# Digital Human Model Reconstructions Using Advanced Deep Learning Methods

**A thesis submitted for the degree of Doctor of Philosophy**

**By**

**Moyu Wang**

**Department of Civil and Environmental Engineering**

**Brunel University London**

**August 2024**

# Abstract

Over the past few decades, 3D digital human modeling has emerged as a vibrant field of research, playing a foundational role in various applications such as film production, sports, medical sciences, and human-computer interaction. Early research efforts predominantly focused on artist-driven modeling techniques or relied on expensive scanning equipment. Our objective, however, is to leverage recent advances in deep learning technology to automatically generate personalized virtual avatars using only low-cost monocular cameras. In this dissertation, we present significant advancements in 3D digital human reconstruction from monocular images. By developing methods that effectively integrate temporal information and realistically reconstruct from sparse data, we address this challenging task. Given images and videos captured from monocular cameras, we have, for the first time, successfully reconstructed not only the 3D pose but also the complete 3D geometry of a person, including facial features, hair, and clothing.

In our initial work, we trained a neural network with a partial attention mechanism to estimate 3D human poses from a single image. The network outputs a 3D mesh model that encapsulates body shape and posture but lacks surface details. This approach yielded promising results. In subsequent work, we advanced the optimization of the nude model obtained from the first study by incorporating multi-view human images to reduce errors caused by occlusions. By predicting the pose for each frame, we re-aligned the standard model and projected it onto each image for further optimization. We then employed shape-from-shading techniques to enhance surface details.

In this dissertation, we explore methods for digital human reconstruction from monocular images and videos, enhanced by deep learning techniques. We present reconstruction approaches that focus on accuracy, simplicity, usability, and visual fidelity, utilizing multi-view image optimization. Through extensive evaluations, we provide a thorough analysis

of key parameters, reconstruction quality, and the robustness of our methods. For the first time, our approach enables camera-based, user-friendly digitization for personal users, opening up exciting new applications such as telepresence and virtual try-on in online fashion shopping.

# Declaration

I hereby declare that this work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

# Acknowledgements

I would like to express my deepest gratitude to my supervisor, Qingping Yang, for his invaluable guidance, encouragement, and support throughout this journey. Your expertise and thoughtful feedback have not only shaped this project but also inspired me to grow as a scholar.

I would also like to extend my heartfelt thanks to my family, whose love and support have been my anchor throughout this process. To my parents, Li Wang and Xiaomin Hu, your constant encouragement and belief in my abilities have been my greatest source of strength. To my uncle Xiaowei Hu, your patience, understanding, and sacrifices have made this achievement possible.

This accomplishment would not have been possible without the support and guidance of my supervisor and the love and encouragement of my family. And finally, thank you to my wife, Xiao Chang, I am truly grateful to have you by my side. I hope we can continue exploring our journey of life together.

# Publications

Wang, Moyu & Yang, Qingping. (2024). From prediction to measurement, an efficient method for digital human model obtainment. International Journal of Metrology and Quality Engineering. 15. 10.1051/ijmqe/2023015.

# Contents

# Acronyms

**VR** Virtual Reality

**AR** Augmented Reality

**3DPW** 3D Poses in the Wild

**SMPL** A Skinned Multi-Person Linear Model

**3D** Three-Dimensional

**SCAPE** Shape Completion and Animation of People

**PCA** Principal Component Analysis

**LBS** Linear Blending Skinning

**SFS** Shape from Shading

**GAN** Generative Adversarial Networks

**CNN** Convolution Neural Networks

**3DMM** 3D Morphable Models

**BANMo** Builder of Animatable 3D Neural Models

**NeRF** Neural Radiance Fields

**SDF** Signed Distance Function

**CSE** Continuous Surface Embeddings

**MLP** Multi-layer Perceptron

**PIFu** Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization

**ICON** Implicit Clothed-humans Obtained from Normals

**NASA** Neural Articulated Shape Approximation

**JIFF** Jointly-aligned Implicit Face Function

**HPS** Human Pose and Shape

**LPIPS** Learned Perceptual Image Patch Similarity

**IS** Inception Score

**FID** Fréchet Inception Distance

**SSIM** Structural Similarity Index

**PSNR** Peak Signal-to-Noise Ratio

# List of Figures

# List of Tables

# 1. Introduction

The task of capturing and modeling the 3D human body from monocular video or photographs represents a fundamental challenge in both the domains of Computer Vision and Computer Graphics. Over the past few decades, the focus has primarily been on estimating the 3D pose of a subject, characterized by the spatial arrangement of distinct body parts or articulated through joint angles. This pursuit has remained a central theme in Computer Vision, continuing to be an active area of research with widespread applications in scene analysis, medical diagnostics, and human-computer interfaces.

In recent times, there has been a noticeable shift towards the automatic 3D reconstruction of the entire human body, marking a significant development in this field. This advancement goes beyond merely estimating a 3D skeleton; instead, the objective is to reconstruct the complete 3D human shape, encompassing intricate details such as hair, clothing, and overall appearance. The ultimate goal is to generate avatars that are virtually indistinguishable from real humans.

The emergence of consumer hardware for Virtual Reality (VR) and Augmented Reality (AR) has laid the groundwork for innovative avenues in entertainment, communication, and online shopping. In these applications, the creation of personalized and highly realistic 3D avatars holds paramount importance. These avatars need to faithfully replicate all the nuances that contribute to our individual identities, including precise body shapes, intricately reconstructed faces, detailed clothing, and realistic hair. Failures in the reconstruction process result in avatars that are not easily identified by others and, more critically, can lead to users feeling disconnected or misrepresented in their virtual self.

It is noteworthy that the acquisition process for these avatars should be streamlined, swift,

and accessible without the need for specialized equipment or training. However, the conventional approach in Computer Graphics for 3D modeling of virtual humans relies heavily on manual effort and expert knowledge. Specially trained artists are typically involved in defining the 3D geometry of the body and clothing, which is then rigged to facilitate animation. The avatar's 3D motion is often dictated by labor-intensive keyframe-based animation or marker-based motion capture. This intricate and time-consuming process poses a practical barrier, particularly for applications in entertainment, communication, and online shopping.

In contrast, the focus of this work is to leverage the ubiquity of cameras in today's environment to develop automatic methods that efficiently harness images and video for realistic 3D avatar creation and animation. This shift towards automation aims to overcome the practical challenges associated with manual processes, ensuring a more accessible and scalable approach to generating lifelike virtual representations of humans.

This thesis delves into the burgeoning field of 3D reconstruction of human shape and pose from monocular images. Within this exploration, we introduce innovative methodologies aimed at reconstructing and tracking mesh-based 3D representations of humans, as captured in both monocular videos and individual photographs. Our research contributes fundamental advancements to the intricate task of 3D human model reconstruction from monocular images.

One key aspect of our work involves the development of methods that efficiently fuse information from multiple points in time. This temporal integration enhances the reconstruction process, allowing for the realistic completion of 3D human avatars from sparse observations. By addressing the challenges associated with sparse data, our methodologies pave the way for significant strides in the realm of 3D reconstruction from monocular images.

A notable achievement of our research is the facilitation of easy acquisition of animatable 3D avatars for a broad audience. By leveraging our novel approaches, individuals can now readily obtain 3D avatars that are not only accurate but also capable of dynamic animation. This democratization of 3D avatar acquisition marks a groundbreaking shift, making sophisticated virtual representations accessible to a wider range of users.

Furthermore, our work opens avenues for various exciting new applications. The ability to effortlessly generate animatable 3D avatars introduces transformative possibilities across diverse fields. From virtual communication to entertainment and beyond, our research sets the stage for the integration of lifelike 3D avatars into applications that were previously constrained by complex acquisition processes.

In essence, this thesis contributes significantly to the evolving landscape of 3D human avatar reconstruction from monocular images. Through our novel methodologies, we overcome challenges associated with sparse data and temporal integration, making the acquisition of animatable 3D avatars accessible to everyone. This research not only pushes the boundaries of technology but also unlocks the potential for innovative applications that harness the power of realistic virtual representations.

## 1.1 Background

In the narrow sense, digitizing the human body primarily involves geometric reconstruction and texture estimation for a specific individual, allowing the faithful representation of their real digital image on a computer. This problem has long been a challenging subject of research in the fields of computer graphics and computer vision, with complexity manifesting in two main aspects: geometric and textural color complexities.

The geometric characteristics of the human body are influenced by various factors, including gender, body posture, race, and stance. Particularly, the variations in posture result in large-scale non-rigid deformations, making it challenging to directly apply linear deformation methods such as BlendShape and Principal Component Analysis (PCA), commonly used for facial shape blending, to the deformation of the entire human body. Additionally, the presence of clothing in real-world scenarios introduces a high level of complexity to the shape of the human body, involving various materials, clothing styles, and intricate interactions between clothing and the body. These complexities impose stringent requirements on the accuracy of reconstruction algorithms and the expressive capabilities of topological changes.

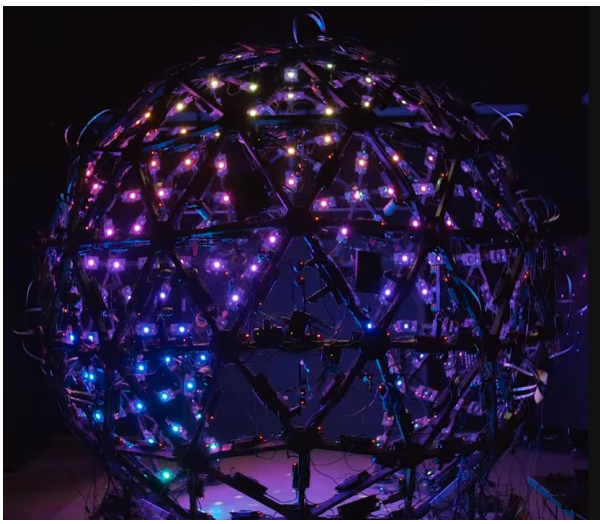


Figure 1.1: Relightable [1]: Google developed a multi-camera system.

On the other hand, the diverse skin tones, as well as the varied colors and material attributes of clothing, make the precise extraction of texture information from the human body considerably challenging. In well-established industries such as film, gaming, and other mature areas of human digitization, complex capture systems and a trade-off between time and precision are typically employed to address the intricate problem of human reconstruction. Traditional methods, as illustrated in Figure 1.1, demand high-quality data processing. Precision-designed multi-camera optical systems require meticulous camera calibration and synchronization. Additionally, captured individuals often need to wear specially designed motion capture suits or carry key point markers for localization. Post-data collection involves time-consuming and complex offline processing to obtain highly accurate digital representations of the target individuals, meeting industrial application requirements. The stringent requirements for high-end equipment and controlled environments make traditional methods challenging to apply to low-end devices and everyday scenarios, hindering the widespread adoption of this technology among the general consumer population.

In recent years, with the continuous development of Internet technology and the gradual popularization of 5G communication technology, the number of Internet users has been steadily increasing. The demand for high-quality digital technology is growing, particularly with the rise of new work and lifestyle trends such as "remote work" and "online communication" sparked by the outbreak of the COVID-19 pandemic in 2020. This has accelerated the development of human digitization technology, gradually revolutionizing the daily lives of ordinary consumers. For example, virtual anchor technology has begun to be applied in news and live broadcasting industries, major companies are gradually introducing their own digital spokespersons, and the gradual application of virtual reality and augmented reality technologies is poised to transform human communication methods through holographic communication. Against this backdrop, the use of mid to low-end devices for high-quality human digitization, especially in the areas of geometric reconstruction and texture estimation, has become

an increasingly important research direction in the fields of computer graphics and computer vision.

With the rise of consumer-grade scanning devices such as Kinect [2], Primesense [3], and iPhone-15 [4], the acquisition of RGB-D data has become more accessible. A wave of methods for capturing the human body based on RGB-D data has emerged. These methods, based on fusion principles, progressively reconstruct the complete shape of a person in real-time during the scanning process by integrating depth information from each frame into a reference space. However, tracking accuracy is a pain point for such methods. Introducing prior knowledge of human body information to achieve faster and more stable tracking is the continued goal of improving such methods.

Although the acquisition of depth information has become easier with the advent of consumer-grade devices, most mobile devices on the market do not widely use depth cameras. This makes single RGB images the truly accessible data form today. How to achieve accurate human body reconstruction from monocular data is an urgent problem to solve. Compared to RGB-D data, information obtained from images is further reduced due to the lack of depth information. For monocular data, the depth ambiguity problem caused by perspective projection theoretically prevents the accurate estimation of the scale information of the human body, leading to potential errors in human body pose estimation. To address this problem, three main approaches are generally employed to reconstruct the human body accurately from monocular data: parameterizing human body shape to regularize spatial resolution, leveraging the powerful fitting capabilities of neural networks to learn reasonable mappings from a large number of images and geometric data, and using video data to impose additional constraints [5].

Parameterized human body models decompose the deformation of the human body into several low-dimensional parameterized representations (such as identity and posture) by learning their statistical distribution from a large amount of human body data. By

embedding the low-dimensional manifold of human body deformation space, the reasonable solution space is significantly reduced to counteract the ambiguity of monocular data. Such methods often establish a mapping between the image and the low-dimensional parameter space of the human body through optimization and regression, enabling human body reconstruction. However, these methods face challenges when applied to clothed human bodies due to the inherently high-dimensional nature of clothing shapes, making parameterized human body approaches difficult to generalize.

Deep neural networks are a model generated by stacking multiple layers of a simple network structure, possessing strong fitting capabilities. They have been widely applied with excellent results in various computer vision tasks [6]. The core of these networks is to train the network by using a large amount of paired data, allowing the network to learn the latent distribution in the data. Neural networks can not only regress the parameters of parameterized models from images but can also directly regress non-parametric representations of the human body, including voxels and signed distance fields. This enables the reconstruction of more complex geometry. However, a major limitation of such methods is that the trained model depends on the dataset, leading to issues of limited generalization and overfitting. Moreover, obtaining high-precision geometric data for clothed human bodies is already challenging, making scarce data a significant obstacle for the practical application of these methods.

Compared to a single image, video data contains more information. Human motion has a certain continuity and semantic meaning over time, and this prior knowledge can help reconstruction algorithms counteract the ambiguity of monocular data. Additionally, regularization based on correspondence between multiple frames can be designed to assist optimization and network fitting. Recently, the rise of implicit neural representations and neural rendering techniques has demonstrated the possibility of self-supervised reconstruction of high-precision geometry and photo-realistic rendering from multiple

image inputs. This presents a novel approach to reduce dependency on data and recover the human body from video data.

In summary, this chapter highlights the challenges and advancements in the digitization of the human body, particularly focusing on the complexities of geometric and textural aspects. It discusses the traditional methods used in mature industries, the emerging trends driven by internet technology, and the research directions in computer graphics and computer vision for achieving high-quality human digitization, especially with the use of mid to low-end devices. In the following chapters, we will cover the various approaches, challenges, and limitations in the field, emphasizing the importance of advancements in capturing devices, data processing, and the application of neural networks for achieving accurate and realistic human body digitization.

## 1.2 Problem Statement

The realm of image-based 3D pose and shape reconstruction of humans is a rich area of investigation, encompassing various approaches and perspectives. Diverse researchers pursue different aspects of this complex task. Some emphasize the 3D skeleton and approximate body proportions [5, 7, 8, 9, 10], while others focus on reconstructing the naked body shape without clothing [11, 12, 13]. There are also those who center their attention on the garments worn by subjects [14, 15], and some concentrate on specific body parts such as the face [16, 17, 18, 19, 20], hands [21, 22, 23, 24], or hair [25, 26].

In contrast to these varied approaches, the primary objective of this thesis is to comprehensively track and model the entire human body, encompassing aspects like hair and clothing. The aim is to achieve detailed reconstructions of the observed subject, not only in terms of body and clothing geometry but also capturing essential information related to coloring and surface structure. In pursuit of this goal, the focus extends beyond 3D shape capture to the reconstruction of surface colors using texture maps. Moreover, the reconstructions are intended to be reusable, emphasizing the importance of a common format that facilitates easy integration, animation, and manipulation by other applications.

Having delineated the desired characteristics of the reconstructions, the second pivotal aspect of this work pertains to the capturing process and equipment. Researchers in Computer Vision have utilized a diverse array of sensors and systems for world capture and analysis, including multi-camera setups, marker-aided capturing, depth sensors, and active scanners [1]. While these systems provide high-resolution 3D data, their limited accessibility—typically confined to laboratories or professional video studios—exposes a constraint.

On the contrary, standard cameras are omnipresent in our daily lives. Many of the devices we interact with regularly, such as smartphones, tablets, or laptops, come equipped with

one or multiple cameras that are easily accessible. Noteworthy advantages of standard cameras include their unobtrusiveness and the simplicity of the capturing process. Unlike complex systems requiring meticulous calibration and setup, cameras enable straightforward recording with minimal setup time. Additionally, they are lightweight, compact, and can be flexibly employed in diverse settings.

In this work, we exclusively rely on monocular image material captured by a standard webcam as the input to our algorithms. This deliberate choice aligns with the unobtrusive and easily accessible nature of standard cameras. Not only does it ensure compatibility with modern devices like phones, tablets, or smart displays, but it also allows seamless integration with a vast repository of existing photo and video material. This approach emphasizes the adaptability and practicality of our work, making it conducive to integration with contemporary technology and leveraging existing visual data resources.

## 1.3 Motivation

The utilization of 3D virtual human avatars has been a prevalent practice in various applications in the past, and their potential central role in future applications is highly anticipated. In the film industry, virtual actors have become a common tool for digital editing and enhancing real-world video footage, and in some cases, they are employed to create entirely computer-generated movies. Producers and designers invest substantial efforts to generate highly realistic and physically plausible virtual counterparts of real-world actors. Similarly, the gaming industry places increasing emphasis on developing realistic characters to enhance the immersive gaming experience. The prospect of fully automatic and widely accessible reconstruction of highly realistic virtual humans holds immense promises for both the entertainment and gaming industries.

Beyond the realms of entertainment, 3D virtual humans have the potential to play crucial roles in various applications. Examples include human understanding for human-computer interfaces, medical diagnostics, virtual assistance, fitness and health tracking, virtual try-on experiences in online shopping, interpretation and understanding of body language, and more. All these applications stand to benefit from more accurate reconstructions and easier acquisition of 3D virtual humans. An emerging and promising area is the application of 3D virtual humans in communications, particularly in Virtual Reality (VR) or Augmented Reality (AR) telepresence. The active research in this field is focused on enabling applications that can significantly impact travel behavior, communication methods, and overall lifestyle [27, 28, 29, 30].

The significance of highly realistic 3D human avatars and accessible reconstruction pipelines extends to numerous scientific subjects and industries. Human communication is inherently visual, and our visual appearance conveys rich information. Through visual inspection of other human beings, we can comprehend their mood, state of health, personal preferences, engagement, and more [31]. This thesis lays the foundation for computers to model and understand the subtle visual cues that facilitate human

understanding of the human body and its language. On the flip side, the scientific findings presented in this work have immediate applicability across various domains, including entertainment, fitness and health, and online shopping. Figure 1.2 shows one application of 3D human avatars use in fitness and health.

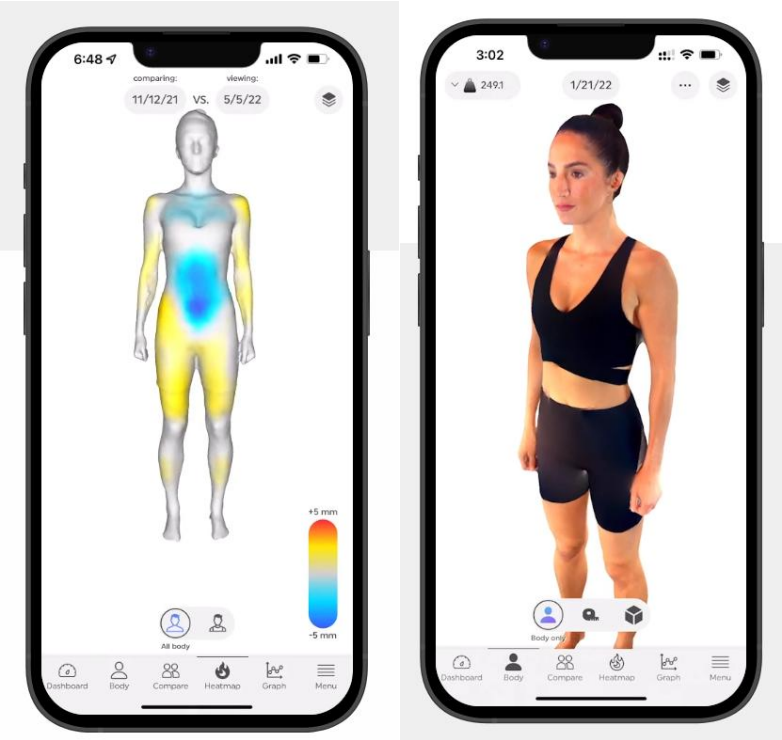



Figure 1.2: 3D human body for fitness and health [32].

In essence, the advancements in 3D virtual human avatars and reconstruction pipelines hold transformative potential for a wide array of applications and industries. The integration of these technologies has the capacity to reshape the way we create content, engage in entertainment, communicate, and understand human behavior. However, existing 3D virtual human technologies fail to achieve a perfect balance between precision and cost, which hinders the expansion of new technologies' impact on daily life and various fields from medicine to online retail. This thesis contributes to the groundwork by leveraging advanced deep learning techniques to eliminate the need for expensive input devices, making extensive, online, and convenient digital human reconstruction a reality for realizing these transformative possibilities.

## 1.4 Challenges

Humans exhibit an extraordinary ability to decipher and predict the intricacies of the human body and its expressions, extracting a wealth of information even from a static 2D photograph. From such images, humans effortlessly discern the 3D body pose, approximate height, body shape, facial expressions, actions performed and even make informed predictions about unseen aspects of the scene. This remarkable proficiency is rooted in our extensive experiential knowledge of how humans look, move, and behave.

In stark contrast, computers and algorithms encounter substantial challenges when tasked with processing monocular videos and photos. The inherent complexity arises from encoded, noisy, or ambiguous information, often requiring intricate computational methods to navigate. A primary challenge emerges from the absence of direct depth information in images. Depth, a crucial dimension in understanding 3D scenes, is instead encoded indirectly through perspective, shading, and semantic cues. Deciphering this indirect information proves significantly more challenging than dealing with explicit depth values.

Compounding the complexity, factors such as the distance to the camera, the actual size of objects, and the focal length of the camera influence the projected size of objects in an image. This gives rise to a situation where multiple 3D skeletons may project onto the same 2D skeleton, introducing ambiguity in recovering the true 3D pose. Lens distortion, uncertainties in recording parameters, perspective distortion, foreshortening effects, occlusion, and self-occlusion further contribute to the intricacies of image interpretation.

Beyond these challenges, images often contain extraneous information that is not directly relevant to the task at hand. Background details, shadows, reflections, sensor noise, compression artifacts, or the introduction of new objects can inadvertently mislead or impede the algorithms. Furthermore, the appearance and shape of the object of interest, particularly when dealing with humans, may undergo significant changes over time.

Factors such as alterations in pose, changes in illumination conditions, variations in camera settings, clothing adjustments, hairstyle modifications, and more, introduce additional layers of complexity.

Reconstructing humans introduces yet another hurdle known as the Uncanny Valley [33], a concept describing the emotional response to the degree of realism in an artificial human. As artificial entities approach human likeness, there is a dip in the curve of familiarity, potentially causing feelings of uncanniness or revulsion. While the Uncanny Valley is a theoretical construct, empirical evidence from studies lends credence to its existence [34].

Acknowledging the multifaceted challenges, this work provides tailored algorithms designed for more or less constrained settings. Striking a careful balance, the setups are crafted to avoid excessive constraints, ensuring reproducibility and facilitating straightforward data acquisition. The specific focus on detailed 3D shape acquisition guides the setup to images capturing individuals in standard A-poses or common standing poses, a practical choice for a variety of applications.

The subsequent sections delve into the nuanced contributions of this work, elucidating the intricate approaches and partial solutions devised to address the myriad challenges outlined. Detailed explanations of the methods, tools, and conceptual frameworks employed shed light on the journey towards overcoming the complexities inherent in the reconstruction of the human form from 2D images.

## 1.5 Contribution to Knowledge

This thesis explores the latest advancements in the reconstruction of digital human models from monocular images and videos. The methods described in this work can be summarized in terms of their input and output modalities: the input consists of monocular videos or photographs, and the output is an animatable 3D mesh that accurately captures the shape, pose, or motion of the individual depicted in the input data.

By addressing the complex task of digital human reconstructions from monocular images, this thesis makes a significant contribution to the field: our work simplifies the digitization of humans by relying solely on regular videos or even photographs. With advanced performance in reconstruction speed and accuracy, it eliminates the need for specialized equipment, enabling the automatic reconstruction of detailed human shapes and the widespread application of virtual humans in emerging technologies.

Our research examines various approaches to 3D human pose and shape reconstruction. We present the advantages of both explicit model-based and learning-based methods, investigate different forms of data representation and supervision losses, and discuss the robustness and limitations of our approach. In the following sections, we briefly summarize the key contributions of each experimental chapter.

Chapter 2 offers a thorough overview of 3D human modeling and reconstruction, covering techniques based on images, depth sensors, and 3D data. Chapter 3 examines the key concepts utilized in the methodologies developed for this dissertation.

In Section 4.1, we propose a novel body part-driven attention framework that leverages pixel-aligned local features for regressing body pose and shape. This method has demonstrated superior performance in benchmark tests across various datasets.

In Section 4.2, while 3D poses and motion estimation gained increasing popularity during

the course of this research, monocular 3D human shape estimation with parametric models remained largely constrained to estimating fully body model. In this chapter, we discuss an advanced method capable of reconstructing the fully digital human model with cloth from video or multi-images. Drawing inspiration from the concept of visual hull, our approach aggregates silhouette information from multiple frames of a video in which the subject is visible from all sides, aligning an initialize model with each frame. Extensive experiments validate our method, demonstrating a reconstruction accuracy of 3.8mm in an opensource dataset and robustness against noisy 3D pose estimates.

Our contribution to knowledge can simply be summarized as:

1. We creatively use deep learning tools to extract information from image to regress parameterized human body model with state of the art results.
2. Our advanced state of the art method can reconstruct the fully digital human model with clothes by optimizing the regressed parameterized models from video or multi-images.
3. Our work simplifies the digitization of humans by relying solely on regular videos or even photographs which are available for widespread online application.

## 1.6 Thesis Structure

This thesis is structured as follows: Chapter 2 provides a comprehensive review of the field of 3D human modeling and reconstruction from images, depth sensors, and 3D data. This chapter traces the evolution of the field from the use of geometric primitives to more sophisticated data-driven models, with a particular focus on recent advancements in Deep Learning techniques. Additionally, it delves into the topic of human pose reconstruction and introduces the foundational concepts of convolutional neural networks. Chapter 3 explores the various concepts employed in the methodologies developed for this dissertation, offering an overview of the essential methods and a detailed explanation of the key tools and algorithms utilized. Chapter 4 discusses a novel body part-driven attention framework that exploits pixel-aligned local features for the regression of body pose and shape. Chapter 4 also presents an advanced method capable of reconstructing fully digital human models, including clothing, from video or multi-image inputs. Chapter 4 constitutes the core contributions of this dissertation. Finally, Chapter 5 concludes the thesis by discussing the results obtained and providing an in-depth analysis of potential future research directions.

# 2. Literature Review

## 2.1 Body Models based on Geometric Primitives

From the early stages of research, scholars recognized the profound advantages of incorporating a model of the human body into their methodologies. The initial endeavors involved representing the human body using geometric primitives, exemplified by Hanavan Jr's pioneering mathematical model in 1964 [35]. This groundbreaking work constructed a personalized body model through the utilization of 15 simple 3D polygonal shapes. In parallel, simpler 2D models emerged and found successful applications in human gait analysis [36, 37, 38].

The evolution of research in 3D human pose estimation and tracking became a catalyst for the development of increasingly sophisticated models of the human body and its kinematic chain [39, 40, 41, 42]. This progression reflected a continuous effort to enhance the fidelity and accuracy of representations capturing the intricate movements and postures of the human form.

The exploration of the human shape itself became a pivotal consideration, marking the introduction of the first comprehensive parametric yet entirely synthetic body models [43, 44, 45, 46]. This shift in focus allowed researchers to delve into the nuanced aspects of human anatomy, enabling the creation of synthetic bodies that encompassed a broad spectrum of shapes and configurations. These parametric models laid the groundwork for more advanced approaches in understanding, simulating, and analyzing human body dynamics.

In summary, the trajectory of research in modeling the human body has traversed various stages, from early geometric primitives to intricate 3D representations. The impetus behind this progression has been the pursuit of more accurate and comprehensive

models that capture the complexity of human anatomy, motion, and shape. The incorporation of parametric synthetic models further exemplifies the commitment to pushing the boundaries of understanding and simulating the human form in diverse applications.

## 2.2 Artist-Driven and Anatomical Models

Concurrently, the realm of Computer Graphics witnessed the inception of digital actors, marking a transformative phase in the movie industry [47]. Analogous to the models emerging in the Computer Vision community, these digital characters were initially crafted from geometric shapes, with an embedded skeleton facilitating animation. In pursuit of heightened realism, researchers embarked on the development of layered models encompassing bones, muscles, and skin, driven either by artistic intuition [48] or inspired by anatomical principles [49, 50]. However, the construction of these intricate models posed significant challenges, and their simulation demanded computationally expensive calculations.

In response to these challenges, researchers turned to the adoption of skinning techniques [51, 52, 53] as a pragmatic solution. Skinning techniques entail the definition of how the surface of a model deforms and moves in tandem with the motion of an embedded skeleton. This approach proved to be instrumental in achieving realistic animations while mitigating the complexity associated with detailed anatomical models. Notably, the use of skinning techniques gained popularity due to their ability to streamline the animation process and enhance computational efficiency.

In the pursuit of more advanced solutions, contemporary parametric models have emerged, representing a paradigm shift in character construction. These models leverage state-of-the-art techniques to derive realistic animations directly from data, ushering in a new era in character modeling and animation. The subsequent sections delve into the intricacies of these data-driven parametric models, shedding light on their

methodologies and contributions to the evolving landscape of Computer Graphics and animation in the academic domain.

## 2.3 Parametric Body Models and Applications

Constructing digital human representations based on meshes fundamentally relies on the human body mesh model. This section focuses on the parameterized model of the human body. A parameterized model of the human body is a mesh model that supports dynamic adjustments of its attributes through parameters. The most classical algorithm for generating parameterized human body models is SCAPE (Shape Completion and Animation of People) [54], which employs PCA (Principal Component Analysis) to extract two independent low-dimensional parameters, body shape and pose, for synthesizing the parameterized human model. Mesh deformation in this context is dependent on the rotational deformation of triangles. Contemporary animation production frequently utilizes mesh vertex deformation, corresponding to the classical skinning technique, which will be elaborated upon subsequently.

A virtual human representation can be conceptualized as consisting of two primary components: the skeleton and the skin. The skeleton is composed of a hierarchy of joints, while the skin comprises surfaces formed by multiple points in three-dimensional space. To construct a human body mesh model, one must first generate a skeleton and then bind the mesh vertices to the joints with specific weights, a process referred to as "skinning" [55]. Human motion can be interpreted as articulated motion within the body, characterized by rotations and translations at the joints. The simulation of human motion is manifested in the computation of the effects on relevant joints due to movement, leading to the determination of joint positions post-motion. The linear blending skinning (LBS) algorithm [56] achieves mesh deformation by performing a weighted summation linear operation based on the specific motion's impact on each bound joint. Traditional LBS algorithms perform linear operations solely on rotations, which results in the "candy-wrapper" effect of limb twisting and potential disjunctions at the joints.

Building upon the LBS algorithm, the SMPL (Skinned Multi-Person Linear Model) [57] has developed a parameterized three-dimensional model of the human body, which ensures smooth transitions at joint connections through parameters learned from data during the blending deformation process. The SMPL model supports the input of pose parameters and body shape parameters from external sources, effectively simulating muscle deformation during limb movement, thereby controlling changes in the human body shape.

In recent years, the SMPL model has evolved into SMPL-X [58], which facilitates the construction of three-dimensional models incorporating body, hand gestures, and facial expressions from single-frame RGB images, thereby enhancing its modeling capabilities to include hands and faces. The SMPL-X model introduces additional details such as facial expressions and hand gestures. Moreover, structured human meshes are utilized for feature extraction [59], providing comprehensive prior knowledge of the human body [60] that supports the synthesis of dynamic virtual humans. This knowledge includes essential posture and body shape data, serving as an initialization tool for constructing 4D virtual human models. Jiang et al. [60] employed point cloud sequences captured frame-by-frame to encode shape, posture, and motion, thereby constructing an initial parameterized human mesh model. An auxiliary encoder is then designed to handle fine details of clothing and hair, facilitating geometric integration to produce a complete human mesh.

Furthermore, Osman et al. [61] proposed STAR (Sparse Trained Articulated Human Body Regressor) as an alternative to SMPL, decomposing pose-related deformations into a set of spatially localized pose-corrective blend shapes, with pose deformations adjusted according to individual body shapes. The SMPL-generated models are vertex-based linear models and are currently the most widely applied parameterized human models. In contrast, models that adopt nonlinear strategies are exemplified by GHUM/GHUML (Generative 3D Human Shape and Articulated Pose Models) [62]. Based on the latent

space representation of Variational Autoencoders (VAE) [63], GHUM/GHUML relies on standard normalizing flows [64] for distribution approximation and inference calculations, generating a nonlinear parameter model that represents skeletal motion.

While mesh-based methods are capable of generating realistic human models with adequate simulation capabilities for articulated motion, they require the modeled object to possess a fixed topology. This requirement results in limited efficacy in modeling detailed structures such as clothing and hair. Current methodologies [65-70] enhance the SMPL model's mesh vertices with displacements to represent the geometry of clothing (SMPL+displacement, SMPL+D), but these enhancements are only suitable for simulating the surfaces of tight-fitting garments and struggle to accurately render clothing boundary details. Loose garments, in particular, present challenges due to identical skinning weights applied post-binding, leading to noticeable artifacts during movement. To address these limitations, Jiang et al. [71] utilized neural networks to establish independent skinning weights for specific types of clothing, allowing the garment mesh to remain independent of the SMPL model while overlaying it. This approach uses a displacement network to represent clothing deformation during motion, mitigating artifacts associated with binding certain clothing types to body mesh vertices, thus refining the mesh-based method for garment transfer solutions. Additionally, recent advancements in mesh-based methods have leveraged these human models as foundational elements for further modeling endeavors [72], exploring alternative, more precise solutions for modeling clothing and hair, or assisting in tasks such as motion inference.

Traditional parametric human body reconstruction methods typically use special devices to obtain dense three-dimensional point cloud data or depth data of the human body. Subsequently, they fit SCAPE parameters through point cloud registration and template deformation, to reconstruct the three-dimensional human body shape. In recent years, many researchers have used depth data captured by Kinect [2] depth cameras and the

SCAPE model to reconstruct three-dimensional human body shapes. Zhang et al. [73] captured multi-view local point cloud data of a rotating human body using a single Kinect camera, performed registration, and then fitted the point cloud using a method similar to SCAPE. Weiss et al. [74], using a single Kinect camera, captured multiple monocular depth images of a person moving in front of Kinect. They optimized the SCAPE body model by minimizing the registration error between the contour reprojection of the SCAPE model and the depth image contours. However, this method's solving process is highly time-consuming (requiring over 1 hour to reconstruct a human body) compared with other methods.

Zhao et al. [75] also proposed a parametric human body reconstruction method using a single Kinect. They first captured two depth images of the front and back of a person using Kinect, then reconstructed the upper body meshes from these two depths images and finally stitched them together. The reconstruction results of this method depend on the quality of the depth images captured by Kinect. However, due to Kinect's hardware limitations, the captured depth images often contain significant noise, severely impacting the reconstruction quality. In addition, some works do not rely on capturing dense three-dimensional point clouds or depth data using special devices for reconstruction input. Instead, they use other forms of data such as human body two-dimensional joint coordinates [77,79], human body contours [76,80,81], and human body descriptor parameters [72–86] to constrain parametric human body geometric shape reconstruction. Guan et al. [77] utilized manually annotated two-dimensional joint positions and the automatically segmented human body contour by GrabCut [87]. They optimized SCAPE parameters by minimizing the registration error between the rendered image and the human body contour using Shape from Shading (SFS).

SMPLify [79] introduced a convolutional neural network (CNN)-based human body two-dimensional pose estimation model. They optimized SMPL parameters (including body shape and pose parameters) by minimizing the reprojection registration error between

the synthesized three-dimensional body pose and the detected two-dimensional joint keypoints. Additionally, they introduced a constraint to reduce the ambiguity in lifting from two dimensions to three dimensions. However, this method does not effectively constrain body shape and is prone to local optima, leading to reconstruction failures. Lassner et al. [88], building on SMPLify, added more constraints from human body landmark points (91 points), obtaining more accurate pose reconstruction results. They also proposed using a Random Forest model to learn the mapping relationship between human body contours and SMPL body shape parameters. However, the predicted quality of human body contours was poor, significantly impacting shape prediction results. In recent years, parametric human body shape reconstruction methods based on deep learning have become popular [89]. Dibra et al. [76] were among the first to use a Convolutional Neural Network (CNN) to estimate human body shape parameters. They directly used a specific view mask of a standing human body as input to the CNN and regressed the SCAPE shape parameters. Compared to manually designed features, CNNs can automatically extract shape features, resulting in more accurate shape predictions.

Subsequently, Dibra et al. [90] further improved the accuracy of shape predictions. They first learned a feature latent space describing the same shape under different views in a fixed pose, then learned a regression model from this latent space to shape parameters. This method can reliably predict shape parameters from human body mask images in other views. Single-view human body mask images often lack some shape information, such as the beer belly of a male, which cannot be displayed on the frontal mask image. To address this issue, Ji et al. [81] designed a novel dual-stream network structure, simultaneously using frontal and side human body masks as input to predict SCAPE shape parameters. Besides predicting human body shape, many researchers use deep learning methods to directly estimate human body shape and pose from images [78,10,92–93], videos [94,95]. HMR [10] added the reprojection registration error of human body keypoints to the loss function, supervising the pose parameters and body shape parameters of SMPL. HMR borrowed ideas from Generative Adversarial Networks (GAN)

[96] and introduced a discriminator into the loss function to supervise the legality of predicting human body parameters. However, this method did not effectively supervise human body shape, resulting in predicted bodies closer to average body shapes, with significant differences in body pose compared to the input images. Pavlakos et al. [93] proposed decoupling pose parameters and shape parameters into two sub-problems for prediction. They used the predicted two-dimensional keypoint heatmaps and human body contours to separately regress pose parameters and shape parameters. Recently, Xu et al. [78] innovatively introduced dense reprojection errors of human body mesh vertices into the loss function. They used the IUV map predicted by DensePose [97] (representing the correspondence between dense mesh vertices and image pixels) as input, regressed the human body mesh, then calculated registration errors between the predicted IUV map and the input IUV map through a Differential Renderer. This method achieved more accurate reconstruction results in both pose and shape.

## 2.4 Image-based Digital Human Reconstruction Method

Image-based methods focus on reconstructing the three-dimensional structure of a scene from multiple two-dimensional images, enabling the development of networks that facilitate image-to-image translation. This approach is particularly prevalent in facial reconstruction. Facial reconstruction pertains to the development of 3D Morphable Models (3DMM) [99], which necessitate extensive datasets encompassing variations in illumination, poses, and expressions. These models are categorized into linear and nonlinear types. Linear 3DMMs are characterized by their low-dimensionality, employing Principal Component Analysis (PCA) to capture texture and facial shape features in a low-dimensional space. Alternatively, learning networks can infer linear facial models, generating realistic data for physical rendering, such as reflectance and normals. However, linear 3DMMs constructed using PCA often fail to reproduce high-frequency details of human texture and geometry, demonstrating limited generalization capabilities for image sets in natural scenes [100].

Nonlinear deformable facial models, generated through unsupervised or weakly supervised learning [101], can process large quantities of images from natural scenes. However, these models are not well-suited for re-illuminating portraits and animations, as environmental lighting conditions and expression data are embedded within the output texture images. A prevalent approach involves using deep learning-based image post-processing to infer linear facial models for re-illumination rendering components [102].

Generative Adversarial Networks (GANs) have significantly improved modeling capabilities through the dynamic interplay between generator and discriminator networks. GANs are now widely utilized for texture extraction in facial reconstruction, often in conjunction with 3DMM to complete the facial reconstruction task. The strength of GANs lies in their ability to handle high-resolution images and facilitate image-to-image translation. In the domain of image processing, the GAN generator actively learns facial features, utilizing Gaussian noise to control variations in facial details, thereby generating synthetic images designed to "deceive" the discriminator. The discriminator, in turn, continuously enhances its ability to distinguish between real and synthetic images. This adversarial training process improves the overall performance and realism of the generated models.

GANFIT (Generative Adversarial Network Fitting) [103] enables the reconstruction of high-quality texture and shape data from a single image taken in a natural scene. This is achieved by training a GAN to produce large-scale, high-resolution texture data while preserving identity features. However, similar to previous methods, the texture data generated by GANFIT incorporates lighting conditions, preventing the reconstruction of high-frequency normal maps and specular reflection data necessary for direct rendering.

Building on the texture and shape data obtained from GANFIT, AvatarMe (Realistically Renderable 3D Facial Reconstruction) [100] processes the input image through a

nonlinear 3DMM. It employs a texture illumination network to extract diffuse reflection data by removing the lighting effects. The generated reliable diffuse reflection data is then used by multiple image translation networks to estimate specular reflection, specular normals, and diffuse normal which are key components for photorealistic rendering. The resulting facial model supports re-illumination, allowing for the simulation of facial appearance under different lighting conditions, making it directly suitable for rendering. Despite the large training dataset used by AvatarMe, it lacks sufficient data for darker skin tones, leading to suboptimal performance in reconstructing faces of individuals with dark skin. Additionally, the method is dependent on the resolution and lighting conditions of the input image.

StyleGAN [104] introduces a style-based GAN algorithm that facilitates the unsupervised and self-learning separation of high-level attributes from input images and generated images, enabling intuitive control over synthesis. In StyleGAN, the style represents the main attributes of the face, such as expression, orientation, and hairstyle. Similar to traditional generator networks, StyleGAN's generator network progressively increases the image resolution at each layer, exhibiting a growth pattern. A notable improvement of StyleGAN over traditional GANs is the decoupling of input feature z, producing an intermediate vector w that is less influenced by the distribution of the training data. This reduces the correlation of specific features with elements in the vector and minimizes the impact of input noise on other features at each layer of the generator network.

StyleRig (Rigging StyleGAN for 3D Control) [105] integrates StyleGAN with 3DMM to achieve facial binding, allowing for the control of semantic parameters such as facial expressions to facilitate facial transformations. However, the transformation capability is highly dependent on the 3DMM and does not allow explicit control over scene attributes not interpreted by the 3DMM.

StyleGAN2 [106] is an improved version of StyleGAN, addressing issues such as water

droplet artifacts and incorporating residual networks to directly map low-resolution features to the final generated results. Luo et al [107] utilized the StyleGAN2 architecture to train a 3DMM that includes 3D geometry and reflectance textures. This model applies perceptual refinement to rendered faces, overcoming challenges posed by extreme lighting conditions, and generating high-resolution, standardized faces with neutral expressions. However, incomplete training data can hinder the model's ability to fully disentangle lighting information from skin tone, resulting in an imperfect separation of lighting and expression information from the face. Beyond facial reconstruction, GANs have also been applied in the domain of virtual clothing [108-109], where they simulate garments with different topologies and use designed mapping networks to position the clothing on various human models.

Image-based human avatars can achieve high resolution, offering visual quality sufficient to "deceive" viewers. However, purely image-based avatars rely on trained image data and are typically suited for frontal face images. Additionally, due to the lack of 3D spatial information, these models struggle to maintain consistency across multiple viewpoints.

## 2.5 Voxel-based Digital Human Reconstruction Method

Voxel-based methods for constructing virtual humans can generate models consistent from multiple viewpoints, requiring operations such as three-dimensional spatial voxelization and two-dimensional projection of three-dimensional objects. The resulting reconstructed images maintain texture and resolution consistent with the original images. Ideally, voxel reconstruction algorithms should possess four key attributes: range uncertainty representation, independent incremental and sequential updates, spatiotemporal efficiency, and unrestricted topology types [110]. Early voxel-based reconstruction methods relied on 3D scan data [110-112], necessitating specific experimental equipment. In recent years, voxel-based reconstruction efforts have been aimed to integrate voxelization concepts into the reconstruction process.

Deep Voxels [113] first extract 2D features from source images and introduce a voxel representation as a fixed viewpoint 3D feature grid. This method elevates 2D features to a 3D space for observation and integration into the feature grid, performing 3D spatial reasoning and 2D feature synthesis sequentially, notably without requiring 3D supervised learning. Ma et al. [114] collects contextual key feature information from all voxels to update the current voxel joint features, constraining limb lengths to estimate 3D poses from a single image. Deep Human [115] voxelizes a human parametric model and proposes an image-guided volume-to-volume transformation network, utilizing multi-scale volume transformations to combine knowledge from 3D volumes and 2D images.

Storing 3D spatial point information incurs high memory costs, and increasing precision significantly raises computational time. Therefore, a feasible voxel-based construction method must address how to scale images to higher resolutions to handle finer details, such as wrinkles. To circumvent the low-resolution issue due to memory constraints, Deep Voxels employs a local resolution exchange strategy, which, however, sacrifices data utilization efficiency and may result in detail loss. Neural Volumes [116] proposes warping fields, allocating storage space preferentially to regions contributing more significantly to the synthesized image. Nonetheless, achieving the fidelity realized by traditional texture mesh surfaces requires further advancements.

## 2.6 Implicit Representations-based Digital Human Reconstruction Method

Implicit representations intuitively define a continuous scalar function in three-dimensional space to represent surfaces, and recent advancements have integrated neural networks for implicit scene representation. In the field of virtual character synthesis, implicit functions use local feature information provided by context to infer overall shape information [117-119], with NeRF (Neural Radiance Fields) often employed for multi-view synthesis. Compared to voxel-based methods, implicit representations are more memory-efficient; and unlike image-based methods, implicit representations can also

infer colors in unobservable areas.

TextureFields [98] proposed a continuous 3D function parameterized by a regression-based neural network to represent texture fields. This approach is independent of the shape representation of 3D objects, learning to transfer the texture of example images onto source meshes to synthesize new views. Sitzamann et al. [120] introduced a continuous, 3D structure-aware scene representation that defines surfaces through a learned directed distance field, allowing for the geometric and appearance modeling of 3D scenes without 3D supervision, while maintaining multi-view consistency. BANMo (Builder of Animatable 3D Neural Models) [121] utilizes implicit functions to implicitly represent objects, combining the concept of NeRF with MLP (Multi-Layer Perceptron) networks to provide the color and volume density of 3D spatial points, alongside training-derived canonical embeddings. These canonical embeddings encode semantic information of 3D spatial points, registering pixel observations across different temporal instances. In this approach, an MLP calculates the Signed Distance Function (SDF) from points to the surface, yielding 3D shapes, and continuous surface embeddings (CSE) [122] initialize pixel embeddings to generate corresponding pixel features. Compared to the parametric models established by SMPL, BANMo requires less data; and compared to NeRF, BANMo is more suited for representing objects with larger movements.

The core concept behind SDF and directed distance fields is to represent an object's surface through a volumetric field, calculating the shortest distance from points within the field to the object's surface. This distance is zero on the surface, negative inside the object, and positive outside. To enhance representation efficiency, DeepSDF [123] combines MLP to achieve a continuous SDF representation of shapes, making it a commonly used implicit representation today. SDFs are widely applied in non-rigid reconstruction [116], with MLP optimization improving the performance of non-rigid reconstruction and deformation tracking tasks [124]. Variants of SDF also provide robust prior knowledge for multimodal 3D reconstruction tasks [125]. Continuous Surface

Embeddings (CSE) predict embedding vectors for corresponding vertices in an object mesh for each pixel in a 2D image, establishing a dense correspondence with the 3D object's geometry.

PIFu (Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization) [126] and PIFuHD [127] introduce a locally aligned pixel-aligned implicit function for 2D images. Unlike other implicit representation methods, PIFu focuses on pixel-level features, maintaining alignment in the output image. In essence, this implicit function projects any given 3D vertex according to camera parameters, obtaining corresponding 2D positional and depth information while learning the feature vector of that point to preserve local details and infer information for occluded regions. Representing the surface as a level set (an implicit representation of curve), PIFu can be extended to multi-image and multi-view inputs, providing a complete, high-resolution 3D model's surface and texture, effectively performing complex clothed human modeling tasks [128].

Implicit representation methods, exemplified by PIFu, offer continuity and support for generating geometries with arbitrary topologies at high memory efficiency, extending even to color synthesis. This memory-efficient nature of implicit representations is used to enhance model construction methods constrained by memory limitations, and their excellent inference capabilities address the shortcomings of human parametric models in clothing modeling.

These implicit representation methods provide specific solutions to occlusion problems in real-life scenarios, requiring advanced inferential capabilities from implicit functions. However, the implicit representation of shapes is limited by the absence of mesh topology, skeleton, and skinning weight structure information, making it incapable of exhibiting new poses and only controlling the shape of dressed avatars from fixed viewpoints. To address this, MVP-Human [129] uses 3D scanning technology to acquire three-dimensional information. BANMo combines neural skinning models and utilizes explicit

3D Gaussian ellipsoids that move with the skeleton to adjust weights, allowing for extensive articulated transformations and incorporating NeRF for multi-view synthesis to present new perspectives. Additionally, NASA (Neural Articulated Shape Approximation) [130] proposes a pose-conditioned implicit occupancy function to replace polygonal human meshes, representing articulated deformable human objects. For single-view facial reconstruction, JIFF (Jointly-aligned Implicit Face Function) [131] leverages shape priors provided by 3DMM, combining spatially aligned 3D features and pixel-aligned 2D features to jointly predict implicit facial functions, thereby improving the quality of implicit functions in facial reconstruction applications.

## 2.7 Mixed-based Digital Human Reconstruction Method

In recent years, significant advancements have been made in the improvement of the aforementioned synthesis technologies, largely due to the widespread application of deep learning, which has enhanced training efficiency. Researchers have also attempted to combine the strengths of various methods to complement each other, aiming to improve the quality of virtual humans. The improvement approaches across different construction methods share similarities, which can be summarized as follows.

In mesh-based training, the parametric naked human models based on the SMPL model remain one of the mainstream methods. Research has focused on improving texture generation techniques to address the limitations in modeling clothing and hair. Image-based methods aim to incorporate 3D spatial information into models and enhance data utilization efficiency. Voxel-based methods primarily focus on reducing memory costs. The performance of implicit representation methods can be improved through unconstrained observation environments and higher resolution training data. By hybridizing these methods, it is possible to complement the shortcomings of each technical approach, effectively improving model quality. Based on these approaches, this section proposes hybrid synthesis methods as an independent category and introduces some existing hybrid construction techniques to provide optimization insights.

Implicit 3D representations possess strong expressiveness and can better capture and reconstruct the shape and appearance of clothed humans when combined with learnable parametric models like SMPL. For instance, a study [132] proposes jointly learning two implicit functions to estimate dressed humans and body part labels, linking implicit functions with parametric models. The inner surface of the body covered by clothing is simulated by SMPL, constrained by the predicted body part information. The inner surface is registered to the outer surface using SMPL+D, optimizing each vertex's displacement D to fit the external implicit representation. The designed implicit functions extend point positions to three categories: inside the body, between the body and clothing, and outside the body.

SCANimate [133] scans dressed humans in 3D and converts them into parameter-driven virtual avatars. It utilizes SMPL to obtain a parametric 3D human model and combines a weakly supervised model for pose correction, designing a local pose-aware implicit function to represent the human model and simulate clothing deformation during movement, thus generating new poses. Experiments demonstrate that the structural information provided by SMPL enhances the performance of implicit representations, improving the generalization capability of human poses. However, SCANimate's model representation is suited for tight-fitting clothing with similar topology to the body and is not applicable to looser garments. Additionally, the constructed model is deterministic, meaning the same pose results in the same degree of clothing wrinkles, limiting its ability to predict all possible variations in clothing deformation.

ICON (Implicit Clothed-humans Obtained from Normals) [134] integrates the SMPL-X model with a custom-designed normal prediction network for iterative optimization. The inferred dressed human normal maps are used to regress the implicit 3D surface of dressed humans, utilizing local features unaffected by global pose transformations for the implicit 3D reconstruction task. ICON can recover 3D dressed human figures from a single image and is applicable to virtual character construction in natural scenes. It can

also be combined with SCANimate to generate dynamic avatars. However, since ICON is trained using orthogonal views, which describe 3D properties from 2D projection images, the perspective effect may not be ideal, leading to potential inconsistencies in limb representation.

Combining various approaches demonstrates that the SMPL parametric model supports controllable deformations of human movements, while NeRF, as an advanced scene representation method, can effectively predict the color and volumetric density of spatial points, facilitating multi-view synthesis. Therefore, combining SMPL with NeRF can provide control over the shape of human models and clothing [134-135]. Xu et al [136] proposes a surface-guided neural radiance field to synthesize controllable human characters, capable of reconstructing a 3D human model based on a small amount of multi-view video and prior knowledge from the SMPL model.

Parametric human models can effectively represent joint movements and integrating them with advanced texture synthesis methods can address the modeling limitations for clothing and hair. The application of GANs can generate high-resolution results, including high-quality textures [137-138]. StylePeople [138] introduces a neural dressing model that utilizes styleGAN2 to learn the neural textures of input images, overlaying them onto the naked human model generated by SMPL-X to create high-quality dressed virtual characters. StylePeople uses a fully convolutional network to generate body part coordinates and stacks assigned to body parts, sampling and mapping body textures to generate RGB images with the weights specified by the stacks. StylePeople extends styleGAN2 from facial reconstruction to full-body character construction, effectively simulating hair and clothing textures, thus addressing the limitations of the SMPL model. However, like image-based methods, the modeling quality depends on a large amount of training data and requires high data utilization efficiency.

Voxel-based methods can effectively present visual effects from multiple angles, while

implicit representation methods are memory efficient and can significantly enhance the resolution of voxel-based outputs. Pixel-Aligned Volumetric Avatars (PVA) [139] combine volume rendering and neural radiance fields for image rendering, leveraging pixel-aligned features introduced by PIFu to retain high-frequency details. This approach is used to adjust parameters of multi-identity neural radiance fields, employing an MLP to transform spatial positions and pixel-aligned features into colors and occupancy. This method mitigates the memory limitations inherent in voxelization. PVA can generate high-fidelity virtual avatars from a small amount of sample data, but it lacks the capability to capture lighting conditions and background variations, limiting its application to natural scene images.

Gafni et al [140] uses implicit functions to represent the geometric appearance of the face and hair with neural radiance fields, combining volume rendering to restore the volumetric feel of the hair and achieve dynamic head changes. Compared to traditional voxel-based volume rendering methods, the integration of neural scene representation networks results in more compact volume rendering, further improving the resolution of the rendered images. However, this experiment is limited to dynamic head image generation and does not extend to full-body dynamic representation, which would involve more complex volumetric models and lighting calculations. This approach collects both 2D and 3D features of the input data, effectively improving data utilization. Such hybrid solutions leverage the structural regularity of human meshes, the expressive power, and memory efficiency of implicit functions, while also achieving multi-view effects.

Zheng et al. [141] employs a non-parametric deep implicit field to represent surfaces, with the SMPL model providing parametric human body regularization. It collects pixel-level and voxel-level features, binding each 3D point to the corresponding value of the implicit function. DeepMultiCap [142] integrates implicit functions with poses and voxelized grids to recover local details from image pixels, enhancing robustness to pose variations. S3 [143] voxelizes input point cloud data into a voxel grid to represent

volumetric features. It combines 2D image feature extraction to represent the shape, pose, and skinning weights of pedestrians as neural implicit functions learned directly from the data, constructing dynamic human models.

These advancements highlight a trend towards combining multiple synthesis techniques, each compensating for the others' limitations, thereby improving the overall quality and versatility of virtual human models. This hybrid approach, drawing from mesh, voxel, and implicit function methods, provides a robust framework for future developments in high-fidelity virtual character construction.

## 2.8 Human Body Pose Reconstruction

Figure 2.1: Optical motion capture system Vicon [144]

Three-dimensional human pose reconstruction typically involves utilizing external devices to recreate the three-dimensional posture of a person. In comparison to dense geometric body shapes, the human skeleton serves as a compact representation of body posture. The industry currently has relatively mature solutions for three-dimensional pose reconstruction, such as contact-based motion capture systems, exemplified by the renowned optical motion capture system Vicon (Figure 2.1). In this system, specially

designed optical markers are attached to key points on the human body, such as joints. Multiple specialized motion capture cameras can detect these markers in real-time from different angles. Subsequently, the spatial coordinates of the markers are precisely calculated using the principles of triangulation. The Inverse Kinematics (IK) algorithm is then applied to compute the joint angles of the human skeleton. Due to scene and equipment limitations, as well as high costs, contact-based motion capture is challenging for ordinary consumers to use. Consequently, researchers have turned their attention to low-cost, non-contact, and markerless motion reconstruction technologies.

Based on RGB-D, three-dimensional pose reconstruction methods can be divided into two categories: discriminative methods and generative methods. Discriminative methods typically attempt to directly infer the three-dimensional human pose from depth images. Some work in this category tries to extract features corresponding to joint positions from depth images. For example, Plagemann et al. [145] use geodesic extrema to identify salient points in the human body and then detect three-dimensional joint positions using local shape descriptors. Other discriminative methods rely on classifiers or regressors trained offline. Shotton et al. [146] first trained a RF (random forest) classifier with a large number of samples to segment different body parts from depth images, and then they used mean-shift algorithm to estimate joint positions. This method requires minimal computational effort for real-time predictions and has been integrated into the Kinect SDK for real-time three-dimensional pose reconstruction. Taylor et al. [147] used a RF method to predict depth pixel regions belonging to human joints and then utilized it for pose optimization. Discriminative methods do not rely on tracking, which can reduce cumulative errors and naturally handle fast movements.

In contrast to discriminative methods, generative methods match observed data using deformed parameterized or non-parameterized templates. Ganapathi et al. [148] used a Dynamic Bayesian Network (DBN) to model motion states and inferred three-dimensional poses within a Maximum a Posterior (MAP) framework. This method requires prior

knowledge of the body shape and cannot effectively handle rapid body movements. Subsequently, Ganapathi et al. [149] improved this method by incorporating an extended ICP measurement model and free space constraints. The new method dynamically adjusts the size of the parameterized body template to fit the captured depth data. Due to hardware limitations, RGB-D-based pose reconstruction methods are susceptible to depth map noise and are applicable only in relatively close-range scenarios.

Thanks to the emergence of large-scale video datasets annotated with three-dimensional human pose (such as Human3.6M [150], Human-Eva [151]), deep learning-based three-dimensional pose reconstruction methods have rapidly developed. They directly utilize deep learning models to extract three-dimensional human joint positions from images or videos [152–158]. Li et al. [152] were the first to introduce deep learning into three-dimensional pose estimation. They designed a multi-task convolutional neural network (CNN) that includes detection and regression, directly learning features from images to regress three-dimensional joint positions, surpassing previous methods relying on manually designed features. Pavlakos et al. [154] proposed a voxel heatmap to describe the likelihood of human joint positions in three-dimensional voxel space and used a coarse-to-fine cascaded strategy to progressively refine the prediction of voxel heatmaps, achieving excellent pose reconstruction accuracy. However, this voxel representation often faces significant storage and computational overhead. Recently, [159] effectively addressed this issue using an encoder-decoder approach.

In addition to directly predicting joint three-dimensional positions, some work predicts bone orientations [162,163], joint angles [164], bone vectors [165,166], and more. The mentioned works all employ strong supervision for training, and since the training data are collected in controlled environments, models trained in this manner usually struggle to generalize to natural scenes. To enhance the model's generalization ability, some work attempts to use weakly supervised methods to supervise images in natural scenes, such as using domain discriminators [167], skeleton length priors [168], and more.

Another category of three-dimensional pose estimation methods treats two-dimensional human pose as an intermediate representation. Initially, two-dimensional human joint positions are obtained in images using manual annotations or automatic detection [169–172]. Subsequently, they are elevated to three-dimensional space through regression methods [155,160,173] or model fitting [174]. Martinez et al. [160] designed a simple but effective fully connected network structure, taking two-dimensional joint positions as input and outputting three-dimensional joint positions. Later, Zhao et al. [173] proposed using semantic graph convolutional layer modules to capture topological correlations between human joints (such as human symmetry), further improving the accuracy of three-dimensional pose reconstruction. However, mapping from two-dimensional pose to three-dimensional pose itself is an ambiguous problem because multiple three-dimensional poses can project to the same two-dimensional pose [175]. Recent works attempt to incorporate more prior knowledge (depth information) to alleviate ambiguity [176–178].

The aforementioned works belong to discriminative models, and the predicted three-dimensional joint positions may not adhere to anatomical constraints (such as symmetry, unreasonable bone length proportions) or kinematic constraints (joint angles exceeding limits). Mehta et al. [161] fitted a template of the human skeleton to the predicted two-dimensional joint positions and three-dimensional joint positions and proposed the first real-time three-dimensional pose reconstruction system VNect based on an RGB camera, achieving accurate pose reconstruction results.

## 2.9 Convolution Neural Network

### 2.9.1 Basic theory of CNN

Research into the visual mechanisms of mammals, particularly through the field of visual neuroscience, has revealed that the human visual system operates in a layered and hierarchical manner when processing and recognizing images [179]. This process begins

when a visual scene is captured by the human eye, where light signals from the scene are converted into electrical signals by the retina. These electrical signals are then transmitted to the visual cortex, the primary visual processing center of the human brain.

A landmark study by David Hubel and Torsten Wiesel in 1959 [180] provided significant insights into how the visual cortex processes these signals. By inserting electrodes into the primary visual cortex (V1) of a cat and presenting the cat with light bars of varying shapes, positions, and orientations, they were able to measure the neuronal responses to these stimuli. Their experiments demonstrated that neurons in the V1 cortex respond most strongly when the light bars are positioned at specific locations and angles, indicating that different neurons have distinct preferences for certain spatial locations and orientations.

Further research has established that the visual cortex is organized into a multi-layered structure. The electrical signals from the retina first reach the primary visual cortex, or V1, where neurons are highly sensitive to particular details of the visual stimuli, such as edges and orientations. From V1, the processed signals are transmitted to the secondary visual cortex (V2), where more complex features such as edges and contours are integrated into simple shapes. The processing then continues to the V4 cortex, which is particularly attuned to color information. Ultimately, the representation of complex objects is formed in the inferior temporal cortex, where higher-level visual processing occurs, allowing for the recognition of complex shapes and objects.

This hierarchical processing model underscores the complexity and efficiency of the mammalian visual system, demonstrating how the brain decomposes and reconstructs visual information at various levels to achieve detailed and accurate perception of the surrounding environment.

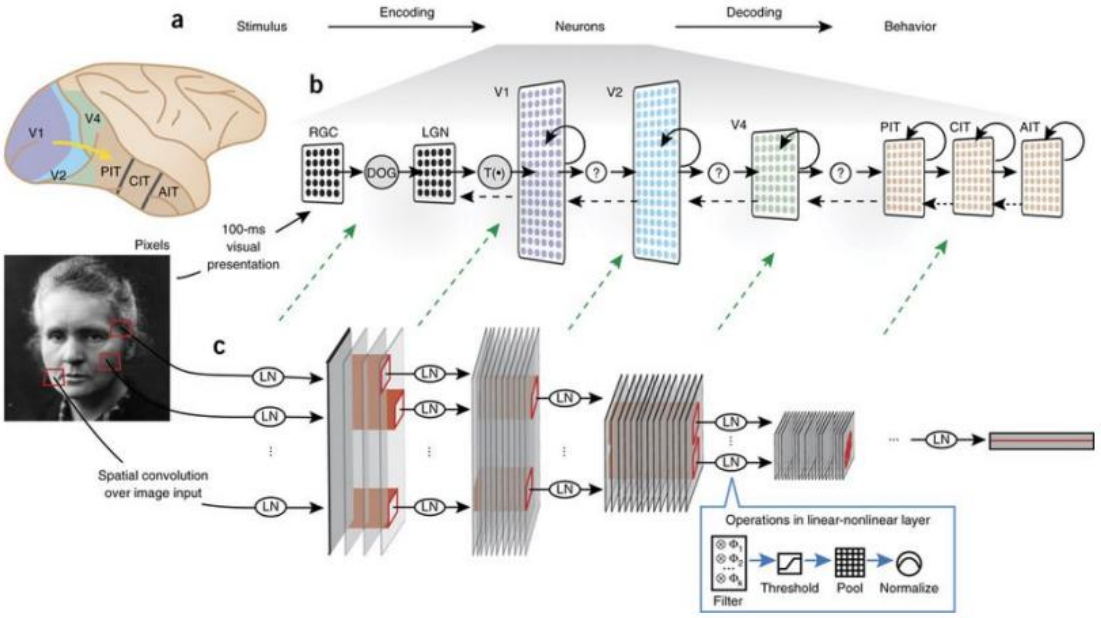



Figure 2.2: Yamins at al. 2016. Visual cortex mechanism and CNN

As illustrated in Figure 2.2, convolutional neural networks (CNNs) can be viewed as a computational model that emulates the hierarchical processing mechanisms of the visual cortex in biological systems [181]. A CNN is composed of multiple convolutional layers, each consisting of several convolutional kernels. These kernels systematically scan all the pixels of an input image, producing a set of matrices known as feature maps, which represent various features detected in the image.

In the early layers of the network, the convolutional layers capture local and detailed information from the image. These layers are characterized by a small receptive field, meaning that each pixel in the output feature map corresponds to a small region of the input image. As the signal progresses through the network, the receptive field of the convolutional layers gradually increases, enabling the network to capture more complex and abstract features. This hierarchical process allows the network to build progressively more sophisticated representations of the input image at multiple scales.

To effectively understand the function of convolution in the context of neural networks, it is essential to have a foundational understanding of signal processing concepts. One

key concept is the representation of a sine wave, which is fundamentally defined by two parameters: amplitude and frequency.

A sine wave is commonly depicted in a time-magnitude coordinate system, where time is represented on the horizontal axis and the magnitude (or amplitude) on the vertical axis. However, this representation is not the only way to describe a sine wave. It can also be transformed into a frequency-magnitude coordinate system, where the horizontal axis represents frequency, and the vertical axis represents the amplitude's intensity. This transformation between time and frequency domains is accomplished through the Fourier Transform, a mathematical tool that maps a signal from the time domain to the frequency domain, as illustrated in Figure 2.3.

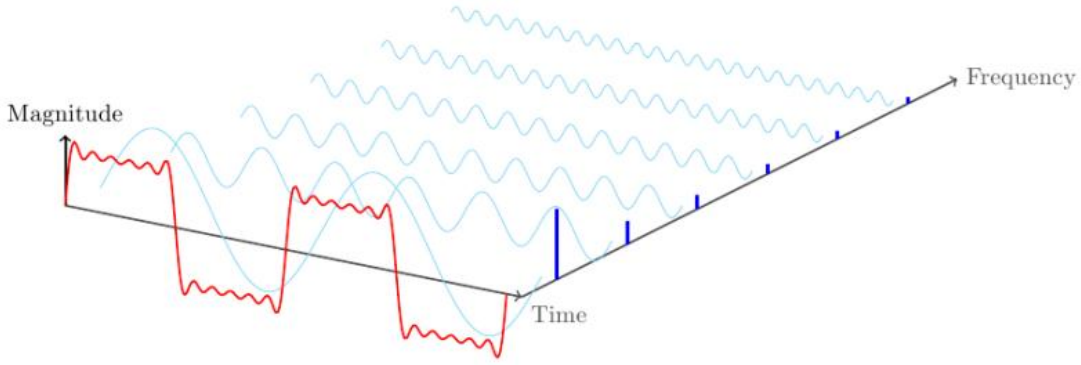



Figure 2.3: Fourier transformation

The convolution between two functions is defined as:

$$x(t) * y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \quad (2.1)$$

Considering Fourier Transform, it can be shown that the convolution has a property which is:

$$x(t) * y(t) \leftrightarrow X(w)Y(w) \quad (2.2)$$

The relationship between time-domain signals and their spectral representations, as described by the Convolution Theorem, is a fundamental concept in signal processing. It states that the convolution of two time-domain signals corresponds to the pointwise

multiplication of their spectral (frequency-domain) representations. This principle implies that when a convolution is performed between a filter and a complex signal, the resulting signal's spectral content will reflect the product of their respective frequency components.

For example, if a filter that includes certain frequency components is convolved with a complex signal that contains additional frequency components that do not present in the filter, the resulting signal will no longer include these excluded frequencies. This is the basis for the function of various types of filters. A low-pass filter, which allows only low-frequency components to pass through while attenuating higher frequencies, effectively removes high-frequency content from the signal. Conversely, a high-pass filter retains only high-frequency components, filtering out the lower frequencies.

The Fourier Transform, a powerful tool in signal processing, can also be applied to two-dimensional functions, such as images. Similar to its one-dimensional counterpart, the 2D Fourier Transform decomposes an image into its frequency components. In this context, a grayscale image can be represented as a discrete two-dimensional function $f(x, y)$, where $x$ and $y$ correspond to the spatial coordinates of the image's pixels, and $f(x, y)$ represents the luminance value (intensity) of each pixel.

In analogy to 1D signals, the high-frequency components of an image correspond to regions where the pixel values change rapidly (e.g., edges or fine details), while low-frequency components correspond to areas with slowly varying pixel values (e.g., smooth gradients or uniform regions). As illustrated in Figure 2.4, when a high-pass filter is convolved with an image, the result highlights areas where the luminance changes rapidly, effectively preserving edges and fine details. In contrast, a low-pass filter applied to an image will smooth out these rapid changes, removing fine details and retaining only the slowly varying components of the image.

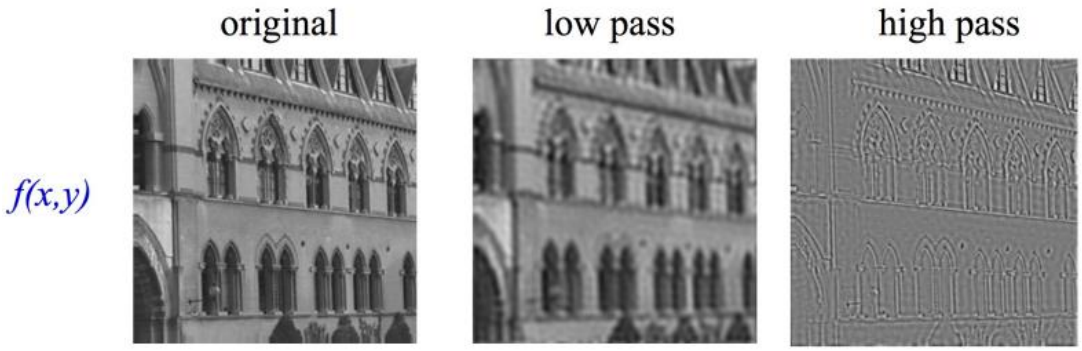



Figure 2.4: Applying different filters.

Thus, when an image is convolved with different filters, the outcome is a set of images that emphasize different frequency components, corresponding to various levels of detail and abstraction. This process lies at the core of convolution operations in image processing and is integral to the functionality of convolutional neural networks. By selectively filtering different frequency components, convolution enables the extraction of relevant features, allowing for a nuanced analysis and interpretation of visual data. This filtering mechanism is essential for tasks such as edge detection, texture analysis, and feature extraction in computer vision and image processing applications.

### 2.9.2 Basic Composition of CNN

In a Convolutional Neural Network (CNN), convolutional layers consist of sets of convolution kernels that interact with the input image to extract various features. The process of convolution, as illustrated in Figure 2.5, involves each convolution kernel sliding over the image matrix from top to bottom and from left to right. During this sliding operation, the elements of the convolution kernel matrix are multiplied by the corresponding elements of the image matrix covered by the kernel, and the results are summed to produce a single value. This operation is repeated across the entire image, resulting in an output matrix that represents the convolution result.

Typically, after convolution, the size of the resulting image is reduced compared to the

original input image. To maintain the same size between the input and output images, padding is often applied. Padding involves adding zero-valued pixels to the periphery of the original image before the convolution operation. A simple example of padding with a padding value of 1 is also shown in Figure 2.5. Additionally, as the convolution kernel slides across the image, the stride (the step size in both horizontal and vertical directions) is usually set to 1, although other stride values can be used depending on the desired level of down sampling.

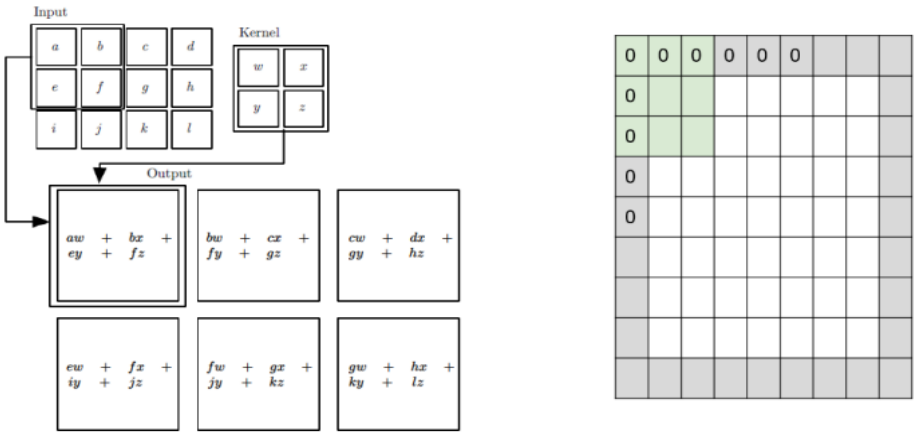



Figure 2.5: Convolution kernel sliding and padding

When dealing with multi-channel inputs, such as RGB images, a single convolution kernel will convolve with all the input channels. The resulting matrices from each channel are then summed up to produce a single output matrix. Consequently, each convolution kernel generates a corresponding feature map, and the number of resulting feature maps is equal to the number of convolution kernels used in the layer. This process enables CNN to capture and represent different aspects of the input image, such as edges, textures, and patterns, across multiple layers of abstraction.

Convolution operations within a neural network represent a linear transformation of the input image based on the filter responses. Specifically, the convolution process calculates a weighted sum of the input pixel values, where the weights are determined by the convolutional kernel. This linear nature of convolutional layers limits their ability to

capture and model complex patterns in the data, as they can only effectively simulate linear relationships between input and output.

To address this limitation and enhance the network's capacity to model more intricate and non-linear relationships, convolutional layers are typically followed by non-linear activation functions. These activation functions introduce non-linearity into the network, allowing it to approximate a broader range of functions beyond simple linear mappings. Consequently, the neural network's ability to model complex features and patterns is significantly improved.

Incorporating non-linear activation functions is crucial for the effective performance of a neural network. Without these non-linear layers, a network composed solely of convolutional layers would essentially perform linear transformations, making it insufficient for tasks that require capturing intricate relationships within the data. Therefore, the depth of a neural network becomes meaningful primarily because each additional non-linear layer enhances the network's ability to approximate increasingly complex functions.

In contrast, a purely linear neural network—regardless of its depth—would not achieve significant performance improvements because adding more linear layers would not increase the network's representational power. The depth in such a network would merely aggregate linear transformations, resulting in no additional benefit.

Figure 2.6 illustrates several different activation functions, such as ReLU (Rectified Linear Unit), sigmoid, and tanh. Each of these functions contributes differently to the non-linear modeling capabilities of the network, influencing how well the network can fit complex patterns in the input data. By combining convolutional layers with appropriate non-linear activation functions, neural networks can leverage both linear and non-linear transformations to achieve superior performance across a variety of tasks.

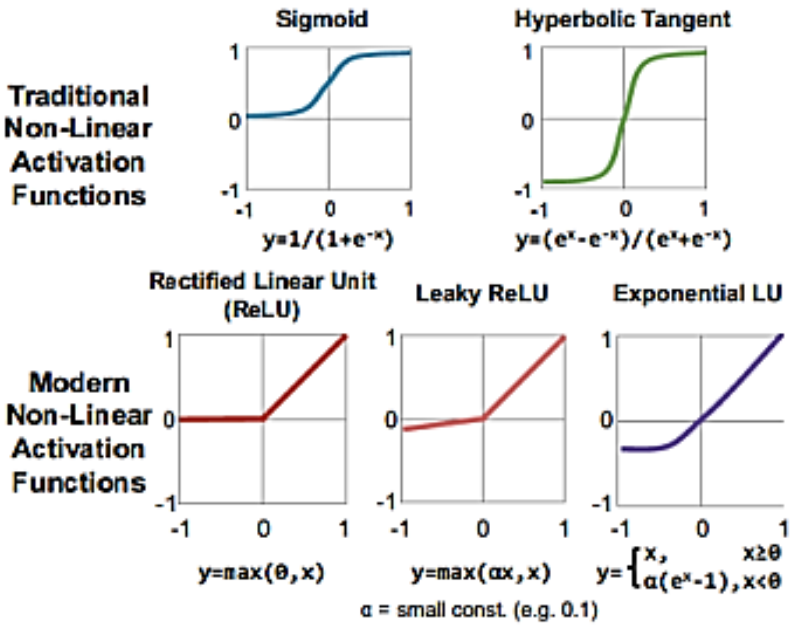



Figure 2.6: Different activation functions.

Through the convolution operation, both dimensionality reduction and feature extraction of the input image are achieved. However, despite these processes, the dimensionality of the resulting feature maps can still be quite high. High-dimensional feature maps pose significant challenges, including increased computational cost and a higher risk of overfitting. To address these issues, down-sampling techniques, commonly referred to as pooling, are employed.

Pooling layers are introduced to further reduce the dimensionality of feature maps while retaining essential information. Compared to convolutional layers, pooling layers are computationally less intensive. Typically, the input to a pooling layer consists of the feature maps that have been processed through activation functions.

The pooling operation involves compressing the sub-matrices within the input feature map matrix. This compression is achieved by applying a pooling function, such as max pooling or average pooling, which reduces the spatial dimensions of the input while preserving the most salient features. By reducing the dimensionality, pooling not only

decreases the computational burden but also helps mitigate the risk of overfitting by simplifying the feature representation and focusing on the most significant patterns in the data.

For example, for 2x2 pooling, every 2*2 sub-matrix which contains 4 parameters will turn to only one element. Considering 3x3 pooling process, then every nine parameters of 3x3 size sub-matrix will turn to one element. Hence, the dimension of an input matrix will be reduced to a much smaller size. In addition to reducing the image size, another benefit of down-sampling is translational, rotational invariance, because the output value is calculated from a region of the image and is not sensitive to translation and rotation. With the increasing of the pooling size, more detailed information could be lost, hence, usually a CNN network selects a small pooling size but with many combinations of convolution layers and pooling layers. In order to implement the pooling process, usually there are two types of pooling, one is max pooling, the other one is average pooling. Max pooling is to select the maximum value of the pooling area as the pooling result. Average pooling is to compute the average value of whole pooling area and take the result as the pooling result. Figure 2.7 shows an example uses a pooling method that takes the maximum value. At the same time, the pooling of 2x2 is processed in other areas of feature map which the pooling stride is 2. The pooling stride indicates the distance between two pooling processes.

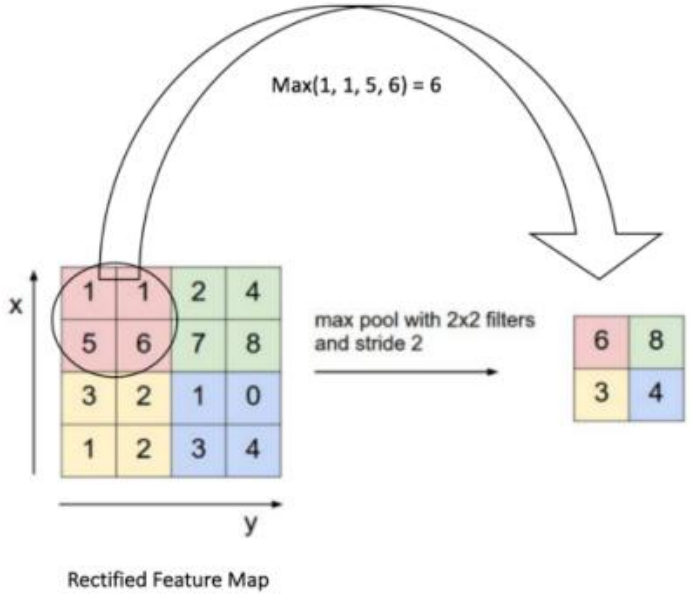



Figure 2.7: Maximum pooling process

First, the red 2x2 area is pooled. Since the maximum value of the 2x2 area is 6. The pooling result is 6 at the corresponding position of the output matrix. Since the stride is 2, the position is moved to the green position for pooling and output. The maximum value is 8. In the same way, the output values of the yellow and blue regions can be obtained. Eventually, our input 4x4 matrix is compressed and becomes a 2x2 matrix after pooling.

In the architecture of Convolutional Neural Networks (CNNs), fully connected (FC) layers play a pivotal role in the final stages of the network. After feature extraction and dimensionality reduction have been accomplished through convolutional and pooling layers, the FC layers serve as the crucial component for classification and decision-making tasks. This transition from convolutional to fully connected layers mark a shift from spatial feature representation to a high-level abstraction of the data.

Fully connected layers, also known as dense layers, are characterized by their architecture in which every neuron is connected to every neuron in the preceding layer. This dense connectivity ensures that the output from the convolutional and pooling layers is comprehensively integrated, enabling the network to make final predictions based on the

entire feature set extracted from the input image. Each neuron in a fully connected layer performs a weighted sum of the inputs followed by a non-linear activation function, allowing the network to model complex, non-linear relationships within the data.

The primary function of fully connected layers is to take the high-dimensional feature maps, flattened into vectors, and transform them into a suitable format for the final output, such as class probabilities in classification tasks. This transformation involves learning a set of weights that maps the input features to the output space, effectively combining and interpreting the features extracted earlier in the network.

Despite their effectiveness, fully connected layers come with their own set of challenges. The high number of parameters in FC layers can lead to increased computational demands and a higher risk of overfitting, especially in networks with many layers and neurons. Regularization techniques such as dropout and weight decay are often employed to address these issues, helping to generalize the model and improve its performance on unseen data.

In summary, fully connected layers are integral to the functionality of CNNs, bridging the gap between feature extraction and final classification. Their ability to integrate and process features from convolutional and pooling layers allows for sophisticated decision-making and accurate predictions, making them a fundamental component in deep learning architectures for a variety of complex tasks.

# 3. Methodology

In the following sections, we will detail the foundational tools, techniques, and principles employed throughout this thesis. In recent years, the emergence of Deep Learning has profoundly transformed the methodologies in Computer Vision research. This shift is evident in the various approaches to 3D human shape and pose reconstruction presented in this work. Despite the diversity of these approaches, they share a commonality: the utilization of a parametric body model to address the inherent complexity of the problem.

Parametric body models are statistical representations that capture the variations in human body shapes and poses. These models play a crucial role in constraining the search space and reducing the dimensionality of tasks related to human body analysis. Essentially, they serve as a foundational template, offering an approximate solution that can be further refined. This refinement is achieved either through the parametrization of the model itself or by using it as a regularization prior. The subsequent sections will introduce the body model in detail, along with the various methods and concepts applied in this dissertation.

## 3.1 Body Model

In this study, we employ the SMPL body model, introduced by Loper et al. in 2015 [57]. The SMPL model is designed as a function $M(\cdot) \in R^{N\times 3}$, which maps pose parameters $\theta \in R^{3K}$ and shape parameters $\beta \in R^{10}$ to a mesh consisting of $N = 6890$ vertices. To generate a watertight mesh, these vertices are connected to $F = 13776$ faces. The pose of the model is determined by $K = 23$ skeleton joints, with their orientations parametrized using the axis-angle representation $\theta$.

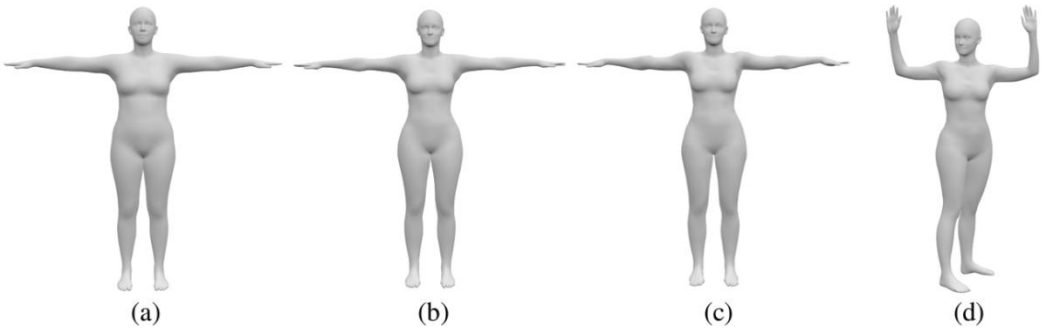



Figure 3.1: Configuring the SMPL model's pose and shape [57].
Starting with a template (a), a new shape is generated (b). Pose-dependent offsets are then applied (c), followed by setting the final pose using blend skinning (d).

The SMPL model has been trained on scans of real human subjects, enabling it to generate highly realistic body shapes and pose-dependent shape deformations. The model is available in three variants: male, female, and neutral, corresponding to male-only, female-only, or mixed-gender subjects, respectively.

The SMPL model generates a posed mesh through a series of transformations. Initially, to create realistic body shapes, a template mesh $T \in R^{N\times 3}$ is deformed using shape deformation offsets $B_S(\beta) \in R^{N\times 3}$. These offsets are derived from a low-dimensional basis of the principal components of body shape variations observed in the SMPL training data. The shape parameters $\beta$ represent a vector of linear coefficients within this shape space. In addition, a linear regressor is used to determine the positions of the skeleton joints $J(\beta) \in R^{K\times 3}$ based on the shape parameters.

Following this, pose-dependent deformations $Bp(\theta) \in R^{N\times 3}$ are applied to the reshaped template (as illustrated in Figure 3.1(c)). The function $Bp(\cdot)$ is a learned linear function parametrized by the desired pose $\theta$, which accounts for muscle and soft-tissue deformations, as well as potential skinning artifacts introduced in subsequent steps.
Finally, the mesh is posed using standard linear blend skinning $W(\cdot) \in R^{N\times 3}$, with blend weights $W \in R^{N\times K}$ (as shown in Figure 3.1(d)). The complete formulation of the SMPL

model is given by:

$$M(\beta,\theta) = W(T(\beta,\theta),J(\beta),\theta,W) \tag{3.1}$$

$$T(\beta,\theta) = T + Bs(\beta) + Bp(\theta). \tag{3.2}$$

As previously mentioned, the body model can serve as a template, prior knowledge, or a representation for methods that work on 3D human body shapes and poses. In Section 4.1, we employ a neural network with an attention mechanism to extract features from the input image and ultimately reconstruct them into a SMPL model.

It is important to note that SMPL only models unclothed subjects, and its shape parametrization does not permit fine-grained personalization. To address this limitation, we extend the standard SMPL formulation by incorporating additional details in many parts of our work. Specifically, we introduce per-vertex offsets $D \in R^{3\times N}$ to enhance the function, following approaches similar to those described in [182, 13, 14]. SMPL+D, an extension of SMPL with added offsets D, is constructed as follows:

$$M(\beta,\theta,D) = W(T(\beta,\theta,D),J(\beta),\theta,W) \tag{3.3}$$

$$T(\beta,\theta,D) = T + Bs(\beta) + Bp(\theta) + D. \tag{3.4}$$

Additionally, we have extended the SMPL model using UV mapping. In this work, we apply textures to the mesh and enhance its surface details using normal maps. The concept of texture encompasses the visible surface details of an object, such as color, roughness, and bumps. In the field of image processing, image texture is used to quantify the perceived characteristics of an image, providing information about the spatial arrangement of color or intensity in the image or a selected region of it.

Textures are represented by arrays of texels, which are the basic units of texture mapping that define the texture space. Texture mapping is the process of associating texture data with a model. A classic method of texture mapping involves using a two-dimensional array to store the texture information of a three-dimensional object. The vertices in 3D

space include not only spatial coordinates but also $u$ and $v$ coordinates, which map to the texture space to generate a UV texture map, thereby linking the texture space information with the 3D model [183-184]. Bump mapping, based on texture mapping, uses height difference information to display detailed bump textures. Similarly, a UV bump texture map needs to be mapped into three-dimensional space to generate a complete mesh model.

UV mapping [185] unfolds the body surface onto a two-dimensional image, such that a given pixel corresponds to a three-dimensional point on the body surface. This mapping is defined over the mesh's faces, where each face, consisting of three 3D vertices, has a corresponding set of three 2D UV coordinates. Here, $u$ and $v$ represent the two axes of the image. The mapping of points within a face is determined through the barycentric interpolation of neighboring coordinates. The 2D image can then be used to enhance the 3D surface. A texture defines the color for each surface point. Similarly, a normal map stores the surface normals, which can add or enhance visual details through shading. A 3D displacement map shifts the surface point in the given direction, allowing for the creation of highly detailed surfaces without altering the resolution of the underlying mesh.

However, some tasks require a higher mesh resolution, which can be achieved by subdividing the SMPL base mesh. This is done by placing a new vertex at the center of each edge of a triangular face. The old face is removed, and four new faces are created by connecting subsets of the six vertices. This process can be repeated. Due to limited computational resources and application constraints, we no longer use this technique to enhance the model's resolution.

In Section 4.2, we use SMPL model with offset to represent detailed shape of human model and further textured model with UV mapping. In next section, we will elaborate on how we utilize the SMPL body model in our work. More importantly, we will introduce the general methods and principles we have consulted to provide effective solutions for

3D human pose and shape reconstruction from monocular images.

## 3.2 Analysis-by-Synthesis

To estimate the 3D body shape of a human, we compare rendered silhouettes with observed silhouettes, employing a method rooted in analysis-by-synthesis. In analysis-by-synthesis, the process involves defining one or more objective functions that are optimized in relation to a scene model. In our context, this scene model is primarily the SMPL model, which may be supplemented with additional components, such as an image formation function. The objective functions are designed to quantify the similarity between the synthesized images generated by the model and the observed images.

Rather than attempting to recreate the images in their entirety, we typically focus on reconstructing abstractions or features of the images, such as segmentation maps or key points. These abstractions exhibit significantly less variation in appearance compared to raw images, making them more straightforward to synthesize and optimize. By focusing on these abstractions, we can achieve a more robust and efficient estimation process, as the reduced variability leads to more stable and reliable optimization.

In the subsequent sections, we will introduce various analysis-by-synthesis techniques that have been employed across different studies within this dissertation. These techniques have been selected and adapted to effectively address the challenges posed by 3D human body shape and pose estimation from monocular images, demonstrating the versatility and power of analysis-by-synthesis in this domain.

### 3.2.1 Image Keypoints

Among the abstractions discussed, the most straightforward are image keypoints. Image keypoints refer to specific 2D coordinates in an image that often carry semantic significance. For instance, in our context, keypoints may represent facial landmarks or skeletal joint positions. The process begins by identifying a corresponding point $l_i \in R^3$

in the 3D scene model for each image keypoint $k_i \in R^2$. The subsequent task during optimization is to determine a scene description such that each $l_i$ projects onto $k_i$ according to a given projection function $\pi(\cdot)$:

$$\sum_i \|\pi(Rl_i + t) - k_i\| = 0 \tag{3.5}$$

Here, $R$ and $t$ represent the rotation and translation parameters of the scene model. As previously mentioned, in the problem scenarios addressed in this work, the scene is described using the SMPL model. This model typically includes global rotation and translation parameters to accurately position the body in the scene.

The 3D points corresponding to the 2D image keypoints are not directly extracted from the image but are instead regressed from the surface of the SMPL body model. This regression is typically achieved through a linear combination of vertices from the model's mesh. One common method for performing this regression is barycentric interpolation, where the 3D point is computed as a weighted sum of the surrounding vertices' positions. This approach allows for an accurate mapping of 2D keypoints to their corresponding 3D locations on the body model, ensuring that the reconstructed 3D scene aligns closely with the observed 2D image data.

This formulation is central to many computer vision tasks, particularly in the domains of human pose estimation and shape reconstruction. By leveraging the SMPL model and the optimization framework outlined above, it is possible to achieve a detailed and accurate 3D representation of the human body from 2D observations, which is crucial for applications in areas such as motion capture, virtual reality, and biometric analysis.

In Section 4.3.2, we use the differences between various key points to numerically compute the reconstruction loss of the model, thereby assessing its quality. In Section 4.2, image keypoints are also involved in body reconstruction optimized process, we will discuss it later.

### 3.2.2 Image Segmentation

Image segmentation is a well-established technique for scene abstraction and is extensively utilized in analysis-by-synthesis frameworks. In image segmentation, each pixel is assigned a specific label, which allows distinguishing different regions within the image. Segmentation can be categorized into binary and multi-part types. Binary segmentation typically separates the foreground from the background, where the definitions of foreground and background are task-specific. For instance, moving objects might be classified as the foreground, while static objects are considered the background. In Section 4.2, we only focus on person, hence the person we define is foreground and the rest of image is background. When optimizing analysis-by-synthesis problems using image segmentation, the goal is to minimize the discrepancy between the predicted silhouette and the observed silhouette. For binary segmentation, this optimization can be expressed as:

$$\min_{R,t} |G(R,t) - S| \tag{3.6}$$

$$G(R,t) = R_c(F(R,t)) \tag{3.7}$$

Here, $S$ represents the observed segmentation image, $F(\cdot)$ denotes an example of a scene function, and $R(\cdot)$ is a binary image formation function under camera $c$.

Although this formulation leads to the expected minimum, it can be challenging and slow to optimize due to potential entrapment in local minima and the fact that gradients only provide information about a one-pixel-wide neighborhood. To address these issues, more sophisticated approaches are often employed. One effective strategy is to perform the optimization simultaneously at different image resolutions. This approach allows the optimization to take larger steps, helping it to avoid local minima.

(A) (B)

Figure 3.2: Reconstruct pose using Chamfer matching.

A specific method we use in Section 4.2.2.3 to improve optimization is Chamfer matching, where the difference between each point in one silhouette and the closest point in the other silhouette is minimized (see Figure 3.2). In Figure 3.2, the predicted silhouette should closely align with the observed silhouette, ensuring it does not extend beyond it (B) while covering it as fully as possible (A). The error is visualized with color coding: red indicates a large error, and blue indicates a small error when compared to the corresponding observed silhouette (grey). The goal is for the predicted silhouette to fully cover the observed silhouette without exceeding it. The distances between all points in silhouette $M$ and the nearest points in silhouette $N$ can be calculated by multiplying silhouette $M$ by the distance transform $C(\cdot)$ of silhouette $N$. The distance transform of a binary image provides the distance of each pixel to the nearest non-zero pixel. The Chamfer matching objective aggregates the errors across all image pixels $p$ as follows:

$$\min_{R,t} \sum_p G_p(R,t) \cdot C_p(S) + S_p \cdot C_p(G(R,t)) \tag{3.8}$$

However, $C(\cdot)$ is not differentiable, which complicates the optimization process. To overcome this, a modified objective function is used:

$$\min_{R,t} \sum_p G_p(R,t) \cdot C_p(S) + (1 - G_p(R,t)) \cdot C_p(1 - S) \tag{3.9}$$

This modification retains the objective's minimum but eliminates the need to compute the distance transform of the predicted silhouette, thereby simplifying the optimization process.

### 3.2.3 Shape-from-shading

Shape from Shading (SFS) [186] is a computer vision technique designed to infer the 3D shape of an object from the shading information in an image. The fundamental principle of this technique is that the variations in shading on the surface of an object are determined by the light source's position, the direction of the surface normals, and the reflective properties of the material. By analyzing these shading cues, the 3D shape of the object can be deduced.

Shape from Shading is based on a key assumption that the brightness value of each pixel in an image is determined by the reflective characteristics of the object's surface and the lighting conditions. Specifically, for a pixel in an image, the brightness $B(x, y)$ can be represented as a function of the surface normal $n(x, y)$ and the scene's reflectance map $P$:

$$B(x, y) = P(n(x, y)) \tag{3.10}$$

In classical SFS models, it is typically assumed that the surface follows a Lambertian reflectance model, meaning that the surface brightness is proportional to the cosine of the angle between the light direction and the surface normal. This relationship can be expressed as:

$$B(x, y) = cos(l, n(x, y)) = \frac{l \cdot n(x,y)}{|l||n(x,y)|} \tag{3.11}$$

where $l$ is the light direction vector, and $n(x, y)$ is the surface normal vector.

The process of Shape from Shading generally involves several key steps:

1. Illumination Model Assumption: Depending on the physical properties of the object's surface, surface reflection can be classified as diffuse (Lambert Reflection), specular

reflection, hybrid reflection, or more complex forms of reflection. In order to simplify the problem, the traditional SFS algorithm assumes that the reflection model is a Lambert body reflection model, where it is assumed that the object's surface is diffusely reflecting light, and the brightness is proportional to the cosine of the angle between the light direction and the surface normal.

2. Normal Direction Estimation: Based on the assumed lighting model, the next step involves analyzing the brightness distribution in the image to calculate the surface normal direction for each pixel. This step typically involves solving partial differential equations, as the normal direction is a three-dimensional vector, while the image brightness value provides only a scalar. Thus, the problem is inherently underdetermined.

3. Depth Map Reconstruction: Once the surface normal direction for each pixel is obtained, the next step is to compute the depth map of the object. This step usually involves integrating the normal direction information to derive the position of the object's surface in 3D space.

4. Shape Optimization: Finally, further optimization of the shape may be necessary to ensure that the reconstructed 3D shape is consistent with the shading information in the original image. This is typically achieved through iterative optimization methods.

In Section 4.2.3, we present our fine-detailed human reconstruction model results optimized using the SFS technique.

Shape from Shading is a valuable technique in computer vision, offering a way to recover 3D shape information from 2D images based on shading. Despite its challenges, such as reliance on accurate lighting models and the complexity of solving for surface normal, it has significant applications in fields like medical imaging, reverse engineering, robotic vision, and object modeling in computer graphics.

## 3.3 Deep Learning

Learning-based methods represent a fundamentally different approach compared to traditional optimization-based methods, particularly those rooted in the analysis-by-synthesis paradigm. Whereas optimization-based techniques iteratively refine parameters to align a model's output with observed data, learning-based methods involve discovering a parameterization for a complex function, typically implemented as a neural network. This neural network is designed to produce the desired output by adjusting its parameters through exposure to a vast dataset of input-output pairs. This adjustment process is known as training the network.

In recent years, convolutional neural networks (CNNs) have brought about a significant transformation in the field of computer vision. These networks often surpass the performance of classical methods by a wide margin and have enabled solutions to problems that were previously out of reach for traditional techniques. Although analysis-by-synthesis methods can contribute during the training phase of neural networks, the operational principles of neural networks differ substantially. Neural networks, particularly deep learning architectures, excel at extracting high-dimensional features from input data such as images and mapping these features to specific task-oriented outputs.

A neural network is fundamentally a collection of interconnected layers, each composed of numerous small computational units known as neurons. Each neuron processes an input vector $x$ by computing a weighted sum using learnable weights $w$ and adding a learnable bias term $b$:

$$y = w^{\top} x + b \tag{3.12}$$

The output $y$ is then passed through a non-linear activation function $h(\cdot)$ to introduce non-linearity into the model:

$$a = h(y) \tag{3.13}$$

Common activation functions include the Rectified Linear Unit (ReLU) and the Sigmoid

function. To construct an entire layer $l$, multiple neurons are organized together, each processing the input and producing an output as follows:

$$y_i^{[l]} = w_i^{[l]\top} a^{[l-1]} + b_i^{[l]} \tag{3.14}$$

$$a_i^{[l]} = h^{[l]}(y_i^{[l]}) \tag{3.15}$$

These neurons can be vectorized for the entire layer:

$$y_i^{[l]} = W^{[l]} a^{[l-1]} + b_i^{[l]} \tag{3.16}$$

$$a^{[l]} = h^l(y^{[l]}) \tag{3.17}$$

When multiple layers are stacked together, they form a neural network. A network with many layers is termed a deep neural network, and the specific configuration of layers, neurons, and activation functions known as the network architecture determines the model's complexity and computational power.

To train the neural network to produce accurate outputs, one must optimize a loss function relative to the network's parameters, namely the weights $W$ and biases $b$. The loss function quantifies the difference between the network's predictions and the actual desired outcomes. By calculating the partial derivatives of the loss function with respect to the network's parameters, one can iteratively adjust these parameters to minimize the loss, thereby improving the network's performance.

In computer vision, CNNs, a specialized type of neural network, have gained prominence. Unlike traditional neural networks, which compute a weighted sum over the entire input vector $x$, CNNs apply a convolution operation over local neighborhoods within the input. This means that each element in the output vector $y$ depends only on a local subset of the input vector $x$. The convolution operation is defined as:

$$(f * g)[n] = \sum_{m=-K}^{K} f[m]g[n-m] \tag{3.18}$$

In this context, $f$ is referred to as the kernel, which in CNNs is composed of learnable weights shared across different parts of the input. CNNs offer several advantages: they

require fewer parameters even for large inputs, they can handle inputs of varying dimensionalities, and they exhibit a degree of translational invariance in their processing.

In this research, deep learning methods are utilized in two principal ways. First, pre-trained models are employed to compute various abstractions, such as foreground segmentation, semantic segmentation, reflectance and shading separation, and keypoint localization. In Section 4.1, a Resnet-50 pre-trained network to extract features from images serves as the backbone for reconstructing human pose and shape. Second, we develop algorithms that incorporate deep learning at their core, leveraging its power to tackle complex vision tasks. We further introduce it in Section 4.2. These applications of deep learning demonstrate its versatility and effectiveness in modern computer vision challenges.

## 3.4 Performance Evaluations

This section begins by introducing the datasets required for the synthesis of virtual humans and the commonly used quantitative metrics for performance evaluation. The performance evaluation results of various synthesis techniques are then presented. We will provide an overview of the common types of datasets, including their primary content and application directions, with a detailed introduction of specific datasets to lay the groundwork for quantitative performance comparisons. Both perceptual and non-perceptual metrics will be discussed. In the results and analysis section, we will compare the performance of our algorithm with several other algorithms.

### 3.4.1 Dataset

In existing research, qualitative comparison methods are typically intuitive and straightforward, often involving user surveys to gather information. However, quantitative comparisons involve more complex calculations and scenarios.

The application of virtual human synthesis technologies spans multiple fields, including

facial modeling, behavior prediction, and virtual clothing, with each type of synthesis experiment requiring datasets with distinct characteristics. Commonly used training datasets consist of images and videos, captured using either single or multi-view camera systems. To enhance model performance, recent efforts have focused on preprocessing raw data to obtain 2D and 3D human data as model inputs, systematically generating datasets tailored for applications such as virtual human facial modeling and behavior prediction. The following sections will summarize and review the forms and characteristics of existing datasets based on specific literature.

Two-dimensional image datasets are typically used to provide 2D information and can also be combined with depth maps to present 3D information from multiple angles. Beyond depth maps, 3D data types include polygonal meshes and point cloud data, with point clouds being particularly suitable for representing sparse structures, which can be converted into standard 3D polygon meshes. For instance, Neural Body [187] created a multi-view video dataset called ZJU-Mocap to evaluate the performance of models in synthesizing novel views from sparse data. This dataset contains nine dynamic human videos captured from 21 synchronized cameras, featuring complex movements such as Tai Chi, warm-up exercises, and boxing, and is widely used for assessing multi-view synthesis quality. The S3 dataset [143] employs 2D image sets and radar-scanned 3D point cloud data, with input formats including single images and voxelized radar scan data.

For virtual human synthesis technologies focusing on posture, the Market-1501 [188] dataset is commonly used for pedestrian re-identification. It contains 32,643 annotated images of 1,501 pedestrians, with each identity captured by up to six cameras. Human3.6M [150] is a larger-scale 3D human posture dataset featuring multi-view videos shot by four cameras, using a marker-based motion capture system. It includes complex actions performed by five female and six male subjects.

In addition to conventional data inputs, 3D human models can also serve as inputs for virtual human synthesis models. The STATE dataset [189] synthesized Human image datasets according to experimental needs, with data sourced from real scanned 3D human models provided by Twindom. Each model was rendered from 496 multi-view images. Recent virtual human datasets include CAPE [69] and AGORA [190]. CAPE is the first dataset to extend dressed 3D human meshes to multiple poses, generating pose- and clothing-conditioned human models from 3D scans. AGORA extends the SMPL-X body model to 3D scans, creating 3D postures and body shapes with various poses and clothing.

The volume of publicly available annotated datasets is vast. In the performance evaluation of virtual human synthesis technologies, only subsets of these datasets are typically analyzed, with input formats still predominantly images and videos. Table 3.1 presents information on datasets used in several studies, covering both 2D and 3D data.

Table 3.1: Brief information of datasets.

| Dataset | Date release | The size of the experimental data | Category | Information dimension |
|---|---|---|---|---|
| **People Snapshot [191]** | 2018 | 24 video sequences, 11 humans. | video | 2D |
| **NeRFace [140]** | 2021 | 2 minutes 6000 frames, 512x512 resolution. | video | 2D |
| **ZJU-MoCap [187]** | 2021 | 9 dynamic human videos. | video | 3D |
| **Market-1501 [188]** | 2015 | 32668 images, 1501 people. | image | 2D |
| **Human [150]** | 2021 | 3D models. | video | 3D |

### 3.4.2 Evaluation Indicators

This section introduces common accuracy metrics used for quantifying the performance of virtual human models, with the selection of metrics depending on the model training methods employed.

L1-Loss (Mean Absolute Error, MAE): This metric represents the average of the absolute differences between the predicted values and the actual values, providing a straightforward measure of prediction accuracy.

LPIPS (Learned Perceptual Image Patch Similarity): Often referred to as "perceptual loss," LPIPS systematically evaluates deep features across different structures and tasks. It is applicable to a wide range of architectures and supervision levels. Given a real image $x$ and a reconstructed image $x_0$, perceptual similarity [192] is defined as:

$$d(x, x_o) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw}) \right\|_2^2 \tag{3.19}$$

Which is calculated by extracting feature stacks from $L$ layers, normalizing them across channels, and computing the $L2$ distance, scaled by a weight vector $W_l$). A lower LPIPS value indicates better modeling performance and greater similarity between the two images.

Inception Score (IS): This score is used to evaluate the performance of GAN models by measuring the divergence between the distribution of generated images and real images when passed through a classification model trained on real images. A smaller distance corresponds to a higher IS, indicating better model performance.

Fréchet Inception Distance (FID): FID calculates the distance between real and generated samples in feature space, commonly used to assess the quality of images produced by GAN models. FID is computed by comparing the mean and covariance of feature vectors from real and generated images [193], which defines as:

$$FID(x,g) = \|\mu_x - \mu_g\| + T_r(X + G - 2\sqrt{XG}) \quad (3.20)$$

Real and generated images are passed through the Inception Net-V3 model, which outputs 2,048-dimensional feature vectors. The covariance matrices of these feature vector sets are denoted by $X$ and $G$, while $\mu_x$ and $\mu_g$ represent the means of these sets. A lower FID score indicates a smaller distance between the distributions of the real and generated images, signifying better model performance.

SSIM (Structural Similarity Index): SSIM [194] measures the structural similarity between two images, defining image information from the perspectives of luminance, contrast, and structure. SSIM values range from -1 to 1, with higher values indicating greater similarity, and a value of 1 denoting identical images. Given two images $x$ and $y$, their luminance is denoted by $l$, contrast by $c$, and structural similarity by $s$:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \quad (3.21)$$

$$c(x,y) = \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \quad (3.22)$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \quad (3.23)$$

$$SSIM(x,y) = l(x,y) \cdot c(x,y) \cdot s(x,y) \quad (3.24)$$

$\mu$ represents the mean intensity of the image. $\sigma_x$and $\sigma_y$ denote the standard deviations of images $x$ and $y$, respectively. $\sigma_{xy}$ represents the covariance between images $x$ and $y$. $c_1$, $c_2$, and $c_3$ are constants introduced to stabilize the division, particularly when the denominators are close to zero.

When $c_3 = \frac{c_2}{2}$, the SSIM index can be expressed as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (3.25)$$

This equation combines the luminance, contrast, and structure components into a single measure, with higher SSIM values indicating greater similarity between the two images.

The Mean Square Error (MSE): also known as L2-Loss, is a metric used to quantify the

difference between a reference image $f$ and a test image $g$. For grayscale images, assuming that both the reference image $f$ and the test image $g$ have dimensions $M \times N$, the MSE [194] is defined as follows:

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (f(i,j) - g(i,j))^2 \tag{3.26}$$

In this equation, $f(i,j)$ and $g(i,j)$ represent the pixel values at position $(i,j)$ in the reference and test images, respectively. The MSE calculates the average squared difference across all pixels, with lower MSE values indicating better similarity between the two images.

PSNR (Peak Signal-to-Noise Ratio): PSNR is commonly used in image and signal processing to measure the fidelity of an image. It is derived from the logarithm of the MSE [194], with higher PSNR values indicating better image quality. For grayscale images, assuming that both the reference image $f$ and the test image $g$ have dimensions $M \times N$, the PSNR is defined as follows:

$$PSNR(f,g) = \frac{10 log_{10} MAX_f^2}{MSE(f,g)} \tag{3.27}$$

For color images, PSNR can be computed for each channel individually, averaged across channels, or by converting the image to a different color space such as YUV and calculating PSNR for the luminance component.

PSNR is generally related to the fidelity of a signal, while SSIM is more aligned with the human visual system, better reflecting the fidelity of reconstruction results. LPIPS, which extracts features using deep learning networks, is considered the most closely aligned with human perception, effectively capturing perceptual similarity between images.

In Chapter 4, these aforementioned datasets and indicators will be utilized in experiments and validation. In the next chapter, we will elaborate on how to extend the various methods discussed in this chapter, innovatively integrating attention mechanisms to construct the model architecture, thereby achieving state-of-the-art performance in human pose and shape reconstruction and detailed human model optimization.

# 4.Experiment and Results

## 4.1 Human Shape and Pose Reconstruction

### 4.1.1 Introduction

Directly regressing 3D human pose and shape (HPS) from RGB images holds significant potential in various advanced fields such as robotics, computer graphics, and augmented/virtual reality (AR/VR). The primary objective of this task is to take a single image or a sequence of video frames as input and predict the parameters of a human body model, such as SMPL [57]. The advent of deep convolutional neural networks (CNNs) has driven rapid advancements in this domain [9, 94, 195, 196].

Building on the success of attention mechanisms in other tasks [197, 198, 199, 200], we have incorporated body part segmentation to supervise the attention masks initially. Subsequently, we transition to end-to-end training using only pose supervision, allowing the attention mechanism to autonomously extract relevant information from both the body and surrounding pixels. This approach enables the network to focus on regions that it identifies as informative in an unsupervised manner, enhancing its ability to handle complex visual inputs where occlusions and other challenges may arise.

### 4.1.2 Method

#### 4.1.2.1 Insights and Body Model

Building on the aforementioned observations, our work is informed by several key insights. First, while state-of-the-art (SOTA) networks [9, 94, 195] are capable of implicitly learning to focus on meaningful regions, they do so with limited spatial information after global average pooling. To more accurately discern whether body parts are visible or occluded, our approach leverages a pixel-aligned structure where each pixel corresponds

to a specific region in the image, storing a pixel-level representation in the form of a feature volume.

Second, recognizing that estimating attention weights and learning end-to-end trainable features for 3D pose estimation are distinct tasks, our method employs two separate feature volumes: one derived from the 2D part branch, which is responsible for estimating attention weights, and another from the 3D body branch, which handles SMPL parameter regression.

Finally, to effectively model the dependencies between body parts, our approach utilizes part segmentations as soft attention masks. These masks modulate the contribution of each feature in the 3D body branch, allowing for joint-specific adjustments that enhance the accuracy of the overall pose estimation.

SMPL [57] represents the body pose and shape by $\theta$, which consists of the pose $\theta \in R^{72}$ and shape $\beta \in R^{10}$ parameters. Here we use the gender-neutral shape model as in previous work [9, 195]. Given these parameters, the SMPL model is a differentiable function that outputs a posed 3D mesh $M(\theta, \beta) \in R^{6890\times 3}$. The 3D joint locations $J_{3D} = WM \in R^{J\times 3}, J = 24$, are computed with a pretrained linear regressor $W$.

4.1.2.2 Architecture and Losses

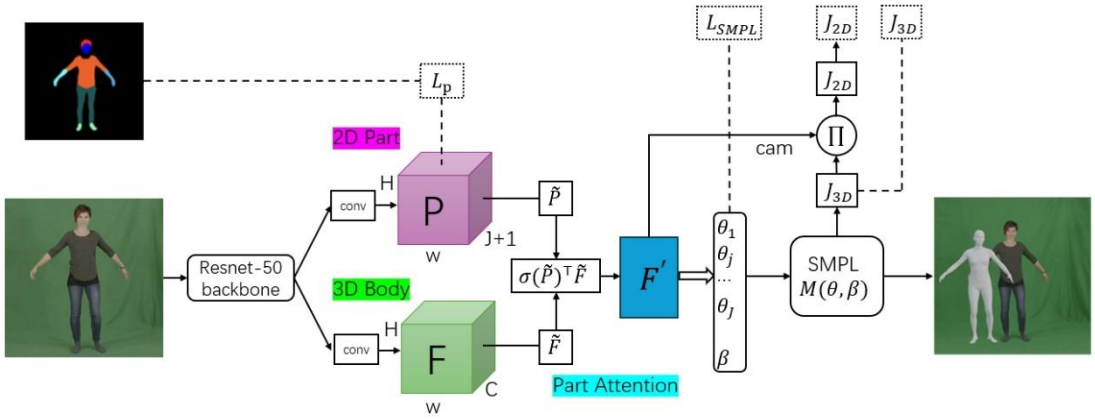



Figure 4.1: Model architecture.

For a given input image, our model first extracts two distinct pixel-level features (P and F). These features are subsequently fused using a part-attention module, resulting in a consolidated feature $F'$. This final feature is used to simultaneously regress the camera parameters and the SMPL body mesh.

Our architecture showing in Figure 4.1 operates as follows: Given an input image $I$ (224*224), we first utilize a CNN backbone to extract volumetric features. ResNet-50 [201] is a deep convolutional neural network consisting of approximately 50 layers. The network begins with standard convolutional and pooling layers, followed by four main stages that progressively extract features. Each stage is composed of multiple stacked "bottleneck residual blocks." Finally, the network outputs classification results through global average pooling and a fully connected layer. Our extracted volumetric features (A dimension of $7 \times 7 \times 2048$) are from a layer preceding the global average pooling layer. This is followed by two distinct feature extraction branches, each performing three rounds of $2 \times$ upsampling to obtain feature maps with a size of $56 \times 56$.

The first branch, which we refer to as the 2D part branch, $P \in R^{H \times W \times (J+1)}$, models $J$ part attention masks and one background mask, where $H$ and $W$ represent the height and width of the feature volume, respectively. Each pixel at location $(h, w)$ within this volume

stores the likelihood of belonging to a specific body part $j$. In the 2D Part Branch, a 1×1 convolution is also applied for dimensionality reduction, resulting in a final channel count of J + 1 = 25 (corresponding to the 24 joints in the SMPL model plus one background channel).

The second branch, denoted as $F \in R^{H\times W\times C}$, is responsible for 3D body parameter estimation. It shares the same spatial dimensions $H \times W$ as $P$, but differs in the number of channels which is $C$. For the 3D Body Branch, the output feature map is a tensor with dimensions $56 \times 56 \times 256$.

Let $P_j \in R^{H\times W}$ and $F_c \in R^{H\times W}$ represent the $j\_th$ and $c\_th$ channels of $P$ and $F$, respectively. The final feature tensor $F' \in R^{J\times C}$ is constructed such that each element in $F_c$ contributes proportionally to $F'$ according to the corresponding elements in $P_j$, following spatial softmax normalization $\sigma$. Formally, the element at location $(j, c)$ in $F'$ is computed as:

$$F'_{j,c} = \sum_{h,w} \sigma(P_j) \odot F_c \tag{4.1}$$

where $\odot$ is the Hadamard product.

In other terms, we utilize the attention map $\sigma(P_j)$ as a soft attention mechanism to aggregate features within the feature map $F_c$. This process can be efficiently executed through a dot product operation, akin to existing attention mechanisms: $F' = \sigma(\tilde{P})^{\top}\tilde{F}$, where $\tilde{P} \in R^{HW\times J}$ and $\tilde{F} \in R^{HW\times C}$ represent the reshaped attention map $P$ (excluding the background mask) and the feature map $F$, respectively. This attention operation indicates that if a specific pixel has a higher attention weight, its corresponding feature will have a more significant contribution to the final representation $F'$.

To guide the attention maps toward the appropriate regions, we supervise the 2D part branch $P$ using ground-truth segmentation labels, enabling the attention maps of visible parts to converge to the corresponding regions. However, for occluded parts, this

supervision method tends to assign zero attention weights to all pixels in $P_j$ since these parts are absent in the ground-truth segmentation labels. An attention map comprising solely of zero weights is both undesirable and impractical, as the spatial softmax operation necessitates that all elements sum to 1. Consequently, we implement a hybrid approach: initially, we supervise the 2D part branch during the early stages of training, followed by unsupervised training. This strategy enables the network to focus on other relevant regions to accurately estimate the poses of occluded joints.

We utilize the complete feature tensor $F'$ to regress the body shape $\beta$ and to estimate a weak-perspective camera model characterized by scale and translation parameters $[s,t], t \in R^2$. Additionally, each row $F_j'$ of the tensor is independently processed by distinct multilayer perceptrons (MLPs) to predict the rotation of each body part. The rotation $\theta_j$ is parameterized as a 6D vector, as described in prior works [88, 188].

In summary, we define total loss:

$$L = \lambda_{2D}L_{2D} + \lambda_{3D}L_{3D} + \lambda_{SMPL}L_{SMPL} + \lambda_P L_P \tag{4.2}$$

Where:

$$L_{2D} = \left\|J_{2D} - \hat{J}_{2D}\right\|_F^2 \tag{4.3}$$

$$L_{3D} = \left\|J_{3D} - \hat{J}_{3D}\right\|_F^2 \tag{4.4}$$

$$L_{SMPL} = \left\|\Theta - \widehat{\Theta}\right\|_2^2 \tag{4.5}$$

$$L_P = \frac{1}{HW}\sum_{h,w} CrossEntropy(\sigma(P_{h,w}), \hat{P}_{h,w}) \tag{4.6}$$

Here, $\hat{x}$ denotes the ground truth for the corresponding variable $x$. To calculate the 2D keypoint loss, the SMPL 3D joint locations $J_{3D}(\theta,\beta) = WM(\theta,\beta)$ are first derived, where they are computed from the body vertices using a pretrained linear regressor $W$. Utilizing the inferred weak-perspective camera model, the 2D projection of the 3D joints $J_{3D}$ is then determined as $J_{2D} \in R^{J\times 2} = s\prod(R\,J_{3D}) + t$, $R \in SO(3)$ is the camera rotation matrix which conditions satisfy: its inverse is equal to its transpose, and the determinant of the rotation matrix is equal to 1. The projection $\prod$ is orthographic.

The scalar coefficient $\lambda$ is employed to balance the various loss terms. Let $P_{h,w} \in R^{1\times1\times(J+1)}$ represent the fiber of $P$ at the location$(h, w)$) and let $\hat{P}_{h,w} \in \{0,1\}^{J+1}$ denote the ground-truth part label at the same location, expressed as a one-hot vector. The part segmentation loss $L_P$ is defined as the cross-entropy loss between $P_{h,w}$ after applying softmax and $\hat{P}_{h,w}$, averaged over all $H \times W$ elements. It is important to note that this softmax operation normalizes along the fiber $P_{h,w}$, whereas the softmax in Equation 4.1 normalizes across the slice $P_j$.

### 4.1.3 Experiments and Results

For all experiments, we utilized a fixed image size of 224 x 224 pixels. The model was optimized using the Adam optimizer with a learning rate of $5 \times 10^{-5}$ and a batch size of 64. Our model was trained on several datasets, including COCO [202], MPII [203], LSPET [204], MPI-INF-3DHP [205], and Human3.6M [150]. Pseudo-ground-truth SMPL annotations for in-the-wild datasets were provided by EFT [92]. Part segmentation labels were generated by rendering segmented SMPL meshes, corresponding to 24 parts associated with the 24 SMPL joints. We employed the PyTorch reimplementation [206] of Neural Mesh Renderer [207] to render the parts.

To accelerate convergence, the backbone used is first pre-trained on MPII [203] for 2D pose estimation. We assign different weight coefficients to each term in the loss function: $\lambda_{2D} = 200, \lambda_{3D} = 200, \lambda_{SMPL} = 40, \lambda_P = 40$. For samples lacking part segmentation labels, the 2D branch was not supervised.

#### 4.1.3.1 Training

Here we briefly introduce details of training datasets:

MPI-INF-3DHP [205] is a multi-view indoor dataset for 3D human pose estimation. The 3D annotations in this dataset are obtained using a commercial markerless motion capture software, which results in less accuracy compared to some other 3D datasets, such as Human3.6M [150]. We utilize all training subjects from S1 to S8, totaling 90,000

images.

Human3.6M [150] is an indoor, multi-view dataset for 3D human pose estimation. Consistent with previous approaches, we use images from 5 subjects (S1, S5, S6, S7, S8) for training, amounting to 292,000 images.

The in-the-wild 2D keypoint datasets COCO [202], MPII [203], and LSPET [204] each contain 2D keypoint annotations. Specifically, MPII has 14,000 instances, COCO has 75,000 instances, and LSPET has 7,000 instances. Additionally, we make use of pseudo-SMPL annotations provided by the EFT [92] method, in conjunction with these 2D keypoint annotations.

To achieve the final optimal model, we adopt the data sampling strategies outlined in EFT [92] and SPIN [195], which employ fixed data sampling ratios for each batch. Initially, we train with 100% COCO-EFT for 175,000 steps. Subsequently, we integrate 50% Human3.6M, 30% In-the-Wild (i.e., [COCO, MPII, LSPET]-EFT), and 20% MPI-INF-3DHP datasets into the training process. We also find that alternative combinations, such as [50% Human3.6M, 30% COCO-EFT, 20% MPI-INF-3DHP] or [20% Human3.6M, 30% COCO-EFT, 50% MPI-INF-3DHP], yield comparable performance on the 3DPW dataset.

In table 4.2, during the ablation experiments, we trained our model and our baseline models on COCO for 175,000 steps and evaluated them on the 3DPW and 3DPW-OCC datasets. Subsequently, we incorporated all training data to compare our work with previous state-of-the-art methods. This pretraining strategy facilitated faster convergence and reduced the overall training time. Training our model to convergence required approximately 120 hours on an Nvidia RTX 3070 GPU.

#### 4.1.3.2 Evaluation

The 3DPW [208] test split, 3DPW-OCC [208, 209], and 3DOH [209] datasets are utilized

for evaluation. We report the Procrustes-aligned mean per joint position error (PA-MPJPE) and the mean per joint position error (MPJPE), both measured in millimeters. Additionally, for the 3DPW dataset, we also report the per vertex error (PVE) in millimeters.

Comparison to the State-of-the-Art. Table 4.1 presents a comparison of our method with previous single-RGB-image human pose and shape (HPS) estimation methods. Our method achieves a significant improvement in PA-MPJPE performance compared to HMR-EFT [92], one of the best-performing methods in recent years.

| Method | | $MPJPE \downarrow$ | $PA-MPJPE \downarrow$ | $PVE \downarrow$ |
|---|---|---|---|---|
| TEMPORAL | HMMR [9] | 116.5 | 72.6 | - |
| | Doersch et al. [210] | - | 74.7 | - |
| | Sun et al. [211] | - | 69.5 | - |
| | VIBE [94] | 93.5 | 56.5 | 113.4 |
| | MEVA [212] | 86.9 | 54.7 | - |
| MULTISTAGE | Pose2Mesh [213] | 89.2 | 58.9 | - |
| | Zanfir et al. [214] | 90.0 | 57.1 | - |
| | I2L-MeshNet [215] | 93.2 | 58.6 | - |
| | LearnedGD [216] | - | 56.4 | - |
| SINGLE STAGE | HMR [9] | 130.0 | 76.7 | - |
| | CMR [217] | - | 70.2 | - |
| | SPIN [195] | 96.9 | 59.2 | 135.1 |
| | HMR-EFT [92] | - | 54.2 | - |
| | Ours | 84.1 | 53.2 | 102.6 |

Table 4.1: Evaluation on the 3DPW dataset.

Ablation Experiments: Table 4.2 provides a detailed summary of our ablation experiments aimed at investigating the role of part attention in our model. Initially, we compared our approach with Neural Body Fitting (NBF) [218], trained under the same conditions. NBF

represents a straightforward combination of part segmentation and human body regression. As demonstrated in Table 4.2, NBF's two-stage approach is outperformed even by the HMR-EFT baseline.

| Method | | 3DPW | | 3DPW-OCC | |
|---|---|---|---|---|---|
| | | $MPJPE \downarrow$ | $PA-MPJPE \downarrow$ | $MPJPE \downarrow$ | $PA-MPJPE \downarrow$ |
| | NBF | 100.4 | 63.2 | 103.5 | 70.4 |
| | HMR-EFT | 99.0 | 59.9 | 97.9 | 64.7 |
| P Supervision | F Sampling | | | | |
| (a) Joints | Pooling | 95.4 | 59.2 | 95.8 | 63.2 |
| (b) Joints | Attention | 95.6 | 59.0 | 99.0 | 63.9 |
| (c) Uns | Attention | 95.0 | 58.2 | 97.0 | 63.0 |
| (d) Parts | Attention | 94.8 | 57.6 | 94.8 | **61.3** |
| (e) Parts/Uns | Attention | **93.7** | **57.4** | **94.1** | 61.9 |
| (f) Parts | Pooling | 98.4 | 59.5 | 100.1 | 65.3 |

Table 4.2: Part attention ablation experiments.

We then explored various supervision strategies for the 2D part branch $P$ and different methods for sampling the final features $F'$ from $F$. In the "Uns" condition, $P$ was not supervised. Inspired by HoloPose [219], we first supervised the 2D branch using keypoints and applied bilinear sampling to pool the 3D features (Table 4.2-a). While this approach resulted in lower errors compared to HMR, the improvement was marginal. This is likely because sparse keypoints do not sufficiently cover the spatial area needed to effectively model body parts.

Given that the 2D branch predicts Gaussian heatmaps, which encompass a broader spatial

area than discrete keypoints, we next investigated the use of soft attention instead of pooling, to achieve a larger effective receptive field (Table 4.2-b). However, this approach did not fully exploit the capabilities of soft attention, which ideally should learn to focus on relevant regions directly from the data. Therefore, we removed supervision from the 2D branch to assess whether soft attention alone could perform as effectively as explicit supervision (Table 4.2-c). Visualization of the resulting attention maps revealed that they did not adequately focus on the body parts.

To introduce greater structure, we then supervised the 2D branch using part segmentation labels (Table 4.2-d). This method yielded significantly better results than the previous attempts. Nonetheless, a limitation remains: supervising with a segmentation loss constrains the attention map to focus solely on predefined parts, whereas purely soft attention has the potential to attend to any region deemed informative by the model. To address this, we employed a mixed supervision strategy, applying the part segmentation loss for approximately 125,000 steps, followed by continued training without supervision (Table 4.2-e). This final approach produced the best results. We also conducted experiments with part segmentation and pooling to investigate the impact of soft attention (Table 4.2-f).

In figure 4.2, we provide People Snapshot [191] test results. Our results are composed of the input images and the corresponding output SMPL human models. The experimental results demonstrate that our algorithm successfully reconstructs human bodies from monocular images while achieving reasonably accurate proportions for different body parts.


Figure 4.2: People Snapshot [191] test results.

In the first row, we present reconstruction results from portrait images with green-screen backgrounds. The second row showcases results from indoor daily scenes, while the third row displays reconstructions from outdoor environments. These successful reconstructions highlight the robustness of our algorithm across varying environmental backgrounds.

Our algorithm was also evaluated on reconstructing human models from images with varying clothing styles: figure 4.2-c and figure 4.2-h wearing a short-sleeved T-shirt, figure 4.2-c and figure 4.2-i wearing shorts, and figure 4.2-b and figure 4.2-f dressed in a loose long-sleeved top. The successful reconstructions in these cases further demonstrate the robustness of our method of handling different clothing conditions.

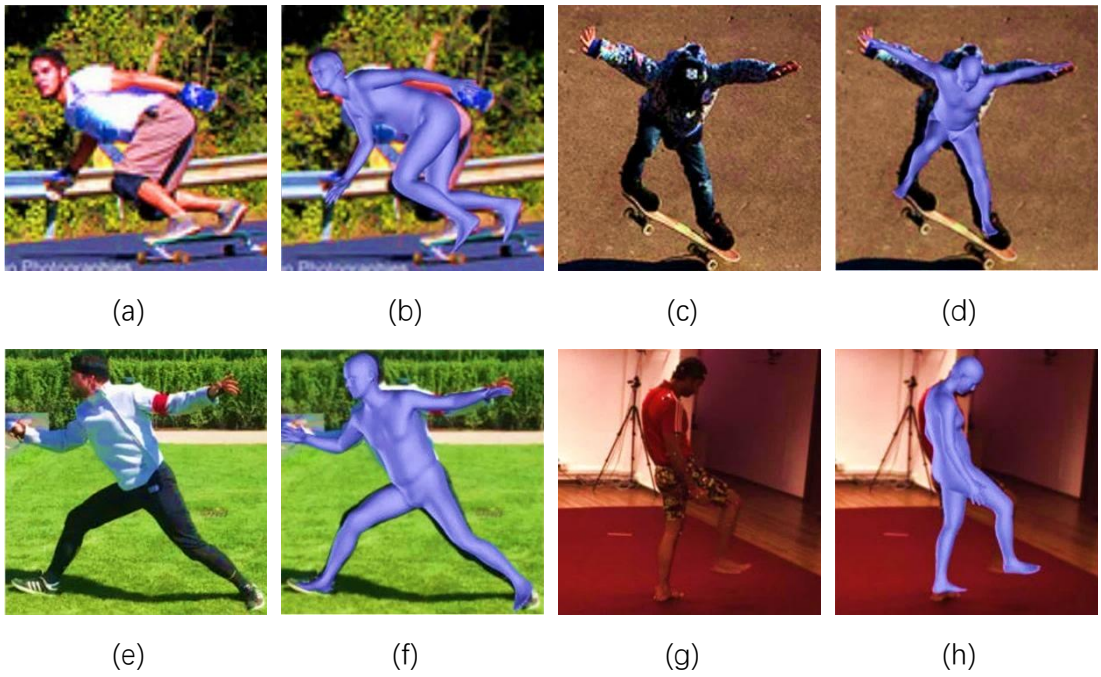



Figure 4.3: Complex poses test results.
Examples come from different datasets, including COCO (a, c), 3DPW (e), and Human3.6M (g).

We further tested the performance of our algorithm under complex conditions. Figure 4.3-a shows a skateboarder bending down to sprint, Figure 4.3-c depicts a skateboarder raising both hands to perform a technical move, Figure 4.3-e presents a fencer executing a fencing action, and Figure 4.3-g illustrates a person lifting a leg under dim lighting. In Figures 4.3-b, d, f and h, our algorithm successfully reconstructed various poses and shapes in these challenging scenarios. The successful reconstructions in these cases further demonstrate the robustness of our method of handling complex poses conditions.

### 4.1.4 Summary

In this chapter, we discussed a novel body part-driven attention framework that leverages pixel-aligned local features for regressing body pose and shape. This method has demonstrated superior performance in benchmark tests across various datasets. The approach developed in this chapter will be further applied in the subsequent chapter, focusing on human body reconstruction tasks, where the effectiveness of this framework will be further validated and expanded upon.

## 4.2 Digital Human Model Obtainment

### 4.2.1 Introduction

Digital humans, which are computer-generated 3D representations of real people, are crucial in creating immersive experiences in a number of recent technological developments, e.g., virtual and augmented reality, and Metaverse which has been gaining a lot of attentions in recent years. They provide a sense of realism and interactivity that is difficult to achieve with traditional computer-generated graphics. As a result, there has been a significant increase in the development of new applications of digital humans in various industries, including manufacturing, gaming, entertainment, education, and healthcare. However, the existing measurement methods of obtaining digital human models are either too expensive or lack accuracy, which presents a challenge for developers looking to create quality and realistic digital humans. The cost of creating a digital human model by using active scanners can be prohibitively expensive, especially for smaller companies and indie developers. This is because the process involves a lot of time and resources, including specialized equipment, software, and skilled personnel, and has special requirements for target people such as standing still for a long time [220, 221]. Additionally, the accuracy of the model can be compromised if the data used to create it is incomplete or of poor quality.

To address these challenges, researchers and developers are exploring new ways to create high-quality and affordable digital human models. A promising approach is to use deep learning algorithms to generate realistic human models from a small amount of data. Recently, there are a lot of learning-based work produced. Considering object or scene representation in 3D learning, those works can be simply categorized as explicit representation-based and implicit representation-based.

Explicit representation-based models. Polygon mesh statistical human body models [222, 223, 224, 57, 62] have been widely used in 3D human reconstruction as an explicit representation model. A polygon mesh is a data structure that represents a polyhedron by defining its surface as a collection of vertices and faces. This representation is useful for conveying topological information about the object's surface and provides a high-quality description of 3D geometric structures. Additionally, polygon meshes are memory-efficient and can be easily textured, making them a versatile tool for various applications in computer graphics and visualization. In [91, 94, 195, 225, 226, 227], those single image-based work estimates a naked human body model from a monocular camera picture. Although those works produced some fine results, they still need further process to dress clothes up. To solve this problem, some other work [14, 66, 68, 70, 191, 228, 229, 230] directly learns a mesh human body model with clothes offset from images. The resulting clothed 3D human models inherit the skeleton and surface covering weights of the based body model, facilitating their animation. However, a significant challenge lies in modelling clothing articles such as skirts and dresses, which exhibit substantial deviations from the body surface. The conventional approach of using body-to-cloth offsets is inadequate in such cases.

Implicit representation-based models. In contrast to meshes, deep implicit functions [118-119] could represent highly detailed 3D shapes with arbitrary topology and are not subject to resolution limitations. Recent research by Saito et al. [126,127] has employed deep implicit functions to reconstruct 3D human shapes from RGB images, achieving high

levels of geometric detail and accurate alignment with image pixels. However, this approach suffers from a lack of regularization, resulting in various artifacts such as broken or missing limbs, incomplete details, and geometric noises. To address this issue, some researchers [231-233] have incorporated additional features, such as coarse-occupancy prediction and depth information from RGB-D cameras, to enhance the accuracy and robustness of the shape estimation. In addition, some [234,235] have proposed efficient volumetric sampling schemes to speed up the inference process. Nevertheless, a major limitation of all these methods is that the resulting 3D human shapes cannot be reposed, as implicit shapes do not possess a consistent mesh topology, a skeleton, or skinning weights that are typically found in statistical models.

In summary of these related work, the learning-based human body reconstruction method provided a significant result with only a few inputs. Although training neural networks may require large, labelled 3D digital human datasets and cost large computation resources and time, it is very convenient and efficient for end users. Consumers may only need to upload a small amount of data and wait for the returned result from the cloud service. But most learning-based works focus on recovering full human body from one image with the powerful prediction ability of neural network. This data-driven prediction method may achieve a great result in pose estimation tasks [236-238], but also lead to an ambiguity problem caused by a lack of unseen body information from only one image. It is hard to guess detailed back information from front body image, despite a strong pre-trained network. Hence, we address our problem of finding a balance and a connection between typical measurement method and the popular learning-based method to generate a digital human from inputs.

In this chapter, we present our prediction-measurement pipeline to reconstruct a detailed human body model from a set of self-rotated target human images captured by a single monocular camera. We estimate an initial human body model from image sequences by a trained neural network and further vertex alignment to optimize it from image to image.

Our research focuses on creating a human body model that is easily modifiable. To achieve this, we have chosen to utilize a parametric representation of an explicit body model known as SMPL (Skinned Multi-Person Linear) [57]. All these related work has shown that the SMPL model possesses excellent expansibility with high-quality open-source resources, which can assist in achieving good results for 3D reconstruction projects. This model allows us to generate body shapes that can be easily modified and adapted to different needs. We begin by collecting data on the SMPL pose and shape parameters, as well as the intrinsic camera parameters from input images. This information is then used to prepare for further optimization.

To create the initial body model, we will use the average pose and shape parameters from the SMPL model. This initial model will serve as a baseline for further modifications and adjustments. This involves projecting the initial SMPL model with the shape and pose of the target image and then minimizing the distance between the projected points and the silhouette of the target image. By doing so, we are able to obtain shape and pose information for every image. This method enables us to create a human body model that is easily adaptable to different needs and requirements. We can modify and adjust the model based on new input data, allowing us to create more accurate and realistic representations of the human body. Overall, our research aims to create a model that can be used in a wide range of applications, from computer graphics to medical simulations.

### 4.2.2 Method

In order to make sure our predicted initial human body model is allowed to modify, we used a parametric representation of the explicit body model SMPL [57], which will be introduced in Section 4.2.2.1. Applying the method we introduced in Section 4.1, we collect the estimate results of SMPL pose and shape parameters and intrinsic camera parameters from input images for the preparation of further vertex aligned optimization. And we will build the initial body model with an average pose and shape parameters in the SMPL model, which will be discussed in Section 4.2.2.2. Section 4.2.2.3 will detail our

optimization method. Since we obtain shape and pose information of every image, we project the initial SMPL model with shape and pose of target image and minimize the distance between projected points to silhouette of target image [244].

#### 4.2.2.1 SMPL Parameterized Human Body Model

The SMPL model [57] is a powerful method for characterizing the human body in terms of both body shape and motion posture. It achieves this through the use of two sets of statistical parameters: body shape parameters and pose parameters.

The body shape parameters, denoted as $\beta$, are used to describe an individual's physique. This 10-dimensional vector allows for the quantification of a person's body shape along various dimensions such as height, weight, and overall body proportions. Each dimension of $\beta$ can be thought of as a specific indicator of a person's physical characteristics, which collectively describe their overall body shape.

On the other hand, the pose parameters, denoted as $\theta$, are used to describe the motion posture of the human body. This set of parameters comprises $24 \times 3$ dimensions, with 24 representing the number of joints and 3 representing the axis-angle representation used to describe rotations. This allows for a detailed and comprehensive description of the human body's motion posture.

To characterize the human body using these parameters, the SMPL model utilizes a base template or mean template $T_m$, which serves as a reference shape. The shape parameters are then linearly superimposed on this base template to produce the final 3D mesh, with the bias for each shape parameter being calculated using the $B_s(\beta)$ function learned from data. This allows for the generation of meshes that accurately reflect the desired body shape.

$$B_s(\beta) = \sum_{n=1}^{|\beta|} \beta_n S_n \tag{4.7}$$

where $S$ is learned through data and has dimensions of $(6890, 3, 10)$.

Similarly, the effect of different pose parameters is determined using the $B_p(\theta)$ function, which is calculated relative to the T-pose state to account for changes in posture. This enables the creation of meshes that accurately reflect the desired motion posture.

$$B_p(\theta) = \sum_{n=1}^{9K}(R_n(\theta) - R_n\theta^*)P_n \quad (4.8)$$

Each pose parameter is represented by a rotation matrix $R$, so there are $9K$ dimensions. P (i.e., the weight matrix) is learned through data and has dimensions of $(6890, 3, 207)$, where $207$ is obtained from $23 \times 9$.

Finally, the SMPL model accounts for skin deformation caused by joint motion through a skinning process. This involves a weighted linear combination of skin nodes that change with the joint, with the weights determined based on the distance of the endpoint from the joint. Closer endpoints are more strongly influenced by joint rotation or translation, resulting in a more realistic and accurate representation of the human body's motion. Here the template is defined as:

$$T(\beta, \theta) = T_m + B_s(\beta) + B_p(\theta) \quad (4.9)$$

Since SMPL body template is a representation of a naked human body, we add an offset S as a detailed cloth supplement:

$$T(\beta, \theta, S) = T_m + B_s(\beta) + B_p(\theta) + S \quad (4.10)$$

A pose and shape driven detailed SMPL model is further defined as:

$$M(\beta, \theta, S) = W(T(\beta, \theta, S) + J(\beta), \theta, \mathcal{W}) \quad (4.11)$$

where $W$ is the Linear Blend Skinning (LBS) function, $J(\beta)$ is the locations of 24 skeleton joints; $\mathcal{W}$ is the learned blend weights.

#### 4.2.2.2 Images Information Extraction

In this section, we extract information from input images with several deep learning technologies. We collect SMPL model shape and pose parameters with a network which introduced in chapter 4 whose main method is to propose a novel deep learning-based

approach for estimating 3D human body shape and pose from a single 2D image. The method is centred around a part attention regressor, which divides the human body into various parts and focuses on each one independently to generate accurate 3D body estimations.

The key components of our work's methodology include:

1. Part Attention: The network utilizes an attention mechanism to focus on specific body parts, enabling it to handle occlusions and varying poses. This mechanism helps the network learn and emphasize individual part features, leading to more precise 3D shape and pose estimations.

2. Multi-stage Estimation: the model employs a multi-stage estimation process, using an initial coarse estimation followed by multiple refinement stages. This hierarchical approach allows the network to progressively refine its predictions, leading to higher accuracy.

3. Joint 2D-3D Representation Learning: our work learns a joint embedding space of 2D and 3D features, enabling it to leverage both 2D and 3D information during the estimation process. This joint learning process allows the model to handle a wide range of poses and improve overall accuracy.

4. Part-based Loss Function: The model uses a part-based loss function, which encourages the network to focus on each body part individually. This loss function helps the network to handle complex poses and occlusions, as well as achieve better generalization across various body shapes.

In summary, our method leverages a part attention mechanism, multi-stage estimation, joint 2D-3D representation learning, and a part-based loss function to achieve the accurate 3D human body shape and pose estimations from a single 2D image.

We simply initialize an SMPL body model with average estimated shape and pose parameters of input images, and further detailed offset optimization will be discussed in the next section.

#### 4.2.2.3 Full Detailed Body Model Optimization

Given that we have acquired the human body pose and camera position data for all input images, we can obtain the projection results of the initialized model concerning angles and poses. By comparing the derived contour images with those of the input images, we can optimize the vertex parameters of the SMPL model. For the $i_{th}$ input image, the associated contour of the human body model is denoted as $S_i$, while the contour of the human body in the input image is represented as $S_i'$. In accordance with a differentiable renderer approach [239], we employ an Intersection-over-Union error metric for the optimization process.

$$L_{sil} = \frac{1}{f}\sum_1^f \left(1 - \frac{\|S_i \otimes S_i'\|_1}{\|S_i \oplus S_i' - S_i \otimes S_i'\|_1}\right) \tag{4.12}$$

Where $\otimes$ is an element wise product and $\oplus$ is a sum operator.

We also add a Laplacian mesh regularizer [240] to ensure the deformation process smoothly. The regularizer is defined as:

$$L_{lp} = \sum_1^N \left\| L(v_i) - L\big(v_i(\beta, 0)\big) \right\|^2 \tag{4.13}$$

where L is a Laplace operator, v is the vertices set.

Similar to [191], we penalize the difference between the optimized detailed body model vertices and the standard SMPL template body model vertices to avoid large differential error.

$$L_{dif} = \sum_{i=1}^N \|v_i(\beta, S) - v_i(\beta, 0)\|^2 \tag{4.14}$$

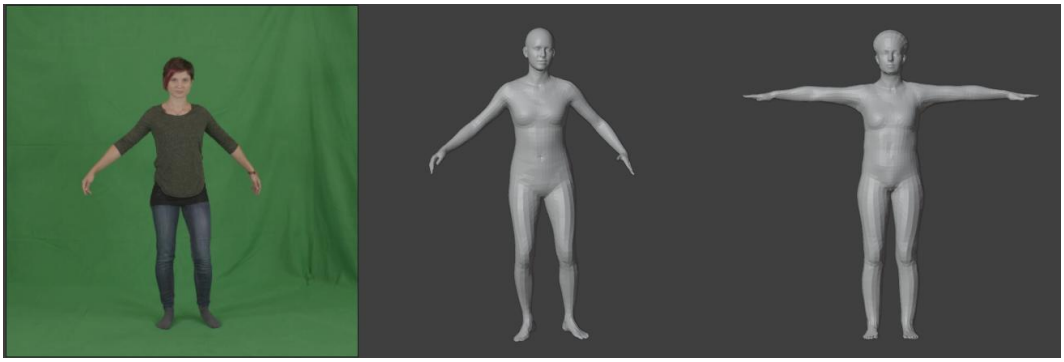
Our joint optimized formula is defined as:

$$L = L_{cham} + L_{sil} + w_{lp}L_{lp} + w_{dif}L_{dif} \quad (4.15)$$

Where $w_{lp}$ and $w_{dif}$ are the balance weights. $L_{cham}$ is Chamfer matching loss introduced in Section 3.2.2. The loss function is optimized by The Dog Leg Method which is a numerical optimization algorithm used for solving nonlinear least squares problems. It falls under the category of trust region methods and enhances computational efficiency by combining the characteristics of the steepest descent method and the Gauss-Newton method.

By minimizing the loss function L, we modify the vertices of SMPL model and finally collect a detailed body model with cloth information offset. Since SMPL is a pose and shape parametric driven model, the result model can be further animated, which is suitable for more applications.

### 4.2.3 Results and Evaluation



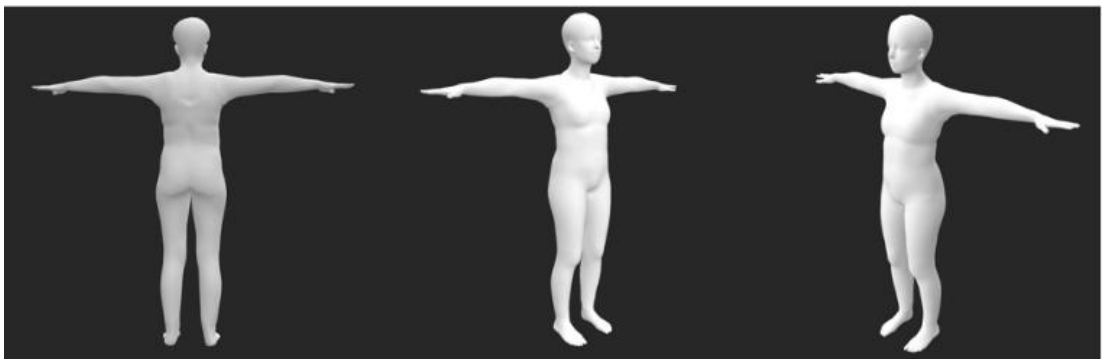
(a) (b) (c)

Figure 4.4: Overview of our method.
From left to right, input images, predicted SMPL model, optimized SMPL-offset model.

We test our method in People-snapshot dataset [191], Figure 4.4 shows the results of every step. The input images (Figure 4.4-a) are captured by a stable camera, and photographed person is self-rotated with a fixed pose. We do not need photographed person keep this pose strictly, a slight change is acceptable. In our method, we extract some frames from the video of dataset, our test used f=100 frames to reconstruct body model.

The mid image (Figure 4.4-b) shows the initial SMPL model reconstructed from information extracted from input images in step one. We take an estimated average pose and shape parameter of images applying to the SMPL template. The main computing cost here is information extraction with deep neural network, also the accuracy is determined by the efficiency of the state-of-the-art network.

Our approach takes about 100 seconds for optimizing every frame. We remove the pose parameter in the result, and a standard T-pose SMPL model is showing in the Figure 4.4-c. And we provide multi-view reconstructed human models in Figure 4.5 to demonstrate that our algorithm effectively operates across the entire 3D orientation of the human body, with both frontal and dorsal anatomical features being accurately characterized. Our result can be further modified and rendered. Compared with the initial model, we successfully recovered some hair, face and cloth details in the SMPL model with offset.



(a) (b) (c)

Figure 4.5: Detailed optimized SMPL-offset model.

In Figure 4.6, we provide People Snapshot [191] test results. Our results are composed of the input images and the corresponding output detailed human models. The experimental results demonstrate that our algorithm successfully reconstructs human bodies from monocular images while achieving reasonably accurate proportions for different body parts and details like hair and clothes.

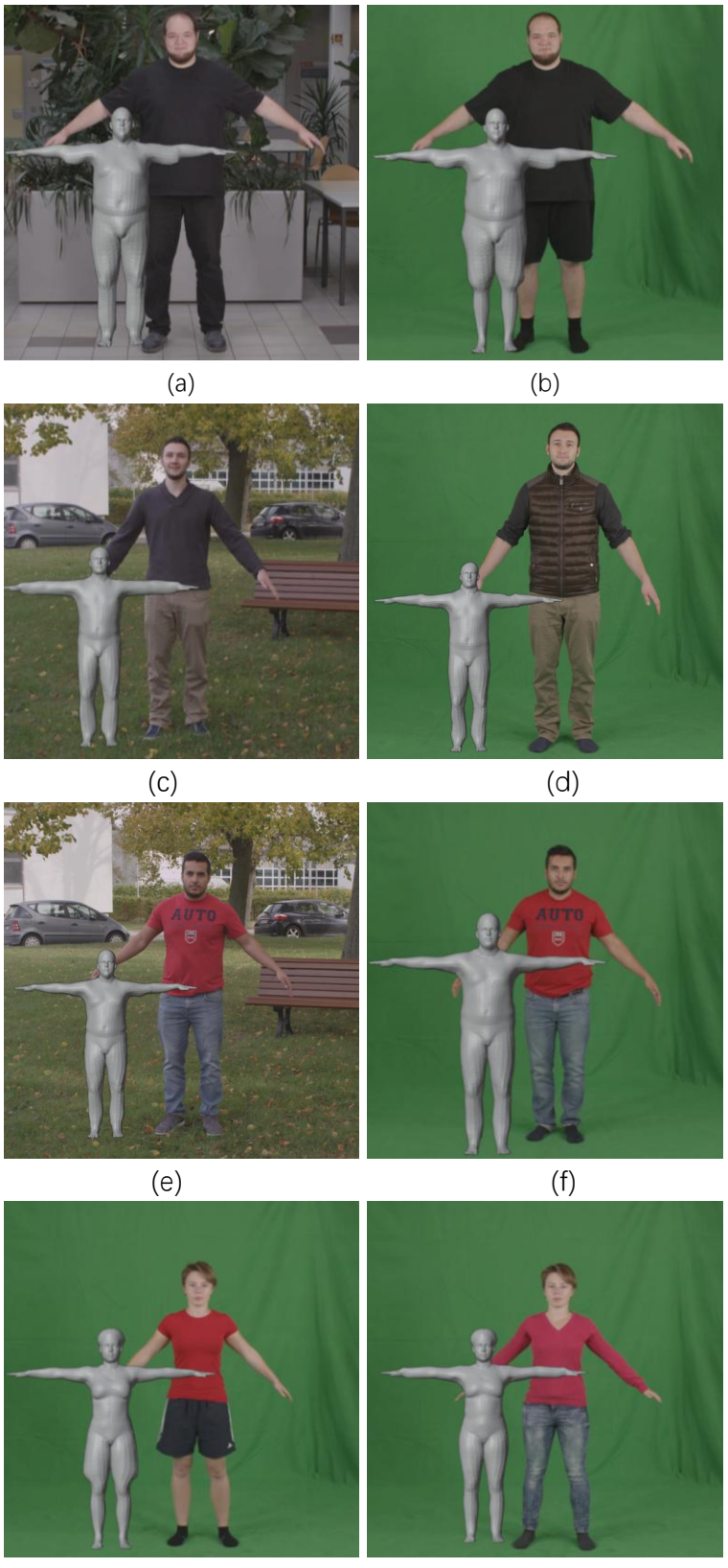


Figure 4.6: More detailed optimized SMPL-offset model results.

We present a set of mixed-result images comprising input photos with varying genders, clothing, and backgrounds, alongside their corresponding reconstructed human models. In the comparative pairs (Figure 4.6- a vs. b, c vs. d, e vs. f), we employed the same subject wearing different outfits under distinct settings—indoor, outdoor, and green-screen environments—to evaluate our algorithm's performance. The results demonstrate that our method robustly handles human model reconstruction across diverse backgrounds while accurately preserving fine clothing details.

We also compared our work with other current state-of-art method in Table 4.3, we have better performance.

| Method | $IS \downarrow$ | $FID \downarrow$ | $LPIPS \downarrow$ | $SSIM \uparrow$ |
|---|---|---|---|---|
| StylePeople [138] | 1.7469 | 272.1 | 0.0836 | 0.9012 |
| LWGAN [137] | 1.7159 | 1771.9 | 0.2727 | 0.2876 |
| 360Degree [97] | 1.8643 | 1383.1 | 0.2123 | 0.8079 |
| Ours | 1.7002 | 254.2 | 0.0644 | 0.9342 |

Table 4.3: Comparation of different methods in People Snapshot dataset.
*IS* (Inception Score), *FID* (Fréchet Inception Distance), *LPIPS* (Learned Perceptual Image Patch Similarity) and *SSIM* (Structural Similarity Index) have been briefly introduced in 3.4.2. ↓ means smaller indicators refer to better performance, ↑ means larger indicators refer to better performance.

We build upon prior research [241] by utilizing the Pablo sequence from their dataset to perform a quantitative analysis. The dataset provides ground truth in the form of surface meshes and 3D joints, which were captured using a multi-view performance capture method [242]. Our method is compared with a state-of-the-art template-based performance capture technique [241] and several single-image human reconstruction approaches [66, 115, 126, 127, 230]. Since body pose estimation is not performed in [66], we apply our estimated body pose to their T-pose results for a fair comparison.

To evaluate our method, we use a surface reconstruction metric. Due to the inherent depth-scale ambiguity in single-view reconstructions, we first compute a global scaling factor to adjust our results in relation to the ground truth. We then align our results to the ground truth through translation to correct for any global depth offsets. The evaluation metric is the average point-to-surface distance between all ground truth vertices within the clothing region and our output mesh. The clothing region, which includes the T-shirt and shorts, is manually segmented from the ground truth surface mesh. This procedure is uniformly applied across all methods under evaluation.

We report the mean surface error averaged across all frames in the middle column of Table 4.4. Our method demonstrates a significantly lower surface error compared to all previous single-image surface reconstruction methods. Moreover, our performance is comparable to the template-based tracking method [241], which relies on a pre-scanned personalized template to provide detailed prior information about the subject's body and clothing shape. In contrast, our method operates without a pre-processed template, making it applicable to a broader range of videos.

| Methods | Surface Error | Joint Error |
|---|---|---|
| MonoPerfCap [242] | 14.6 | 118.7 |
| HMD [230] | 31.9 | |
| Tex2Shape [66] | 27.7 | |
| DeepHuman [115] | 24.2 | |
| PIFu [126] | 30.5 | |
| PIFuHD [127] | 26.4 | |
| Ours | 17.8 | 77.1 |

Table 4.4: Quantitative comparison.

Quantitative comparison with prior research on the Pablo sequence is performed using the mean point-to-surface error and the mean joint error across frames. All

measurements are reported in mm. Some methods do not provide a SMPL-like model to compute joint error.

We evaluate our method against the state-of-the-art RGB-D based approach [243] using their dataset, referred to as KinectCap in the main chapter. To ensure a fair comparison, we align the scale of their results with the ground truth. The original study evaluated performance based on the distance between the scan and the reconstructed mesh. Due to noise in the scan data, the authors filtered out points exceeding a certain threshold. To facilitate a fair comparison, we report their result obtained using this method, which was 2.54 mm.

Since the appropriate threshold for noise filtering in the scan data was unknown and different sampling densities can lead to varying results, we adopted the strategy outlined in the main chapter and also used in [13]. This involves first performing non-rigid registration, regularized by the body model, to achieve a ground truth registration free from scan noise. We then compute a bidirectional surface-to-surface distance between the ground truth registration and the reconstructed shape. Using this approach, their method achieved an accuracy of 3.3 mm, while our method achieved 3.8 mm.

Although our monocular approach does not match the accuracy of depth camera-based methods [243], it provides competitive results despite relying solely on a single RGB camera.

To generate the texture, we first warp our estimated canonical model back to each frame, then back-project the image color onto all visible vertices. Finally, a texture image is created by calculating the median of the most orthogonal texels from all viewpoints. In Figure 4.7, we present one example of texture map and textured model.

(a) (b)

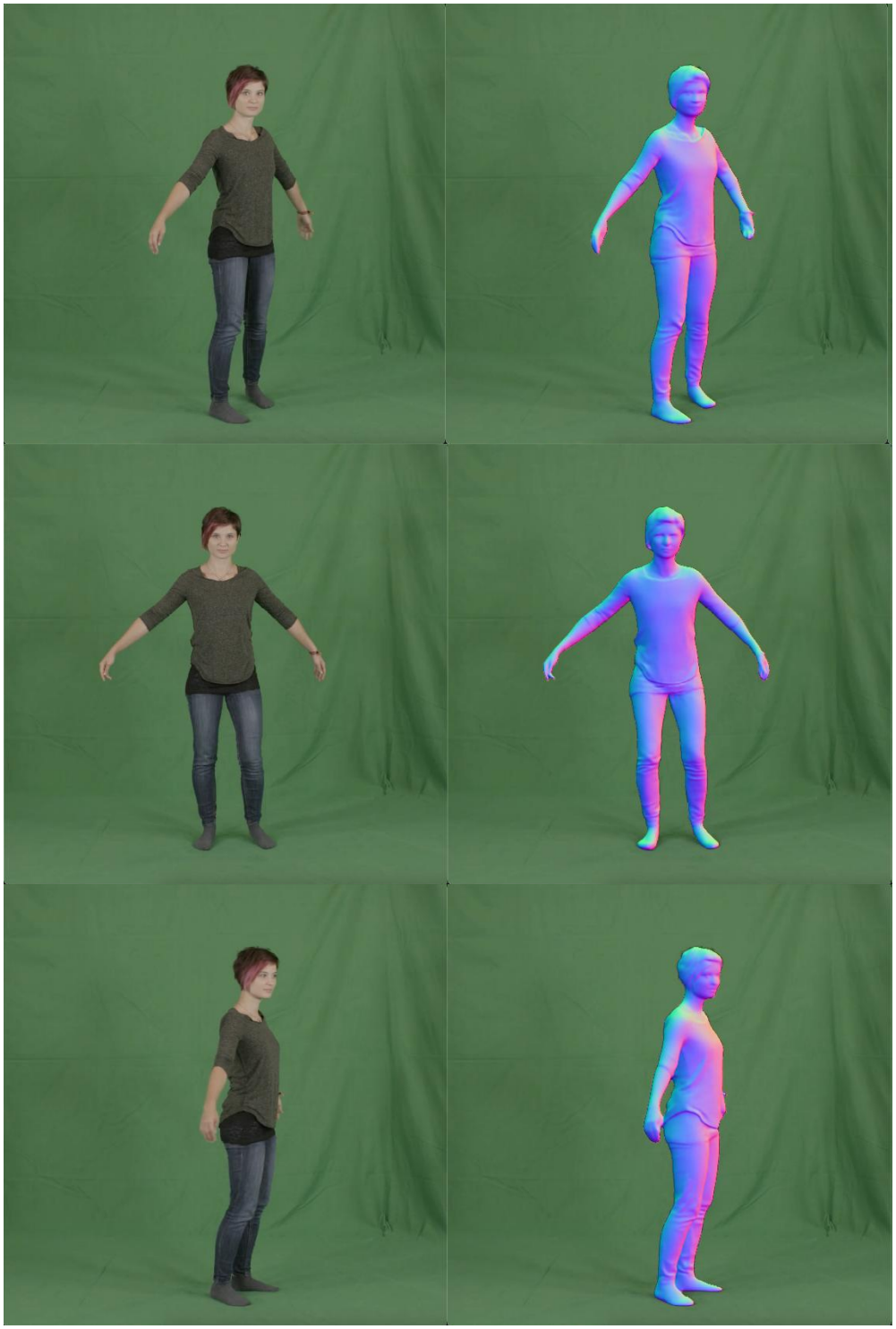
Figure 4.7: Texture map and textured human model.

As the results generated by our method still have shortcomings in terms of detail representation. We tried a normal map aligned method to refine more details in our result. Traditionally, more refined details have been captured using Shape from Shading (SFS) [191]. However, for monocular clothing capture in unconstrained environments, we have empirically found it challenging to reliably extract such refined details using SFS due to the complexity of garment albedo, wide variations in lighting conditions, and self-shadowing effects. Recently, the success of learning-based approaches [126,127] in estimating accurate surface normals for human appearance using neural networks has been observed. These estimated surface normals provide robust and direct indications for incorporating wrinkles into our clothing capture results to achieve better alignment with the original images. To merge the estimated normals, we use SMPL's inverse pose function to transform them into a canonical T-pose. We then optimize the surface geometry to fit these merged normals. The Shape-from-Shading loss is defined as:

$$L_{SFS} = \sum_{f=k-1}^{f=k+1} \sum_{i \in \nu} \left\| n_i - \tilde{n}_i^f \right\|^2 \tag{4.16}$$

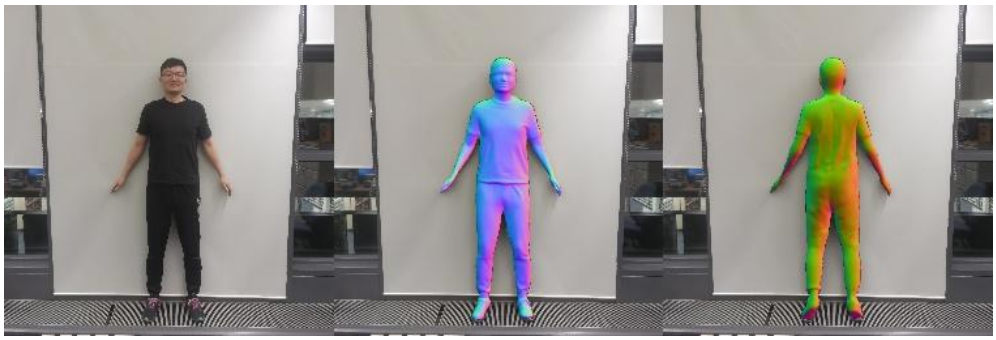
where $\nu$ denotes the subset of visible vertices, where $k$ is the current key-frame. $\tilde{n}_i^f$ denotes the auxiliary normal of vertex $i$ calculated from frame $f$. All normals are in T-pose space.

Our results shown in Figure 4.8 and a more generalized test of a daily indoor work environment images shown in Figure 4.9.



(a) Input image (b) Refined model

Figure 4.8: Detail-refined with normal maps result.



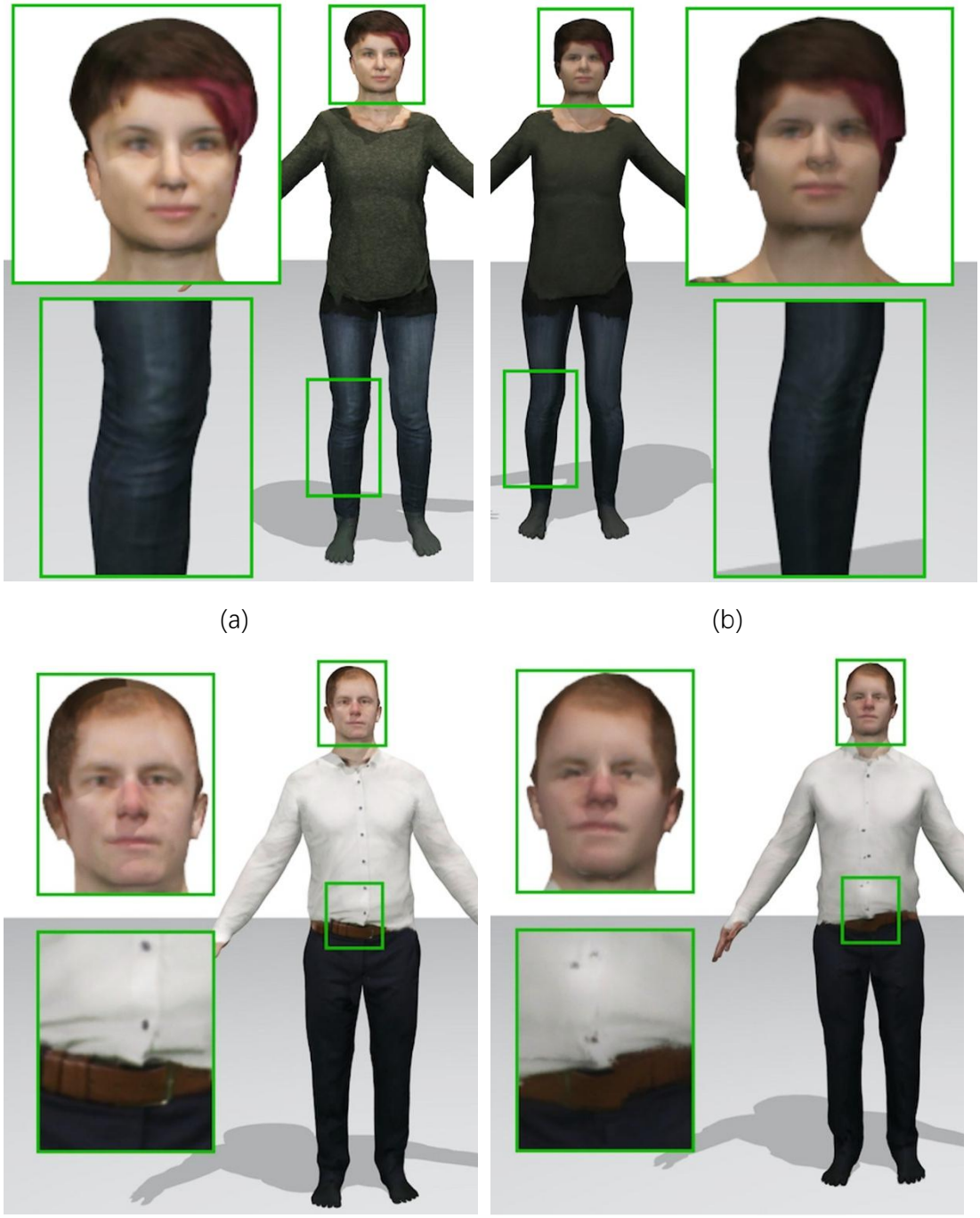
(a) Input image (b) Front of model (c) Back of model

Figure 4.9: Daily scene test result.

In Figure 4.8, compared with Figure 4.4 and Figure 4.5, our reconstructed details such as hair, face and clothes have been significantly improved by normal map refinement.

In Figure 4.9, we show our reconstructed result from a target person standing in front of a green screen. We also test our method in a simple and daily environment. And the result reveals our method is adaptable.

In Figure 4.10, we compared the textured model optimized using normal maps (a and c) with the unoptimized results (b and d). The comparison clearly demonstrates improvements in the model's facial and clothing details. Figure 4.10-a and c not only exhibit more intricate clothing wrinkles but also show enhanced accuracy and refinement in facial structure modeling.



(a) (b)

(c) (d)

Figure 4.10: Comparation of textured models.

### 4.2.4 Summary

In this section, we discussed a vertices-pixels aligned method jointly using deep learning method and the key idea of traditional computer 3D graphics to achieve a fine level of digital human geometry reconstruction from images. Our method relies on several deep

learning-based methods such as pose, and shape estimate from single images. Compared with related deep learning-based methods, our method eliminates the inherent ambiguity of predicting the complete body model from a single image. With the assistance of deep learning techniques such as pose estimation and human parameter model prediction, we have improved computational speed and reduced experimental conditions compared to traditional optical measurement techniques for obtaining human models. Despite some shortcomings in our work, we have successfully demonstrated the possibility and potential of combining deep learning with traditional techniques.

Although significant progress has been made in deep learning-based methods for 3D human body reconstruction from 2D images, several challenges and limitations still need to be addressed. We will discuss these achievements, challenges, limitations and future work in next chapter.

# 5. Discussion and future work

## 5.1 Discussion

In this thesis, we first introduce a novel body part-driven attention framework that leverages pixel-aligned local features for regressing body pose and shape based on the SMPL body model. Subsequently, we shift our focus to the detailed reconstruction of human shape. Drawing inspiration from a traditional 3D reconstruction method, visual hull, we present an advanced approach for digital human model reconstruction from monocular images of a moving person, extending beyond the parameter space of a parametric model. Our method enables the estimation of animatable 3D human avatars, including detailed elements such as hair, clothing, and surface texture. We achieve state-of-the-art performance with this approach compared to other similar works. This thesis makes a significant contribution to the field: our work simplifies the digitization of humans by relying solely on regular videos or even photographs. With advanced performance, our method eliminates the need for specialized equipment, enabling the automatic reconstruction of detailed human models and the widespread application of virtual humans in emerging technologies.

Nevertheless, three critical observations emerged from our work: First, although leveraging silhouettes and semantic segmentation makes the problem more manageable and applicable to real-world data, it simultaneously abstracts away valuable information. Second, regressing 3D vertex locations from 2D images presents an alignment challenge, as 2D images and 3D meshes can only be accurately compared by projecting and rasterizing the meshes through rendering. Consequently, 3D pose plays a pivotal role in supervision, with incorrect poses leading to diminished result quality. Furthermore, by relying on the SMPL body model, our methods are limited in representing shapes with topologies different from the human body, such as skirts, dresses, long hair, braids, open jackets, ties, and scarves. In the subsequent chapter, we explore potential approaches to

address this limitation and discuss additional extensions to our current methods. Finally, we provide an outlook on potential research directions that extend beyond our current methodology.

## 5.2 Limitation and Future Work

The research presented in this thesis adheres to the methodology of model-based reconstruction. As previously outlined, this approach relies on a robust prior—a parametric statistical body model. This model is meticulously tracked and extensively personalized to generate virtual avatars that closely resemble the individuals depicted in the input images. Despite the significant advancements contributed to the field through this work, several challenges and limitations still need to be addressed:

1. Handling of Complex Clothing and Occlusions

Most current methods rely on the SMPL model, which primarily represents the human body with minimal clothing. Incorporating complex clothing, accessories, and occlusions remains a significant challenge. Future research could explore the integration of garment-specific models, leveraging semantic information, or employing unsupervised learning techniques to improve the reconstruction of clothed human bodies.

2. Robustness to Lighting and Shadows

Deep learning models may struggle to generalize varying lighting conditions and shadows, which can significantly impact on the accuracy of 3D reconstruction. Developing methods that are more robust to these factors, such as incorporating illumination-invariant features, is an essential direction for future work.

3. Evaluation Metrics and Benchmarks

Evaluating the performance of 3D human body reconstruction methods is non-trivial due to the lack of ground truth data and the subjectivity of visual quality. Developing standardized evaluation metrics and benchmarks, including datasets with accurate ground truth 3D annotations, is crucial for enabling a fair comparison of methods and

guiding future research.

4. Real-Time Performance and Computational Efficiency

Many deep learning-based methods for 3D human body reconstruction require significant computational resources, limiting their applicability in real-time scenarios or on resource-constrained devices. Future research should focus on developing efficient algorithms and network architectures that can deliver high-quality reconstructions with minimal computational overhead.

In summary, while deep learning has shown tremendous potential in the domain of 3D human body reconstruction from images, there is still ample room for improvement and exploration. Addressing the challenges and limitations discussed in this section will pave the way for more accurate, robust, and efficient 3D human body reconstruction techniques, ultimately benefiting a wide range of applications, from entertainment and virtual reality to healthcare and sports analytics.

Our method is limited by the accuracy and precision of some of the deep learning techniques used. Although we have employed multi-angle image optimization to minimize the inherent ambiguity of the prior prediction model method as much as possible, we still need to spend a considerable amount of computational power and time to optimize our loss function. Therefore, in order to achieve faster and higher-precision human body model reconstruction, more work needs to be done to optimize the method. One approach is to train a deep learning network with multi-angle view priors, allowing the network to learn more 3D human body knowledge. Another approach is to improve the speed of the multi-view optimization process.

Creating a realistic avatar that can be controlled through low-cost sensors represents just one intriguing aspect of virtual human development. Certain applications, such as virtual assistants, may need to function autonomously, without a real human directly controlling

the avatar. For instance, virtual assistants like Siri, or Cortana could be embodied by avatars that resemble real humans. Even in scenarios where a real person controls the virtual avatar, sensors may fail to capture the subtle micro-expressions and social cues essential for human communication. Yet, we expect these avatars to replicate such nuances, particularly in collaborative, multi-user applications. Therefore, virtual humans must transcend being mere 3D templates of their real counterparts. Our ultimate goal is to achieve full immersion in virtual environments, where avatars not only act, move, and speak like their real-world counterparts but may also eventually simulate cognitive processes. Achieving this requires a deeper understanding of human behavior, including the identification and modeling of social cues and unique motion patterns.

Constructing a virtual avatar does not signify the completion of this endeavor. To produce convincing performances, the avatar must adapt to the real human it represents. On a broad scale, this adaptation includes mirroring the same clothing, hairstyle, and makeup. On a finer scale, the avatar should reflect the real person's current state—whether they are tired or refreshed, healthy or ill, happy or depressed. These subtle changes are crucial for accurately portraying the real human. Additionally, the virtual human must age alongside its real-world counterpart.

As realistic virtual humans have the potential to revolutionize how we live and communicate, several considerations extend beyond 3D reconstruction and modeling. These include issues related to security, ethics, social sciences, and advancements in display and sensor technologies. In this chapter, we have outlined potential research directions for the development of 3D virtual humans and 3D reconstruction more broadly. We have discussed the critical aspects that will gain importance as we approach the ability to convincingly and indistinguishably digitize ourselves and highlight the research avenues that warrant further exploration.

# Reference

[1] Guo, Kaiwen , et al. "The relightables: volumetric performance capture of humans with realistic relighting." ACM Transactions on Graphics 38.6(2019):1-19.

[2] http://k4w.cn/

[3] https://www.pixart.com/index/

[4] https://www.apple.com

[5] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. PonsMoll, and C. Theobalt, “Single-shot multi-person 3D pose estimation from monocular rgb,” in International Conference on 3D Vision. IEEE, 2018.

[6] Schmidhuber, Jürgen. Deep Learning in Neural Networks: An Overview[J]. Neural Netw, 2015, 61: 85-117.DOI: 10.1016/j.neunet.2014.09.003.

[7] ——, “3D reconstruction of human motion from monocular image sequences,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[8] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image,” in European Conference on Computer Vision. Springer, 2016.

[9] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, “Neural body fitting: Unifying deep learning and model based human pose and shape estimation,” in International Conference on 3D Vision. IEEE, 2018.

[10] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[11] A. O. Balan and M. J. Black, “The naked truth: Estimating body shape ˇ under clothing,” in European Conference on Computer Vision, 2008, pp. 15–29.

[12] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang, “Estimation of human body shape and posture under clothing,” Computer Vision and Image Understanding, vol. 127, pp. 31–42, 2014

[13] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, “Detailed, accurate, human shape

estimation from clothed 3D scan sequences," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017.

[14] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," ACM Transactions on Graphics, vol. 36, no. 4, 2017

[15] Z. Lahner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in European Conference on Computer Vision, 2018, pp. 667–684.

[16] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in Conference on Computer graphics and interactive techniques, 1999, pp. 187–194.

[17] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3D face recognition with a morphable model," in 8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008, pp. 1–6.

[18] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016, pp. 5543–5552.

[19] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," ACM Transactions on Graphics, vol. 36, no. 6, pp. 194:1–194:17, 2017.

[20] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H. P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: face model learning from videos," in IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019, pp. 10 812–10 822.

[21] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision based hand pose estimation: A review," Computer Vision and Image Understanding, vol. 108, no. 1-2, pp. 52–73, 2007.

[22] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei et al., "Accurate, robust, and flexible real-time hand tracking," in Proceedings of the ACM Human Factors in Computing Systems, 2015, pp. 3633–3642.

[23] S. Khamis, J. Taylor, J. Shotton, C. Keskin, S. Izadi, and A. Fitzgibbon, "Learning an efficient model of hand shape variation from depth images," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015, pp. 2540–2548.

[24] J. Romero, D. Tzionas, and M. J. Black, “Embodied hands: Modeling and capturing hands and bodies together,” ACM Transactions on Graphics, vol. 36, no. 6, p. 245, 2017.

[25] L. Luo, H. Li, S. Paris, T. Weise, M. Pauly, and S. Rusinkiewicz, “Multiview hair capture using orientation fields,” in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.

[26] G. Nam, C. Wu, M. H. Kim, and Y. Sheikh, “Strand-accurate multi-view hair capture,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019.

[27] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, “Livecap: Real-time human performance capture from monocular video,” ACM Transactions on Graphics, vol. 38, no. 2, pp. 14:1–14:17, 2019.

[28] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Facevr: Real-time gaze-aware facial reenactment in virtual reality,” ACM Transactions on Graphics, 2018.

[29] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh, “Deep appearance models for face rendering,” ACM Transactions on Graphics, vol. 37, no. 4, p. 68, 2018.

[30] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou et al., “Holoportation: Virtual 3D teleportation in real-time,” in Symposium on User Interface Software and Technology, 2016, pp. 741–754.

[31] W. Paul, B. Janet, and D. Jackson Don, “Pragmatics of human communication. a study of interactional patterns, pathologies, and paradoxes,” New York and London: WW Norton & Co, 1967.

[32] https://shapescale.com/

[33] M. Mori, K. F. MacDorman, and N. Kageki, “The uncanny valley [from the field],” IEEE Robotics & Automation Magazine, vol. 19, no. 2, pp. 98–100, 2012.

[34] M. Mustafa, S. Guthe, J.-P. Tauscher, M. Goesele, and M. Magnor, “How human am I? EEG-based evaluation of animated virtual characters,” in Proceedings of the ACM Human Factors in Computing Systems, May 2017, pp. 5098–5108.

[35] E. P. Hanavan Jr, “A mathematical model of the human body,” Air Force Aerospace Medical Research Lab Wright-Patterson AFB OH, Tech. Rep., 1964.

[36] J. O'rourke and N. I. Badler, "Model-based image analysis of human motion using constraint propagation," IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 6, pp. 522–536, 1980.

[37] D. Hogg, "Model-based vision: a program to see a walking person," Image and Vision computing, vol. 1, no. 1, pp. 5–20, 1983.

[38] H. Ning, L. Wang, W. Hu, and T. Tan, "Model-based tracking of human walking in monocular image sequences," in IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, vol. 1, 2002, pp. 537–540.

[39] D. Metaxas and D. Terzopoulos, "Shape and nonrigid motion estimation through physics-based synthesis," IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 6, pp. 580–591, 1993.

[40] K. Rohr, "Towards model-based recognition of human movements in image sequences," CVGIP: Image understanding, vol. 59, no. 1, pp. 94– 115, 1994.

[41] D. M. Gavrila and L. S. Davis, "3-d model-based tracking of humans in action: a multi-view approach," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 1996, pp. 73–80.

[42] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, "Tracking loose limbed people," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2004, pp. I–421.

[43] D. Thalmann, J. Shen, and E. Chauvineau, "Fast realistic human body deformations for animation and VR applications," in Proceedings of CG International'96, 1996, pp. 166–174.

[44] R. Plankers and P. Fua, "Articulated soft objects for video-based body modeling," in IEEE International Conference on Computer Vision, no. CVLAB-CONF-2001-005. IEEE, 2001, pp. 394–401.

[45] C. Sminchisescu and A. Telea, "Human pose estimation from silhouettes. a consistent approach using distance level sets," in 10th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG '02), 2002.

[46] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2003,

pp. I–I.

[47] N. Magnenat-Thalmann and D. Thalmann, “The direction of synthetic actors in the film rendez-vous à montréal,” IEEE Computer Graphics and applications, vol. 7, no. 12, pp. 9–19, 1987.

[48] J. E. Chadwick, D. R. Haumann, and R. E. Parent, “Layered construction for deformable animated characters,” in ACM Siggraph Computer Graphics, vol. 23, no. 3, 1989, pp. 243–252.

[49] F. Scheepers, R. E. Parent, W. E. Carlson, and S. F. May, “Anatomy based modeling of the human musculature,” in Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, 1997, pp. 163–172.

[50] L. P. Nedel and D. Thalmann, “Modeling and deformation of the human body using an anatomically-based approach,” in Proceedings Computer Animation, 1998, pp. 34–40.

[51] J. P. Lewis, M. Cordner, and N. Fong, “Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation,” in Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, 2000, pp. 165–172.

[52] L. Kavan and J. Žára, “Spherical blend skinning: a real-time deformation of articulated models,” in Proceedings of the 2005 symposium on Interactive 3D graphics and games, 2005, pp. 9–16.

[53] L. Kavan, S. Collins, J. Žára, and C. O’Sullivan, “Geometric skinning with approximate dual quaternion blending,” ACM Transactions on Graphics, vol. 27, no. 4, p. 105, 2008.

[54] ANGUELOV D, SRINIVASAN P, KOLLER D, et al. SCAPE: shape completion and animation of people[J]. ACM Transactions on Graphics, 2005, 24(3): 408-416.

[55] KAVAN L, COLLINS S, ŽÁRA J, et al. Geometric skinning with approximate dual quaternion blending[J]. ACM Transactions on Graphics, 2008, 27(4): 1-23.

[56] JACOBSON A, BARAN I, POPOVIĆ J, et al. Bounded biharmonic weights for realtime deformation[J]. ACM Transactions on Graphics, 2011, 30(4): 1-8.

[57] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: a skinned multi-person linear model[J]. ACM Transactions on Graphics, 2015, 34(6): 1-16.

[58] PAVLAKOS G, CHOUTAS V, GHORBANI N, et al. Expressive body capture: 3D hands, face,

and body from a single image[C]//Proceedings of 2019 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10967-10977.

[59] WU S Z, JIN S, LIU W T, et al. Graphbased 3D multi-person pose estimation using multi-view images[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 11128-11137.

[60] JIANG B Y, ZHANG Y D, WEI X K, et al. H4D: human 4D modeling by learning neural compositional representation[C]// Proceedings of 2022 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 19333-19343.

[61] OSMAN A A A, BOLKART T, BLACK M J. STAR: sparse trained articulated human body regressor[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 598-613.

[62] XU H Y, BAZAVAN E G, ZANFIR A, et al. GHUM & GHUML: generative 3D human shape and articulated pose models[C]//Proceedings of 2020 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6183-6192.

[63] KINGMA D P, WELLING M. Autoencoding variational bayes[J]. arXiv preprint, 2013, arXiv: 1312.6114.

[64] REZENDE D J, MOHAMED S. Variational inference with normalizing flows[J]. arXiv preprint, 2015, arXiv: 1505. 05770.

[65] BHATNAGAR B, TIWARI G, THEOBALT C, et al. Multi-garment Net: learning to dress 3D people from images[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 5419-5429.

[66] ALLDIECKT, PONS - MOLLG, THEOBALT C, et al. Tex2Shape: detailed full human body geometry from a single image[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 2293-2303.

[67] WENGCY, CURLESSB, KEMELMACHER-SHLIZERMAN I. Photo wake-up: 3D character animation from a single photo[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 5901-5910.

[68] ALLDIECT, M AGNOR M, XU WP, et al. Detailed human avatars from monocular video[C]//Proceedings of 2018 International Conference on 3D Vision. Piscataway: IEEE Press,

2018: 98-109.

[69] MA Q L, YANG J L, RANJAN A, et al. Learning to dress 3D people in generative clothing[C]//Proceedings of 2020 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6468-6477.

[70] ALLDIECK T, MAGNOR M, BHATNAGAR B L, et al. Learning to reconstruct people in clothing from a single RGB camera[C]//Proceedings of 2019 I EEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1175-1186.

[71] JIANG B Y, ZHANG J Y, HONG Y, et al. BCNet: learning body and cloth shape from a single image[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 18-35.

[72] WEI W L, LIN J C, LIU T L, et al. Capturing humans in motion: temporal attentive 3D human pose and shape estimation from monocular video[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 13201-13210.

[73] ZHANG Q, FU B, YE M, et al. Quality dynamic human body modeling using a single lowcost depth camera[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:676-683.

[74] WEISS A, HIRSHBERG D, BLACKM J. Home 3dbody scans from noisy image and range data[C]//2011 International Conference on Computer Vision. IEEE, 2011: 1951-1958.

[75] ZHAO T, LI S, NGAN K N, et al. 3-dreconstruction of human body shape from a single commodity depth camera[J]. IEEE Transactions on Multimedia, 2018, 21(1):114-123.

[76] DIBRA E, JAIN H, OZTIRELI C, et al. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks[C]//2016 fourth international conference on 3D vision (3DV). IEEE,2016: 108-117.

[77] GUAN P, WEISS A, BALAN A O, et al. Estimating human shape and pose from a single image[C]//IEEE International Conference on Computer Vision. 2009: 1381-1388.

[78] XUY, ZHUS C, TUNGT. Denserac: Joint 3dpose and shape estimation by dense render-and compare[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019:7760 - 7770.

[79] BOGO F, KANAZAWA A, LASSNER C, et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image[C]//European Conference on Computer Vision. 2016: 561-578.

[80] SIGAL L, BALAN A, BLACK M. Combined discriminative and generative articulated pose and non-rigid shape estimation[J]. Advances in neural information processing systems, 2007,20:1337-1344.

[81] JI Z, QI X, WANG Y, et al. Shape-from-mask: A deep learning based human body shape reconstruction from binary mask images[J]. arXiv preprint arXiv:1806.08485, 2018.

[82] STREUBER S, QUIROS-RAMIREZ M A, HILL MQ, et al. Body talk: Crowd shaping realistic 3d avatars with words[J]. ACM Transactions on Graphics (TOG), 2016, 35(4):1-14.

[83] SEO H, MAGNENAT-THALMANN N. An example-based approach to human body manipulation[J]. Graphical Models, 2004,66(1):1-23.

[84] WUHRER S, SHU C. Estimating 3d human shapes from measurements[J]. Machine vision and applications, 2013,24(6):1133-1147.

[85] ALLEN B, CURLESS B, POPOVIĆ Z. The space of human body shapes: reconstruction and parameterization from range scans[J]. ACM transactions on graphics (TOG), 2003,22 (3):587-594.

[86] Xie Haoyang. High precision 3D human body reconstruction and its application in virtual fitting [D]. Donghua University, 2020

[87] ROTHER C, KOLMOGOROVV, BLAKE A." grabcut" interactive foreground extraction using iterated graphcuts[J]. ACM transactions on graphics (TOG), 2004, 23(3):309-314.

[88] LASSNER C, ROMERO J, KIEFEL M, et al. Unite the people: Closing the loop between 3D and 2D human representations[C]//IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017: 6050-6059.

[89] Xu Haocan, Li Jituo, Lu Guodong. Reconstructing 3D human body from a single dressing image using LeNet-5 [J]. Journal of Zhejiang University, 2021, 55(1):153-161.

[90] DIBRA E, JAIN H, OZTIRELI C, et al. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4826-4836.

[91] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE : Fast and accurate scans from an image in less than a second. In International Conference on Computer Vision (ICCV). 5330–5339 (2019).

[92] Joo H, Neverova N, Vedaldi A. Exemplar fine-tuning for 3d human pose fitting towards in the-wild 3d human pose estimation[J]. arXiv preprint arXiv:2004.03686, 2020.

[93] ZIMMERMANN C, BROX T. Learning to estimate 3D hand pose from single RGB images [C]//IEEE International Conference on Computer Vision. 2017: 4903-4911.

[94] KOCABAS M, ATHANASIOUN, BLACKM J. Vibe: Video inference for human body pose and shape estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:5253-5263.

[95] KANAZAWA A, ZHANG J Y, FELSEN P, et al. Learning 3d human dynamics from video [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5614-5623.

[96] GOODFELLOWI J, POUGET-ABADIE J, MIRZAM, et al. Generative adversarial networks [J]. arXiv preprintarXiv:1406.2661, 2014.

[97] GÜLER R A, NEVEROVA N, KOKKINOS I. Densepose: Dense human pose estimation in the wild[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7297-7306.

[98] OECHSLE M, MESCHEDER L, NIEMEYER M, et al. Texture fields: learning texture representations in function space[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 4530-4539.

[99] BLANZ V, VETTER T. A morphable model for the synthesis of 3D faces[C]// Proceedings of the 26th annual conference on Computer graphics and interactive techniques. New York: ACM Press, 1999: 187-194.

[100] LATTAS A, MOSCHOGLOU S, GECER B, et al. AvatarMe: realistically renderable 3D facial reconstruction “In-the wild” [C]//Proceedings of 2020 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 757-766.

[101] ZHENG M W, YANG H Y, HUANG D, et al. ImFace: a nonlinear 3D morphable face model with implicit neural representations[C]//Proceedings of 2022IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 20311-20320.

[102] ZHENG Y F, ABREVAYA V F, BÜHLER M C, et al. I M avatar: implicit morphable head avatars from videos[C]//Proceedings of 2 022 I EEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 13535- 13545.

[103] GECER B, PLOUMPIS S, KOTSIA I, et al. GANFIT: generative adversarial network fitting for high fidelity 3D face reconstruction[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1155-1164.

[104] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]// Proceedings of 2019 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4396-4405.

[105] TEWARI A, ELGHARIB M, BHARAJ G, et al. StyleRig: rigging StyleGAN for 3D control over portrait images[C]// Proceedings of 2020 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6141-6150.

[106] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 8107-8116.

[107] LUO H W, NAGANO K, KUNG H W, et al. Normalized avatar synthesis using StyleGAN and perceptual refinement[C]// Proceedings of 2021 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11657-11667.

[108] SHEN Y, LIANG J B, LIN M C. GAN based garment generation using sewing pattern images[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 225-247.

[109] RAFFIEE A H, SOLLAMI M. GarmentGAN: photo-realistic adversarial fashion transfer[C]//Proceedings of 2020 25th International Conference on Pattern Recognition. Piscataway: IEEE Press, 2021: 3923-3930.

[110] CURLESS B, LEVOY M. A volumetric method for building complex models from range images[C]//Proceedings of the 23rd Annual Conference on Computer graphics and Interactive Techniques. New York: ACM Press, 1996: 303-312.

[111] IZADI S, KIM D, HILLIGES O, et al. Kinect Fusion: real - time 3 D reconstruction and interaction using a moving depth camera[C]//Proceedings of the 24th annual ACM

symposium on User Interface Software and Technology. New York: ACM Press, 2011: 559-568.

[112] DAI A, NIEßNER M, ZOLLHÖFER M, et al. BundleFusion: real-time globally consistent 3D reconstruction using on the-fly surface reintegration[J]. ACM Transactions on Graphics, 2017, 36(4): 76a.

[113] SITZMANN V, THIES J, HEIDE F, et al. DeepVoxels: learning persistent 3D feature embeddings[C]//Proceedings of 2 019 I EEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2432- 2441.

[114] MA X X, SU J J, WANG C Y, et al. Context modeling in 3D human pose estimation: a unified perspective[C]// Proceedings of 2021 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 6234-6243.

[115] ZHENG Z R, YU T, WEI Y X, et al. DeepHuman: 3D human reconstruction from a single image[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 7738-7748.

[116] LOMBARDI S, SIMON T, SARAGIH J, et al. Neural volumes: learning dynamic renderable volumes from images[J]. ACM Transactions on Graphics, 2019, 38(4): 1-14.

[117] MESCHEDER L, OECHSLE M, NIEMEYER M, et al. Occupancy networks: learning 3D reconstruction in function space[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4455-4465.

[118] D. R. Canelhas, E. Schaffernicht, T. Stoyanov, A. J. Lilienthal, and A. J. Davison. An eigenshapes approach to compressed signed distance fields and their utility in robot mapping. arXiv preprint arXiv:1609.02462, 2016.

[119] CHENZQ, ZHANGH. Learning implicit fields for generative shape modeling[C]//Proceedings of 2019 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 5932-5941.

[120] SITZMANN V, ZOLLHÖFER M, WETZSTEIN G. Scene representation networks: continuous 3D-structureaware neural scene representations[J]. arXiv preprint, 2019, arXiv: 1906.01618.

[121] YANG G S, VO M, NEVEROVA N, et al. BANMo: building animatable 3D neural models

from many casual videos[C]//Proceedings of 2022 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 2853-2863.

[122] NEVEROVA N, NOVOTNY D, KHALIDOV V, et al. Continuous surface embeddings[J]. arXiv preprint, 2020, arXiv: 2011.12438.

[123] PARK J J, FLORENCE P, STRAUB J, et al. DeepSDF: learning continuous signed distance functions for shape representation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 165-174.

[124] BOŽIČ A, PALAFOX P, ZOLLHÖFER M, et al. Neural deformation graphs for globally consistent non-rigid reconstruction[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 1450-1459.

[125] MITTAL P, CHENG Y C, SINGH M, et a l. Auto SDF: shape prior s for 3D completion, reconstruction and generation[C]//Proceedings of 2022 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 306-315.

[126] SAITO S, HUANG Z, NATSUME R, et al. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 2304-2314.

[127] SAITO S, SIMON T, SARAGIH J, et al. PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3D human digitization[C]//Proceedings of 2 02 0 I EEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 81-90.

[128] JADHAV O, PATIL A, SAM J, et al. Virtual dressing using augmented reality[J]. ITM Web of Conferences, 2021, 40.

[129] ZHU X, LIAO T, LYU J, et al. MVPhuman dataset for 3D human avatar reconstruction from unconstrained frames[J]. arXiv preprint, 2022, arXiv: 2204.11184.

[130] DENG B Y, LEWIS J P, JERUZALSKI T, et al. NASA Neural articulated shape approximation[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 612-628.

[131] CAO Y K, CHEN G Y, HAN K, et al. JIFF: jointly-aligned implicit face function for high quality single view clothed human reconstruction[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 2719-

2729.

[132] BHATNAGAR B L, SMINCHISESCU C, THEOBALT C, et al. Combining implicit function learning and parametric models for 3D human reconstruction[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 311-329.

[133] SAITO S, YANG J L, MA Q L, et al. SCANimate: weakly supervised learning of skinned clothed avatar networks[C]// Proceedings of 2021 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 2885-2896.

[134] XIU Y L, YANG J L, TZIONAS D, et al. ICON: implicit clothed humans obtained from normals[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 13286-13296.

[135] ZHENG Z R, HUANG H, YU T, et al. Structured local radiance fields for human avatar modeling[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 15872-15882.

[136] XU T H, FUJITA Y, MATSUMOTO E. Surface-aligned neural radiance fields for controllable 3D human synthesis[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 15862-15871.

[137] LIU W, PIAO Z X, MIN J, et al. Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis[C]// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 5903-5912.

[138] GRIGOREV A, ISKAKOV K, IANINA A, et al. StylePeople: a generative model of full body human avatars[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5147- 5156.

[139] RAJ A, ZOLLHÖFER M, SIMON T, et al. Pixel-aligned volumetric avatars[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11728-11737.

[140] GAFNI G, THIES J, ZOLLHÖFER M, et al. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 8645-8654.

[141] ZHENG Z R, YU T, LIU Y B, et al. PaMIR: parametric model-conditioned implicit representation for image-based human reconstruction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3170- 3184.

[142] ZHENG Y, SHAO R Z, ZHANG Y X, et al. DeepMultiCap: performance capture of multiple characters using sparse multi-view cameras[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 6219-6229.

[143] YANG Z, WANG S L, MANIVASAGAM S, et al. S3: neural shape, skeleton, and skinning fields for 3D huma n modeling[C]//Proceedings of 2021 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13279-13288.

[144] https://www.vicon.com/hardware/

[145] PLAGEMANN C, GANAPATHI V, KOLLER D, et al. Real-time identification and localization of body parts from depth images[C]//2010 IEEE International Conference on Robotics and Automation. IEEE,2010: 3108-3113.

[146] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images[C]//CVPR2011. 2011: 1297-1304.

[147] TAYLOR J, SHOTTON J, SHARP T, et al. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 103-110.

[148] GANAPATHI V, PLAGEMANN C, KOLLER D, et al. Real time motion capture using a single time-of-flight camera[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE,2010: 755-762.

[149] GANAPATHI V, PLAGEMANN C, KOLLER D, et al. Real-time human pose tracking from range data[C]//European conference on computer vision. Springer, 2012: 738-751.

[150] IONESCU C, PAPAVA D, OLARUV, et al. Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7):1325-1339.

[151] SIGAL L, BALAN A O, BLACK M J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International journal of computer vision, 2010, 87(1-2):4.

[152] LI S, CHANA B. 3d human pose estimation from monocular images with deep convolutional neural network[C]//Asian Conference on Computer Vision. Springer, 2014: 332-347.

[153] POPA A I, ZANFIR M, SMINCHISESCU C. Deep multitask architecture for integrated 2d and 3d human sensing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6289-6298.

[154] PAVLAKOS G, ZHOU X, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017:7025-7034.

[155] FANG H S, XU Y, WANG W, et al. Learning pose grammar to encode human body configuration for 3d pose estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.

[156] SUN X, XIAO B, WEI F, et al. Integral human pose regression[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 529-545.

[157] LEE K, LEE I, LEE S. Propagating lstm:3d pose estimation based on joint interdependency [C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 119-135.

[158] HABIBIE I, XU W, MEHTA D, et al. In the wild human pose estimation using explicit 2d features and intermediate 3d representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10905-10914.

[159] FABBRI M, LANZI F, CALDERARA S, et al. Compressed volumetric heatmaps for multi-person 3d pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7204-7213.

[160] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3D human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2640-2649.

[161] MEHTA D, SRIDHAR S, SOTNYCHENKO O, et al. VNect: Real-time 3D human pose estimation with a single RGB camera[J]. ACM Transactions on Graphics (TOG), 2017, 36 (4):44.

[162] LUO C, CHU X, YUILLE A. Orinet: A fully convolutional network for 3d human pose estimation[J]. arXiv preprintarXiv:1811.04989, 2018.

[163] JOO H, SIMON T, SHEIKH Y. Totalcapture: A 3D deformation model for tracking faces, hands, and bodies[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8320-8329.

[164] HABERMANN M, XU W, ZOLLHOEFER M, et al. Deepcap: Monocular human performance capture using weak supervision[J]. arXiv: Computer Vision and Pattern Recognition, 2020.

[165] SUN X, SHANG J, LIANG S, et al. Compositional human pose regression[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2602-2611.

[166] SUN X, LI C, LIN S. Explicit spatiotemporal joint relation learning for tracking human pose[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.

[167] YANGW, OUYANGW, WANGX, et al. 3Dhumanpose estimation in the wild by adversarial learning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 52555264.

[168] ZHOU X, HUANG Q, SUN X, et al. Towards3D human pose estimation in the wild: a weakly-supervised approach[C]//IEEE International Conference on Computer Vision. 2017: 398-407.

[169] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[170] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation [C]//European conference on computer vision. 2016: 483-499.

[171] CHEN Y, WANG Z, PENG Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[172] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]//The European Conference on Computer Vision (ECCV). 2018.

[173] ZHAO L, PENG X, TIANY, et al. Semantic graph convolutional networks for 3d human pose regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:3425-3435.

[174] CHENC H, RAMANAND. 3D human pose estimation= 2D pose estimation+

matching[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7035-7043.

[175] HOSSAINMR I, LITTLE J J. Exploiting temporal information for 3d human pose estimation [C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 68-84.

[176] TEKIN B, MÁRQUEZ-NEILA P, SALZMANN M, et al. Learning to fuse 2d and 3d image cues for monocular body pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3941-3950.

[177] WANG J, HUANG S, WANG X, et al. Not all parts are created equal: 3d pose estimation by modeling bi-directional dependencies of body parts[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7771-7780.

[178] PAVLAKOS G, ZHOU X, DANIILIDIS K. Ordinal depth supervision for 3D human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7307-7316.

[179] A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference, volume 2, pages 1033–1038. IEEE, 1999.

[180] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pages 341–346. ACM, 2001.

[181] Daniel Yamins, James J Dicarlo. Using goal-driven deep learning models to understand sensory cortex. 2016, Nature Neuroscience.

[182] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, “Dyna: a model of dynamic human shape in motion,” ACM Transactions on Graphics, vol. 34, p. 120, 2015.

[183] JOEYDEVRIES. Textures[Z]. 2022.

[184] ZENG W, OUYANG W L, LUO P, et al. 3D human mesh regression with dense correspondence[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7052-7061.

[185] J. F. Blinn and M. E. Newell, “Texture and reflection in computer generated images,” Communications of the ACM, vol. 19, no. 10, pp. 542–547, 1976.

[186] B. K. Horn, “Shape from shading: A method for obtaining the shape of a smooth opaque object from one view,” Ph.D. dissertation, Massachusetts Inst. of Technology, 1970.

[187] PENG S D, Z H A NG Y Q, XU Y H, et al. Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C]//Proceedings of 2021 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 9050-9059.

[188] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: a benchmark[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 1116-1124.

[189] JING X Y, FENG Q, LAI Y K, et al. STATE: learning structure and texture representations for novel view synthesis[C]//Proceedings of IEEE International Conference on Computer Vision. [S. l.: s. n.], 2022.

[190] PATEL P, HUANG C H P, TESCH J, et al. AGORA: avatars in geography optimized for regression analysis[C]//Proceedings of 2 021 I EEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13463- 13473.

[191] ALLDIECK T, MAGNOR M, XU W P, et al. Video based reconstruction of 3D people models[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8387-8397.

[192] Z H A NG R, IS OL A P, EFRO S A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of 2018 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 586-595.

[193] BROWNLEE J. How to implement the frechet inception distance (FID) for evaluating GANs[Z]. 2019.

[194] SETIADI D R I M. PSNR vs SSIM: imperceptibility quality assessment for image steganography[J]. Multimedia Tools and Applications, 2021, 80(6): 8423- 8444.

[195] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In International Conference on Computer Vision, pages 2252–2261, 2019. 1, 2, 3, 4, 5, 6, 8, 7

[196] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to

estimate 3D human pose and shape from a single color image. In IEEE Conference on Computer Vision and Pattern Recognition, pages 459–468, 2018. 1, 2

[197] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In European Conference on Computer Vision, pages 20–40, 2020. 2

[198] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In IEEE Conference on Computer Vision and Pattern Recognition, pages 7779–7788, 2020. 2

[199] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsuper vised video object segmentation with co-attention siamese networks. In IEEE Conference on Computer Vision and Pat tern Recognition, 2019. 2

[200] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In IEEE Conference on Computer Vision and Pattern Recognition, pages 7794 7803, 2018. 2

[201] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision, 2016. 3, 5

[202] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In European Conference on Computer Vision, 2014. 6, 1

[203] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In IEEE Conference on Computer Vision and Pattern Recognition, 2014. 6, 1

[204] Sam Johnson and Mark Everingham. Learning effective hu man pose estimation from inaccurate annotation. In IEEE Conference on Computer Vision and Pattern Recognition, 2011. 6, 1

[205] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In International Conference on 3DVision, 2017. 6, 1

[206] Nikos Kolotouros. Pytorch implementation of the neural mesh renderer, 2018. 6

[207] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In IEEE

Conference on Computer Vision and Pattern Recognition, 2018. 6

[208] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using IMUs and a moving camera. In European Conference on Computer Vision, pages 614–631, 2018. 2, 3, 6

[209] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object occluded human shape and pose estimation from a single color image. In IEEE Conference on Computer Vision and Pattern Recognition, pages 7374–7383, 2020. 1, 2, 3, 6, 5

[210] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D pose estimation: motion to the rescue. In Advances in Neural Information Processing, 2019. 6, 2

[211] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In International Conference on Computer Vision, 2019. 6, 2

[212] Zhengyi Luo, S. Golestaneh, and Kris M. Kitani. 3D human motion estimation via motion compression and refinement. In Asian Conference on Computer Vision, pages 324–340, 2020. 1, 6, 2

[213] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In Euro pean Conference on Computer Vision, pages 769–787, 2020. 5, 6

[214] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In European Conference on Computer Vision, pages 465–481, 2020. 2, 5, 6

[215] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In Euro pean Conference on Computer Vision, pages 752–768, 2020. 5, 6

[216] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In European Conference on Computer Vision, 2020. 6

[217] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In IEEE Conference on Computer

Vision and Pattern Recognition, pages 4501–4510, 2019. 6

[218] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In International Conference on 3DVision, 2018. 2, 6, 7

[219] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In IEEE Conference on Computer Vision and Pattern Recognition, pages 10884 10894, 2019. 2, 5, 6

[220] S. Fuhrmann, F. Langguth, and M. Goesele. Mve-a multiview reconstruction environment. In EUROGRAPHICS Workshops on Graphics and Cultural Heritage. 11– 18 (2014).

[221] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In IEEE International Conf. on Computer Vision. 2320–2327 (2011).

[222] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture : A 3D deformation model for tracking faces, hands, and bodies. In Computer Vision and Pattern Recognition (CVPR).8320–8329 (2018)

[223] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture : 3D hands, face, and body from a single image. In Computer Vision and Pattern Recognition (CVPR). 10975–10985 (2019).

[224] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands : Modeling and capturing hands and bodies together. Transactions on Graphics (TOG). 36(6) :245 :1–245 :17(2017).

[225] Vasileios Choutas, Lea Muller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression via linguistic attributes and anthropometric measurements. In Computer Vision and Pattern Recognition (CVPR). (2022).

[226] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place : Monocular regression of 3D people in depth. In Computer Vision and Pattern Recognition (CVPR). (2022).

[227] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for

visual environment reconstruction. In Computer Vision and Pattern Recognition (CVPR). (2022).

[228] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-Degree textures of people in clothing from a single image. In International Conference on 3D Vision (3DV). 643–653 (2019).

[229] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica K. Hodgins. MonoClothCap: Towards temporally coherent clothing capture from monocular RGB video. In International Conference on 3D Vision (3DV). 322–332 (2020).

[230] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In Computer Vision and Pattern Recognition (CVPR). 4491–4500 (2019).

[231] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In Conference on Neural Information Processing Systems (NeurIPS). (2020).

[232] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3D self-portraits in Seconds. In Computer Vision and Pattern Recognition (CVPR). 1341–1350 (2020).

[233] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. PINA : Learning a personalized implicit neural avatar from a single RGB-D video sequence. In Computer Vision and Pattern Recognition (CVPR). (2022).

[234] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In ACM SIGGRAPH 2020 Real-Time Live. 1–1 (2020).

[235] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In European Conference on Computer Vision (ECCV). **12368**, 49–67 (2020).

[236] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL : Automatic estimation of 3D human pose and shape from a single image. In European Conf. on Computer Vision. Springer International Publishing. (2016).

[237] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE : Part attention regressor for 3D human body estimation. In International Conference on

Computer Vision (ICCV). 11127–11137 (2021).

[238] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2019).

[239] S. Liu, T. Li, W. Chen, and H. Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In Proceedings of the IEEE International Conference on Computer Vision. 7708–7717 (2019).

[240] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rossl, and H.P. Seidel. Laplacian surface editing. In Eurographics/ACM SIGGRAPH symposium on Geometry processing. 175–184 (2004).

[241] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. ACM Transactions on Graphics (ToG), 37(2):1–15, 2018. 1, 2, 3, 5, 6, 7, 8

[242] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt. Model-based outdoor performance capture. In 2016 Fourth International Conference on 3D Vision (3DV), pages 166–175. IEEE, 2016. 3, 6

[243] F. Bogo, M. J. Black, M. Loper, and J. Romero, “Detailed full-body reconstructions of moving people from monocular RGB-D sequences,” in IEEE International Conference on Computer Vision. IEEE, 2015, pp. 2300–2308.

[244] Wang, Moyu & Yang, Qingping. (2024). From prediction to measurement, an efficient method for digital human model obtainment. International Journal of Metrology and Quality Engineering. 15. 10.1051/ijmqe/2023015.