pose heatmap through two stages: the first is trained with a pixelwise ℓ_1

[2] `https://vision.rwth-aachen.de/warp3d_reposing`

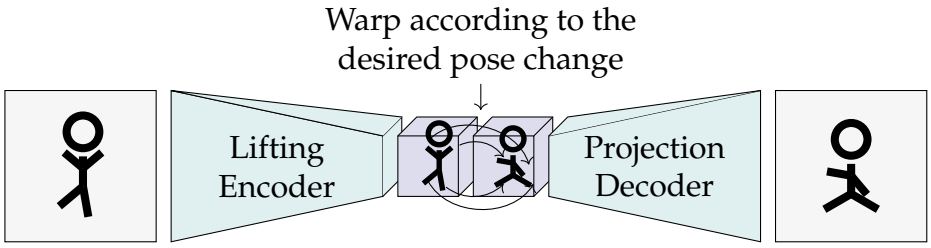


Figure 9.1: We use a fully convolutional encoder network to generate a volumetric representation of the input person, which can be warped in 3D and decoded into a 2D output image to achieve reposing.

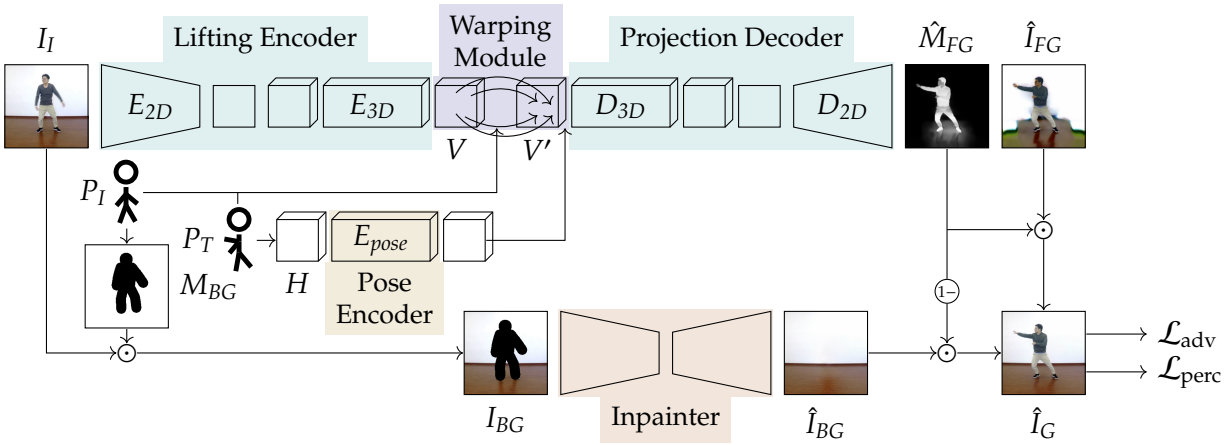


Figure 9.2: A detailed look at our image generation architecture. The foreground stream produces a 3D voxel grid of features from a 2D image and applies 3D feature warping. The warped features are combined with the target pose and projected to an RGBA image through a decoder module. Alpha blending with an inpainted background yields the final output.

loss, the second with an adversarial loss. Lakhal *et al.* (2018) use two encoders in both stages, distinguishing between aligned and misaligned input in Stage I and between pose and images in Stage II. Similar subdivisions are used in other works (Si *et al.*, 2018; Zhu *et al.*, 2019).

Explicit Warping. Several approaches tackle the misalignment of the input and the target by explicit warping. Siarohin *et al.* (2018) use affine transformations on the skip connections of a U-Net architecture (Ronneberger *et al.*, 2015). They mask out features corresponding to the various body parts based on the input pose, and warp these parts to align with the target pose. Horiuchi *et al.* (2019) propose an extension, adding self-attention layers, spectral normalization and a relativistic discriminator to

the architecture. Balakrishnan *et al.* (2018) apply a similar transformation directly on the input image, using learned soft masks.

Dense Conditioning. Some methods guide the warping based on body part segmentation or use DensePose (Güler *et al.*, 2018) to encode the input and the target poses, instead of using keypoint representations (Dong *et al.*, 2018; Neverova *et al.*, 2018; Grigorev *et al.*, 2019). Such a fine-grained conditioning input tells the network the exact shape of the target person, making the task simpler, but a dense target pose is not always available.

Mesh-Based Approaches. Zanfir *et al.* (2018) fit a parametric 3D body mesh model to the given person, back-project image pixels onto the mesh, then transform the mesh to the target pose, which can be rendered to yield the final image. Unseen parts of the texture are inpainted with a neural net. Other methods use mesh models to compute a flow field from the input to the target pose, and use this flow to transform network features in 2D (Li *et al.*, 2019a; Liu *et al.*, 2019c).

Volumetric Heatmaps. A line of works in 3D human pose estimation (Pavlakos *et al.*, 2017; Luvizon *et al.*, 2018; Sun *et al.*, 2018a; Sárándi *et al.*, 2020) has shown that it is feasible to predict depth-related information from images in a volumetric representation (in that case volumetric body joint heatmaps), by a tensor reshaping operation. We take this as inspiration to predict volumetric feature maps of humans in our work.

Novel View Synthesis. Human reposing can be viewed as a generalization of novel view synthesis (NVS) from rigid pose to articulated pose. As volumetric prediction has also been successfully applied for NVS (Nguyen-Phuoc *et al.*, 2019; Sitzmann *et al.*, 2019), we take this as further motivation to investigate the usefulness of a similar representation in reposing.

In contrast to the volumetric approach, a sparse 3D representation is used in Rhodin *et al.* (2018b) to learn NVS. Their encoder outputs an appearance feature vector and a 3D point list representing the pose. After rotating the point cloud, the decoder transforms the resulting point cloud and the appearance features to an image seen from the novel view. The implicitly learned point list is then given as input to a shallow human pose estimation network, thereby reducing the amount of labeled pose estimation data needed. Rhodin *et al.*'s idea of explicitly transforming an implicitly learned 3D structure is one of the inspirations of our work. However, instead of transforming a list of points, which only encodes the pose, we transform rich volumetric features, which also contain appearance information. Furthermore, we consider full articulation instead of only rigid transformation.

Unpaired Training. Most methods for human reposing (including ours) rely on paired training data, *i.e.*, pairs of images depicting the same person two different poses. However, we note that some approaches can learn the reposing task from unpaired images as well. Pumarola *et al.* (2018) use a CycleGAN (Zhu *et al.*, 2017) framework for this, Esser *et al.* (2018) use variational autoencoders, while Ma *et al.* (2018) construct disentangled streams for background, body parts and pose. These research directions are orthogonal to our goal in this chapter, which is to investigate the value in volumetric 3D representations over 2D ones in this task.

9.3 Method

Given an input image I_I of a person and a target pose P_T, we aim at generating an image $\hat{I}_G$ of the same person in pose P_T. We use a two-stream generator network to tackle this problem, where the first stream reposes the person using our novel volumetric feature warping approach, while the second stream inpaints missing parts of the background. To perform volumetric warping, the model first needs to estimate the depths of different body parts, such that it can then lift the corresponding features into to a 3D volume. Our model learns this volumetric feature generation and depth estimation implicitly. We neither give depth information about the input pose to our model, nor do we apply any explicit supervision with respect to the input pose. Instead, since the generated volumetric features are explicitly warped during training, the model learns to place relevant features at each position in the volume, in order to have them shuttled to the correct place in the result. In this, we make use of the fact that the warping module is differentiable.

9.3.1 Architecture

Our architecture consists of a lifting encoder, a 3D warping module, a projection decoder and a background inpainter, as shown in Figure 9.2.

The **lifting encoder** maps a 2D input image to 3D volumetric features. The 2D input image I_I is passed to a convolutional network E_{2D} which outputs 2D feature maps $E_{2D}(I_I) \in \mathbb{R}^{H \times W \times D \cdot C}$. A reshape operation splits the channel dimension of the resulting tensor into different depth layers, yielding the feature volume $F \in \mathbb{R}^{H \times W \times D \times C}$. This is similar to how joint heatmaps are estimated in Pavlakos *et al.* (2017), but instead of heatmaps, we produce a latent feature volume. E_{2D} thus learns that different features in its output belong to different depths. To further process these volumetric features, a 3D convolutional network (E_{3D}) is applied to yield $V \in \mathbb{R}^{H \times W \times D \times C}$.

The key element of our approach is our novel **3D warping module**, whose purpose is to shuttle voxel features to their target location. The warping module gets a feature volume $V \in \mathbb{R}^{H \times W \times D \times C}$, together with the 3D input and target pose $P_I, P_T \in \mathbb{R}^{J \times 3}$ which are given as 3D joint coordinates. The input pose P_I is used to create ten masks

$M_i \in \{0,1\}^{H \times W \times D}$, one per body part. Masks are generated by drawing capsular shapes (cylinder capped with two half spheres) between the joints corresponding to that body part. *E.g.*, the lower left leg's mask is based on the left ankle and the left knee joints and the mask of the torso depends on the hips and shoulders. We then create ten copies V_i of the feature volume and apply the corresponding mask by voxelwise multiplication, giving us ten volumes, again one per body part. Next, we fit a transformation T_i for each body part based on input and target joints. We assume that each part moves rigidly, but as the scale of the person in pixel space may change, we also add a scaling parameter. The result is a seven-parameter Helmert transformation, which we estimate by least squares. When a body part has only two joints, as for leg and arm parts, we use a third joint to specify the rotation around the body part's own axis. For example, the left lower arm's movement would otherwise only depend on the left wrist and the left elbow, which alone do not determine a unique Helmert transformation. We therefore additionally use the left shoulder's position as an anchor. The masked body part features are then warped according to the respective transformation with trilinear interpolation and the warped features are combined by taking the maximum value across the ten body parts. Given M_i and T_i, the output feature volume of the warping module is

$$V' = \max_i T_i(M_i \odot V). \tag{9.1}$$

The **target pose encoder** E_{pose} feeds the target pose into our model. Its inputs are Gaussian volumetric heatmaps $H \in \mathbb{R}^{H \times W \times D \times J}$, one per body joint. Its output is concatenated with the warped volumetric features and are processed by the projection decoder to produce the foreground result.

Mirroring the lifting encoder, our **projection decoder** contains two parts: D_{3D} and D_{2D}. Since the warping module cuts and pastes different parts of the volumetric feature maps to new positions, its output can contain some artifacts at the borders of body parts or at voxels where body parts overlap. The purpose of the 3D convolutional network D_{3D} is to clean up and enhance these features and to combine them with the output of the target pose encoder. The 3D feature volume is then reshaped back to 2D, by combining the depth axis and and volumetric channel axis into a single channel dimension using tensor reshaping. We then apply the second decoder network D_{2D}, which yields the generated foreground RGB image $\hat{I}_{FG}$ together with a soft mask $\hat{M}_{FG}$.

We apply a **background inpainter** stream, since our warping module only copies masked body parts to the decoder, thus losing background information. We first remove the foreground person from the inpainter module's input, based on the body part masks as given in our warping module. Pixels that are not included in any of the projected body part masks become part of the background mask M_{BG}. The inpainting itself is performed using partial convolutional layers (PartialConv; Liu *et al.*, 2018). The final result is a weighted combination (alpha blending) of the inpainted background $\hat{I}_{BG}$ and the generated person $\hat{I}_{FG}$ using $\hat{M}_{FG}$ as the weights.

Architectural Details. All our subnetworks, except the background inpainter, but including the discriminator, are based on the ResNet architecture (He *et al.*, 2016a,b). We use GroupNorm (Wu and He, 2018) instead of BatchNorm (Ioffe and Szegedy, 2015) due to its better performance with small batch sizes (our batch size is only 2, due to memory costs). In E_{2D} and D_{2D}, we use bottleneck residual blocks to reduce computational cost. Our 3D convolutional networks E_{3D}, D_{3D} and E_{pose} do not use bottlenecks, because the number of features is already comparatively low.

9.3.2 Training

We use two losses, a perceptual and an adversarial one. The perceptual loss $\mathcal{L}_{\text{perc}}$ (Johnson *et al.*, 2016) compares the generated image with the target image, by passing both through an ImageNet-pretrained VGGNet (Simonyan and Zisserman, 2015), and computing the ℓ_1 distance on multiple feature maps. The adversarial loss $\mathcal{L}_{\text{adv}}$ uses a discriminator network as in any GAN. Our discriminator takes as input either the generated or the ground truth target image, along with the source image and the 3D target heatmap and performs a real *vs*. fake classification.

We jointly optimize a weighted combination of these two losses:

$$\mathcal{L} = \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \tag{9.2}$$

We use data augmentation with rotation, scaling, translation, horizontal flip and color distortion. We train with the Adam optimizer (Kingma and Ba, 2015) for 150 000 steps with batch size 2 and learning rate 2×10^{-4}. We empirically set $\lambda_{\text{adv}} = 1$ and $\lambda_{\text{perc}} = 3$.

9.4 Experiments

9.4.1 Datasets

Commonly used in related work, the In-shop Clothes Retrieval Benchmark of the DeepFashion dataset (Fashion; Liu *et al.*, 2016) has almost 50 000 images and 8 000 sets of clothes.

The newer Impersonator dataset (iPER; Liu *et al.*, 2019c) contains videos of 30 people and 103 clothing styles in total. The dataset provides two videos per clothing style, filmed from a static camera. In one video, the person turns around in an A-pose, the other one shows more complex, arbitrary movements.

As these benchmark datasets do not supply 3D poses, we generate the input and target poses P_I and P_T using our MeTRAbs 3D human pose estimator from Chapter 7. This particular model has a ResNet152V2 backbone and we train it on Human3.6M, MPI-INF-3DHP, CMU-Panoptic and 3DPW for 3D supervision, as well as on COCO and MPII for 2D supervision.

9.4.2 Evaluation Metrics

Generated image quality is somewhat subjective, and therefore several quantitative metrics have been used in related work to compare methods.

SSIM. The structural similarity index (SSIM; Wang *et al.*, 2004) compares patches of the generated image to patches of the ground truth according to luminance, contrast and structure.

Inception Score. While many related works in person reposing evaluate their approach using the Inception Score (IS; Salimans *et al.*, 2016), we argue that the IS is not appropriate for this task. The IS was originally proposed for evaluating unconditioned GANs whose output distribution is supposed to cover diverse classes, *e.g.*, all ImageNet classes. The IS combines two aspects, the realism of individual output images and the diversity of a large set of generated images. To compute the IS, the result image is first passed through the Inception network (Szegedy *et al.*, 2016). If the generated image x is realistic, it should be confidently assigned to a single class the Inception network, so the output class y's posterior distribution $p(y|x)$ should have a prominent peak. At the same time, an unconditioned GAN should cover many classes in its generations, therefore the marginal $p(y)$ should be rather uniform. In other words $p(y|x)$ and $p(y)$ should be different for a good unconditioned GAN. The IS is the Kullback–Leibler divergence of these two distributions, meaning that the score is high if the distributions are dissimilar and therefore the GAN is working well. However, in the case of human reposing, we only have one output class: *person*. Therefore, $p(y)$ and $p(y|x)$ should be the same distribution, the Inception network should classify every generated image as a person. This makes the use of the IS for this task invalid and we therefore do not report it. The use of the IS can also be misleading in other contexts, as noted in Barratt and Sharma (2018)

LPIPS. We further use the learned perceptual image patch similarity (LPIPS; Zhang *et al.*, 2018), which compares deep features between generated image and ground truth, similar to perceptual losses (Johnson *et al.*, 2016).

Pose AUC. To evaluate high-level structure as opposed to low-level texture quality, we compare the response of a pretrained MeTRo 3D pose estimator as described in Chapter 6, when applied to the generated and the true image. This evaluator model has a ResNet50V2 backbone and is pretrained on the iPER 3D pseudo–ground truth, which we originally obtained with the MeTRAbs model described in Section 9.4.1, as well as COCO and MPII for 2D supervision.

We use the area under the PCK (percentage of correct keypoints) curve (AUC@150mm), a standard pose metric (Mehta *et al.*, 2017a).

3D warping	3D target pose	SSIM ↑	$SSIM_{fg}$ ↑	Pose AUC ↑
–	–	0.872	0.566	0.698
–	✓	0.875	0.578	0.749
✓	–	0.877	0.607	0.749
✓	✓	**0.883**	**0.626**	**0.777**

Table 9.1: Ablation study on iPER. $SSIM_{fg}$ is masked to evaluates only the foreground pixels.

	iPER		Fashion	
	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓
PG2 (Ma *et al.*, 2017)	0.854	0.135	0.762	–
SHUP (Balakrishnan *et al.*, 2018)	0.823	0.099	–	–
DSC (Siarohin *et al.*, 2018)	0.829	0.129	0.756	–
LWB (Liu *et al.*, 2019c)	0.840	0.087	–	–
SGW (Dong *et al.*, 2018)	–	–	0.793	–
UPIS (Pumarola *et al.*, 2018)	–	–	0.747	–
VUNET (Esser *et al.*, 2018)	–	–	0.786	0.196
BodyROI7 (Ma *et al.*, 2018)	–	–	0.614	–
DPT (Neverova *et al.*, 2018)	–	–	0.796	–
CTI (Grigorev *et al.*, 2019)	–	–	0.791	**0.169**
Li *et al.* (2019a)	–	–	0.778	–
Ours	**0.863**[3]	**0.077**[3]	**0.800**	0.186

Table 9.2: Comparison to prior work. iPER scores for competing methods are taken from Liu *et al.* (2019c).

9.4.3 Ablation Study

In contrast to prior work on person reposing, we propose to perform two different aspects in 3D: first, we use a 3D target pose and second, we perform 3D feature warping in the center of our model. Architectural differences make it hard to directly compare our results to prior works, so we define ablation models to investigate these two aspects while keeping the exact same architecture otherwise.

To drop the depth information from the 3D target pose heatmaps, we project the pose to the image plane and replicate it to all depth layers. Similarly, to perform

[3]Compared to our published work in Knoche *et al.* (2020), the scores here have been updated after Liu *et al.* (2019c) shared their detailed evaluation protocol with us. Our originally published results, using our own split, were 0.883 for SSIM and 0.081 for LPIPS.

Figure 9.3: Comparison with Siarohin *et al.*'s (2018) 2D feature-warping method using deformable skip connections. (Note that the target image is only used for visualization here, it not used as input to the networks, only the pose.)

warping only in 2D, we project the body part masks to the image plane and copy them to all depths and apply 2D affine warping to all depths independently.

The results on iPER are shown in Table 9.1. Both of our 3D enhancements improve the scores compared to the 2D baseline, and the results get even better when the two 3D aspects are combined. The qualitative results shown in Figure 9.4 further support this observation. In the first row, the 2D pose models wrongly generate the right hand in front of the body, while the second row shows that a combination of both 3D aspects achieves the best results.

9.4.4 Comparison to Prior Work

Quantitatively, our model achieves state-of-the-art scores on both iPER and Fashion, as shown in Table 9.2.

Figure 9.3 presents a qualitative example on Fashion. Our model generates the overlapping arms of the right person better than the 2D feature warping approach of Siarohin *et al.* (2018). A qualitative comparison to Liu *et al.* (2019c) on iPER is presented in Figure 9.4, showing that our model is able to transform the features of the left arm independently from the body features. In the upper row the hand correctly appears behind the body and the blue jacket in the lower row does not have a white stain as residue from the arm color. Further qualitative comparisons can be found in Figures 9.5–9.7.

Limitations. As seen in these results, a downside of our approach is that certain fine details are lost in some cases. For example, the buttons on the shirt in the first row of Figure 9.4 are missing in two ablation models and replaced by a zipper and shirt pockets in the other two. This is likely because our architecture produces low-resolution feature maps internally.

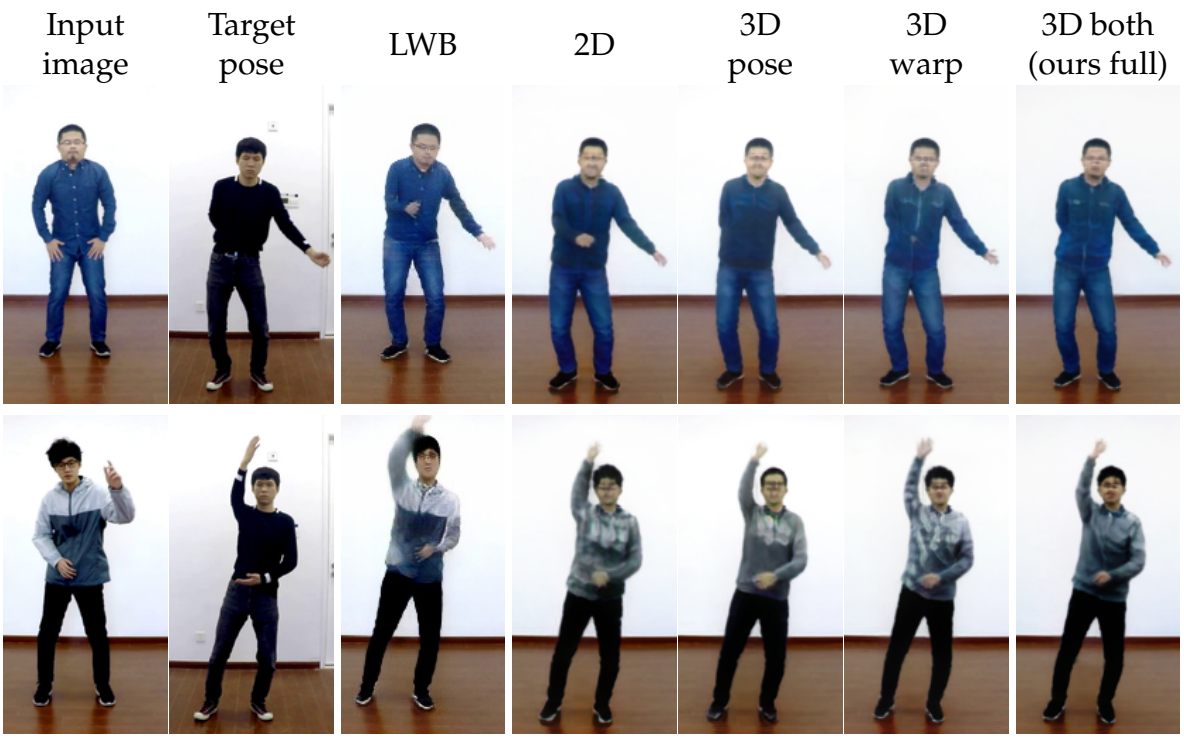


Figure 9.4: Comparison to the mesh-based method using the Liquid Warping Block (LWB; Liu *et al.*, 2019c), as well as to our ablation baseline models.

9.5 Conclusion

We presented a novel architecture for person reposing, which relies on 3D warping of implicitly learned volumetric features. Different from prior work, our approach is neither limited by approximating 3D motion with 2D transformations nor is an explicit 3D human mesh model required.

The ablation study and the comparison to related approaches showed that our method outperforms 2D warping methods by a significant margin. This indicates that volumetric representations and 3D warping are a promising way to tackle reposing. In the broader context of the thesis, it also demonstrates that rich appearance features can be learned in a structurally similar way to volumetric pose heatmaps.

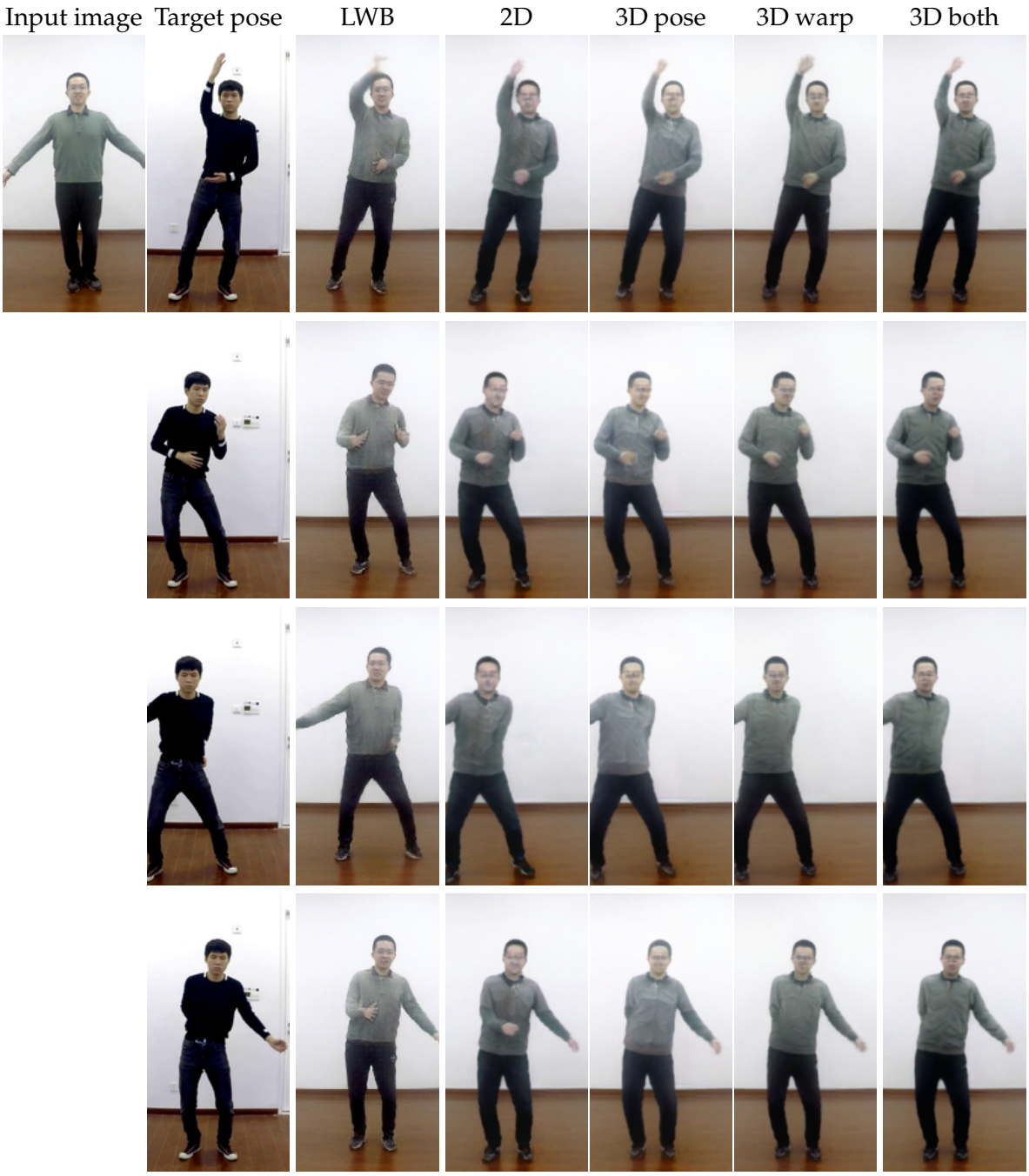


Figure 9.5: Qualitative comparison between Liu *et al.*'s (2019c) LWB and our main and ablation models.

Figure 9.6: Qualitative comparison between Liu *et al.*'s (2019c) LWB and our main and ablation models.

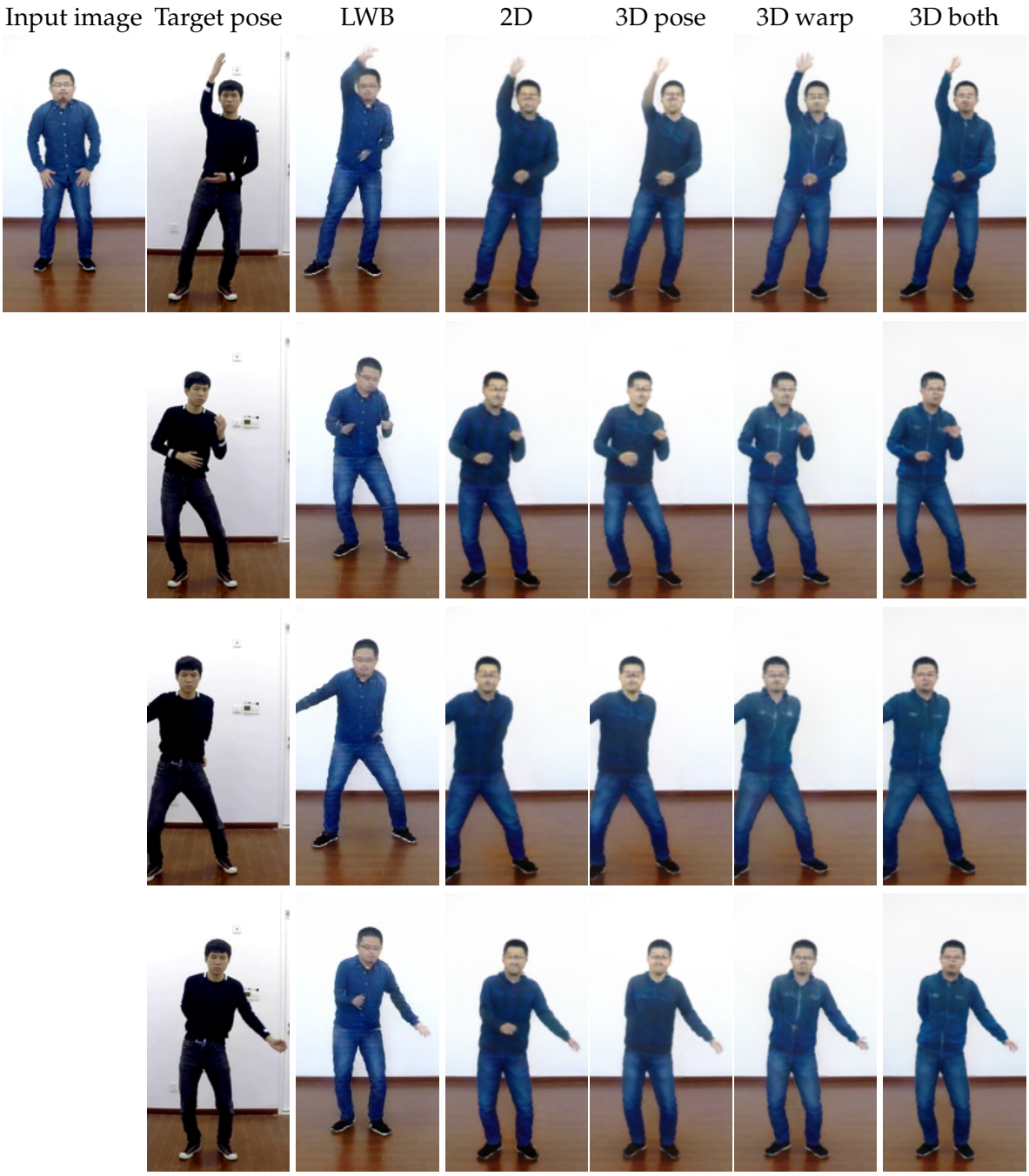


Figure 9.7: Qualitative comparison between Liu *et al.*'s (2019c) LWB and our main and ablation models.

10 Conclusion

In this thesis, we have presented several novel approaches for improving the state of the art in 3D human pose estimation. To conclude the thesis, we now summarize our contributions presented in the technical chapters, and discuss exciting open challenges for this field as well as potential future research directions to tackle them.

10.1 Summary and Contributions

After introducing related work and fundamentals in Chapters 2 and 3, we described our study of the occlusion robustness of 3D human pose estimation in Chapter 4. Through detailed experiments on the Human3.6M benchmark, we found out that the shape of the occluder object matters. We analyzed the effects of occluding with a single rectangle, several rectangles, circles, oriented bars and realistically segmented objects. Out of these, rectangles turned out to be the least problematic, even though prior occlusion research (in other tasks) typically used rectangles for this purpose. On the other hand, circular occluders were most detrimental. To mitigate the issue, we proposed to apply occlusion augmentation on the images during training. Here, we saw that augmenting with simple geometric shapes did very little to increase robustness against realistic occlusions at test time. In contrast, augmentation by pasting complex objects over the image was much more versatile and generalized to all 5 occlusion types, both realistic and simple geometric ones. We have also noted that synthetic occlusion augmentation with realistic segmented objects improved prediction quality also on non-occluded test images. This can be explained as a regularization effect similar to dropout. Making use of this observation, we presented a case study in Chapter 5, adapting our approach to the ECCV 2018 PoseTrack Challenge, where it achieved first place.

Then, in Chapter 6, we introduced our metric-scale truncation-robust (MeTRo) volumetric heatmap representation for 3D pose estimation. By defining the heatmap space at a fixed metric size, we decoupled it from the input image space, which

resulted in both truncation robustness and learnable scale recovery. This stands in contrast to typical 2.5D heatmap methods that tie two axes to the image space, and do not learn scale recovery. Our representation allowed for a similar flexibility to direct coordinate regression methods, while keeping the structural benefits of heatmap estimation, *i.e.*, encoding positions via high-activation locations, instead of activation values. We confirmed the effectiveness of the method by comparing it to a typical 2.5D heatmap baseline and by achieving state-of-the-art scores on Human3.6M and MPI-INF-3DHP. We also introduced the centered striding mechanism to allow an even distribution of the receptive fields of the last-layer units across the image, instead of a top-left bias. This allowed us to dynamically change the striding to different values at test time compared to the training configuration. We further verified that the occlusion augmentation remained effective in our new approach, on both datasets, and performed experiments to disentangle the effects of the irregular outlines and the realistic textures of synthetic occlusions.

In Chapter 7, we then extended the MeTRo approach with a 2D heatmap branch and a differentiable absolute pose reconstruction module. This allowed us to make predictions in the camera-relative 3D space, and to train the network end-to-end with an absolute pose loss. We kept architectural simplicity as a key priority in our design, and used a simple backbone network without any high-resolution decoder at the end. Instead, we showed that soft-argmax enables high-quality output even at a low (8×8×8) heatmap resolution. With this approach, we outperformed the prior state of the art on the MuPoTS-3D benchmark and also won the 3DPW Challenge at ECCV 2020. From the perspective of efficiency, we also evaluated our method on low-powered embedded hardware and found it to be capable of real-time performance.

Next, we presented our contributions to multi-dataset learning of 3D pose estimation. We called attention to the often overlooked problem of different skeleton formats used across datasets. To address the problem, we proposed a geometric autoencoding method for discovering the relations and redundancies among the different skeleton formats. For this, we first had to assemble a pseudo-labeled parallel corpus of examples annotated with every skeleton format. We then used the pseudo–ground truth to train our novel affine-combining autoencoder to discover latent 3D keypoints in its information bottleneck. Since we defined both the encoder and the decoder to perform simple affine combinations, the whole architecture is equivariant to rotation and translation. This has the benefit that the skeletal relations observed in the 2D image plane can be transferred to the more challenging depth axis. With this approach, we trained high-quality models on 28 datasets simultaneously, which is a much larger scale than prior experiments. We demonstrated improved results with our autoencoder-based regularization compared to baselines. We further showed that it is also possible to directly predict the latent keypoints, for cases when the prediction of a large set of keypoints would be too computationally expensive.

Finally, we presented our work on the image-generation task of human reposing. Our main inspiration for this research was the success of volumetric heatmap prediction

in 3D human pose estimation. Hence, we formulated reposing as volumetric feature prediction, piecewise affine warping and feature decoding. While typical prior work operated using 2D image features (or parametric meshes), we showed that a volumetric 3D representation can be very effective for learning reposing. We used 3D representations in two ways in this work. First, we replaced 2D feature prediction and warping with volumetric 3D features and warping; second, we replaced the 2D heatmap representation of the target pose with 3D volumetric heatmaps. Both of these contributions were found effective over the corresponding baselines, and our full approach achieved state-of-the-art quantitative scores on the DeepFashion and iPER benchmarks.

Overall, with these approaches we have made significant contributions to improve the state of the art in 3D human pose estimation.

10.2 Perspectives

While improvements in visual human analysis have been impressive in recent years—to which we hope to have contributed with the approaches described above—many limitations and open challenges remain. Perceptual tasks, such as human pose estimation should be more tightly integrated with higher-level reasoning, "common sense" priors, world knowledge and physical constraints. Such top-down influences can help disambiguate poses, to reconstruct plausible actions and to forecast possible next actions. Long-term progress will probably require better modeling of various contextual cues, intentions and goals, object affordances and functionality, implicit social knowledge, as well as exploiting knowledge learned from other modalities such as language and audio. In the following, we discuss some more specific exciting future research possibilities and emerging directions.

Humans in Context. Most current approaches in human analysis focus on the human and consider everything else a distraction. However, in many cases hand-held objects, furniture and other scene components are crucial for interpreting what the person is doing or how they can be assisted by *e.g.*, a robot. Context cues can also help resolve ambiguities, *e.g.*, the scene layout can constrain the absolute poses. This needs to be flexible enough, without overly strong assumptions such as the presence of a single ground plane.

Joint scene–human reconstruction has already started to emerge as an important area of research, with promising initial approaches (*e.g.*, Huang *et al.*, 2022). A challenge for the coming years will be to generalize such approaches to arbitrary objects and environments.

Uncertainty and Probabilistic Modeling. Advances of the last decade have been largely fueled by strong representations learned via deep learning. However, while deep learning yields impressive results in aggregate and on average, it can also fail

unexpectedly, and automatically detecting when this happens is challenging. Modeling uncertainties will therefore be an important topic of future research. This includes out-of-distribution detection (*i.e.*, "I don't know" answers), uncertainty quantification, or full-scale modeling of the complex multi-modal predictive probability distribution. Research on the conformal calibration of neural networks is promising in this regard, and could also be applied to calibrating (volumetric) heatmaps in a conformal sense. Normalizing flows have recently been introduced for modeling complicated, multi-modal probability distributions, and some initial studies have applied them for 3D human pose estimation as well. We expect such approaches to become more common.

Complex Interactions. While multi-person pose estimation has made much progress, close human-to-human interactions (*e.g.*, dancing, hugging, wrestling) are still challenging to model due to occlusions and complex interdependencies. Methods are currently constrained by the lack of training and test data in this regard. Current large-scale multi-person 3D pose datasets, such as CMU-Panoptic (Joo *et al.*, 2019), contain little physical interaction. Action recognition datasets like NTU-RGB+D (Liu *et al.*, 2019b) or PKU-MMD (Liu *et al.*, 2017) do depict a few types of interactions like hugging or pushing, but the clips are relatively simple, with Kinect-based reference poses of limited accuracy. The recent CHI3D (Fieraru *et al.*, 2020) has more kinds of interactions, and could bring more focus to interaction modeling.

A second limitation of current interaction datasets is their scripted nature, consisting of short disconnected clips. Learning from longer, unscripted interactions could pave the way towards more socially aware AI.

Forecasting with a Theory of Mind. A key reason for performing human pose estimation in the first place is that poses contain cues about what the person is doing and intends to do. Current pose forecasting methods typically only work well on a time horizon of about one second, where extrapolation based on inertia is sufficient. On longer time horizons, it is crucial to model people as goal-oriented agents, *i.e.*, the forecasting system requires a theory of mind. Discovery of plausible goals in unstructured novel environments (*e.g.*, based on gestures, gaze and "common sense") will be important for helpful collaborative robots.

Perspectives in Data Collection. Many recent successes in AI have relied on ingesting Internet-scale uncurated data and performing some form of self- or unsupervised training. We have seen in Chapter 8 that data scale is also important for 3D human pose estimation quality. Further research along this path could make use of orders of magnitude more unlabeled data collected from *e.g.*, movies or YouTube, where rare poses and clothing can also be observed (*e.g.*, extreme sports, contortion, yoga), which are unlikely to be recorded in typical labeled datasets. We believe that a strong initialization trained on as much labeled data as possible, as we did in Chapter 8, can play a key role in bootstrapping such a self-supervised approach.

However, since current 3D pose estimation models are near saturation for common poses and appearances, collecting more such "easy" data is wasteful. This may

diminish the impact of "casting a wide net" as discussed above. Instead, the data collection process should focus specifically on covering poses and appearances where current models struggle. Smaller-scale but well-targeted data in this manner could help fill remaining blind spots of the models. The active learning paradigm can be explored for this.

With the rise of more and more realistic real-time rendering engines, synthetic data generation will also remain an important way to scale the available data.

Continual Learning. The strict separation into a training phase and an inference phase is artificial and very unlike how natural learning happens. Its necessity today is largely a consequence of the much higher computational cost of training updates compared to forward passes in inference. Ideally, future methods would allow robots to self-supervise themselves on post-deployment experiences, adapting their models to new environments, new people and actions over time. Recent research results on continual learning could be integrated into 3D human understanding to explore this aspect.

Task Fusions. Computer vision tasks are often tackled and evaluated separately, in isolation, although they need to be solved jointly in practical applications. However, multi-task learning has been a growing research topic as deep learning provides a common foundation for most vision tasks. Still, in many cases, innovations in one task (or research community) can take a long time to find adoption in others. For example, segmentation-based tracking has made quick progress recently, and the underlying ideas can be also applicable to 3D pose tracking.

We also observe that body mesh estimation methods and keypoint-based traditional 3D pose estimation have started to converge and more synergies could be exploited in this direction.

Furthermore, while we handled human pose estimation (Chapters 4–8) and reposing (Chapter 9) separately, there are further potential interactions to be exploited in the analysis-by-synthesis paradigm. Human pose estimation, reposing and motion modeling could all play important roles in self-supervisory cognitive loops (*cf*. Gong *et al.*, 2022). *E.g.*, when human pose estimation is uncertain, motion models could generate plausible continuations, which can in turn be rendered with a reposing model for comparison with the input image. When single-frame human pose estimation is certain, the motion model can be updated to learn about a potentially novel movement pattern, which can be recalled and used later when the image happens to be noisier, and so on. The recent quality improvements in image-generation methods (diffusion models, NeRFs) could further motivate such research directions.

Dynamic Level of Detail. The right level of human representation (*e.g.*, point, box, skeleton, body model, textured and clothed mesh *etc*.) is task-dependent. In the current paradigm, we need to choose it at the time of model design. However, for a goal-oriented intelligent agent with constraints on compute, this is not ideal. Ideally, the agent could pay more detailed and fine-grained attention to people (and objects)

who are nearby or relevant for other reasons. Meanwhile, more distant people, or people already walking away from the robot can probably be represented in less detail. This will become increasingly relevant if low-powered mobile robots become more widely adopted in dense crowds and urban environments.

Impact of Wider Trends. The development of the Vision Transformer has inspired researchers to tackle various vision tasks with Transformers as well, including human pose estimation. It is currently unclear whether ViTs have clear advantages over CNNs as feature-extractor backbones, but applying Transformer layers on the extracted features has been a fruitful line of recent research.

Most recently, following their sweeping success in image generation, diffusion models have started to see applications in human motion modeling, as well as human pose estimation, but it is not possible to say yet if this will result in more widespread adoption.

Another recent success story has been the merging vision models with language models (as in CLIP). We can therefore also expect more use of language in human-related vision tasks as well.

*
* *

Bibliography

van der Aa, N. P.; Luo, X.; Giezeman, G. J.; Tan, R. T.; and Veltkamp, R. C. (2011). Utrecht multi-person motion (UMPM) benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *IEEE International Conference on Computer Vision – Workshops (ICCVW)*. (→ 48, 124)

Agarwal, Ankur and Triggs, Bill (2006). Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **28**(1). (→ 11)

Aggarwal, Jake K. and Cai, Quin (1999). Human motion analysis: A review. *Computer Vision and Image Understanding (CVIU)*, **73**(3). (→ 11, 26)

Agrawal, Akshay; Amos, Brandon; Barratt, Shane; Boyd, Stephen; Diamond, Steven; and Kolter, J. Zico (2019). Differentiable convex pptimization layers. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 86)

Aho, Alfred V.; Lam, Monica S.; Sethi, Ravi; and Ullman, Jeffrey D. (1986). *Compilers: Principles, Techniques, and Tools*. Pearson Education, Inc. (→ 20)

Akhter, Ijaz and Black, Michael J (2015). Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 22)

Akita, Koichiro (1984). Image sequence analysis of real world human motion. *Pattern Recognition*, **17**(1), 73–83. (→ 10)

Alexiadis, Dimitrios S.; Chatzitofis, Anargyros; Zioulis, Nikolaos; Zoidi, Olga; Louizis, Georgios; Zarpalas, Dimitrios; and Daras, Petros (2017). An integrated platform for live 3D human reconstruction and motion capturing. *IEEE Transactions on Circuits and Systems for Video Technology*, **27**(4), 798–813. (→ 13, 49)

Alldieck, Thiemo; Magnor, Marcus; Xu, Weipeng; Theobalt, Christian; and Pons-Moll, Gerard (2018). Video based reconstruction of 3D people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 74)

Alpaydin, Ethem (2021). *Machine learning*. MIT Press, 4th edition. (→ 36)

Amos, Brandon and Kolter, J. Zico (2017). OptNet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning (ICML)*. (→ 86)

Anderson, James A. and Rosenfeld, Edward (2000). *Talking Nets: An Oral History of Neural Networks*. MIT Press. (→ 30)

Andriluka, Mykhaylo; Pishchulin, Leonid; Gehler, Peter; and Schiele, Bernt (2014). 2D human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 14, 51, 82, 86, 124)

Andriluka, Mykhaylo; Iqbal, Umar; Insafutdinov, Eldar; Pishchulin, Leonid; Milan, Anton; Gall, Juergen; and Schiele, Bernt (2018). PoseTrack: A benchmark for human pose estimation and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 51, 124)

Anguelov, Dragomir; Srinivasan, Praveen; Koller, Daphne; Thrun, Sebastian; Rodgers, Jim; and Davis, James (2005). SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, **24**(3). (→ 11, 17)

Ba, Jimmy Lei; Kiros, Jamie Ryan; and Hinton, Geoffrey E. (2016). Layer normalization. *arXiv:1607.06450*. (→ 34)

Baak, Andreas; Helten, Thomas; Müller, Meinard; Pons-Moll, Gerard; Rosenhahn, Bodo; and Seidel, Hans-Peter (2010). Analyzing and evaluating markerless motion tracking using inertial sensors. In *European Conference on Computer Vision – Workshops (ECCVW)*. (→ 49)

Badler, Norman I. (1975). *Temporal scene analysis: Conceptual descriptions of object movements*. Ph.D. thesis, University of Pennsylvania. (→ 10)

Badler, Norman I. and Smoliar, Stephen W. (1979). Digital representations of human movement. *ACM Computing Surveys (CSUR)*, **11**(1), 19–38. (→ 10)

Badrinarayanan, Vijay; Kendall, Alex; and Cipolla, Roberto (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **39**(12), 2481–2495. (→ 14)

Balakrishnan, Guha; Zhao, Amy; Dalca, Adrian V.; Durand, Fredo; and Guttag, John (2018). Synthesizing images of humans in unseen poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 152, 154, 159)

Balan, Alexandru O.; Sigal, Leonid; Black, Michael J.; Davis, James E.; and Haussecker, Horst W. (2007). Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 11, 17)

Baldi, Pierre (2021). *Deep Learning in Science*. Cambridge University Press. (→ 30, 36)

Barratt, Shane and Sharma, Rishi (2018). A note on the Inception Score. *arXiv:1801.01973*. (→ 158)

Barron, Carlos and Kakadiaris, Ioannis A (2000). Estimating anthropometry and pose from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 11, 18)

Bartol, Kristijan; Bojanić, David; Petković, Tomislav; D'Apuzzo, Nicola; and Pribanic, Tomislav (2020). A review of 3D human pose estimation from 2D images. In *International Conference and Exhibition on 3D Body Scanning and Processing Technologies (3DBODY.TECH)*. (→ 26)

Bartol, Kristijan; Bojanić, David; Petković, Tomislav; and Pribanić, Tomislav (2022). Generalizable human pose triangulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 24)

Bashirov, Renat; Ianina, Anastasia; Iskakov, Karim; Kononenko, Yevgeniy; Strizhkova, Valeriya; Lempitsky, Victor; and Vakhitov, Alexander (2021). Real-time RGBD-based extended body pose estimation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 16)

Bazavan, Eduard Gabriel; Zanfir, Andrei; Zanfir, Mihai; Freeman, William T.; Sukthankar, Rahul; and Sminchisescu, Cristian (2021). HSPACE: Synthetic parametric humans animated in complex environments. *arXiv:2112.12867*. (→ 50, 124)

Belagiannis, Vasileios; Wang, Xinchao; Shitrit, Horesh Beny Ben; Hashimoto, Kiyoshi; Stauder, Ralf; Aoki, Yoshimitsu; Kranzfelder, Michael; Schneider, Armin; Fua, Pascal; Ilic, Slobodan; Feussner, Hubertus; and Navab, Nassir (2016). Parsing human skeletons in an operating room. *Machine Vision and Applications*, **27**(7), 1035–1046. (→ 26, 74)

Ben-Shabat, Yizhak; Yu, Xin; Saleh, Fatemeh; Campbell, Dylan; Rodriguez-Opazo, Cristian; Li, Hongdong; and Gould, Stephen (2021). The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 48, 124)

Benzine, Abdallah; Luvison, Bertrand; Pham, Quoc Cuong; and Achard, Catherine (2021). Single-shot 3D multi-person pose estimation in complex images. *Pattern Recognition*, **112**, 107534. (→ 114, 115)

Bergman, Alexander W.; Kellnhofer, Petr; Wang, Yifan; Chan, Eric R.; Lindell, David B.; and Wetzstein, Gordon (2022). Generative neural articulated radiance fields. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 18)

Bhatnagar, Bharat Lal; Xie, Xianghui; Petrov, Ilya; Sminchisescu, Cristian; Theobalt, Christian; and Pons-Moll, Gerard (2022). BEHAVE: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 22, 49, 124)

Bieler, Didier; Günel, Semih; Fua, Pascal; and Rhodin, Helge (2019). Gravity as a reference for estimating a person's height from video. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 24, 109)

Bigalke, Alexander; Hansen, Lasse; Diesel, Jasper; and Heinrich, Mattias P. (2021). Domain adaptation through anatomical constraints for 3D human pose estimation under the cover. In *Medical Imaging with Deep Learning (MIDL)*. (→ 26)

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Springer. (→ 36)

Bochkovskiy, Alexey; Wang, Chien-Yao; and Liao, Hong-Yuan Mark (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv:2004.10934*. (→ 139)

Bogo, Federica; Kanazawa, Angjoo; Lassner, Christoph; Gehler, Peter; Romero, Javier; and Black, Michael J. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*. (→ 18, 22)

Boser, Bernhard E.; Guyon, Isabelle M.; and Vapnik, Vladimir N. (1992). A training algorithm for optimal margin classifiers. In *Annual Workshop on Computational Learning Theory (COLT)*. (→ 31)

Bouazizi, Arij; Wiederer, Julian; Kressel, Ulrich; and Belagiannis, Vasileios (2021). Self-supervised 3D human pose estimation with multiple-view geometry. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. (→ 22)

Bourlard, Hervé and Kabil, Selen Hande (2022). Autoencoders reloaded. *Biological Cybernetics*, **116**(4), 1–18. (→ 129)

Bourlard, Hervé and Kamp, Yves (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**(4), 291–294. (→ 129)

Bradski, Gary (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*. (→ 38)

Brand, Matthew (1999). Shadow puppetry. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 10)

Bray, Joseph (2001). Markerless based human motion capture: A survey. Technical report, Vision and VR Group, Dept. Systems Engineering, Brunel University, London, UK. (→ 26)

Bregler, Christoph and Malik, Jitendra (1998). Tracking people with twists and exponential maps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 10)

Brock, Andrew; De, Soham; and Smith, Samuel L (2021). Characterizing signal propagation to close the performance gap in unnormalized ResNets. In *International Conference on Learning Representations (ICLR)*. (→ 34)

Brubaker, Marcus A; Sigal, Leonid; and Fleet, David J (2009). Physics-based human motion modeling for people tracking: A short tutorial. (→ 24)

Burgos-Artizzu, Xavier P.; Perona, Pietro; and Dollár, Piotr (2013). Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 55)

Cai, Yujun; Ge, Liuhao; Liu, Jun; Cai, Jianfei; Cham, Tat-Jen; Yuan, Junsong; and Thalmann, Nadia Magnenat (2019). Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 88)

Cai, Zhongang; Zhang, Mingyuan; Ren, Jiawei; Wei, Chen; Ren, Daxuan; Lin, Zhengyu; Zhao, Haiyu; Yang, Lei; and Liu, Ziwei (2021). Playing for 3D human recovery. *arXiv:2110.07588*. (→ 50)

Cai, Zhongang; Ren, Daxuan; Zeng, Ailing; Lin, Zhengyu; Yu, Tao; Wang, Wenjia; Fan, Xiangyu; Gao, Yang; Yu, Yifan; Pan, Liang; *et al.* (2022). HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. In *European Conference on Computer Vision (ECCV)*. (→ 48)

Cao, Jiale; Pang, Yanwei; Xie, Jin; Khan, Fahad Shahbaz; and Shao, Ling (2021). From handcrafted to deep features for pedestrian detection: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **44**(9). (→ 20)

Cao, Zhe; Simon, Tomas; Wei, Shih-En; and Sheikh, Yaser (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 14, 21)

Cao, Zhe; Gao, Hang; Mangalam, Karttikeya; Cai, Qizhi; Vo, Minh; and Malik, Jitendra (2020). Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*. (→ 23, 50)

Carion, Nicolas; Massa, Francisco; Synnaeve, Gabriel; Usunier, Nicolas; Kirillov, Alexander; and Zagoruyko, Sergey (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*. (→ 25)

Carissimi, Nicolo; Rota, Paolo; Beyan, Cigdem; and Murino, Vittorio (2018). Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In *European Conference on Computer Vision – Workshops (ECCVW)*. (→ 23)

Carreira, Joao; Agrawal, Pulkit; Fragkiadaki, Katerina; and Malik, Jitendra (2016). Human pose estimation with iterative error feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 14)

Cauchy, Augustin-Louis (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, **25**, 536–538. (→ 35)

Chapelle, Olivier and Wu, Mingrui (2010). Gradient descent optimization of smoothed information retrieval metrics. *Information Retrieval*, **13**(3), 216–235. (→ 15)

Chatzitofis, Anargyros; Saroglou, Leonidas; Boutis, Prodromos; Drakoulis, Petros; Zioulis, Nikolaos; Subramanyam, Shishir; Kevelham, Bart; Charbonnier, Caecilia; Cesar, Pablo; Zarpalas, Dimitrios; *et al.* (2020). HUMAN4D: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, **8**, 176241–176262. (→ 48, 124)

Chen, Ching-Hang and Ramanan, Deva (2017). 3D human pose estimation=2D pose estimation + matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 19, 78)

Chen, Ching-Hang; Tyagi, Ambrish; Agrawal, Amit; Drover, Dylan; Mv, Rohith; Stojanov, Stefan; and Rehg, James M (2019a). Unsupervised 3D pose estimation with geometric self-supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Chen, Kenny; Gabriel, Paolo; Alasfour, Abdulwahab; Gong, Chenghao; Doyle, Werner K.; Devinsky, Orrin; Friedman, Daniel; Dugan, Patricia; Melloni, Lucia; Thesen, Thomas, David Gonda; Sattar, Shifteh; Wangs, Sonya; and Gilja, Vikash (2018). Patient-specific pose estimation in clinical environments. *IEEE Journal of Translational Engineering in Health and Medicine*, **6**, 1–11. (→ 26)

Chen, Liang-Chieh; Papandreou, George; Kokkinos, Iasonas; Murphy, Kevin; and Yuille, Alan L. (2017a). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **40**(4), 834–848. (→ 81)

Chen, Liang-Chieh; Papandreou, George; Schroff, Florian; and Adam, Hartwig (2017b). Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*. (→ 139)

Chen, Yucheng; Tian, Yingli; and He, Mingyi (2020). Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding (CVIU)*, **192**, 102897. (→ 26)

Chen, Zerui; Guo, Yiru; Huang, Yan; and Wang, Liang (2019b). Learning depth-aware heatmaps for 3D human pose estimation in the wild. In *British Machine Vision Conference (BMVC)*. (→ 79, 87, 88, 89)

Cheng, Yu; Yang, Bo; Wang, Bo; Yan, Wending; and Tan, Robby T (2019). Occlusion-aware networks for 3D human pose estimation in video. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 23)

Cheng, Yu; Yang, Bo; Wang, Bo; and Tan, Robby T. (2020). 3D human pose estimation using spatio-temporal networks with explicit occlusion training. In *AAAI Conference on Artificial Intelligence*. (→ 23)

Cheng, Yu; Wang, Bo; Yang, Bo; and Tan, Robby T. (2021). Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 123)

Cheng, Yu; Wang, Bo; and Tan, Robby (2023). Dual networks based 3D multi-person pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **45**(2), 1636–1651. (→ 143)

Choi, Jeongjun; Shim, Dongseok; and Kim, H Jin (2022). DiffuPose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. *arXiv:2212.02796*. (→ 25)

Chollet, François (2021). *Deep Learning with Python*. Manning, 2nd edition. (→ 13, 36)

Cireșan, Dan Claudiu; Meier, Ueli; Masci, Jonathan; Gambardella, Luca Maria; and Schmidhuber, Jürgen (2011). Flexible, high performance convolutional neural networks for image classification. In *International Joint Conference on Artificial Intelligence (IJCAI)*. (→ 32)

CMU (2003). Carnegie Mellon University Motion Capture Database. `mocap.cs.cmu.edu`. (→ 49)

Colyer, Steffi L.; Evans, Murray; Cosker, Darren P.; and Salo, Aki I. T. (2018). A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Medicine – Open*, **4**(1). (→ 26)

Cong, Peishan; Xu, Yiteng; Ren, Yiming; Zhang, Juze; Xu, Lan; Wang, Jingya; Yu, Jingyi; and Ma, Yuexin (2022). Weakly supervised 3D multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. *arXiv:2211.16951*. (→ 16, 26)

Cormier, Mickael; Clepe, Aris; Specker, Andreas; and Beyerer, Jürgen (2022). Where are we with human pose estimation in real-world surveillance? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 26)

Cornman, Hannah L.; Stenum, Jan; and Roemmich, Ryan T. (2021). Video-based quantification of human movement frequency using pose estimation: A pilot study. *PLOS One*, **16**(12), e0261450. (→ 26)

Corona, Enric; Pons-Moll, Gerard; Alenyà, Guillem; and Moreno-Noguer, Francesc (2022). Learned vertex descent: A new direction for 3D human model fitting. In *European Conference on Computer Vision (ECCV)*. (→ 18)

Cortes, Corinna and Vapnik, Vladimir (1995). Support-vector networks. *Machine learning*, **20**(3), 273–297. (→ 11, 31)

CROWDBOT (2021). Local sensing system. Technical Report D2.3, CROWDBOT EU H2020 Project. `https;//doi.org/10.3030/779942`. (→ 118)

Csurka, Gabriella; Dance, Christopher; Fan, Lixin; Willamowski, Jutta; and Bray, Cédric (2004). Visual categorization with bags of keypoints. In *European Conference on Computer Vision – Workshops (ECCVW)*. (→ 12)

Cunningham, John P. and Ghahramani, Zoubin (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research (JMLR)*, **16**(1), 2859–2900. (→ 129)

Dabral, Rishabh; Gundavarapu, Nitesh B.; Mitra, Rahul; Sharma, Abhishek; Ramakrishnan, Ganesh; and Jain, Arjun (2019). Multi-person 3D human pose estimation from monocular images. In *International Conference on 3D Vision (3DV)*. (→ 109, 114)

Dabral, Rishabh; Shimada, Soshi; Jain, Arjun; Theobalt, Christian; and Golyanik, Vladislav (2021). Gravity-aware monocular 3D human–object reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 22, 24)

Dalal, Navneet and Triggs, Bill (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 12)

Dang, Qi; Yin, Jianqin; Wang, Bin; and Zheng, Wenqing (2019). Deep learning based 2D human pose estimation: A survey. *Tsinghua Science and Technology*, **24**(6), 663–676. (→ 26)

Dantcheva, Antitza; Bremond, Francois; and Bilinski, Piotr (2018). Show me your face and I will tell you your height, weight and body mass index. In *International Conference on Pattern Recogntion (ICPR)*. (→ 79)

Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; and Fei-Fei, Li (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 13)

Desmarais, Yann; Mottet, Denis; Slangen, Pierre; and Montesinos, Philippe (2021). A review of 3D human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding (CVIU)*, **212**, 103275. (→ 26)

DeVries, Terrance and Taylor, Graham W. (2017). Improved regularization of convolutional neural networks with Cutout. *arXiv:1708.04552*. (→ 55, 58)

Dong, Haoye; Liang, Xiaodan; Gong, Ke; Lai, Hanjiang; Zhu, Jia; and Yin, Jian (2018). Soft-gated warping-GAN for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 152, 154, 159)

Dong, Zijian; Song, Jie; Chen, Xu; Guo, Chen; and Hilliges, Otmar (2021). Shape-aware multi-person pose estimation from multi-view images. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 24)

Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob; and Houlsby, Neil (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. (→ 25, 34)

Drover, Dylan; MV, Rohith; Chen, Ching-Hang; Agrawal, Amit; Tyagi, Ambrish; and Phuoc Huynh, Cong (2018). Can 3D pose be learned from 2D projections alone? In *European Conference on Computer Vision – Workshops (ECCVW)*. (→ 21)

Dubey, Shradha and Dixit, Manish (2022). A comprehensive survey on human pose estimation approaches. *Multimedia Systems*. (→ 26)

Dunn, Timothy W; Marshall, Jesse D; Severson, Kyle S; Aldarondo, Diego E; Hildebrand, David GC; Chettih, Selmaan N; Wang, William L; Gellis, Amanda J; Carlson, David E; Aronov, Dmitriy; *et al.* (2021). Geometric deep learning enables 3D kinematic profiling across species and environments. *Nature Methods*, **18**(5), 564–573. (→ 26)

Dvornik, Nikita; Mairal, Julien; and Schmid, Cordelia (2018). Modeling visual context is key to augmenting object detection datasets. In *European Conference on Computer Vision (ECCV)*. (→ 55)

Dwibedi, Debidatta; Misra, Ishan; and Hebert, Martial (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 55)

Ebadi, Salehe Erfanian; Jhang, You-Cyuan; Zook, Alex; Dhakad, Saurav; Crespi, Adam; Parisi, Pete; Borkman, Steven; Hogins, Jonathan; and Ganguly, Sujoy (2021). Peoplesanspeople: A synthetic data generator for human-centric computer vision. *arXiv:2112.09290*. (→ 50)

Ebadi, Salehe Erfanian; Dhakad, Saurav; Vishwakarma, Sanjay; Wang, Chunpu; Jhang, You-Cyuan; Chociej, Maciek; Crespi, Adam; Thaman, Alex; and Ganguly, Sujoy (2022). PSP-HDRI+: A synthetic dataset generator for pre-training of human-centric computer vision models. *arXiv:2207.05025*. (→ 50)

Elsken, Thomas; Metzen, Jan Hendrik; and Hutter, Frank (2019). Neural architecture search: A survey. *Journal of Machine Learning Research (JMLR)*, **20**(1), 1997–2017. (→ 33)

Esser, Patrick; Sutter, Ekaterina; and Ommer, Björn (2018). A variational U-Net for conditional appearance and shape generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 155, 159)

Everingham, Mark; Gool, Luc Van; Williams, Christopher K. I.; Winn, John; and Zisserman, Andrew (2012). The PASCAL visual object classes challenge 2012 (VOC2012) results. `http://host.robots.ox.ac.uk/pascal/VOC/voc2012/`. Accessed 2018-07-20. (→ 58, 68, 87)

Fabbri, Matteo; Lanzi, Fabio; Calderara, Simone; Palazzi, Andrea; Vezzani, Roberto; and Cucchiara, Rita (2018). Learning to detect and track visible and occluded body joints in a virtual world. In *European Conference on Computer Vision (ECCV)*. (→ 50, 124)

Fabbri, Matteo; Lanzi, Fabio; Calderara, Simone; Alletto, Stefano; and Cucchiara, Rita (2020). Compressed volumetric heatmaps for multi-person 3D pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 19, 123)

Fang*, Hao-Shu; Xu*, Yuanlu; Wang, Wenguan; Liu, Xiaobai; and Zhu, Song-Chun (2018). Learning pose grammar to encode human body configuration for 3D pose estimation. In *AAAI Conference on Artificial Intelligence*. (→ 88)

Fang, Qi; Shuai, Qing; Dong, Junting; Bao, Hujun; and Zhou, Xiaowei (2021). Reconstructing 3D human pose by watching humans in the mirror. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Felzenszwalb, Pedro F. and Huttenlocher, Daniel P. (2000). Efficient matching of pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 11)

Felzenszwalb, Pedro F. and Huttenlocher, Daniel P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, **61**(1). (→ 11)

Ferrari, Vittorio; Marín-Jiménez, Manuel; and Zisserman, Andrew (2009). 2D human pose estimation in tv shows. In *Statistical and Geometrical Approaches to Visual Motion Analysis*. (→ 50)

Fieraru, Mihai; Zanfir, Mihai; Oneata, Elisabeta; Popa, Alin-Ionut; Olaru, Vlad; and Sminchisescu, Cristian (2020). Three-dimensional reconstruction of human interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 48, 109, 124, 168)

Fieraru, Mihai; Zanfir, Mihai; Pirlea, Silviu-Cristian; Olaru, Vlad; and Sminchisescu, Cristian (2021a). AIFit: Automatic 3D human-interpretable feedback models for fitness training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 26, 48, 124)

Fieraru, Mihai; Zanfir, Mihai; Oneata, Elisabeta; Popa, Alin-Ionut; Olaru, Vlad; and Sminchisescu, Cristian (2021b). Learning complex 3D human self-contact. In *AAAI Conference on Artificial Intelligence*. (→ 48, 124)

Fieraru, Mihai; Zanfir, Mihai; Szente, Teodor; Bazavan, Eduard; Olaru, Vlad; and Sminchisescu, Cristian (2021c). REMIPS: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 22)

Fisch, Martin and Clark, Ronald (2021). Orientation keypoints for 6d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **44**(12), 10145–10158. (→ 17)

Fischler, Martin A. and Elschlager, Robert A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*. (→ 11)

Forsyth, David A.; Arikan, Okan; Ikemoto, Leslie; O'Brien, James; and Ramanan, Deva (2006). Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, **1**(2–3), 77–254. (→ 26)

Fukushima, Kunihiko (1969). Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, **5**(4), 322–333. (→ 31)

Fukushima, Kunihiko (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**, 193–202. (→ 30, 31)

Fukushima, Kunihiko and Miyake, Sei (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, **15**(6), 455–469. (→ 31)

Gavrila, Dariu M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU)*, **73**(1). (→ 11, 26)

Gavrila, D. M. and Davis, L. S. (1996). 3-D model-based tracking of humans in action: a multi-view approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 10)

Geng, Jiaqi; Huang, Dong; and De la Torre, Fernando (2022). DensePose from WiFi. *arXiv:2301.00250*. (→ 16)

Georgakis, Georgios; Mousavian, Arsalan; Berg, Alexander C.; and Kosecka, Jana (2017). Synthesizing training data for object detection in indoor scenes. In *Robotics: Science and Systems (RSS)*. (→ 55)

Ghiasi, Golnaz and Fowlkes, Charless C. (2014). Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 55)

Ghorbani, Saeed; Mahdaviani, Kimia; Thaler, Anne; Kording, Konrad; Cook, Douglas James; Blohm, Gunnar; and Troje, Nikolaus F. (2021). MoVi: A large multipurpose human motion and video dataset. *PLOS One*, **16**(6), e0253157. (→ 48, 124)

Girdhar, Rohit; Gkioxari, Georgia; Torresani, Lorenzo; Paluri, Manohar; and Tran, Du (2018). Detect-and-track: Efficient pose estimation in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23)

Glorot, Xavier and Bengio, Yoshua (2010). Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. (→ 32)

Gong, Kehong; Zhang, Jianfeng; and Feng, Jiashi (2021). PoseAug: A differentiable pose augmentation framework for 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 143)

Gong, Kehong; Li, Bingbing; Zhang, Jianfeng; Wang, Tao; Huang, Jing; Mi, Michael Bi; Feng, Jiashi; and Wang, Xinchao (2022). PoseTriplet: Co-evolving 3D human pose estimation, imitation, and hallucination under self-supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 169)

Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; and Bengio, Yoshua (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 152)

Goodfellow, Ian; Bengio, Yoshua; and Courville, Aaron (2016). *Deep Learning*. MIT Press. (→ 30, 31, 36)

Göransson, Rasmus; Aydemir, Alper; and Jensfelt, Patric (2014). Kinect@Home: A crowdsourced RGB-D dataset. In *International Conference on Intelligent Autonomous Systems*. (→ 13)

Goyal, Anirudh and Bengio, Yoshua (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, **478**(2266), 20210068. (→ 30)

Grenander, Ulf (1978). *Lectures in Pattern Theory: Volume 2: Pattern Analysis*. Springer Science Business Media. (→ 36)

Grigorev, Artur; Sevastopolsky, Artem; Vakhitov, Alexander; and Lempitsky, Victor (2019). Coordinate-based texture inpainting for pose-guided human image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 152, 154, 159)

Gu, Kerui; Yang, Linlin; and Yao, Angela (2021). Removing the bias of integral pose regression. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 19)

Gu, Kerui; Yang, Linlin; and Yao, Angela (2022). Dive deeper into integral pose regression. In *International Conference on Learning Representations*. (→ 19)

Guerra-Urzola, Rosember; Van Deun, Katrijn; Vera, Juan C.; and Sijtsma, Klaas (2021). A guide for sparse pca: model comparison and applications. *Psychometrika*, **86**(4), 893–919. (→ 129)

Güler, Rıza Alp; Neverova, Natalia; and Kokkinos, Iasonas (2018). DensePose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18, 154)

Günel, Semih; Rhodin, Helge; and Fua, Pascal (2019). What face and body shapes can tell about height. In *IEEE International Conference on Computer Vision – Workshops (ICCVW)*. (→ 79)

Guo, Hengkai; Tang, Tang; Luo, Guozhong; Chen, Riwei; Lu, Yongchen; and Wen, Linfu (2018). Multi-domain pose network for multi-person pose estimation and tracking. In *European Conference on Computer Vision – Workshops (ECCVW)*. (→ 128)

Guo, Wen; Corona, Enric; Moreno-Noguer, Francesc; and Alameda-Pineda, Xavier (2021). PI-Met: Pose interacting network for multi-person monocular 3D pose estimation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 21)

Habibie, Ikhsanul; Xu, Weipeng; Mehta, Dushyant; Pons-Moll, Gerard; and Theobalt, Christian (2019). In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 89)

Han, Jungong; Shao, Ling; Xu, Dong; and Shotton, Jamie (2013). Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Transactions on Cybernetics*, **43**(5), 1318–1334. (→ 13)

Hanson, Stephen and Pratt, Lorien (1988). Comparing biases for minimal network construction with back-propagation. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 35)

Hartley, Richard and Zisserman, Andrew (2004). *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, UK. (→ 24, 37)

Hassan, Mohamed; Choutas, Vasileios; Tzionas, Dimitrios; and Black, Michael J (2019). Resolving 3D human pose ambiguities with 3D scene constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 22)

Hastie, Trevor; Tibshirani, Robert; and Friedman, Jerome (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd edition. (→ 36)

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; and Sun, Jian (2016a). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 32, 55, 157)

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; and Sun, Jian (2016b). Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*. (→ 32, 66, 81, 157)

He, Kaiming; Gkioxari, Georgia; Dollár, Piotr; and Girshick, Ross (2017). Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 14, 20)

Hedlin, Eric; Rhodin, Helge; and Yi, Kwang Moo (2022). A simple method to boost human pose estimation accuracy by correcting the joint regressor for the Human3.6M dataset. In *Conference on Computer and Robot Vision*. (→ 128)

Henderson, Leah (2022). The problem of induction. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2022 edition. (→ 30)

Henning, Dorian F; Laidlow, Tristan; and Leutenegger, Stefan (2022). BodySLAM: joint camera localisation, mapping, and human motion tracking. In *European Conference on Computer Vision (ECCV)*. (→ 19)

Hentout, Abdelfetah; Aouache, Mustapha; Maoudj, Abderraouf; and Akli, Isma (2019). Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017. *Advanced Robotics*, **33**(15-16), 764–799. (→ 26)

Hernandez, Alejandro; Gall, Jurgen; and Moreno-Noguer, Francesc (2019). Human motion prediction via spatio-temporal inpainting. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 24)

Hirschorn, Or and Avidan, Shai (2022). Normalizing flows for human pose anomaly detection. *arXiv:2211.10946*. (→ 25, 26)

Ho, Jonathan; Jain, Ajay; and Abbeel, Pieter (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 25)

Hoffer, Elad; Hubara, Itay; and Soudry, Daniel (2017). Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 35, 137)

Hogg, David (1983). Model-based vision: a program to see a walking person. *Image and Vision Computing*, **1**(1), 5–20. (→ 10)

Holden, Daniel; Saito, Jun; and Komura, Taku (2016). A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, **35**(4), 1–11. (→ 128)

Holmquist, Karl and Wandt, Bastian (2022). DiffPose: Multi-hypothesis human pose estimation using diffusion models. *arXiv:2211.16487*. (→ 25)

Horiuchi, Yusuke; Iizuka, Satoshi; Simo-Serra, Edgar; and Ishikawa, Hiroshi (2019). Spectral normalization and relativistic adversarial training for conditional pose generation with self-attention. In *Machine Vision and Applications*. (→ 152, 153)

Hormann, Kai and Sukumar, N., editors (2017). *Generalized barycentric coordinates in computer graphics and computational mechanics*. CRC Press. (→ 131)

Hossain, Mir Rayat Imtiaz and Little, James J. (2018). Exploiting temporal information for 3D human pose estimation. In *European Conference on Computer Vision (ECCV)*. (→ 23)

Howard, Andrew; Sandler, Mark; Chu, Grace; Chen, Liang-Chieh; Chen, Bo; Tan, Mingxing; Wang, Weijun; Zhu, Yukun; Pang, Ruoming; Vasudevan, Vijay; Le, Quoc V.; and Adam, Hartwig (2019). Searching for mobilenetv3. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 33, 118)

Howard, Andrew G.; Zhu, Menglong; Chen, Bo; Kalenichenko, Dmitry; Wang, Weijun; Weyand, Tobias; Andreetto, Marco; and Adam, Hartwig (2017). MobilenNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*. (→ 33)

Hu, Jie; Shen, Li; and Sun, Gang (2018). Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 33, 34)

Hu, Ming-Kuei (1962). Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, **8**(2). (→ 11)

Hu, Yuan-Ting; Chen, Hong-Shuo; Hui, Kexin; Huang, Jia-Bin; and Schwing, Alexander G. (2019). SAIL-VOS: Semantic amodal instance level video object segmentation – a synthetic dataset and baselines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 50, 117, 124)

Hua, Gang; Yang, Ming-Hsuan; and Wu, Ying (2005). Learning to estimate human pose with data driven belief propagation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 12)

Huang, Chun-Hao P.; Yi, Hongwei; Höschle, Markus; Safroshkin, Matvey; Alexiadis, Tsvetelina; Polikovsky, Senya; Scharstein, Daniel; and Black, Michael J. (2022). Capturing and inferring dense full-body human–scene contact. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 49, 124, 167)

Huang, Gao; Li, Yixuan; Pleiss, Geoff; Liu, Zhuang; Hopcroft, John E.; and Weinberger, Kilian Q. (2017). Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations (ICLR)*. (→ 69)

Huang, Jia-Bin and Yang, Ming-Hsuan (2009). Estimating human pose from occluded images. In *Asian Conference on Computer Vision (ACCV)*. (→ 55)

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, **160**(1), 106–154. (→ 30)

Huber, Eric (1996). 3-D real-time gesture recognition using proximity spaces. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 10)

Hutter, Frank; Kotthoff, Lars; and Vanschoren, Joaquin (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature. (→ 33)

Insafutdinov, Eldar; Pishchulin, Leonid; Andres, Bjoern; Andriluka, Mykhaylo; and Schiele, Bernt (2016). DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision (ECCV)*. (→ 14)

Ioffe, Sergey (2017). Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 34)

Ioffe, Sergey and Szegedy, Christian (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*. (→ 32, 34, 157)

Ionescu, Catalin; Li, Fuxin; and Sminchisescu, Cristian (2011). Latent structured models for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 45)

Ionescu, Catalin; Papava, Dragos; Olaru, Vlad; and Sminchisescu, Cristian (2014). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **36**(7), 1325–1339. (→ 12, 14, 42, 45, 54, 56, 65, 79, 86, 124, 139)

Ionescu, Catalin; Papava, Dragos; Olaru, Vlad; and Sminchisescu, Cristian (2018). Results of the 2018 eccv posetrack challenge on 3D human pose estimation. `http://vision.imar.ro/human3.6m/ranking.php`. Accessed 2022-02-23. (→ 67)

Iqbal, Umar; Garbade, Martin; and Gall, Juergen (2017). Pose for action – action for pose. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. (→ 26)

Iqbal, Umar; Molchanov, Pavlo; Breuel, Thomas; Gall, Juergen; and Kautz, Jan (2018). Hand pose estimation via latent 2.5D heatmap regression. In *European Conference on Computer Vision (ECCV)*. (→ 75)

Iqbal, Umar; Molchanov, Pavlo; and Kautz, Jan (2020). Weakly-supervised 3D human pose learning via multi-view images in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Iqbal, Umar; Xie, Kevin; Guo, Yunrong; Kautz, Jan; and Molchanov, Pavlo (2021). KAMA: 3D keypoint aware body mesh articulation. In *International Conference on 3D Vision (3DV)*. (→ 18)

Iskakov, Karim; Burkov, Egor; Lempitsky, Victor; and Malkov, Yury (2019). Learnable triangulation of human pose. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 24, 49)

Islam, Md Amirul; Jia, Sen; and Bruce, Neil DB (2020). How much position information do convolutional neural networks encode? In *International Conference on Learning Representations (ICLR)*. (→ 83)

Isola, Phillip; Zhu, Jun-Yan; Zhou, Tinghui; and Efros, Alexei A. (2017). Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 152)

Jahangiri, Ehsan and Yuille, Alan L (2017). Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections. In *IEEE International Conference on Computer Vision – Workshops (ICCVW)*. (→ 25)

Jain, Arjun; Tompson, Jonathan; Andriluka, Mykhaylo; Taylor, Graham W.; and Bregler, Christoph (2014). Learning human pose estimation features with convolutional networks. In *International Conference on Learning Representations (ICLR)*. (→ 13)

Jakab, Tomas; Gupta, Ankush; Bilen, Hakan; and Vedaldi, Andrea (2018). Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 129)

Jakab, Tomas; Gupta, Ankush; Bilen, Hakan; and Vedaldi, Andrea (2020). Self-supervised learning of interpretable keypoints from unlabelled videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 129)

Jakab, Tomas; Tucker, Richard; Makadia, Ameesh; Wu, Jiajun; Snavely, Noah; and Kanazawa, Angjoo (2021). KeypointDeformer: Unsupervised 3D keypoint discovery for shape control. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 129)

Jang, Eric; Gu, Shixiang; and Poole, Ben (2017). Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR)*. (→ 19)

Jegou, Herve; Douze, Matthijs; and Schmid, Cordelia (2008). Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision (ECCV)*. (→ 87)

Ji, Xiaopeng; Fang, Qi; Dong, Junting; Shuai, Qing; Jiang, Wen; and Zhou, Xiaowei (2020). A survey on monocular 3D human pose estimation. *Virtual Reality & Intelligent Hardware (VRIH)*, **2**(6), 471–500. (→ 26)

Jiang, Hao and Grauman, Kristen (2017). Seeing invisible poses: Estimating 3D body pose from egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 16)

Jiang, Wen; Kolotouros, Nikos; Pavlakos, Georgios; Zhou, Xiaowei; and Daniilidis, Kostas (2020). Coherent reconstruction of multiple humans from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 19, 22, 109)

Jin, Lei; Xu, Chenyang; Wang, Xiaojuan; Xiao, Yabo; Guo, Yandong; Nie, Xuecheng; and Zhao, Jian (2022). Single-stage is enough: Multi-person absolute 3D pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Johansson, Gunnar (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, **14**(2), 201–211. (→ 10)

Johnson, Justin; Alahi, Alexandre; and Fei-Fei, Li (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*. (→ 32, 157, 158)

Johnson, Kerri and Shiffrar, Maggie (2012). *People watching: Social, perceptual, and neurophysiological studies of body perception*. Oxford University Press. (→ 27)

Johnson, Sam and Everingham, Mark (2010). Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*. (→ 50, 124)

Johnson, Sam and Everingham, Mark (2011). Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 50)

Joo, Hanbyul; Simon, Tomas; and Sheikh, Yaser (2018). Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 17)

Joo, Hanbyul; Simon, Tomas; Li, Xulong; Liu, Hao; Tan, Lei; Gui, Lin; Banerjee, Sean; Godisart, Timothy; Nabbe, Bart C.; Matthews, Iain A.; Kanade, Takeo; Nobuhara, Shohei; and Sheikh, Yaser (2019). Panoptic Studio: A massively multiview system

for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **41**(1), 190–204. (→ 13, 47, 117, 124, 168)

Joo, Hanbyul; Neverova, Natalia; and Vedaldi, Andrea (2021). Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*. (→ 22, 49, 123)

Kanazawa, Angjoo; Black, Michael J.; Jacobs, David W.; and Malik, Jitendra (2018). End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18)

Kar, Abhishek; Tulsiani, Shubham; Carreira, Joao; and Malik, Jitendra (2015). Amodal completion and size constancy in natural scenes. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 79, 110)

Kaufmann, Manuel; Aksan, Emre; Song, Jie; Pece, Fabrizio; Ziegler, Remo; and Hilliges, Otmar (2020). Convolutional autoencoders for human motion infilling. In *International Conference on 3D Vision (3DV)*. (→ 24)

Ke, Lipeng; Chang, Ming-Ching; Qi, Honggang; and Lyu, Siwei (2018). Multi-scale structure-aware network for human pose estimation. In *European Conference on Computer Vision (ECCV)*. (→ 55, 74)

Kearney, Sinead; Li, Wenbin; Parsons, Martin; Kim, Kwang In; and Cosker, Darren (2020). Rgbd-dog: Predicting canine pose from rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 26)

Khan, Salman; Naseer, Muzammal; Hayat, Munawar; Zamir, Syed Waqas; Khan, Fahad Shahbaz; and Shah, Mubarak (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, **54**(10s), 1–41. (→ 25)

Khirodkar, Rawal; Chari, Visesh; Agrawal, Amit; and Tyagi, Ambrish (2021). Multi-instance pose networks: Rethinking top-down pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 21)

Khirodkar, Rawal; Tripathi, Shashank; and Kitani, Kris (2022). Occluded human mesh recovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23)

Kinchla, Ronald A. and Wolfe, Jeremy M. (1979). The order of visual processing: "top-down," "bottom-up," or "middle-out". *Perception & Psychophysics*, **25**(3), 225–231. (→ 20)

Kingma, Diederik P. and Ba, Jimmy (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. (→ 35, 68, 133, 157)

Kissos, Imry; Fritz, Lior; Goldman, Matan; Meir, Omer; Oks, Eduard; and Kliger, Mark (2020). Beyond weak perspective for monocular 3D human pose estimation. In *European Conference on Computer Vision – Workshops (ECCVW)*. (→ 39, 113)

Klette, Reinhard and Tee, Garry (2008). Understanding human motion: A historic review. In B. Rosenhahn, R. Klette, and D. Metaxas, editors, *Human Motion: Understanding, Modelling, Capture, and Animation*, pages 1–22. Springer. (→ 10)

Knoche, Markus; Sárándi, István; and Leibe, Bastian (2020). Reposing humans by warping 3D features. In *IEEE Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW)*. (→ 7, 8, 151, 159)

Kobyzev, Ivan; Prince, Simon J. D.; and Brubaker, Marcus A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **43**(11), 3964–3979. (→ 25)

Kocabas, Muhammed; Karagoz, Salih; and Akbas, Emre (2019). Self-supervised learning of 3D human pose using multi-view geometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Kocabas, Muhammed; Athanasiou, Nikos; and Black, Michael J. (2020). VIBE: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23, 123)

Kocabas, Muhammed; Huang, Chun-Hao P.; Hilliges, Otmar; and Black, Michael J. (2021a). PARE: Part attention regressor for 3D human body estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 22)

Kocabas, Muhammed; Huang, Chun-Hao P.; Tesch, Joachim; Müller, Lea; Hilliges, Otmar; and Black, Michael J. (2021b). SPEC: Seeing people in the wild with an estimated camera. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 39, 50, 124)

Kolotouros, Nikos; Pavlakos, Georgios; Jayaraman, Dinesh; and Daniilidis, Kostas (2021). Probabilistic modeling for human mesh recovery. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 25)

Kreiss, Sven; Bertoni, Lorenzo; and Alahi, Alexandre (2019). PifPaf: Composite fields for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Krizhevsky, Alex; Sutskever, Ilya; and Hinton, Geoffrey E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 13, 31, 35)

Kumar, Pranjal; Chauhan, Siddhartha; and Awasthi, Lalit Kumar (2022). Human pose estimation using deep learning: review, methodologies, progress and future research directions. *International Journal of Multimedia Information Retrieval (IJMIR)*, pages 1–33. (→ 26)

Kundu, Jogendra Nath; Buckchash, Himanshu; Mandikal, Priyanka; Jamkhandi, Anirudh; and Radhakrishnan, Venkatesh Babu (2020). Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 23)

Kundu, Jogendra Nath; Seth, Siddharth; YM, Pradyumna; Jampani, Varun; Chakraborty, Anirban; and Babu, R. Venkatesh (2022). Uncertainty-aware adaptation for self-supervised 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 123)

van Laarhoven, Twan (2017). L2 regularization versus batch and weight normalization. *arXiv:1706.05350*. (→ 137)

Laban, Rudolf (1928). Grundprinzipien der Bewegungsschrift. *Schrifttanz*, **1**(1). (→ 10)

Lakhal, Mohamed Ilyes; Lanz, Oswald; and Cavallaro, Andrea (2018). Pose guided human image synthesis by view disentanglement and enhanced weighting loss. In *European Conference on Computer Vision (ECCV)*. (→ 153)

Lan, Gongjin; Wu, Yu; Hu, Fei; and Hao, Qi (2022). Vision-based human pose estimation via deep learning: A survey. *IEEE Transactions on Human-Machine Systems*. (→ 26)

Lassner, Christoph; Romero, Javier; Kiefel, Martin; Bogo, Federica; Black, Michael J; and Gehler, Peter V. (2017). Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 49)

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, **1**(4), 541–551. (→ 31)

LeCun, Yann; Bottou, Léon; Bengio, Yoshua; and Haffner, Patrick (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324. (→ 31)

LeCun, Yann; Kavukcuoglu, Koray; and Farabet, Clément (2010). Convolutional networks and applications in vision. In *IEEE International Symposium on Circuits and Systems*. (→ 31)

LeCun, Yann; Bengio, Yoshua; and Hinton, Geoffrey (2015). Deep learning. *Nature*, **521**(7553), 436–444. (→ 13)

Lee, Hsi-Jian and Chen, Zen (1985). Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing (CVGIP)*, **30**(2). (→ 10, 18)

Levine, Sergey; Finn, Chelsea; Darrell, Trevor; and Abbeel, Pieter (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research (JMLR)*, **17**(39), 1–40. (→ 15, 55, 67, 77, 78)

Li, Chen and Lee, Gim Hee (2019). Generating multiple hypotheses for 3D human pose estimation with mixture density network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 25)

Li, Jiefeng; Bian, Siyuan; Zeng, Ailing; Wang, Can; Pang, Bo; Liu, Wentao; and Lu, Cewu (2021a). Human pose regression with residual log-likelihood estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 25)

Li, Jiefeng; Xu, Chao; Chen, Zhicun; Bian, Siyuan; Yang, Lixin; and Lu, Cewu (2021b). HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18)

Li, Jiefeng; Chen, Tong; Shi, Ruiqi; Lou, Yujing; Li, Yong-Lu; and Lu, Cewu (2021c). Localization with sampling-argmax. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 19)

Li, Jianwei; Hu, Haing; Li, Jinyang; and Zhao, Xiaomei (2022a). 3d-yoga: A 3D yoga dataset for hierarchical sports action analysis. In *Asian Conference on Computer Vision (ACCV)*. (→ 49)

Li, Jialian; Zhang, Jingyi; Wang, Zhiyong; Shen, Siqi; Wen, Chenglu; Ma, Yuexin; Xu, Lan; Yu, Jingyi; and Wang, Cheng (2022b). LiDARCap: Long-range marker-less 3D human motion capture with LiDAR point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 16, 25)

Li, Ke; Wang, Shijie; Zhang, Xiang; Xu, Yifan; Xu, Weijian; and Tu, Zhuowen (2021d). Pose recognition with cascade transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 25)

Li, Ruilong; Yang, Shan; Ross, David A.; and Kanazawa, Angjoo (2021e). AI choreographer: Music conditioned 3D dance generation with AIST++. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 48, 124)

Li, Sijin and Chan, Antoni B. (2014). 3D human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*. (→ 14)

Li, Sijin; Zhang, Weichen; and Chan, Antoni B. (2015). Maximum-margin structured learning with deep networks for 3D human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 14)

Li, Yining; Huang, Chen; and Loy, Chen Change (2019a). Dense intrinsic appearance flow for human pose transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 154, 159)

Li, Yang; Li, Kan; Jiang, Shuai; Zhang, Ziyue; Huang, Congzhentao; and Da Xu, Richard Yi (2020). Geometry-driven self-supervised method for 3D human pose estimation. In *AAAI Conference on Artificial Intelligence*. (→ 22)

Li, Yanjie; Zhang, Shoukui; Wang, Zhicheng; Yang, Sen; Yang, Wankou; Xia, Shu-Tao; and Zhou, Erjin (2021f). TokenPose: Learning keypoint tokens for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 25)

Li, Zongmian; Sedlar, Jiri; Carpentier, Justin; Laptev, Ivan; Mansard, Nicolas; and Sivic, Josef (2019b). Estimating 3D motion and forces of person-object interactions from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 24)

Li, Zhengqi; Dekel, Tali; Cole, Forrester; Tucker, Richard; Snavely, Noah; Liu, Ce; and Freeman, William T (2019c). Learning the depths of moving people by watching frozen people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 22)

Liang, Yangwen; Ranade, Rohit; Wang, Shuangquan; Bai, Dongwoon; Lee, Jungwon; Valtonen Ornhag, Marcus; Olsson, Carl; Heyden, Anders; Gao, Zhongpai; Zhang, Juyong; *et al.* (2020). The "vertigo effect" on your smartphone: Dolly zoom via single shot view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW)*. (→ 110)

Lim, Hyunchul; Li, Yaxuan; Dressa, Matthew; Hu, Fang; Kim, Jae Hoon; Zhang, Ruidong; and Zhang, Cheng (2022). BodyTrak: Inferring full-body poses from body silhouettes using a miniature camera on a wristband. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, **6**(3), 1–21. (→ 16)

Lin, Kevin; Wang, Lijuan; and Liu, Zicheng (2021a). End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18, 19, 25, 73, 123)

Lin, Kevin; Wang, Lijuan; and Liu, Zicheng (2021b). Mesh graphormer. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 25, 143)

Lin, Tsung-Yi; Maire, Michael; Belongie, Serge; Hays, James; Perona, Pietro; Ramanan, Deva; Dollár, Piotr; and Zitnick, C. Lawrence (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*. (→ 14, 51, 124)

Liu, Chunhui; Hu, Yueyu; Li, Yanghao; Song, Sijie; and Liu, Jiaying (2017). PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In *ACM Multimedia Conference (MM) – Workshop on Visual Analysis in Smart and Connected Communities (VSCC)*. (→ 13, 49, 168)

Liu, Chenchen; Li, Yongzhi; Ma, Kangqi; Zhang, Duo; Bao, Peijun; and Mu, Yadong (2021). Learning 3-D human pose estimation from catadioptric videos. In *International Joint Conference on Artificial Intelligence (IJCAI)*. (→ 21)

Liu, Ding; Zhao, Zixu; Wang, Xinchao; Hu, Yuxiao; Zhang, Lei; and Huang, Thomas (2019a). Improving 3D human pose estimation via 3D part affinity fields. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 21, 88)

Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; and Catanzaro, Bryan (2018). Image inpainting for irregular holes using partial convolutions. In *European Conference on Computer Vision (ECCV)*. (→ 156)

Liu, Jun; Shahroudy, Amir; Perez, Mauricio; Wang, Gang; Duan, Ling-Yu; and Kot, Alex C. (2019b). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **42**(10), 2684–2701. (→ 26, 49, 168)

Liu, Qihao; Zhang, Yi; Bai, Song; and Yuille, Alan (2022a). Explicit occlusion reasoning for multi-person 3D human pose estimation. In *European Conference on Computer Vision*. (→ 23)

Liu, Ruixu; Shen, Ju; Wang, He; Chen, Chen; Cheung, Sen-ching; and Asari, Vijayan (2020). Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 123)

Liu, Wu and Mei, Tao (2022). Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. *ACM Computing Surveys (CSUR)*, **55**(4). (→ 26)

Liu, Wen; Piao, Zhixin; Min, Jie; Luo, Wenhan; Ma, Lin; and Gao, Shenghua (2019c). Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 152, 154, 157, 159, 160, 161, 162, 163, 164)

Liu, Ziwei; Luo, Ping; Qiu, Shi; Wang, Xiaogang; and Tang, Xiaoou (2016). DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 157)

Liu, Zhuang; Mao, Hanzi; Wu, Chao-Yuan; Feichtenhofer, Christoph; Darrell, Trevor; and Xie, Saining (2022b). A ConvNet for the 2020s. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 34)

Long, Jonathan; Shelhamer, Evan; and Darrell, Trevor (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 13)

Loper, Matthew; Mahmood, Naureen; Romero, Javier; Pons-Moll, Gerard; and Black, Michael J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia)*, **34**(6), 248:1–248:16. (→ 14, 17, 49, 50, 124, 128)

Loper, Matthew M.; Mahmood, Naureen; and Black, Michael J. (2014). MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia)*, **33**(6), 220:1–220:13. (→ 129)

Loshchilov, Ilya and Hutter, Frank (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. (→ 36, 83, 135)

Lowe, David G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, **60**(2). (→ 12)

Luo, Chenxu; Chu, Xiao; and Yuille, Alan (2018a). OriNet: A fully convolutional network for 3D human pose estimation. In *British Machine Vision Conference (BMVC)*. (→ 75, 90)

Luo, Chenxu; Chu, Xiao; and Yuille, Alan (2018b). OriNet-demo. `https://github.com/chenxuluo/OriNet-demo`. Accessed 2018-11-16. (→ 90)

Luo, Ping; Wang, Xinjiang; Shao, Wenqi; and Peng, Zhanglin (2019). Towards understanding regularization in batch normalization. In *International Conference on Learning Representations (ICLR)*. (→ 34)

Luo, Yiyue; Li, Yunzhu; Foshey, Michael; Shou, Wan; Sharma, Pratyusha; Palacios, Tomás; Torralba, Antonio; and Matusik, Wojciech (2021). Intelligent carpet: Inferring 3D human pose from tactile signals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 16)

Luvizon, Diogo; Picard, David; and Tabia, Hedi (2020). Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **43**(8), 2752–2764. (→ 26, 88)

Luvizon, Diogo C.; Picard, David; and Tabia, Hedi (2018). 2D/3D pose estimation and action recognition using multitask deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 15, 19, 26, 71, 75, 77, 82, 88, 100, 154)

Luvizon, Diogo C.; Tabia, Hedi; and Picard, David (2019). Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, **85**, 15–22. (→ 77, 78)

Ma, Liqian; Jia, Xu; Sun, Qianru; Schiele, Bernt; Tuytelaars, Tinne; and Van Gool, Luc (2017). Pose guided person image generation. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 152, 159)

Ma, Liqian; Sun, Qianru; Georgoulis, Stamatios; Van Gool, Luc; Schiele, Bernt; and Fritz, Mario (2018). Disentangled person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 155, 159)

Ma, Yi; Soatto, Stefano; Košecká, Jana; and Sastry, S. Shankar (2004). *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer. (→ 37)

Magnenat-Thalmann, Nadia and Thalmann, Daniel, editors (2004). *Handbook of virtual humans*. John Wiley & Sons. (→ 26)

Mahmood, Naureen; Ghorbani, Nima; Troje, Nikolaus F.; Pons-Moll, Gerard; and Black, Michael J. (2019). AMASS: Archive of motion capture as surface shapes. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 128)

Mandery, Christian; Terlemez, Ömer; Do, Martin; Vahrenkamp, Nikolaus; and Asfour, Tamim (2016). Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, **32**(4), 796–809. (→ 128)

Manesco, João Renato Ribeiro and Marana, Aparecido Nilceu (2022). A survey of recent advances on two-step 3D human pose estimation. In *Brazilian Conference on Intelligent Systems (BRACIS)*. (→ 26)

Mao, Weian; Ge, Yongtao; Shen, Chunhua; Tian, Zhi; Wang, Xinlong; Wang, Zhibin; and den Hengel, Anton van (2022). Poseur: Direct human pose regression with transformers. In *European Conference on Computer Vision (ECCV)*. (→ 25)

von Marcard, Timo; Henschel, Roberto; Black, Michael J.; Rosenhahn, Bodo; and Pons-Moll, Gerard (2018). Recovering accurate 3D human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*. (→ 16, 48, 108, 117, 139)

Marr, David (1976). Early processing of visual information. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **275**(942), 483–519. (→ 10)

Marr, David and Nishihara, Herbert Keith (1976). Representation and recognition of the spatial organization of three dimensional shapes. Technical Report 377, Massachusetts Institute of Technology, Artificial Intelligence Laboratory. (→ 10, 17)

Marr, David and Nishihara, Herbert Keith (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **200**(1140), 269–294. (→ 10)

Martin-Martin, Roberto; Patel, Mihir; Rezatofighi, Hamid; Shenoi, Abhijeet; Gwak, JunYoung; Frankel, Eric; Sadeghian, Amir; and Savarese, Silvio (2021). JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Early access. (→ 51, 124)

Martinez, Julieta; Black, Michael J.; and Romero, Javier (2017a). On human motion prediction using recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23)

Martinez, Julieta; Hossain, Rayat; Romero, Javier; and Little, James J. (2017b). A simple yet effective baseline for 3D human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 15, 19, 71, 74, 78, 88, 89)

Matan, Ofer; Burges, Christopher J.; LeCun, Yann; and Denker, John (1991). Multi-digit recognition using a space displacement neural network. *Advances in Neural Information Processing Systems (NIPS)*, **4**. (→ 13)

McCulloch, Warren S and Pitts, Walter (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, **5**(4), 115–133. (→ 30)

McGinty, Stephen (2013). Grand theft auto v: Scottish game conquering world. *The Scotsman*. (→ 50)

McLaughlin, Niall and Martinez del Rincon, Jesus (2018). Refining the pose: Training and use of deep recurrent autoencoders for improving human pose estimation. In *International Conference on Articulated Motion and Deformable Objects (AMDO)*. (→ 24)

Mehta, Dushyant; Rhodin, Helge; Casas, Dan; Fua, Pascal; Sotnychenko, Oleksandr; Xu, Weipeng; and Theobalt, Christian (2017a). Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3D Vision (3DV)*. (→ 14, 42, 47, 86, 87, 90, 109, 110, 124, 139, 158)

Mehta, Dushyant; Sridhar, Srinath; Sotnychenko, Oleksandr; Rhodin, Helge; Shafiei, Mohammad; Seidel, Hans-Peter; Xu, Weipeng; Casas, Dan; and Theobalt, Christian (2017b). VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics (TOG)*, **36**(4), 44. (→ 15, 22, 57, 74, 78, 90, 91)

Mehta, Dushyant; Sotnychenko, Oleksandr; Mueller, Franziska; Xu, Weipeng; Sridhar, Srinath; Pons-Moll, Gerard; and Theobalt, Christian (2018). Single-shot multi-person 3D pose estimation from monocular RGB. In *International Conference on 3D Vision (3DV)*. (→ 15, 21, 22, 48, 90, 111, 112, 114, 124, 139)

Mehta, Dushyant; Sotnychenko, Oleksandr; Mueller, Franziska; Xu, Weipeng; Elgharib, Mohamed; Fua, Pascal; Seidel, Hans-Peter; Rhodin, Helge; Pons-Moll, Gerard; and Theobalt, Christian (2019). XNect: Real-time multi-person 3D human pose estimation with a single RGB camera. *arXiv:1907.00837*. (→ 109, 114)

Mihajlovic, Marko; Saito, Shunsuke; Bansal, Aayush; Zollhoefer, Michael; and Tang, Siyu (2022). COAP: Compositional articulated occupancy of people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18)

Mildenhall, Ben; Srinivasan, Pratul P; Tancik, Matthew; Barron, Jonathan T; Ramamoorthi, Ravi; and Ng, Ren (2021). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM (CACM)*, **65**(1), 99–106. (→ 18)

Minsky, Marvin and Papert, Seymour (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press. (→ 30)

Mitchell, Tom (1997). *Machine Learning*. McGraw Hill. (→ 29, 36)

Moeslund, Thomas B. and Granum, Erik (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding (CVIU)*, **81**(3). (→ 26)

Moeslund, Thomas B.; Hilton, Adrian; and Krüger, Volker (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, **104**(2-3), 90–126. (→ 26)

Moeslund, Thomas B.; Hilton, Adrian; Krüger, Volker; and Sigal, Leonid, editors (2011). *Visual analysis of humans: Looking at people*. Springer, London, UK. (→ 26)

Moon, Gyeongsik and Lee, Kyoung Mu (2020). I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*. (→ 18)

Moon, Gyeongsik; Chang, Ju Yong; and Lee, Kyoung Mu (2019). Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 19, 109, 114, 115, 123)

Moreno-Noguer, Francesc (2017). 3D human pose estimation from a single image via distance matrix regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 19)

Mori, Greg and Malik, Jitendra (2002). Estimating human body configurations using shape context matching. In *European Conference on Computer Vision (ECCV)*. (→ 11)

Munea, Tewodros Legesse; Jembre, Yalew Zelalem; Weldegebriel, Halefom Tekle; Chen, Longbiao; Huang, Chenxi; and Yang, Chenhui (2020). The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access*, **8**, 133330–133348. (→ 26)

Murphy, Kevin P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press. (→ 36)

Nair, Vinod and Hinton, Geoffrey E. (2010). Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*. (→ 31)

Neverova, Natalia; Güler, Rıza Alp; and Kokkinos, Iasonas (2018). Dense pose transfer. In *European Conference on Computer Vision (ECCV)*. (→ 74, 152, 154, 159)

Newell, Alejandro; Yang, Kaiyu; and Deng, Jia (2016). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*. (→ 14, 54, 74, 78, 85)

Newell, Alejandro; Huang, Zhiao; and Deng, Jia (2017). Associative embedding: End-to-end learning for joint detection and grouping. *Advances in Neural Information Processing Systems (NIPS)*. (→ 14, 21)

Nguyen-Phuoc, Thu; Li, Chuan; Theis, Lucas; Richardt, Christian; and Yang, Yong-Liang (2019). HoloGAN: Unsupervised learning of 3D representations from natural images. In *IEEE International Conference on Computer Vision – Workshops (ICCVW)*. (→ 152, 154)

Ni, Bingbing; Wang, Gang; and Moulin, Pierre (2011). RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *IEEE International Conference on Computer Vision – Workshops (ICCVW)*. (→ 13)

Nibali, Aiden; He, Zhen; Morgan, Stuart; and Prendergast, Luke (2018). Numerical coordinate regression with convolutional neural networks. *arXiv:1801.07372*. (→ 15, 19, 55, 67, 77, 78, 98)

Nibali, Aiden; He, Zhen; Morgan, Stuart; and Prendergast, Luke (2019). 3D human pose estimation with 2D marginal heatmaps. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 54, 75, 78, 79, 88, 89, 90, 100)

Nibali, Aiden; Millward, Joshua; He, Zhen; and Morgan, Stuart (2021). ASPset: An outdoor sports pose video dataset with 3D keypoint annotations. *Image and Vision Computing*, **111**, 104196. (→ 49, 124)

Nie, Bruce Xiaohan; Wei, Ping; and Zhu, Song-Chun (2017). Monocular 3D human pose estimation by predicting depth on joints. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 89)

Nieto, Jesús R. and Susín, Antonio (2013). *Deformation Models: Tracking, Animation and Applications*, chapter Cage Based Deformations: A Survey", pages 75–99. Springer Netherlands. (→ 131)

Nilsson, Nils J. (2009). *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge University Press, Cambridge, UK. (→ 9)

Ning, Huazhong; Xu, Wei; Gong, Yihong; and Huang, Thomas (2008). Discriminative learning of visual words for 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. (→ 12)

Noh, Hyeonwoo; Hong, Seunghoon; and Han, Bohyung (2015). Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 14)

Ofli, Ferda; Chaudhry, Rizwan; Kurillo, Gregorij; Vidal, René; and Bajcsy, Ruzena (2013). Berkeley MHAD: A comprehensive multimodal human action database. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 13, 48, 124)

Okawa, Yoshikuni and Hanatani, Shingo (1992). Recognition of human body motions by robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. (→ 10)

Olazaran, Mikel (1996). A sociological study of the official history of the perceptrons controversy. *Social Studies of Science*, **26**(3), 611–659. (→ 30)

O'Rourke, Joseph and Badler, Norman I. (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **2**(6). (→ 10, 49)

Osman, Ahmed A. A.; Bolkart, Timo; and Black, Michael J. (2020). STAR: Sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*. (→ 17)

Osman, Ahmed A A; Bolkart, Timo; Tzionas, Dimitrios; and Black, Michael J. (2022). SUPR: A sparse unified part-based human body model. In *European Conference on Computer Vision (ECCV)*. (→ 17)

Papandreou, George; Zhu, Tyler; Chen, Liang-Chieh; Gidaris, Spyros; Tompson, Jonathan; and Murphy, Kevin (2018). PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *European Conference on Computer Vision (ECCV)*. (→ 21)

Park, Soonchan; Lee, Sang-baek; and Park, Jinah (2020). Data augmentation method for improving the accuracy of human pose estimation with cropped images. *Pattern Recognition Letters*, **136**, 244–250. (→ 79)

Patel, Priyanka; Huang, Chun-Hao P.; Tesch, Joachim; Hoffmann, David T.; Tripathi, Shashank; and Black, Michael J. (2021). AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 50, 124)

Pavlakos, Georgios; Zhou, Xiaowei; Derpanis, Konstantinos G.; and Daniilidis, Kostas (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 15, 16, 54, 55, 59, 60, 63, 67, 71, 75, 76, 77, 78, 79, 80, 81, 85, 86, 88, 89, 108, 152, 154, 155)

Pavlakos, Georgios; Zhou, Xiaowei; and Daniilidis, Kostas (2018). Ordinal depth supervision for 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21, 71, 88)

Pavlakos, Georgios; Choutas, Vasileios; Ghorbani, Nima; Bolkart, Timo; Osman, Ahmed A. A.; Tzionas, Dimitrios; and Black, Michael J. (2019). Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 17, 22, 50, 124)

Pavllo, Dario; Feichtenhofer, Christoph; Grangier, David; and Auli, Michael (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 19, 24)

Pearson, Karl (1901). On lines and planes of closest fit to systems of points in space. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**(11), 559–572. (→ 129)

Peelen, Marius V. and Downing, Paul E. (2007). The neural basis of visual body perception. *Nature Reviews Neuroscience*, **8**(8), 636–648. (→ 1)

Perez-Sala, Xavier; Escalera, Sergio; Angulo, Cecilio; and Gonzalez, Jordi (2014). A survey on model based approaches for 2D and 3D visual human pose recovery. *Sensors*, **14**(3), 4189–4210. (→ 26)

Pfeiffer, Kilian; Hermans, Alexander; Sárándi, István; Weber, Mark; and Leibe, Bastian (2019). Visual person understanding through multi-task and multi-dataset learning. In *DAGM German Conference on Pattern Recognition (GCPR)*. (→ 34)

Pirinen, Aleksis; Gärtner, Erik; and Sminchisescu, Cristian (2019). Domes to drones: Self-supervised active triangulation for 3D human pose reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, **32**. (→ 24)

Pishchulin, Leonid; Insafutdinov, Eldar; Tang, Siyu; Andres, Bjoern; Andriluka, Mykhaylo; Gehler, Peter V.; and Schiele, Bernt (2016). DeepCut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 14)

Polyak, Boris T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, **4**(5), 1–17. (→ 35)

Pons-Moll, Gerard; Baak, Andreas; Helten, Thomas; Müller, Meinard; Seidel, Hans-Peter; and Rosenhahn, Bodo (2010). Multisensor-fusion for 3D full-body human motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 16, 49)

Pons-Moll, Gerard; Fleet, David J; and Rosenhahn, Bodo (2014). Posebits for monocular human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Poppe, Ronald (2007). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding (CVIU)*, **108**(1-2), 4–18. (→ 26)

Prince, Simon J.D. (2022). *Understanding Deep Learning*. MIT Press. Draft version. (→ 36)

Pumarola, Albert; Agudo, Antonio; Sanfeliu, Alberto; and Moreno-Noguer, Francesc (2018). Unsupervised person image synthesis in arbitrary poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 155, 159)

Pumarola, Albert; Sanchez, Jordi; Choi, Gary; Sanfeliu, Alberto; and Moreno-Noguer, Francesc (2019). 3DPeople: Modeling the geometry of dressed humans. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 50, 124)

Qammaz, Ammar and Argyros, Antonis (2021). Occlusion-tolerant and personalized 3D human pose estimation in RGB images. In *International Conference on Pattern Recogntion (ICPR)*. (→ 23)

Rafi, Umer; Gall, Juergen; and Leibe, Bastian (2015). A semantic occlusion model for human pose estimation from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW)*. (→ 55)

Rajasegaran, Jathushan; Pavlakos, Georgios; Kanazawa, Angjoo; and Malik, Jitendra (2022). Tracking people by predicting 3D appearance, location and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 20, 24)

Ramanan, Deva and Forsyth, David A. (2003). Finding and tracking people from the bottom up. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 12)

Rapczyński, Michał; Werner, Philipp; Handrich, Sebastian; and Al-Hamadi, Ayoub (2021). A baseline for cross-database 3D human pose estimation. *Sensors*, **21**(11), 3769. (→ 125, 127)

Rashid, Richard F. (1980). Towards a system for the interpretation of moving light displays. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **2**(6), 574–581. (→ 10)

Reddy, N Dinesh; Guigues, Laurent; Pishchulin, Leonid; Eledath, Jayan; and Narasimhan, Srinivasa G. (2021). TesseTrack: End-to-end learnable multi-person articulated 3D pose tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23)

Redmon, Joseph and Farhadi, Ali (2018). YOLOv3: An incremental improvement. *arXiv:1804.02767*. (→ 66, 111)

Remelli, Edoardo; Han, Shangchen; Honari, Sina; Fua, Pascal; and Wang, Robert (2020). Lightweight multi-view 3D pose estimation through camera-disentangled representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 24)

Rempe, Davis; Guibas, Leonidas J.; Hertzmann, Aaron; Russell, Bryan; Villegas, Ruben; and Yang, Jimei (2020). Contact and human dynamics from monocular video. In *European Conference on Computer Vision (ECCV)*. (→ 24)

Ren, Pengzhen; Xiao, Yun; Chang, Xiaojun; Huang, Po-yao; Li, Zhihui; Chen, Xiaojiang; and Wang, Xin (2022). A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, **54**(4). (→ 33)

Ren, Shaoqing; He, Kaiming; Girshick, Ross; and Sun, Jian (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 15)

Rhodin, Helge; Spörri, Jörg; Katircioglu, Isinsu; Constantin, Victor; Meyer, Frédéric; Müller, Erich; Salzmann, Mathieu; and Fua, Pascal (2018a). Learning monocular 3D human pose estimation from multi-view images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 22, 82)

Rhodin, Helge; Salzmann, Mathieu; and Fua, Pascal (2018b). Unsupervised geometry-aware representation for 3D human pose estimation. In *European Conference on Computer Vision (ECCV)*. (→ 22, 129, 154)

Roberts, Lawrence G (1963). *Machine perception of three-dimensional solids*. Ph.D. thesis, Massachusetts Institute of Technology. (→ 10)

Robinson, Nicole; Tidd, Brendan; Campbell, Dylan; Kulić, Dana; and Corke, Peter (2022). Robotic vision for human–robot interaction and collaboration: A survey and systematic review. *ACM Transactions on Human-Robot Interaction*. (→ 26)

Roburin, Simon; de Mont-Marin, Yann; Bursuc, Andrei; Marlet, Renaud; Pérez, Patrick; and Aubry, Mathieu (2022). Spherical perspective on learning with normalization layers. *Neurocomputing*, **487**(C), 66–74. (→ 137)

Rogez, Gregory; Weinzaepfel, Philippe; and Schmid, Cordelia (2017). LCR-Net: Localization-classification-regression for human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 15, 79, 90, 112, 114)

Rogez, Gregory; Weinzaepfel, Philippe; and Schmid, Cordelia (2019). LCR-Net++: Multi-person 2D and 3D pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **42**(5), 1146–1161. (→ 15, 114)

Rohr, Karl (1994). Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, **59**(1), 94–115. (→ 10)

Rombach, Robin; Blattmann, Andreas; Lorenz, Dominik; Esser, Patrick; and Ommer, Björn (2022). High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 32)

Ronneberger, Olaf; Fischer, Philipp; and Brox, Thomas (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. (→ 14, 153)

Rosales, Rómer and Sclaroff, Stan (2000). Inferring body pose without tracking body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 11)

Rosenblatt, Frank (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**(6), 386. (→ 30)

Rosenblatt, Frank (1962). *Principles of neurodynamics perceptrons and the theory of brain mechanisms*. Spartan Books, Washington, DC. (→ 11, 30)

Rosenhahn, Bodo; Klette, Reinhard; and Metaxas, Dimitris, editors (2008). *Human Motion: Understanding, Modelling, Capture, and Animation*. Springer, Dordrecht, the Netherlands. (→ 26)

Roy, Soumava Kumar; Citraro, Leonardo; Honari, Sina; and Fua, Pascal (2022). On triangulation as a form of self-supervision for 3D human pose estimation. *arXiv:2203.15865*. (→ 22)

Ruder, Sebastian (2016). An overview of gradient descent optimization algorithms. *arXiv:1609.04747*. (→ 35)

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. (1986). *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press. (→ 31)

Russakovsky, Olga; Deng, Jia; Su, Hao; Krause, Jonathan; Satheesh, Sanjeev; Ma, Sean; Huang, Zhiheng; Karpathy, Andrej; Khosla, Aditya; Bernstein, Michael; Berg, Alexander C.; and Fei-Fei, Li (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, **115**(3), 211–252. (→ 13, 31)

Saito, Shunsuke; Huang, Zeng; Natsume, Ryota; Morishima, Shigeo; Kanazawa, Angjoo; and Li, Hao (2019). PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 18)

Saito, Shunsuke; Simon, Tomas; Saragih, Jason; and Joo, Hanbyul (2020). PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18)

Salimans, Tim; Goodfellow, Ian; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec; and Chen, Xi (2016). Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 32, 158)

Sambasivan, Nithya; Kapania, Shivani; Highfill, Hannah; Akrong, Diana; Paritosh, Praveen Kumar; and Aroyo, Lora Mois (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *ACM Conference on Human Factors in Computing Systems (CHI)*. (→ 12)

Sampieri, Alessio; di Melendugno, Guido Maria D'Amely; Avogaro, Andrea; Cunico, Federico; Setti, Francesco; Skenderi, Geri; Cristani, Marco; and Galasso, Fabio (2022). Pose forecasting in industrial human–robot collaboration. In *European Conference on Computer Vision (ECCV)*. (→ 26)

Sanakoyeu, Artsiom; Khalidov, Vasil; McCarthy, Maureen S; Vedaldi, Andrea; and Neverova, Natalia (2020). Transferring dense pose to proximal animal classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 26)

Sandler, Mark; Howard, Andrew; Zhu, Menglong; Zhmoginov, Andrey; and Chen, Liang-Chieh (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 33)

Santurkar, Shibani; Tsipras, Dimitris; Ilyas, Andrew; and Madry, Aleksander (2018). How does batch normalization help optimization? In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 34)

Sapp, Benjamin and Taskar, Ben (2013). MODEC: Multimodal decomposable models for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 50)

Sarafianos, Nikolaos; Boteanu, Bogdan; Ionescu, Bogdan; and Kakadiaris, Ioannis A. (2016). 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, **152**, 1–20. (→ 12, 26)

Sárándi, István; Linder, Timm; Arras, Kai O.; and Leibe, Bastian (2018a). How robust is 3D human pose estimation to occlusion? In *IEEE/RSJ International Conference on Intelligent Robots and Systems – Workshops (IROSW)*. (→ 6, 8, 39, 53)

Sárándi, István; Linder, Timm; Arras, Kai O.; and Leibe, Bastian (2018b). Synthetic occlusion augmentation with volumetric heatmaps for the 2018 ECCV PoseTrack Challenge on 3D human pose estimation. *arXiv:1809.04987*. (→ 6, 8, 65, 109)

Sárándi, István; Linder, Timm; Arras, Kai O.; and Leibe, Bastian (2020). Metric-scale truncation-robust heatmaps for 3D human pose estimation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. (→ 6, 8, 74, 91, 154)

Sárándi, István; Linder, Timm; Arras, Kai O.; and Leibe, Bastian (2021). MeTRAbs: Metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, **3**(1), 16–30. (→ 7, 8, 107)

Sárándi, István; Hermans, Alexander; and Leibe, Bastian (2023). Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 7, 8, 123)

Schmidhuber, Jürgen (2015). Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117. (→ 30)

Schölkopf, Bernhard and Smola, Alexander J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, USA. (→ 11, 31)

Schönemann, Peter H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, **31**(1), 1–10. (→ 41)

Schops, Thomas; Larsson, Viktor; Pollefeys, Marc; and Sattler, Torsten (2020). Why having 10,000 parameters in your camera model is better than twelve. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 38)

Seethapathi, Nidhi; Wang, Shaofei; Saluja, Rachit; Blohm, Gunnar; and Kording, Konrad P (2019). Movement science needs different pose tracking algorithms. *arXiv:1907.10226*. (→ 26)

Sermanet, Pierre; Eigen, David; Zhang, Xiang; Mathieu, Michaël; Fergus, Rob; and LeCun, Yann (2013). OverFeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR)*. (→ 13)

Shahroudy, Amir; Liu, Jun; Ng, Tian-Tsong; and Wang, Gang (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 13)

Shakhnarovich, Gregory; Viola, Paul; and Darrell, Trevor (2003). Fast pose estimation with parameter-sensitive hashing. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 11)

Shalev-Shwartz, Shai and Ben-David, Shai (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. (→ 36)

Shao, Jie; Hu, Kai; Wang, Changhu; Xue, Xiangyang; and Raj, Bhiksha (2020). Is normalization indispensable for training deep neural network? In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 34)

Sharma, Saurabh; Varigonda, Pavan Teja; Bindal, Prashast; Sharma, Abhishek; and Jain, Arjun (2019). Monocular 3D human pose estimation by generation and ordinal ranking. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 88)

Shepard, Roger N. and Metzler, Jacqueline (1971). Mental rotation of three-dimensional objects. *Science*, **171**(3972), 701–703. (→ 2)

Shetty, Karthik; Birkhold, Annette; Jaganathan, Srikrishna; Strobel, Norbert; Kowarschik, Markus; Maier, Andreas; and Egger, Bernhard (2022). PLIKS: A pseudo-linear inverse kinematic solver for 3D human body estimation. *arXiv:2211.11734*. (→ 18)

Shimada, Soshi; Golyanik, Vladislav; Xu, Weipeng; and Theobalt, Christian (2020). PhysCap: Physically plausible monocular 3D motion capture in real time. *ACM Transactions on Graphics (TOG)*, **39**(6), 1–16. (→ 24)

Shotton, Jamie; Girshick, Ross; Fitzgibbon, Andrew; Sharp, Toby; Cook, Mat; Finocchio, Mark; Moore, Richard; Kohli, Pushmeet; Criminisi, Antonio; Kipman, Alex; *et al.* (2013). Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **35**(12), 2821–2840. (→ 13)

Si, Chenyang; Wang, Wei; Wang, Liang; and Tan, Tieniu (2018). Multistage adversarial losses for pose-based human image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 153)

Siarohin, Aliaksandr; Sangineto, Enver; Lathuilière, Stéphane; and Sebe, Nicu (2018). Deformable GANs for pose-based human image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 152, 153, 159, 160)

Sifre, Laurent (2014). *Rigid-motion scattering for image classification*. Ph.D. thesis, École Polytechnique. (→ 33)

Sigal, Leonid and Black, Michael J. (2006a). HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity Technical Report*, **120**(2). (→ 12)

Sigal, Leonid and Black, Michael J. (2006b). Predicting 3D people from 2D pictures. In *International Conference on Articulated Motion and Deformable Objects (AMDO)*. (→ 11, 17)

Sigal, Leonid; Isard, Michael; Sigelman, Benjamin; and Black, Michael J. (2003). Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 12, 17)

Sigal, Leonid; Bhatia, Sidharth; Roth, Stefan; Black, Michael J.; and Isard, Michael (2004). Tracking loose-limbed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 12)

Sigal, Leonid; Balan, Alexandru O.; and Black, Michael J. (2010). HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, **87**(1), 4–27. (→ 12, 47)

Silberman, Nathan and Guadarrama, Sergio (2016). TensorFlow-Slim image classification model library. `https://github.com/tensorflow/models/tree/master/research/slim`. Accessed 2018-07-20. (→ 58, 68)

Simo-Serra, Edgar; Ramisa, Arnau; Alenya, Guillem; Torras, Carme; and Moreno-Noguer, Francesc (2012). Single image 3D human pose estimation from noisy observations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 12)

Simo-Serra, Edgar; Quattoni, Ariadna; Torras, Carme; and Moreno-Noguer, Francesc (2013). A joint model for 2D and 3D pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 12)

Simonyan, Karen and Zisserman, Andrew (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. (→ 32, 157)

Sitzmann, Vincent; Thies, Justus; Heide, Felix; Nießner, Matthias; Wetzstein, Gordon; and Zollhofer, Michael (2019). DeepVoxels: Learning persistent 3D feature embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 154)

Sivaprasad, Prabhu Teja; Mai, Florian; Vogels, Thijs; Jaggi, Martin; and Fleuret, François (2020). Optimizer benchmarking needs to account for hyperparameter tuning. In *International Conference on Machine Learning (ICML)*. (→ 36)

Sminchisescu, Cristian; Kanaujia, Atul; and Metaxas, Dimitris (2006). Learning joint top-down and bottom-up processes for 3D visual inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 20)

Smith, Leslie N. (2017). Cyclical learning rates for training neural networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 69)

Sobel, Irwin and Feldman, Gary (1968). A 3x3 isotropic gradient operator for image processing. Talk at the Stanford Artificial Project. (→ 10)

Sohl-Dickstein, Jascha; Weiss, Eric; Maheswaranathan, Niru; and Ganguli, Surya (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*. (→ 25)

Song, Jie; Chen, Xu; and Hilliges, Otmar (2020). Human body model fitting by learned gradient descent. In *European Conference on Computer Vision (ECCV)*. (→ 18)

Srivastav, Vinkle; Issenhuth, Thibaut; Kadkhodamohammadi, Abdolrahim; de Mathelin, Michel; Gangi, Afshin; and Padoy, Nicolas (2018). MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation. In *International Conference on Medical Image Computing and Computer Assisted Intervention – Workshops (MICCAIW)*. (→ 74)

Stewart, Lawrence; Bach, Francis; Berthet, Quentin; and Vert, Jean-Philippe (2022). Regression as classification: Influence of task formulation on neural network features. *arXiv:2211.05641*. (→ 15)

Summers, Cecilia and Dinneen, Michael J. (2020). Four things everyone should know to improve batch normalization. In *International Conference on Learning Representations (ICLR)*. (→ 35, 137)

Sun, Xiao; Shang, Jiaxiang; Liang, Shuang; and Wei, Yichen (2017). Compositional human pose regression. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 15, 22, 63, 71, 88, 89)

Sun, Xiao; Xiao, Bin; Liang, Shuang; and Wei, Yichen (2018a). Integral human pose regression. In *European Conference on Computer Vision (ECCV)*. (→ 15, 19, 54, 55, 60, 63, 67, 71, 75, 77, 78, 79, 80, 81, 82, 83, 85, 87, 88, 89, 154)

Sun, Xiao; Xiao, Bin; Liang, Shuang; and Wei, Yichen (2018b). Integral human pose regression (code repository). `https://github.com/JimmySuen/integral-human-pose`. Accessed 2018-04-28. (→ 79, 87)

Sun, Xiao; Li, Chuankang; and Lin, Stephen (2018c). An integral pose regression system for the ECCV2018 PoseTrack Challenge. *arXiv:1809.06079*. (→ 76, 79)

Sun, Xiao; Li, Chuankang; and Lin, Stephen (2019). Explicit spatiotemporal joint relation learning for tracking human pose. In *IEEE International Conference on Computer Vision – Workshops (ICCVW)*. (→ 23)

Sun, Yu; Bao, Qian; Liu, Wu; Fu, Yili; and Mei, Tao (2020). CenterHMR: a bottom-up single-shot method for multi-person 3D mesh recovery from a single image. *arXiv:2008.12272v1*. (→ 113)

Sun, Yu; Bao, Qian; Liu, Wu; Fu, Yili; Michael J., Black; and Mei, Tao (2021). Monocular, one-stage, regression of multiple 3D people. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 21, 123, 143)

Sutton, Richard (2019). The bitter lesson. `http://incompleteideas.net/IncIdeas/BitterLesson.html`. Accessed: 2022-11-25. (→ 29)

Szegedy, Christian; Liu, Wei; Jia, Yangqing; Sermanet, Pierre; Reed, Scott; Anguelov, Dragomir; Erhan, Dumitru; Vanhoucke, Vincent; and Rabinovich, Andrew (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 32)

Szegedy, Christian; Vanhoucke, Vincent; Ioffe, Sergey; Shlens, Jon; and Wojna, Zbigniew (2016). Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 158)

Tan, Mingxing and Le, Quoc (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*. (→ 33)

Tan, Mingxing and Le, Quoc (2021). EfficientNetV2: Smaller models and faster training. In *International Conference on Machine Learning (ICML)*. (→ 33, 118, 135)

Taylor, Camillo J. (2000). Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 11, 18)

Tekin, Bugra; Rozantsev, Artem; Lepetit, Vincent; and Fua, Pascal (2016). Direct prediction of 3D body poses from motion compensated sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23, 63, 71)

Tekin, Bugra; Márquez-Neila, Pablo; Salzmann, Mathieu; and Fua, Pascal (2017). Fusing 2D uncertainty and 3D cues for monocular body pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 15)

Terzopoulos, D. and Metaxas, D. (1991). Dynamic 3D models with local and global deformations: deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **13**(7). (→ 10)

Tian, Yating; Zhang, Hongwen; Liu, Yebin; and Wang, Limin (2022). Recovering 3D human mesh from monocular images: A survey. *arXiv:2203.01923*. (→ 26)

Tieleman, Tijmen; Hinton, Geoffrey; Srivastava, Nitish; and Swersky, Kevin (2012). Lecture 6.5 – rmsprop: normalize the gradient. Part of the course "Neural networks for machine learning". (→ 35)

Tiwari, Garvita; Antic, Dimitrije; Lenssen, Jan Eric; Sarafianos, Nikolaos; Tung, Tony; and Pons-Moll, Gerard (2022). Pose-NDF: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*. (→ 22)

Tome, Denis; Russell, Chris; and Agapito, Lourdes (2017). Lifting from the deep: Convolutional 3D pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 19, 78)

Tome, Denis; Alldieck, Thiemo; Peluse, Patrick; Pons-Moll, Gerard; Agapito, Lourdes; Badino, Hernan; and De la Torre, Fernando (2020). SelfPose: 3D egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Early access. (→ 16)

Tompson, Jonathan; Jain, Arjun; LeCun, Yann; and Bregler, Christoph (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 13)

Toshev, Alexander and Szegedy, Christian (2014). DeepPose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 13, 78, 79)

Toshpulatov, Mukhiddin; Lee, Wookey; Lee, Suan; and Haghighian Roudsari, Arousha (2022). Human pose, hand and mesh estimation using deep learning: a survey. *The Journal of Supercomputing*, **78**(6), 7616–7654. (→ 26)

Trumble, Matt; Gilbert, Andrew; Malleson, Charles; Hilton, Adrian; and Collomosse, John (2017). Total capture: 3D human pose estimation fusing video and inertial sensors. In *British Machine Vision Conference (BMVC)*. (→ 48, 124)

Trumble, Matthew; Gilbert, Andrew; Hilton, Adrian; and Collomosse, John (2018). Deep autoencoder for combined human pose estimation and body model upscaling. In *European Conference on Computer Vision (ECCV)*. (→ 18)

Tsochantaridis, Ioannis; Joachims, Thorsten; Hofmann, Thomas; Altun, Yasemin; and Singer, Yoram (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, **6**(9). (→ 14)

Tsuchida, Shuhei; Fukayama, Satoru; Hamasaki, Masahiro; and Goto, Masataka (2019). AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *International Society for Music Information Retrieval Conference (ISMIR)*. (→ 48, 124)

Vapnik, Vladimir N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. (→ 31)

Varol, Gul; Romero, Javier; Martin, Xavier; Mahmood, Naureen; Black, Michael J.; Laptev, Ivan; and Schmid, Cordelia (2017). Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 49, 117, 124)

Varol, Gül; Ceylan, Duygu; Russell, Bryan; Yang, Jimei; Yumer, Ersin; Laptev, Ivan; and Schmid, Cordelia (2018). BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*. (→ 18)

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; and Polosukhin, Illia (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 25)

Véges, Márton and Lőrincz, András (2019). Absolute human pose estimation with depth prediction network. In *International Joint Conference on Neural Networks (IJCNN)*. (→ 19, 79, 109, 110, 111, 112, 113)

Véges, Márton and Lőrincz, András (2020a). Multi-person absolute 3D human pose estimation with weak depth supervision. In *International Conference on Artificial Neural Networks (ICANN)*. (→ 19, 112, 113, 114, 115)

Véges, Márton and Lőrincz, András (2020b). Temporal smoothing for 3D human pose estimation and localization for occluded people. In *International Conference on Neural Information Processing*. (→ 24)

Veit, Andreas; Wilber, Michael J; and Belongie, Serge (2016). Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 32)

Vendrow, Edward; Le, Duy Tho; and Rezatofighi, Hamid (2022). JRDB-Pose: A large-scale dataset for multi-person pose estimation and tracking. *arXiv:2210.11940*. (→ 51)

Villani, Valeria; Pini, Fabio; Leali, Francesco; and Secchi, Cristian (2018). Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, **55**, 248–266. (→ 26)

Vogels, Rufin (2022). More than the face: Representations of bodies in the inferior temporal cortex. *Annual Review of Vision Science*, **8**. (→ 1)

Vosoughi, Saeid and Amer, Maria A. (2018). Deep 3D human pose estimation under partial body presence. In *IEEE International Conference on Image Processing (ICIP)*. (→ 79, 82, 91)

Waldmann, Urs; Naik, Hemal; Máté, Nagy; Kano, Fumihiro; Couzin, Iain D; Deussen, Oliver; and Goldlücke, Bastian (2022). I-MuPPET: Interactive multi-pigeon pose estimation and tracking. In *DAGM German Conference on Pattern Recognition (GCPR)*. (→ 26)

Wandt, Bastian; Ackermann, Hanno; and Rosenhahn, Bodo (2018). A kinematic chain space for monocular motion capture. In *European Conference on Computer Vision – Workshops (ECCVW)*. (→ 22)

Wandt, Bastian; Rudolph, Marco; Zell, Petrissa; Rhodin, Helge; and Rosenhahn, Bodo (2021). Canonpose: Self-supervised monocular 3D human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 139)

Wandt, Bastian; Little, James J; and Rhodin, Helge (2022). ElePose: Unsupervised 3D human pose estimation by predicting camera elevation and learning normalizing flows on 2D poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Wang, Chunyu; Wang, Yizhou; Lin, Zhouchen; Yuille, Alan L.; and Gao, Wen (2014a). Robust estimation of 3D human poses from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 12)

Wang, Dongkai and Zhang, Shiliang (2022). Contextual instance decoupling for robust multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 21)

Wang, Jiang; Nie, Xiaohan; Xia, Yin; and Wu, Ying (2014b). Mining discriminative 3D poselet for cross-view action recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (→ 13, 26, 49)

Wang, Jinbao; Tan, Shujie; Zhen, Xiantong; Xu, Shuo; Zheng, Feng; He, Zhenyu; and Shao, Ling (2021a). Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding (CVIU)*, **210**, 103225. (→ 26)

Wang, Keze; Lin, Liang; Jiang, Chenhan; Qian, Chen; and Wei, Pengxu (2019a). 3D human pose machines with self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **42**(5), 1069–1082. (→ 22)

Wang, Liang; Hu, Weiming; and Tan, Tieniu (2003). Recent developments in human motion analysis. *Pattern Recognition*, **36**(3), 585–601. (→ 26)

Wang, Tao; Zhang, Jianfeng; Cai, Yujun; Yan, Shuicheng; and Feng, Jiashi (2021b). Direct multi-view multi-person 3D human pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 25)

Wang, Zhou; Bovik, Alan C.; Sheikh, Hamid R.; and Simoncelli, Eero P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**(4), 600–612. (→ 158)

Wang, Zhe; Chen, Liyan; Rathore, Shuarya; Shin, Daeyun; and Fowlkes, Charless (2019b). Geometric pose affordance: 3D human pose with scene constraints. *arXiv:1905.07718*. (→ 23, 48, 124)

Wang, Zhe; Yang, Jimei; and Fowlkes, Charless (2022). The best of both worlds: Combining model-based and nonparametric approaches for 3D human body estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18)

Wehrbein, Tom; Rudolph, Marco; Rosenhahn, Bodo; and Wandt, Bastian (2021). Probabilistic monocular 3d human pose estimation with normalizing flows. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 25)

Wei, Shih-En; Ramakrishna, Varun; Kanade, Takeo; and Sheikh, Yaser (2016). Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 14, 48)

Wen, Guo; Xiaoyu, Bie and Xavier, Alameda-Pineda; and Francesc, Moreno-Noguer (2022). Multi-person extreme motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23)

Weng, Chung-Yi; Curless, Brian; Srinivasan, Pratul P.; Barron, Jonathan T.; and Kemelmacher-Shlizerman, Ira (2022). HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 18)

Werbos, Paul J (1982). Applications of advances in nonlinear sensitivity analysis. In *System modeling and optimization*, pages 762–770. Springer. (→ 31)

Wolf, Ralph and Platt, John (1993). Postal address block location using a convolutional locator network. In *Advances in Neural Information Processing Systems (NIPS)*. (→ 13)

Wolpert, David H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, **8**(7), 1341–1390. (→ 30)

Workman, Scott; Greenwell, Connor; Zhai, Menghua; Baltenberger, Ryan; and Jacobs, Nathan (2015). DEEPFOCAL: A method for direct focal length estimation. In *IEEE International Conference on Image Processing (ICIP)*. (→ 110)

Wren, Christopher Richard; Azarbayejani, Ali; Darrell, Trevor; and Pentland, Alex Paul (1997). Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **19**(7). (→ 10)

Wu, Cunlin; Xiao, Yang; Zhang, Boshen; Zhang, Mingyang; Cao, Zhiguo; and Zhou, Joey Tianyi (2022). C3P: Cross-domain pose prior propagation for weakly supervised 3D human pose estimation. In *European Conference on Computer Vision (ECCV)*. (→ 16)

Wu, Yuxin and He, Kaiming (2018). Group normalization. In *European Conference on Computer Vision (ECCV)*. (→ 34, 157)

Xie, Kevin; Wang, Tingwu; Iqbal, Umar; Guo, Yunrong; Fidler, Sanja; and Shkurti, Florian (2021). Physics-based human motion estimation and synthesis from videos. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 24)

Xu, Hongyi; Bazavan, Eduard Gabriel; Zanfir, Andrei; Freeman, William T.; Sukthankar, Rahul; and Sminchisescu, Cristian (2020a). GHUM & GHUML: Generative 3D human shape and articulated pose models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 17, 50, 124)

Xu, Jingwei; Yu, Zhenbo; Ni, Bingbing; Yang, Jiancheng; Yang, Xiaokang; and Zhang, Wenjun (2020b). Deep kinematics analysis for monocular 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 87, 88, 89)

Xu, Weipeng; Chatterjee, Avishek; Zollhoefer, Michael; Rhodin, Helge; Fua, Pascal; Seidel, Hans-Peter; and Theobalt, Christian (2019). Mo^2Cap^2 : Real-time mobile 3D motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, **25**(5). (→ 16)

Xu, Yufei; Zhang, Jing; Zhang, Qiming; and Tao, Dacheng (2022). ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 25)

Yamamoto, Masanobu and Koshikawa, Kazutada (1991). Human motion analysis based on a robot arm model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 10)

Yang, Sen; Quan, Zhibin; Nie, Mu; and Yang, Wankou (2021). TransPose: Keypoint localization via transformer. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 25)

Yang, Wei; Li, Shuang; Ouyang, Wanli; Li, Hongsheng; and Wang, Xiaogang (2017). Learning feature pyramids for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 74)

Yang, Wei; Ouyang, Wanli; Wang, Xiaolong; Ren, Jimmy; Li, Hongsheng; and Wang, Xiaogang (2018). 3D human pose estimation in the wild by adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 88)

Ye, Hang; Zhu, Wentao; Wang, Chunyu; Wu, Rujie; and Wang, Yizhou (2022). Faster VoxelPose: Real-time 3D human pose estimation by orthographic projection. In *European Conference on Computer Vision (ECCV)*. (→ 24)

Yeh, Raymond; Hu, Yuan-Ting; and Schwing, Alexander (2019). Chirality nets for human pose regression. In *Advances in Neural Information Processing Systems (NeurIPS)*. (→ 127, 133)

Yu, Baosheng and Tao, Dacheng (2021). Heatmap regression via randomized rounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. (→ 19)

Yu, Frank; Salzmann, Mathieu; Fua, Pascal; and Rhodin, Helge (2020a). PCLs: Geometry-aware neural reconstruction of 3D pose with perspective crop layers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 39)

Yu, Zhixuan; Yoon, Jae Shin; Lee, In Kyu; Venkatesh, Prashanth; Park, Jaesik; Yu, Jihun; and Park, Hyun Soo (2020b). HUMBI: A large multiview dataset of human body expressions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 48, 124)

Yuan, Ye; Wei, Shih-En; Simon, Tomas; Kitani, Kris; and Saragih, Jason (2021). SimPoE: Simulated character control for 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 24)

Yuan, Ye; Iqbal, Umar; Molchanov, Pavlo; Kitani, Kris; and Kautz, Jan (2022a). GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 19, 23, 24)

Yuan, Ye; Song, Jiaming; Iqbal, Umar; Vahdat, Arash; and Kautz, Jan (2022b). PhysDiff: Physics-guided human motion diffusion model. *arXiv:2212.02500*. (→ 24)

Yuen, Kevan and Trivedi, Mohan M. (2017). An occluded stacked hourglass approach to facial landmark localization and occlusion estimation. *IEEE Transactions on Intelligent Vehicles*, **2**(4), 321–331. (→ 55)

Zanfir, Andrei; Bazavan, Eduard Gabriel; Xu, Hongyi; Freeman, William T.; Sukthankar, Rahul; and Sminchisescu, Cristian (2020). Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*. (→ 25)

Zanfir, Andrei; Zanfir, Mihai; Gorban, Alex; Ji, Jingwei; Zhou, Yin; Anguelov, Dragomir; and Sminchisescu, Cristian (2022). HUM3DIL: Semi-supervised multi-modal 3D human pose estimation for autonomous driving. In *Conference on Robot Learning*. (→ 26)

Zanfir, Mihai; Popa, Alin-Ionut; Zanfir, Andrei; and Sminchisescu, Cristian (2018). Human appearance transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 152, 154)

Zanfir, Mihai; Zanfir, Andrei; Bazavan, Eduard Gabriel; Freeman, William T; Sukthankar, Rahul; and Sminchisescu, Cristian (2021). THUNDR: Transformer-based 3D human reconstruction with markers. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 18)

Zeng, Ailing; Yang, Lei; Ju, Xuan; Li, Jiefeng; Wang, Jianyi; and Xu, Qiang (2022). SmoothNet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision (ECCV)*. (→ 24)

Zhang, Dejun; Wu, Yiqi; Guo, Mingyue; and Chen, Yilin (2021). Deep learning methods for 3D human pose estimation under different supervision paradigms: A survey. *Electronics*, **10**(18), 2267. (→ 26, 123)

Zhang, Jason Y.; Pepose, Sam; Joo, Hanbyul; Ramanan, Deva; Malik, Jitendra; and Kanazawa, Angjoo (2020a). Perceiving 3D human–object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*. (→ 22)

Zhang, Richard; Isola, Phillip; Efros, Alexei A.; Shechtman, Eli; and Wang, Oliver (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 32, 158)

Zhang, Tianshu; Huang, Buzhen; and Wang, Yangang (2020b). Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 23, 48, 124)

Zhang, Weichen; Liu, Zhiguang; Zhou, Liuyang; Leung, Howard; and Chan, Antoni B. (2017). Martial arts, dancing and sports dataset: a challenging stereo and multi-view dataset for 3D human pose estimation. *Image and Vision Computing*, **61**, 22–39. (→ 48, 124)

Zhang, Yifu; Wang, Chunyu; Wang, Xinggang; Liu, Wenyu; and Zeng, Wenjun (2022). VoxelTrack: Multi-person 3D human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. (→ 23)

Zhao, Mingmin; Li, Tianhong; Abu Alsheikh, Mohammad; Tian, Yonglong; Zhao, Hang; Torralba, Antonio; and Katabi, Dina (2018). Through-wall human pose estimation using radio signals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 16)

Zhao, Ruiqi; Wang, Yan; and Martinez, Aleix M. (2017). A simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **40**(12). (→ 19)

Zhao, Shu; Salah, Almıla Akdağ; and Salah, Albert Ali (2022). Automatic analysis of human body representations in western art. *arXiv:2210.08860*. (→ 26)

Zheng, Ce; Zhu, Sijie; Mendieta, Matias; Yang, Taojiannan; Chen, Chen; and Ding, Zhengming (2021). 3d human pose estimation with spatial and temporal transformers. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 25)

Zheng, Jingxiao; Shi, Xinwei; Gorban, Alexander; Mao, Junhua; Song, Yang; Qi, Charles R; Liu, Ting; Chari, Visesh; Cornman, Andre; Zhou, Yin; *et al.* (2022a). Multi-modal 3D human pose estimation with 2D weak supervision in autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW)*. (→ 26)

Zheng, Xu; Zheng, Yali; and Yang, Shubing (2022b). Generating multiple hypotheses for 3D human mesh and pose using conditional generative adversarial nets. In *Asian Conference on Computer Vision (ACCV)*. (→ 25)

Zhong, Zhun; Zheng, Liang; Kang, Guoliang; Li, Shaozi; and Yang, Yi (2020). Random erasing data augmentation. In *AAAI Conference on Artificial Intelligence*. (→ 55, 56, 58, 60, 87, 96, 99)

Zhou, Kun; Han, Xiaoguang; Jiang, Nianjuan; Jia, Kui; and Lu, Jiangbo (2019). Hemlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 90)

Zhou, Xiaowei; Zhu, Menglong; Leonardos, Spyridon; Derpanis, Konstantinos G.; and Daniilidis, Kostas (2015). Sparseness meets deepness: 3D human pose estimation from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 63, 71)

Zhou, Xingyi; Sun, Xiao; Zhang, Wei; Liang, Shuang; and Wei, Yichen (2016). Deep kinematic pose regression. In *European Conference on Computer Vision (ECCV)*. (→ 63, 71)

Zhou, Xingyi; Huang, Qixing; Sun, Xiao; Xue, Xiangyang; and Wei, Yichen (2017). Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *IEEE International Conference on Computer Vision (ICCV)*. (→ 15, 21, 71, 74, 77, 82, 86, 88, 90, 91, 101)

Zhu, Jun-Yan; Park, Taesung; Isola, Phillip; and Efros, Alexei A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 155)

Zhu, Zhen; Huang, Tengteng; Shi, Baoguang; Yu, Miao; Wang, Bofei; and Bai, Xiang (2019). Progressive pose attention transfer for person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 153)

Zimmermann, Christian; Welschehold, Tim; Dornhege, Christian; Burgard, Wolfram; and Brox, Thomas (2018). 3D human pose estimation in RGBD images for robotic task learning. In *IEEE International Conference on Robotics and Automation (ICRA)*. (→ 13, 16, 74)

Zoph, Barret and Le, Quoc V (2017). Neural architecture search with reinforcement learning. In *International Conference on Learning Representations (ICLR)*. (→ 33)

Zou, Shihao; Zuo, Xinxin; Qian, Yiming; Wang, Sen; Xu, Chi; Gong, Minglun; and Cheng, Li (2020a). 3D human shape reconstruction from a polarization image. In *European Conference on Computer Vision (ECCV)*. (→ 16)

Zou, Shihao; Zuo, Xinxin; Qian, Yiming; Wang, Sen; Guo, Chuan; Xu, Chi; Gong, Minglun; and Cheng, Li (2020b). Polarization human shape and pose dataset. *arXiv:2004.14899*. (→ 49)

Zou, Zhengxia; Shi, Zhenwei; Guo, Yuhong; and Ye, Jieping (2019). Object detection in 20 years: A survey. *arXiv:1905.05055*. (→ 20)

Zuffi, Silvia and Black, Michael J. (2015). The stitched puppet: A graphical model of 3D human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (→ 17)

Zuffi, Silvia; Kanazawa, Angjoo; Berger-Wolf, Tanya; and Black, Michael J. (2019). Three-D Safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *IEEE International Conference on Computer Vision (ICCV)*. (→ 26)

Curriculum Vitae

István Sárándi

E-Mail	sarandi@vision.rwth-aachen.de
Date of Birth	05.08.1989
Place of Birth	Budapest, Hungary
Citizenship	Hungarian

Education

Apr. 2017 – present	Doctoral student at the Computer Vision Group RWTH Aachen University Supervisor: Prof. Dr. Bastian Leibe
Oct. 2012 – Mar. 2016	Student of Computer Science RWTH Aachen University Degree: Master of Science
Sep. 2008 – Jan. 2012	Student of Computer Engineering Budapest University of Technology and Economics Degree: Bachelor of Science

Professions

Nov. 2013 – May 2014	Student Research Assistant at the Computer Vision Group, RWTH Aachen University
Dec. 2012 – Oct. 2013	Student Research Assistant at the Department of Medical Informatics, Uniklinik RWTH Aachen

Publications

István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023.

István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. MeTRAbs: Metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 3(1):16–30, 2021.

István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. Metric-scale truncation-robust heatmaps for 3D human pose estimation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020.

Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3D features. In *IEEE Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW)*, 2020.

Kilian Pfeiffer, Alexander Hermans, István Sárándi, Mark Weber, and Bastian Leibe. Visual person understanding through multi-task and multi-dataset learning. In *DAGM German Conference on Pattern Recognition (GCPR)*, 2019.

István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. Synthetic occlusion augmentation with volumetric heatmaps for the 2018 ECCV PoseTrack Challenge on 3D human pose estimation. *arXiv:1809.04987*, 2018.

István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. How robust is 3D human pose estimation to occlusion? In *IEEE/RSJ International Conference on Intelligent Robots and Systems – Workshops (IROSW)*, 2018.

István Sárándi, Dan Philipp Claßen, Anatoli Astvatsatourov, Oliver Pfaar, Ludger Klimek, Ralph Mösges, and Thomas M. Deserno. Quantitative conjunctival provocation test for controlled clinical trials. *Methods of Information in Medicine*, 53(4):238–244, 2014.

Thomas M. Deserno, István Sárándi, Abin Jose, Daniel Haak, Stephan Jonas, Paula Specht, and Vincent Brandenburg. Towards quantitative assessment of calciphylaxis. In *Medical Imaging: Computer-Aided Diagnosis*, volume 9035, page 90353C, 2014.

Suman Raj Bista, István Sárándi, Serkan Dogan, Anatoli Astvatsatourov, Ralph Mösges, and Thomas M. Deserno. Automatic conjunctival provocation test combining hough circle transform and self-calibrated color measurements. In *Medical Imaging: Computer-Aided Diagnosis*, volume 8670, page 86702J, 2013.