# MONOCULAR 3D HUMAN BODY RECONSTRUCTION

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in*
*Computer Science and Engineering*
*by Research*

*by*

*ABBHINAV VENKAT*
*201302063*
ABBHINAV.VENKAT@RESEARCH.IIIT.AC.IN

*International Institute of Information Technology*
*Hyderabad - 500 032, INDIA*
*NOV 2020*

International Institute of Information Technology
Hyderabad, India

# CERTIFICATE

It is certified that the work contained in this thesis, titled **"Monocular 3D Human Body Reconstruction"** by Abbhinav Venkat, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Dr. Avinash Sharma,
Centre for Visual Information Technology,
Kohli Center on Intelligent Systems,
IIIT Hyderabad.

To My Late Grandfather

# Acknowledgments

Research is a beautiful journey of self-discovery. It parallels life - from the constant challenges to momentary success. Along this arduous voyage, I've had the privilege of meeting several special people who've left an everlasting impact on me, without whose support and constant encouragement this thesis wouldn't have existed. I'm forever grateful for the memories they've given me during these three remarkable years of my life.

While there are many to thank, I must begin with my family. They've ensured that they put no burden upon me - emotionally or financially, thereby giving me a platform to take risks and achieve my dreams. I must particularly thank my parents for providing me constant motivation during phases where I doubted myself. No amount of gratitude is sufficient to repay their efforts and sacrifices.

Next is my advisor, Dr. Avinash Sharma, who always gave me the freedom to steer my journey as per my liking and find my feet in this new world. The researcher I am today is because of his support and belief in me. From daily discussions on research to philosophical ones on life, he has been a mentor in the true sense of the word, guiding me to develop my philosophy to deal with research and life. I am truly indebted to him for allowing me to grow in my professional and personal life.

In CVIT, I found a great group of friends who were in a similar boat. The camaraderie and constant support were unparalleled. Specifically, I'd like to acknowledge Govinda, Abhijeet, Sukesh, Joyneel, Neeraj, and Chetan. From philosophical talks about the meaning of life to ordering every sweet-dish under the sun, these were my go-to people. By exploring mutual interests via taking swimming lessons, long-distance running, yoga, table tennis, and much more, we kept each other in good spirits. I'd also like to extend a special thanks to Sagar for being a part of my journey, helping me out in times of turmoil, and sharing responsibility for the group. Further, I'm grateful for the confidence and belief in my ideas shown by my co-authors Chaitanya and Yudhik. I've learned a lot from them and hope they do very well in life.

This acknowledgment is incomplete without mentioning friends from my undergraduate days. In particular, Vishal, my gym buddy; Debayan, my confidante and go-to person; Gaurav, for all my binge-watching needs and Arjun, for things of the soul. I am grateful for meeting these people and shall always cherish the moments we've shared.

# Abstract

Monocular 3D human reconstruction is a very relevant problem due to numerous applications to the entertainment industry, e-commerce, health care, mobile-based AR/VR platforms, etc. However, it is severely ill-posed due to self-occlusions from complex body poses and shapes, clothing obstructions, lack of surface texture, background clutter, single view, etc. Conventional approaches address these challenges by using different sensing systems - marker-based, marker-less multi-view cameras, inertial sensors, and 3D scanners. Although effective, such methods are often expensive and have limited wide-scale applicability. In an attempt to produce scalable solutions, a few have focused on fitting statistical body models to monocular images, but are susceptible to the costly optimization process.

Recent efforts focus on using data-driven algorithms such as deep learning to learn priors directly from data. However, they focus on template model recovery, rigid object reconstruction, or propose paradigms that don't directly extend to recovering personalized models. To predict accurate surface geometry, our first attempt was VolumeNet, which predicted a 3D occupancy grid from a monocular image. This was the first of its kind model for non-rigid human shapes at that time. To circumvent the ill-posed nature of this problem (aggravated by an unbounded 3D representation), we follow the ideology of providing maximal training priors with our unique training paradigms, to enable testing with minimal information. As we did not impose any body-model based constraint, we were able to recover deformations induced by free-form clothing. Further, we extended VolumeNet to PoShNet by decoupling Pose and Shape, in which we learn the volumetric pose first, and use it as a prior for learning the volumetric shape, thereby recovering a more accurate surface.

Although volumetric regression enables recovering a more accurate surface reconstruction, they do so without an animatable skeleton. Further, such methods yield reconstructions of low resolution at higher computational cost (regression over the cubic voxel grid) and often suffer from an inconsistent topology via broken or partial body parts. Hence, statistical body models become a natural choice to offset the ill-posed nature of this problem. Although theoretically, they are low dimensional, learning such models has been challenging due to the complex non-linear mapping from the image to the relative axis-angle representation. Hence, most solutions rely on different projections of the underlying mesh (2D/3D keypoints, silhouettes, etc.). To simplify the learning process, we propose the CR framework that uses classification as a prior for guiding the regression's learning process. Although recovering personalized models with high-resolution meshes isn't a possibility in this space, the framework shows that learning such template models can be difficult without additional supervision.

As an alternative to directly learning parametric models, we propose HumanMeshNet to learn an "implicitly structured point cloud", in which we make use of the mesh topology as a prior to enable better learning. We hypothesize that instead of learning the highly non-linear SMPL parameters, learning its corresponding point cloud (although high dimensional) and enforcing the same parametric template topology on it is an easier task. This proposed paradigm can theoretically learn local surface deformations that the body model based PCA space can't capture. Further, going ahead, attempting to produce high-resolution meshes (with accurate geometry details) is a natural extension that is easier in 3D space than in the parametric one.

In summary, in this thesis, we attempt to address several of the aforementioned challenges and empower machines with the capability to interpret a 3D human body model (pose and shape) from a single image in a manner that is non-intrusive, inexpensive and scalable. In doing so, we explore different 3D representations that are capable of producing accurate surface geometry, with a long-term goal of recovering personalized 3D human models.

# Contents

# List of Figures

# List of Tables

# Chapter *1*

# Introduction

Recovering the 3D human shape from a monocular image is a severely ill-posed problem. Yet, humans can imagine unseen parts of a known object and perform this estimation successfully. In addition to the 3D shape and pose, we also possess the ability to predict the texture on the object, its 3D orientation, and the viewpoint from which the image was captured. Interestingly, there is an elegant experiment in psychology which introduces the concept of "mental rotation", according to which humans visualize 2D images (of objects) to be in the three-dimensional space to perform mental operations (rotations, translation, etc.) on them [73].

However, this is not the case with computers. Several of the existing solutions [8, 13, 22, 50, 65, 78, 91] are extremely limited, especially for non-rigid shapes with high articulations (such as humans), due to limitations in sensing as well as interpretation. Specifically, concerning optical capture systems, while marker-based capture [50] often restricts the performer's motion and requires skin-tight clothing, markerless multi-view systems [22] are expensive and limited to studio environments. Along similar lines, 3D body scanners [78], although accurate, are expensive and tedious to use. Such systems that require dedicated cameras to track and capture the human often face issues in recording human motion in dynamic outdoor settings (for example, while biking or skiing). Although systems with inertial sensors [8, 65, 91] address this with pose estimation from specialized sensors attached to the body, they are often intrusive and have limited wide-scale applicability. Further, in an attempt to produce scalable solutions, a few [13] have focused on fitting statistical body models to monocular images, but succumb to the costly optimization process.

Recently, there has been a resurgence of data-driven algorithms such as deep-learning due to the surge in graphical computing power (GPUs) and the availability of large-scale synthetic datasets [85]. This enables solving ill-posed problems such as monocular reconstruction by learning priors directly from the data. However, most of the efforts have been focused on template model recovery [41], rigid-object reconstruction [21], or propose paradigms [76, 94] that are not directly extendable to learn personalized human body shapes.

Therefore, in this thesis, we strive to empower machines with the capability to interpret 3D human poses and shapes from a single image, with an eventual aim of obtaining personalized 3D models, by

exploring 3D representations suitable for obtaining accurate surface geometry. Our designs try to address several of the aforementioned bottlenecks and propose models with wide-scale applicability.

## 1.1 Motivation - "Dissecting The Problem"

(a) Input Image (b) 3D Human - View 1 (c) 3D Human - View 2

Figure 1.1: **An illustration of the Problem of Monocular 3D Human Reconstruction**. The figure shows (a) the input RGB image and (b) the corresponding 3D Human aligned to the input view and (c) rotated to another view (for aesthetic reasons).

To understand the motivation behind solving "Monocular 3D Human Reconstruction", we shall dissect each word of the problem to reveal several exciting prospects. Each "perspective" offered below talks about the relevance of that facet of the problem -

**The "3D" Perspective -** We live in a three-dimensional world. Therefore, for problems related to the shape and structure of an object, lifting the representation space to 3D is a natural step. Although it is more computationally expensive, this allows for a more accurate representation of the real world.

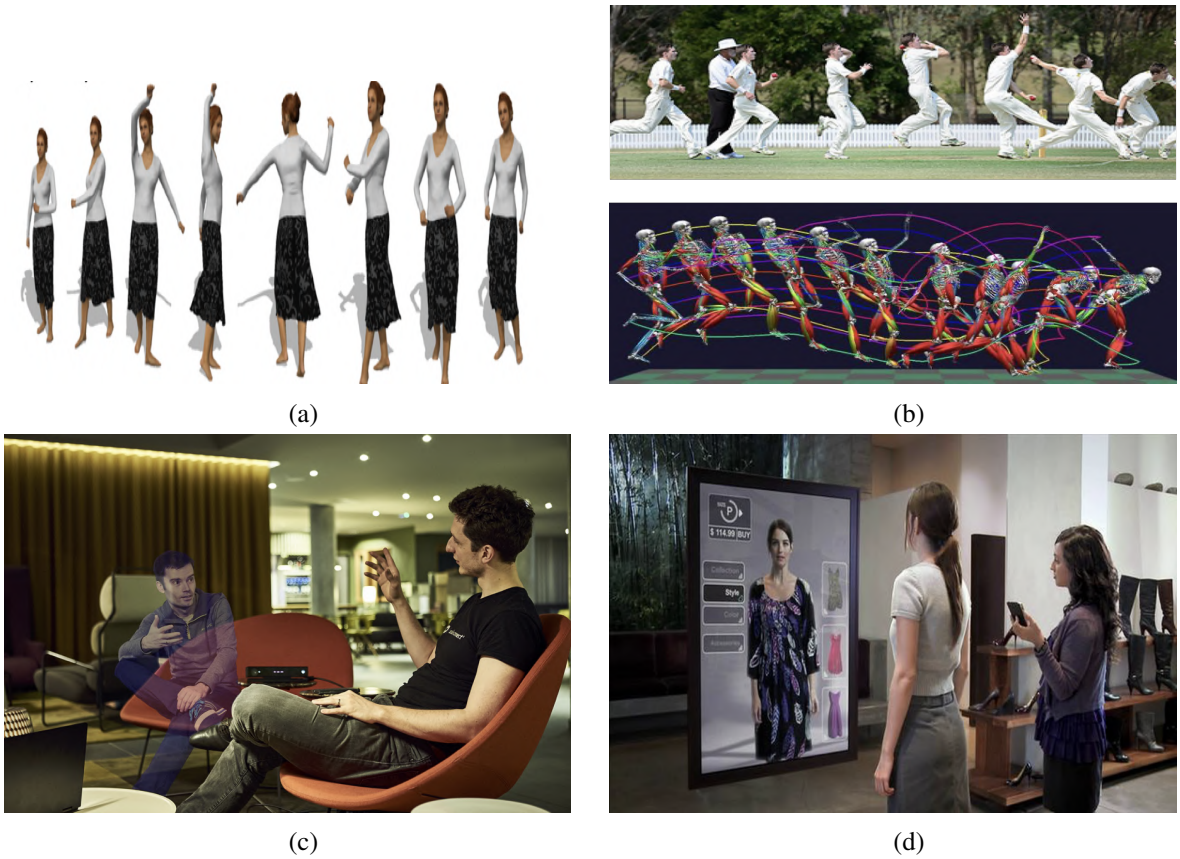
**The "Human" Perspective -** To effectively interact with our 3D world, modeling its constituent elements are of paramount importance. Although there exist several such real-world objects, one of the most common, yet, vital constituents of our world tend to be humans. Humans are widely found in most of our scenes of interest and recovering their 3D model is a vital step for humans/human-like agents to effectively interpret and understand our world, and perform tasks such as recognition, navigation, prediction, and much more.

**The "Reconstruction" Perspective -** Successful 3D inference paves the way to solving several higher-order problems in 3D Computer Vision such as 3D visual question answering, modeling human-

human/human-robot interactions, temporal prediction of human actions, etc. Hence, the potential impact of learning to represent the structure of a human in 3D is one to look forward to.

**The "Monocular" Perspective -** Usage of calibrated multi-camera setups is not only expensive but also limits the applicability of the problem to those in controlled lab setups. On the other hand, although challenging and severely ill-posed, performing monocular reconstruction makes it non-intrusive, affordable, and scalable to real-world environments. As shown in Figure 1.2, this problem is of high practical importance in real-world scenarios, with applications in the animation, gaming, entertainment industry, AR/VR, e-commerce, healthcare, robotics and many more.

Therefore, by putting all these elements together, we aim to reconstruct an accurate 3D representation of a human in a fashion that makes it non-intrusive, cost-efficient, and applicable to real-world scenarios. In doing so, we open a world of opportunities that blur the lines between the real and virtual world. Further, this paves the way for solving several higher-order problems in 3D Computer Vision.



(a) (b) (c) (d)

Figure 1.2: **A wide range of applications possible because of 3D Human Shape Recovery** - (a) Procuring animatable 3D shape for gaming and animation, (b) Analysis of Sports Bio-mechanics, (c) Augmented Reality (AR) based Chat rooms, (d) Virtual Reality (VR) based Dressing Rooms.

## 1.2 Challenges

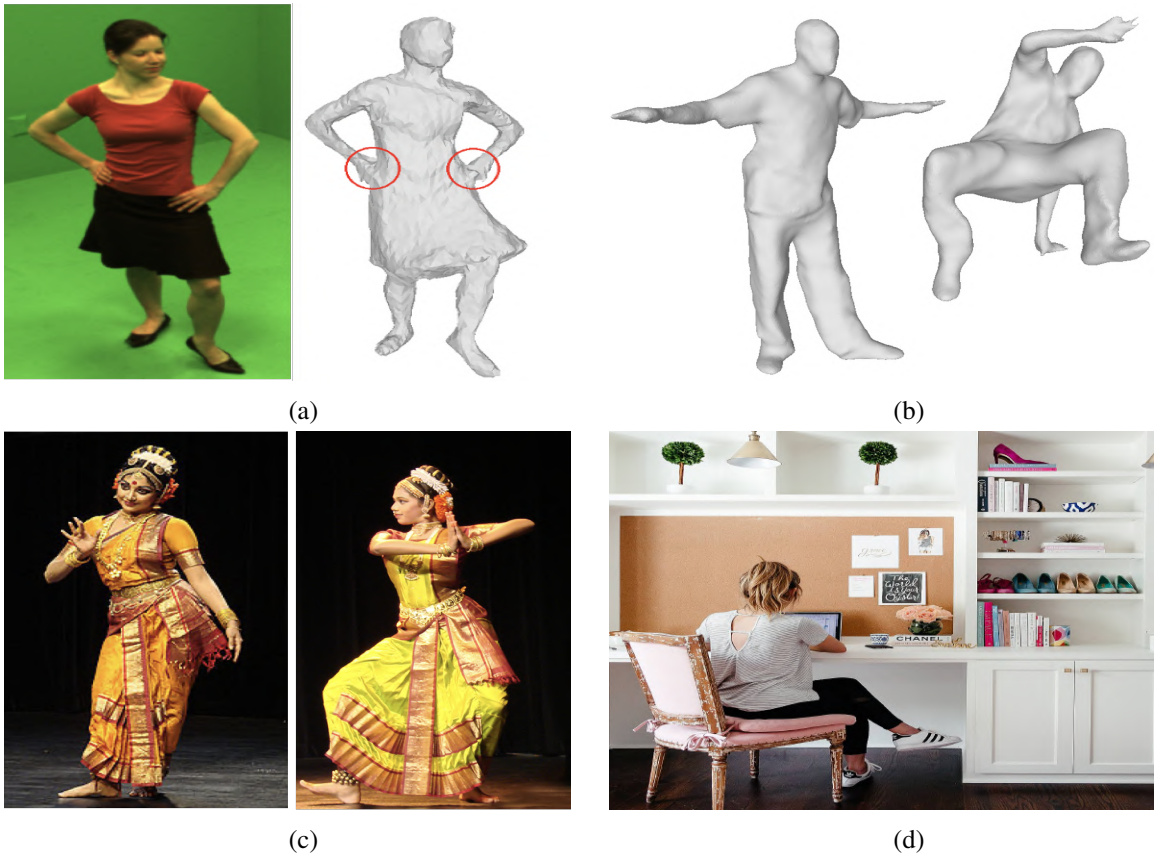


Figure 1.3: **An illustration of the various challenges in Monocular 3D Human Reconstruction** - (a) Topological Noise due to limited views and self-occlusion, (b) Non-rigid deformations, resulting in a large space of pose-shape space to learn in, (c) Deformations induced by free-form clothing and (d) Background clutter

Figure 1.3 shows the challenges in Monocular 3D Human Reconstruction. An explanation of the same is given below -

- **Ill-posed nature of the problem.** In addition to performing depth estimation for every visible point in the image, any proposed model should be capable of estimating the depth of unseen parts of the human as well. This requires the proposed model to implicitly or explicitly understand what the shape and appearance of a typical human is. If not, it will result in noisy reconstructions such as Figure 1.3a, where the hands are merged with the body of the person.

- **Topological Noise.** The absence of a sufficient number of overlapping views as well as self-occlusions results in inaccurate surface estimation, thereby causing topological noise, as seen in Figure 1.3a.

- **Non-rigid Deformations.** "Humans" exhibit non-rigid deformations which means that the position and orientation of points/triangles on the surface can be changed relative to both an internal and external frame of reference. Further, unlike rigid objects such as chairs, tables, etc., the object geometry of non-rigid human shapes evolve over time, yielding a large space of complex body poses as well as shape variations. An illustration of the same is seen in Figure 1.3b.

- **Clothing.** Deformations induced by free-form clothing make it particularly challenging to estimate the true surface of the underlying naked-body. Further, these elastic deformations can cause self-occlusions, resulting in errors in surface estimation. A gold-standard of complex clothing is illustrated by the frills in the saris of the dancers in Figure 1.3c.

- **Background Clutter.** Learning from complex environments such as Figure 1.3d with abundant background clutter is a critical challenge, further made difficult by the lack of sufficient data "in-the-wild".

- **Lack of Large-Scale Real-World Datasets.** With a paradigm shift to data-driven algorithms, data has become the new currency. Current datasets with image-3D pairs are either synthetic or are captured in studio environments. Therefore, recovering 3D human shapes "in-the-wild" without large-scale real-world datasets proves to be a challenge.

## 1.3 The Scope and Contributions

The central problem of this thesis is "Monocular 3D Human Body Reconstruction", broadly defined as the process of recovering the 3D "structure" of a human (Figure 1.1b) from the corresponding input RGB image, captured from any random location (Figure 1.1a). Here, we limit the "structure" to encompass only the pose and shape of the human body while maintaining accurate surface geometry. Recovering facial expressions, hand articulations, hair features, clothing, etc. are separate challenging problems themselves and are beyond the scope of this thesis.

In this thesis, we explore and analyze different 3D representations (volumetric, a new implicit point-cloud, template-based) and propose transformation models (VolumeNet, HumanMeshNet, and CR Framework) to exploit them. The 3D representation used is of primal importance due to the complexity of learning and representation's capability to capture an accurate surface. Therefore, through this thesis, we progressively move to more accurate, efficient, easy-to-learn, and scalable models. To that end, the following are the major contributions of this thesis -

- First, we exploit the volumetric grid representation for 3D human reconstruction using a novel deep learning model - "VolumeNet". Our model can handle non-rigid deformations induced by free-form clothing, as a result of not imposing any body-model based constraint. At the time of publication, this was the first work to learn an occupancy grid for non-rigid shapes. We build

upon VolumeNet to design PoShNet, which simplifies the learning process by decoupling pose and shape estimation and providing volumetric pose as a prior for better surface prediction.

- Second, we facilitate learning in VolumeNet with a training ideology of showing maximum information at training time to enable prediction from minimal information. Specifically, we propose co-training of RGB and depth, along with multi-view information while training to facilitate reconstruction from a single RGB image.

- Third, we propose the CR framework to explore and simplify the SMPL template-based 3D reconstruction by using classification as a prior for better parametric regression.

- Fourth, we propose a simple end-to-end multi-branch, multi-task deep network - "HumanMeshNet" that exploits a "structured point cloud" to recover a smooth and fixed topology mesh model from a monocular image. The proposed paradigm can theoretically learn local surface deformations induced by body shape variations which the PCA space of parametric body models can't capture. The simplicity of the model makes it efficient in terms of network size as well as feed-forward time yielding significantly high frame-rate reconstructions, while simultaneously achieving comparable accuracy in terms of surface and joint error, as shown on three publicly available datasets.

- Fifth, we also show the generalizability of our proposed paradigm, HumanMeshNet, for a similar task of reconstructing the hand mesh models from a monocular image.

- Finally, we collected a real dataset of textured 3D human body models and their corresponding multi-view RGBD, that can be used in solving a variety of other problems such as human tracking, segmentation, etc.

## 1.4 Thesis Roadmap

The remainder of this thesis is organized as follows -

- Chapter 2 discusses the necessary background required for understanding the context of our solutions. While the first half presents traditional approaches, their shortcomings, and a paradigm shift in algorithmic solutions, the second half talks about relevant 3D representations, their suitability, and adoption in the context of this paradigm shift.

- Chapter 3 explores the volumetric representation for extracting better surface geometry and proposes a method, VolumeNet to exploit it. The importance of providing maximal training priors with unique training paradigms to enable testing with minimal information is explored.

- Chapter 4 presents the CR framework to provide a "lighter model" for exploiting template body models. It explores the idea of using classification as a prior for easing the highly non-linear parametric learning process.

- Chapter 5 proposes an "implicitly-structured" point-cloud representation along with a multi-task network, HumanMeshNet to learn it. It explores methods to use the mesh topology as a prior, to effectively address the monocular point-cloud regression problem.
- Chapter 6 presents the concluding thoughts and future directions to look out for.

# Chapter *2*

# Background

Broadly speaking, the scope of any problem in Computer Vision is determined by the sensing modality, class of algorithms used, and intended application. Traditionally, 3D Human Reconstruction has been studied in controlled studio environments with different sensory inputs. However, in addition to being tremendously expensive, these approaches tend to be restrictive as they do not capture the 3D pose and shape of the human in real-world scenarios.

Recently, to cater to a larger range of applications, the field has begun focusing on non-intrusive and scalable sensing by estimating the 3D model from just a monocular RGB image. This shift has been supported by the resurgence of data-driven algorithms (such as deep learning) in which task-specific features are directly learned from the data, rather than designed by hand. Although there is only limited availability of 3D Human Model data outside of lab environments, the creation of large scale synthetic datasets such as SURREAL [85] has made this transition possible. Further, a key component of this problem is the 3D representation used because of the high time and space complexity required to learn it. Therefore, significant efforts have been put in this direction as well.

In this section, we shall review how the field of 3D Human Reconstruction has evolved over the years - moving from restricted lab environments to monocular reconstruction.

## 2.1 Traditional Methods

Conventional approaches can be characterized based on their sensing modalities into systems that use markers [50, 61], marker-less multi-view cameras [22, 90], inertial sensors [8, 65, 69, 91] and 3D scanners [48, 55, 78]. Below, we shall review a few path breaking seminal work in detail.

**Marker based Capture.** Placing markers or sensors on the human body has been one of the most common ways to estimate 3D skeletal motion [18, 37]. These systems are most often used for estimating skeleton proxies from motion capture (mocap) marker sets. Although they tend to be accurate, they have typically been critiqued for producing lifeless animations. To address this, some rely on dense marker

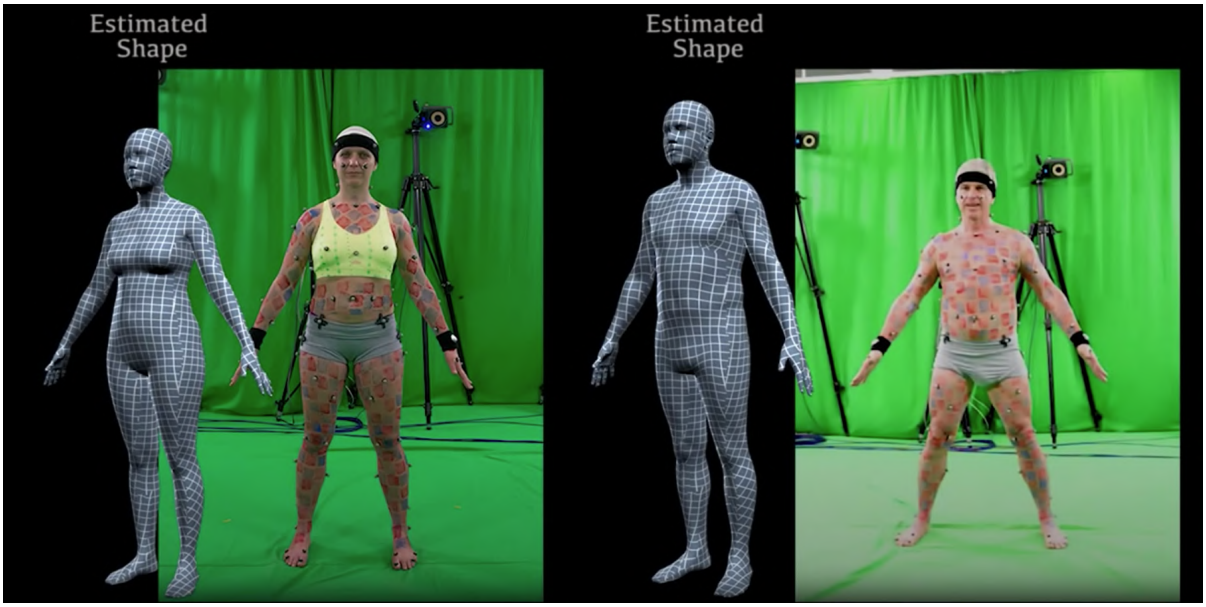



Figure 2.1: **Results of a Marker based Motion Capture setup - MoSh [50]**. Given a marker set of 47 placed on the performer, MoSh estimates the pose and shape in addition to soft-tissue deformations.

sets [60, 61]. However, they tend to be time-consuming and difficult to label. Moreover, they don't apply to archival mocap data.

MoSh [50] was a seminal method to estimate both the pose and shape directly from sparse mocap data, thereby converting a mocap system into an approximate body scanner. They argue that the relevant surface information is already present in standard marker sets but is lost in going from a non-rigid surface to a rigid skeleton. Even though these markers are placed at locations that move as rigidly as possible, the surrounding soft-tissue motion often affects the surface marker motion. They show that these perturbations, typically considered as noise, is important for realistic animation. MoSh uses a learned statistical body model, SCAPE [9], and optimizes a cost function to infer the 3D body pose (in terms of the observed marker locations relative to the body model) and the 3D body shape that best explains the marker data. Further, they were also able to capture soft tissue motion by allowing the body shape to vary over time. An overview of the results can be seen in Figure 2.1.

Although effective, several times, such sensors restrict motion and are intrusive for the performer. Further, to ensure that the markers move rigidly with the associated limb, they require skin-tight clothing.

**Markless Multi-view Capture.** Markerless methods address several of the aforementioned limitations of marker-based capture by estimating the 3D directly from multi-view cameras via triangulation or voxel carving [22, 24, 90]. Typically, as shown in Figure 2.2, they make use of expensive high-end equipment in large studio environments - calibrated cameras, controlled lighting, green screens, and much more to produce high-quality results.

Figure 2.2: **An image of a typical multi-view capture setup [22]** in a studio with a central stage surrounded by RGB and IR cameras, green screen and static IR laser light sources.

Vlasic et. al. [90] presents one such system to record complex human performances such as dancing and acrobatics. They reconstruct meshes by capturing the motion of the skeleton and the shape from synchronized multi-view video and an articulated template of the performer. As indicated in Figure 2.3, they follow a two-stage process - (a) geometric skeleton tracking and optimizing the fit of the skeleton (of the template) to the visual hull, and (b) surface refinement by deforming the template to match the observed silhouettes. The resultant output mesh had correspondences and captured high-frequency details such as clothing information from garments. Further, to enable temporal consistency, they make use of a bilateral filter. Liu [49] extends this concept to recover the 3D of multiple closely interacting people with the help of a body template and multi-view segmentation.

Although impressive, these multi-camera approaches are seldom real-time. Even though real-time 3D reconstruction has improved due to the ubiquity of RGBD cameras [56, 104], yet, the majority of systems focus on static scenes due to algorithmic challenges in reconstructing non-rigid motion. In addition to the challenges faced by rigid body reconstruction, the changing scene topology in the non-rigid case makes it complicated. As shown in Figure 2.4, Fusion4D [24] is a seminal work in this direction that produces consistent temporal reconstruction in real-time. Borrowing from the idea of anchor frames in non-rigid tracking [22, 35], they use a voxel grid as a reference model (called as “key-volumes”) to deal with radically different surface topologies over time.

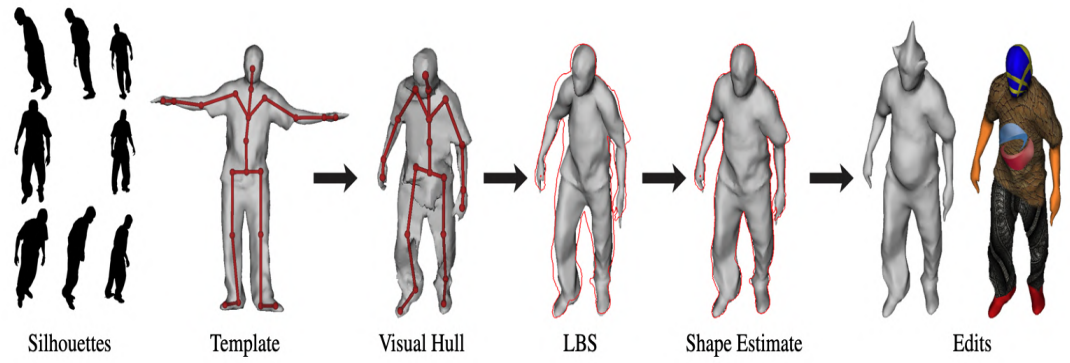



Figure 2.3: **Gives the methodology proposed by Vlasic et. al [90] for markerless multi-view 3D capture**. Given a stream of silhouette videos and an articulated template mesh, at every frame, they fit the skeleton to the visual hull, followed by deforming the template with linear blend skinning (LBS) and refining the surface to fit the silhouettes. Finally, the user can edit the geometry or texture of the sequence.

Despite the advancements in multi-view systems, they yield reconstructions with severe topological noise [72] and often require manual clean-up. Although recent attempts replace/augment the capture setup with high-resolution depth sensors [24, 105], making it more accurate, the fundamental limitation of these techniques is the requirement of a calibrated multi-sensor setup that restricts their applicability to studio environments.

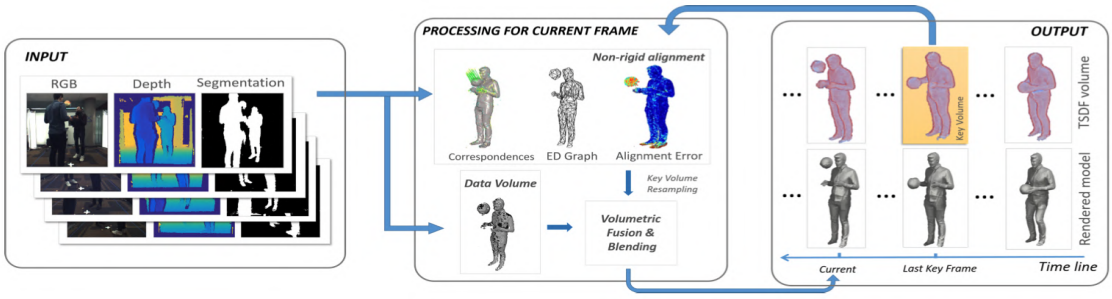



Figure 2.4: **An Illustration of Fusion4D - a real-time 4D reconstruction method [24]**. Given multiple RGBD frames, a per-camera segmentation mask is first calculated. This is followed by estimating a per-frame correspondence field that is used to initialize the non-rigid alignment. The final step involves performing the non-rigid alignment by warping the current key volume to the data volume. The system is made more responsive to new data by additionally fusing the currently accumulated model into the data volume. Non-rigid alignment error and the estimated correspondence fields are used to guide the fusion process.

In the absence of a calibrated setup, the fusion between the view specific point clouds (2.5D) is done by the registration of non-rigid point sets. We refer to [30] for an extensive literature survey. Weber [93]

doesn't assume any order for the input 2.5D and uses an overlap heuristic to perform the registration. Golyanik [31] provides the Extended Coherent Point Drift (ECPD) algorithm which is a probabilistic approach for point set registration that allows embedding prior matches. Some methods [28, 90] make use of a template scan of the performer and use it to guide the registration procedure. Further, extending mesh registration to 4D is important for temporal consistency in performance capture. To that end, DFAUST [16] proposes a new 4D registration algorithm that uses the 3D geometry and texture information to register all scans of the sequence to a template topology. It is to be noted that most registration algorithms require sufficient overlap between the point clouds and are susceptible to sensor noise, therefore, once again, limiting their applicability.

**3D Scanners.** There exist several different types of scanners, each with their scanning methodology and associated use cases. To do scene/environment capture, a time of flight [20, 57] scanner such as Lidar is commonly used. It uses a laser to probe and a range finder to estimate the depth of each point. Triangulation based 3D laser scanners [25] are similar, except that they use a camera instead of a range finder to detect points where the laser interacts with the subject. While time-of-flight scanners have very high operating distances and are less accurate, triangulation based scanners have low operating ranges and are more accurate.

Structured light scanners [48, 55] scan multiple points or their entire field of view simultaneously by projecting a pattern of light on the subject and use a camera to interpret the deformation of the pattern on the subject. Therefore, they're capable of capturing real-time non-static scenes as well, unlike time-of-flight scanners. It is common to find several hand-held scanners [78] to use triangulation or structured light setups. Although 3D scanners are extremely accurate (see Figure 2.5), they are expensive and tedious to use. Therefore, they have limited applicability to commercial, real-world situations.

(a) (b)

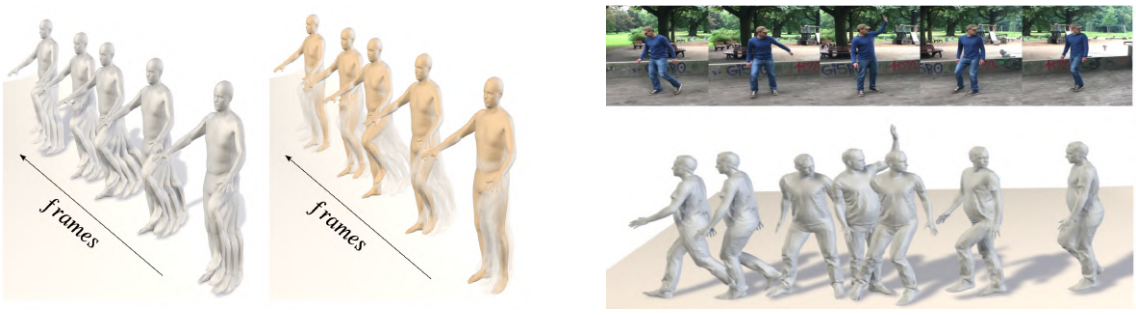
Figure 2.5: **A depiction of 3D Scanning Technology.** (a) Structured Lighting Hand held scanner, Artec 3D Evo [1] and (b) a sample output human mesh captured by it.

Most of the methods discussed thus far are optical vision-based systems i.e., they require dedicated cameras to track and capture the subject. Given this requirement, they may not be able to capture human-motion in dynamic settings (such as skiing, cycling, etc) or everyday activities (such as cooking, praying, etc).

**Inertial Sensors.** Systems built with Inertial Measurement Units (IMUs) address this fundamental limitation of optical systems. IMUs consist of built-in sensors such as gyroscopes, accelerometers, and magnetometers to detect position and movement, thereby making them capable of tracking human pose without cameras. Therefore, they are more suitable for scenarios with occlusions, baggy clothing, or outdoor scenes where tracking with a camera isn't possible.

A good part of the early literature that uses IMUs is focused on database retrieval. While Liu [47] uses online local models to regress the pose by querying the database with 6 IMUs, Schwarz [71] uses the Gaussian process regression to directly regress the full pose and uses 4 IMUs for the query. Such database based approaches are limited as they cannot capture activities not present in the dataset.

To recover arbitrary poses directly from IMUs, a sparse IMU setup wouldn't suffice as it will result in weak constraints on human motion and noise due to acceleration data. Therefore, the majority of the early setups use a large number of sensors for stabilization and accuracy. For example, Xsens' Moven [69] consists of 17 IMUs and Vlasic's [89] system consists of 18 sensors. However, such systems are extremely intrusive to the performer and time-consuming to setup/reproduce. Therefore, to make it less intrusive, a few have constrained the sparse IMUs with optical inputs [8, 65]. Although interesting, optical data introduces challenges such as occlusions or limitations in outdoor tracking.




(a) (b)

Figure 2.6: **An Illustration of Sparse Inertial Pose's [91] (SIP) optimization and results.** (a) A depiction of the joint optimization of SIP. Multiple poses fit the IMU readings in each frame (shown by models in grey). By optimizing over all poses in the sequence, we can see a pose trajectory over frames (in orange). (b) A depiction of the recovered SMPL models (in grey) by using 6 IMUs and the joint optimization procedure. Note that the RGB images are shown for reference and not used in the optimization.

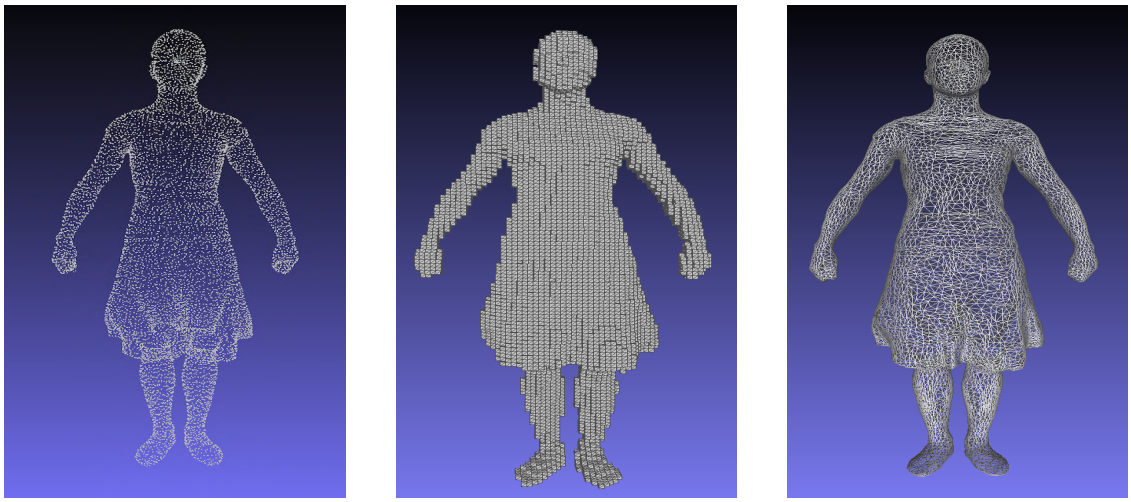
Sparse Inertial Poser [91] is a seminal work that uses only 6 IMUs and constrains the problem by using the SMPL statistical body model [12], therefore resulting in a minimally intrusive yet accurate

system. As shown in Figure 2.6a, they optimize all the poses of a sequence simultaneously in a single objective function and enforce coherency between the IMU readings and body model's orientation and acceleration.

## 2.2 Progression towards learn-able representations

With the resurgence of deep learning, there has been an emphasis on moving towards learn-able representations. Below, we shall first review various 3D representations and assess their capability to be integrated with these learn-able models, followed by the state-of-the-art methods that have adopted this paradigm.

### 2.2.1 3D Representations



(a) (b) (c)

Figure 2.7: **A visual depiction of various 3D Representations** with a human in a canonical pose - (a) Point Cloud, (b) Voxel-grid and (c) Mesh.

**Point Cloud.** A point cloud is a set of data points in space (3D). They're are capable of capturing objects/scenes whose structure is not known. Although extremely powerful, this unstructured nature of theirs makes them computationally expensive to operate on. Further, from a deep learning point of view, it was not until recently [66] that the community was able to operate on them in a fashion that captures the global and local features of 3D objects/scenes.

**Voxel-grid.** A voxel is the smallest, most basic unit of 3D space. A voxel-grid is a 3D occupancy grid made up of such voxels. Therefore, each cube in the grid represents a binary number - 0 for free and 1 for occupied. This representation can be interpreted as a discretized version of a point cloud; therefore,

still capable of learning objects with unknown structures, with an added advantage of being easier to operate on. However, for those applications interested in capturing only the surface information [39], they can be inefficient as they typically fill the cubes within the surface boundary as well. Octree based methods [92] improve upon traditional voxel-grids by providing higher granularity in regions containing high-resolution details and lower granularity in low-resolution regions.

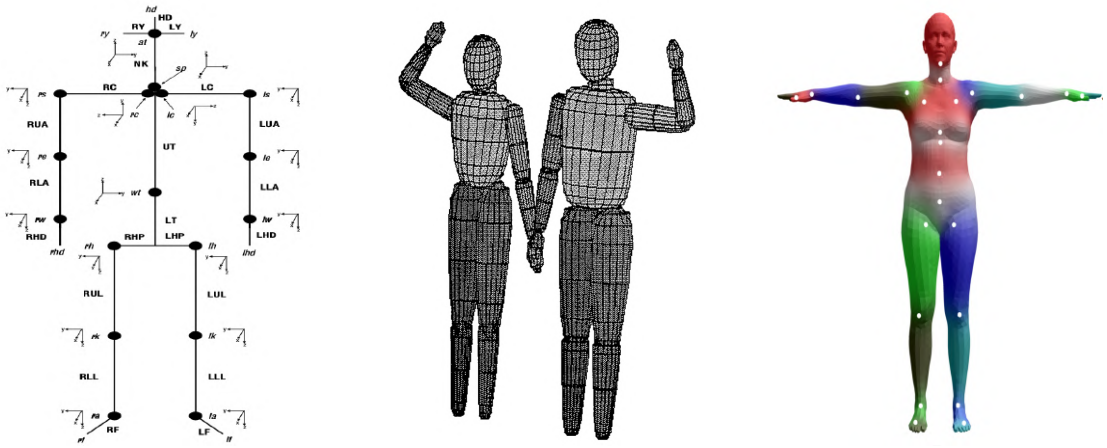
**Mesh.** A mesh consists of a point cloud with a well-defined topology imposed on it, therefore, giving each 3D point a neighborhood. It is a collection of vertices, edges, and faces that characterize the shape of the object in 3D. Very recently, graph-based CNNs [44] and MeshCNNs [36] are being explored to make use of this neighborhood constraint by defining network operations such as convolutions, upsampling, and downsampling on them.

The aforementioned representations (visually depicted in Figure 2.7) can be considered as a sampling of the underlying manifold. A manifold is a parametric surface and learning such functions have recently come into focus with models such as AtlasNet [33]. Recovering the parametric surface is a more challenging task and a separate research area in itself, hence, out of the scope of this thesis.

**Statistical Body Models.** When the structure of the object of interest is known, it makes sense of utilizing this information in creating a 3D representation for it. This is the idea behind parametric human-body models. As shown in Figure 2.8, some of the earliest parameterizations were based on stick models [10], which consisted of labeled landmarks (to represent joints) and edges between them (to indicate connectivity). These evolved into models based on geometric primitives [26] (such as cubes, spheres, etc). Modern-day state-of-the-art statistical models are learned from thousands of 3D scans of people. Therefore, they exhibit realism due to inherently modeling anthropomorphic constraints such as limb/bone proportions and their associated symmetry. Although such parameterizations are typically low dimensional (in comparison with voxel-grids or point-clouds), they do not sacrifice on the quality of reconstruction. An added advantage is that they easily integrate with existing graphics pipelines, thereby catering to a wide range of applications.

SCAPE [9] was one of the first statistical models built from real scans and it decomposed the human into pose and shape. However, due to being a model built on triangle deformations, it presented challenges in optimization and fitting. SMPL [12], on the other hand, is a vertex-based model and is, therefore, easier to use in optimization. Some other models such as [61] model even low-level details such as skin deformations as a function of the skeletal motion.

Since SMPL is used in our work, we shall describe it in further detail. The SMPL model decomposes the human body shape into identity-dependent shape and non-rigid pose-dependent shape. It uses a vertex-based skinning method that incorporates corrective blend shapes, in which each blend shape is represented as a vector of concatenated vertex offsets. The skeletal structure of the human body is modeled as a kinematic tree. The model consists of naked meshes of $N = 6890$ vertices, with pose




(a) (b) (c)

Figure 2.8: **Evolution of Body Models** - (a) Skeleton Based [10], (b) Geometric Primitive Based [26] and (c) Statistical Body Model, SMPL [12].

parameters $\theta \in \mathbb{R}^{72}$ (of which the first three are for the global rotation) and shape parameters $\beta \in \mathbb{R}^{10}$. The pose parameters are represented in terms of the relative angles

As given in Equation 2.2, the shape offset ($B_s(\beta)$) and pose offset ($B_p(\theta)$) are applied to the base template $T$, which is the statistical mean shape in the training scans $T_\mu$. Then, as per Equation 2.1, each body part is rotated around skeleton joints J($\beta$) using a skinning function $W$. Therefore, the SMPL model $M(\beta, \theta)$ is given by -

$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \theta, W) \tag{2.1}$$

$$T(\beta, \theta) = T_\mu + B_s(\beta) + B_p(\theta) \tag{2.2}$$

where $T(\beta, \theta)$ outputs an intermediate mesh in a T-pose after the pose deformations are applied.

### 2.2.2 Monocular Reconstruction

The capability of these over-fitting machines (deep learning), availability of large scale datasets with image-3D pairs as well as a requirement for scalable and non-intrusive methods has catapulted the field to focus on monocular 3D reconstruction.

**Iterative Optimizations.** Some of the early work in literature focuses on optimization-based techniques, beginning with the SCAPE body model [9]. Guan [34] uses silhouettes, edges, and shading as cues during the fitting process and required an initialization through a user-specified 2D skeleton. Sigal [74] was among the first to automatically fit the parametric SCAPE to ground truth silhouettes without user intervention. They use a discriminative method to initialize a generative model for more fine detail

recovery. However, due to difficulties in optimization with triangle-based models such as SCAPE, the focus shifted to vertex-based models such as SMPL [51].

Figure 2.9: **Illustration of SMPLify [13]**, an optimization based method. Given a monocular RGB, they use a CNN to predict 2D joint locations (hot colours denote high confidence). This is followed by fitting a 3D model using an iterative optimization (overlayed over the input image in orange)

Further, SMPL provided the right balance between high anatomical flexibility and realism, therefore became a popular model choice for 3D human recovery. As shown in Figure 2.9, SMPLify [13] was amongst the first and they estimate the 2D joint locations followed by an optimization framework to fit it to the SMPL model. Similarly, [46] predicts 91 landmarks on the human body in the image and uses an extended version of SMPLify to fit it to the parametric model. While [6] recovers the textured SMPL model from a monocular RGB video, [102] recovers the naked shape under clothing given static 3D scans or 3D scan sequences.

Despite showing compelling results, the aforementioned methods are over-reliant on error-prone 2D observations. Further, solving an iterative optimization problem is susceptible to highly complex and costly; not enabling real to semi-real time performance.

**Learning Body Models.** Deep learning provided an alternate solution that learned priors automatically from the data and addressed several of the issues with iterative optimizations. Dibra [23] presents one of the first approaches using this paradigm by learning the shape parameters from silhouettes using a CNN, but assume a frontal view. Subsequent methods focused on learning the parameters from an image [41, 59, 63, 80, 83].

Tan [80] indirectly predicts the model parameters from the bottleneck layer of an encoder-decoder architecture that is trained on silhouette prediction. As shown in Figure 2.10, HMR [41] proposes a seminal work with an iterative regression using 3D and 2D joint loss as feedback and adversarial supervision for each joint. However, this architecture has a large number of networks and takes 5 days to train. [59] predicts a color-coded body segmentation that is used as a prior for predicting the parameters. Similarly, in [63], 2D heatmaps, and silhouettes are predicted first, which are then used to predict the

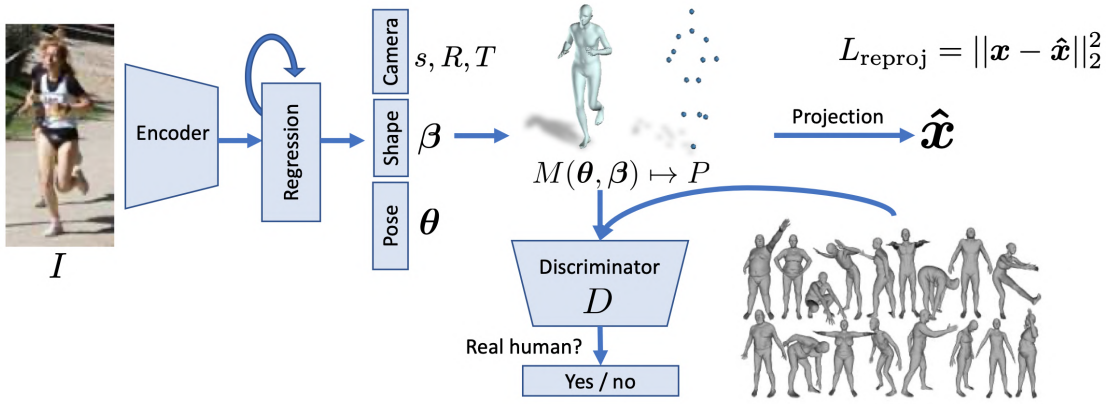



Figure 2.10: **Overview of Human Mesh Recovery [41], a parametric human model recovery method**. An image 'I' is sent to a CNN encoder, followed by. an iterative 3D regression module that infers the SMPL parameters and minimizes the joint re-projection error. The 3D parameters are then sent to a discriminator D to limit the parametric space to produce only valid poses and shapes.

pose and shape parameters. Although showcasing impressive results, all of the above methodologies calculate the loss on 2D keypoints or silhouette projections of the rendered mesh, which significantly slows down training time (due to model complexity), in addition to requiring additional supervision.

**Learning Surface Representations.** Although parametric body models are effective, they lack high-resolution details (accurate geometrical information), and in an attempt to recover personalized models, learning the surface representation for humans shapes became an area of interest.

As discussed in Section 2.2.1, the volumetric representation is a discretized form of a point cloud. Therefore, it became a popular representation to capture the surface of an object. This was first adopted for learning the 3D of rigid objects (e.g. cars, chairs, rooms), in which they learned a class-specific 3D structure of objects using large scale datasets of synthetic 3D models [21, 82, 95, 96, 100]. ShapeNet[96] proposed a deep network representation with a convolutional deep belief network to give a probabilistic representation of the voxel-grid. Along similar lines, 3D Generative Adversarial Networks (GAN's) were proposed to learn a probabilistic latent space of rigid objects (such as chairs, tables) in [95]. [100] proposed an encoder-decoder network that utilizes observations from the 2D space, and without supervision from 3D, performs reconstruction for a few classes of objects. This relationship between 2D and 3D is further exploited in [82] where they define a loss function in terms of ray consistency and train for single-view reconstruction.

Regarding non-rigid reconstruction, SurfNet [76] proposed to directly achieve the surface reconstruction by learning a mapping between 3D shapes and the geometry image representation (introduced in [75]). One of the key limitations of their method is that it is only suitable for genus-0 meshes. This constraint is frequently violated in real-world human body shapes due to topological noise [72] induced

by complex poses, clothing, etc. Another very recent work proposed in SilNet [94] uses multi-view silhouette images to obtain reconstructions of free form blob-like objects/sculptures. However, the use of silhouettes limits the application to scenarios where background subtraction is assumed to be trivial. All these initial efforts are focused on textureless 3D reconstruction and do not seem directly extend-able to non-rigid human body shapes.

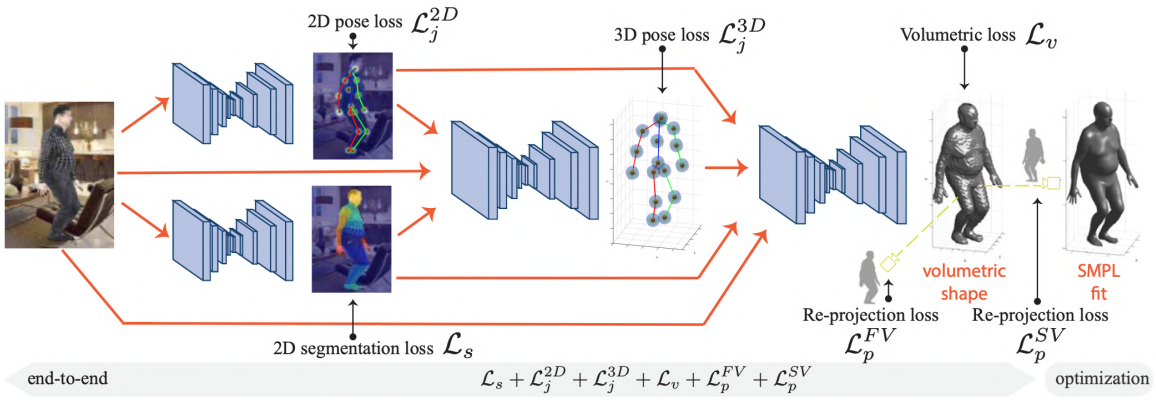



Figure 2.11: **Overview of Bodynet [84], a volumetric human recovery method**. The input RGB image is given to two networks that predict the 2D pose and 2D body part segmentation. Both of these are combined with the input RGB and fed to another network that regresses the 3D pose. All the sub-networks' outputs are fed to an encoder-decoder model to predict the volumetric shape, trained using reprojection losses. The final model is fit to predict the SMPL parameters for evaluation.

Non-rigid shapes pose an additional challenge of an ever-changing topology over time in comparison with rigid shapes. As a parallel work to ours (described in Chapter 3) for the volumetric reconstruction of humans, Bodynet [84] proposes a complex multi-task to do the same. As shown in Figure 2.11, they had a total of 4 networks (having respective losses computed on 2D and 3D joint locations, 2D segmentation mask, volumetric grid, and silhouette reprojection of volumetric and SMPL model). In addition to recovering the voxel-grid, they fit the recovered volume to the closest SMPL parameters, thereby recovering an animatable model. Further, similar to our early attempts to exploit the mesh topology (described in Chapter 3), another concurrent method [45] proposes a Graph Neural Network(GCN) to recover the mesh corresponding to the SMPL body model.

Recently, early attempts have been made to learn full-body models [62] from monocular images and enforce temporal consistency [42] as well. These are interesting fields of study which might have more emphasis moving ahead.

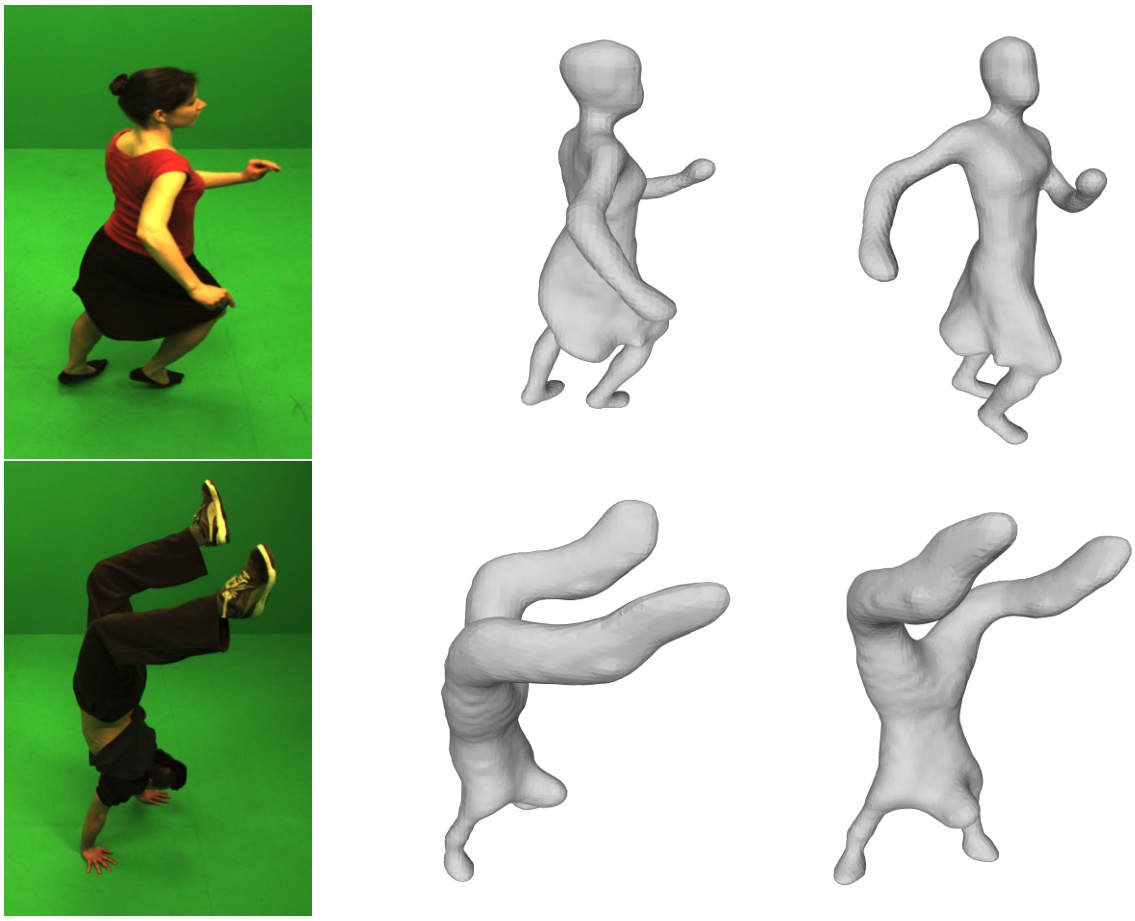
**Extension to Other Non-rigid Shapes.** Learning of other non-rigid shapes such as faces [29, 64], hands [17, 27, 53], clothing [5, 11], animals [106, 107] etc. have also been active fields of study.

We shall discuss hand model recovery in a bit more detail because of our initial efforts in that direction (see Section 5.3.4). While most of the hand recovery methods typically estimate the 3D pose from

one or multiple RGB/Depth images, hand shape estimation hasn't been extensively explored. For a detailed survey of the field, we refer to [79, 101]. Most of the early models only approximate the shape of the hand [58, 81]. While personalized model such as [40] works better at tracking, they do not scale well to a large number of users. Recent effort in [53] was the first attempt to predict both the pose and the vertex-based full 3D mesh representation (surface shape) from a single depth image. The recently proposed MANO [70] model is an SMPL-like model that describes both the shape and pose, and is learned from thousands of high-resolution scans. [17] predicted the MANO parameters from a monocular RGB image, but, they don't show much shape variations. Similar to our attempts to exploit the mesh topology (see Section 5.3.4), [27] uses a graph CNN to recover the hand surface from monocular RGB image of the hand.

# Chapter *3*

# Volumetric Reconstruction



(a) (b) (c)

Figure 3.1: **An Illustration of Volumetric Reconstruction of a Human from a Monocular image**. Given a monocular input image (a), we recover the corresponding volumetric 3D of the human (b,c). Here, (b) shows the recovered 3D aligned to the input and (c) shows it aligned to another view, for aethetic reasons. Note that the results shown here are after surface smoothing [43].

Capturing dynamic scenes in a calibration-free environment via a non-intrusive method requires innovation along both lines - sensing and interpretation. Therefore, following a ubiquitous approach, we focus on developing efficient learning algorithms from monocular cameras. However, learning non-rigid shapes from a monocular image is a severely ill-posed problem. Unlike rigid shapes, this non-rigid nature of human body shapes results in additional challenges due to an object geometry that evolves over time, resulting in a large space of pose and shape deformations.

We overcome these challenges with algorithm level novelty. A key part of the algorithm is the underlying 3D representation used. While parametric model-based techniques [9, 12, 16] are effective, they are non-ideal for developing personalized models with clothing details, facial features, etc. due to following a template geometry. In this chapter, we explore the volumetric representation, in an attempt to learn more accurate surface information via a novel deep learning model.

Further, since this learning process isn't straightforward, a key algorithmic trick we employ is to use additional priors that provide cues about the relative depth and symmetry of human bodies and indirectly guide the learning process. We do so in a unique fashion by showing maximum information at training time, to successfully predict from minimum information at test time. Specifically, we co-learn RGB with Depth cues as well as provide multi-view information at training time, to enable reconstruction from just single-view RGB at test time. The shared filter weights in this co-learning process enable the model to capture the wide range of poses and shapes, in addition, to help address the challenges caused by cluttered background, shape variations, and free form clothing. Further, we extend our model by decoupling pose and shape to ease the learning process.

*Note that the work showcased in this chapter has been published in BMVC, 2018 [87].* The full pipeline with texture recovery isn't showcased here as it is out of the scope of this thesis. Also, note that qualitative results are given at the end of the chapter.

## 3.1 Contributions

The key contributions of this work are:

- First, we introduce a novel deep learning model, VolumeNet, to obtain 3D models of non-rigid human body shapes from a single image. Obtaining the reconstruction of non-rigid shapes in a volumetric form had not been attempted in the literature (at the time of publication).

- Second, we introduce the ideology of showing maximum information while training, to enable prediction with minimum information. Specifically, we demonstrate the importance of depth cues for this task by co-learning RGB and Depth. This, coupled with showing multi-view information while training enables prediction from monocular RGB.

- Third, we show that our model can partially handle non-rigid deformations induced by free form clothing (as a result of not imposing any model-based constraint while training), thus moving one step closer towards recovering personalized models.

- Fourth, we collected a real dataset (that shall be publicly released) of textured 3D human body models and their corresponding multi-view RGBD, that can be used in solving a variety of other problems such as human tracking, cloth modeling, etc. This is the first of its kind dataset that captures the Indian demographic. We show impressive quantitative and qualitative results on four publicly available datasets as well as our proposed dataset.

- Finally, we use VolumeNet to build PoShNet, as an attempt to simplify the learning procedure. Here, we decouple pose and shape, and use the volumetric pose as a prior for better shape estimation.

## 3.2 Proposed Method: VolumeNet

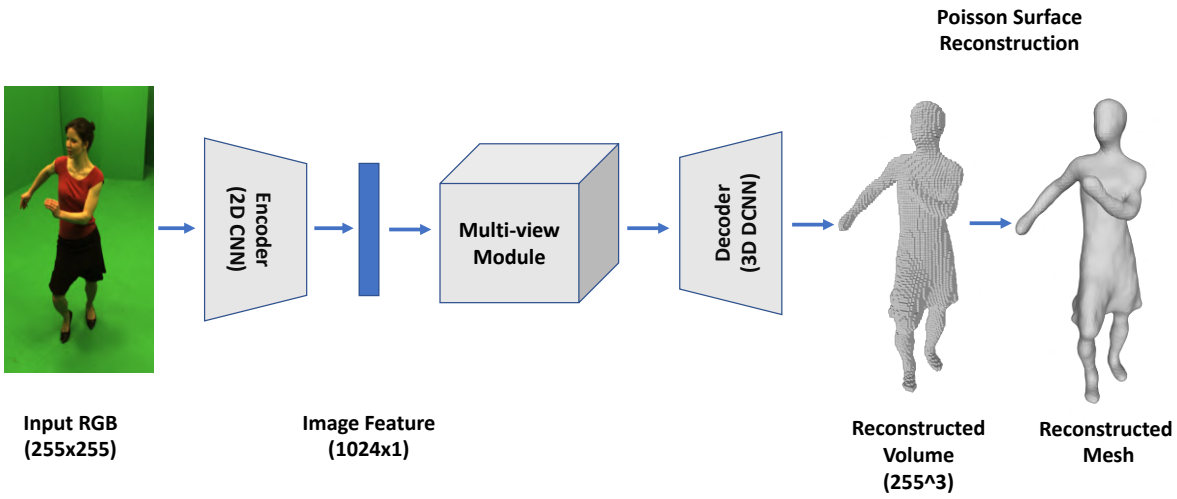



Figure 3.2: **Test time flow of VolumeNet, our proposed model for Volumetric 3D Human Reconstruction from a single image**. Given a monocular input RGB image, we first obtain the image features via a 2D CNN Encoder. This is forwarded to a multi-view module, that was taught to map the input view to a certain part of the output space via our unique training methodology. The Decoder, a 3D DCNN then upsamples the output of the multi-view module to produce a reconstructed volume of size $128 \times 128 \times 128$. Using Poisson's surface reconstruction algorithm [43], we then smooth the voxel-grid to obtain the reconstructed mesh.

As shown in Figure 3.2, we propose an encoder-decoder network with a multi-view module sandwiched between the two. The multi-view module is used to propagate the ideology of providing as

much information as possible while training (for the network to automatically learn priors) to enable reconstruction from minimum information at test time. Specifically, at train time, we provide a random number of input views (anywhere from one to eight) in each iteration, thereby allowing the network to auto-correlate and predict from just one view at test time. Following the same ideology, as explained in "Input Modes" below, we co-learn RGB and Depth information (RGB/D mode) by optimizing the network to predict equally as well from both, thereby exploiting the coherence in both modalities and enabling better prediction from only RGB. Putting both these ideas together, we thus enable reconstruction from single view RGB at test time. The details of the network architecture are given below.

### 3.2.1 Network Architecture

- **Encoder -** The encoder consists of a 2D CNN (ResNet-18) that takes an image of size 255x255 in one of the input modes (see below) and produces a 1024 dimensional feature vector. Each view produces one such feature vector, which is combined in the multi-view module. We use Leaky ReLU as the activation function in both the encoder and decoder.

- **Multi-view Module -** The CNN feature vectors produced by the encoder (for each view) are combined using either a view-pooling operation [94] or a 3D-GRU [21]. The outputs are resized to $4^3$ and fed to the decoder. While the view-pooling operation extracts the most prominent features from each view and derives a global shape, the 3D-GRU correlates the CNN image features to the right location of the 3D grid (as shown in Figure 3.4).

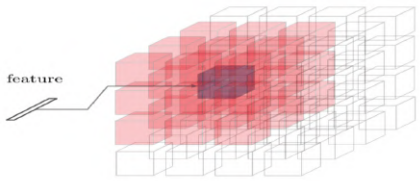



Figure 3.3: **Spatially distributed 3D-GRU Grid [21]** consisting of $4 \times 4 \times 4$ 3D-GRU units. The purple cell receives the 2D CNN feature vector along with the hidden states of its neighbours (red) via a $3 \times 3 \times 3$ convolution.

The 3D-GRU units are spatially distributed in a 3D grid structure. Inside the 3D grid, there are $N \times N \times N$ 3D-GRU units, where $N$ is the spatial resolution of the 3D-GRU grid (in our case, N=4). As shown in Figure 3.3, in each view, each unit (purple) in the 3D-GRU receives the same feature vector from the encoder in addition to the hidden states from its neighbors (red) by a $3 \times 3 \times 3$ convolution as inputs. For vanilla GRUs, all elements in the hidden layer $h_{t-1}$ affect the current hidden state $h_t$, whereas a spatially structured 3D GRU only allows its hidden states

$h_t(i, j, k)$ to be affected by its neighboring 3D-GRU units for all $i$, $j$, and $k$, where the neighbours are defined by a convolutional kernel size of $3 \times 3 \times 3$.

The multi-view module sends a 3D grid of hidden states to the decoder, that is upsampled to reach the target volumetric representation. Since at training time, each voxel-grid is reconstructed from all of 1 to V views, even though the order of the views aren't maintained, this configuration forces a 3D-GRU unit to handle the mismatches between a particular region of the predicted reconstruction and the ground truth model such that each GRU unit learns to reconstruct one part of the voxel space (see Figure 3.4).

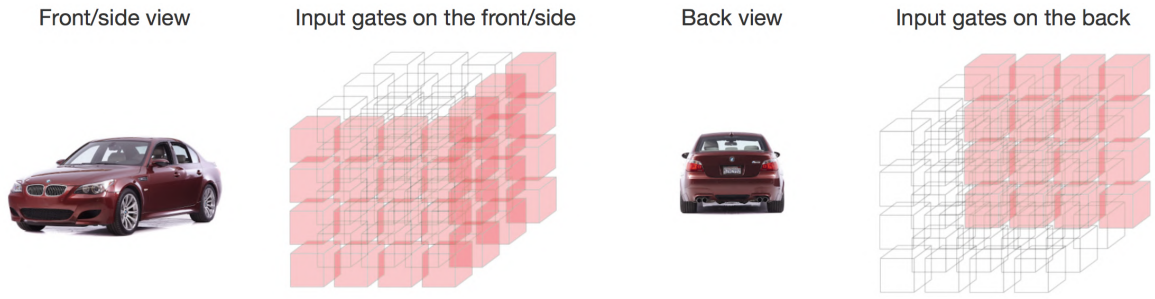



Figure 3.4: **An Illustration of correlation of image features with locations in the 3D GRU grid.**If the input image is taken from the front/side view (for e.g.), the input gates corresponding to the front and side view respectively activate (opens). If the view of an object taken from the back is fed into the network, the input gate will open up for the voxels on the back. This operation allows the network to put image features to the right position [21].

- **Decoder -** The decoder consists of 3D DCNN (Deconvolutional ResNet-18) that up-samples the output of the multi-view module to $128^3$. We stop at this resolution because of the associated computational complexity with higher resolutions. Finally, a mesh is obtained from the voxel-grid via Poisson's surface reconstruction algorithm [43].

### 3.2.2 Input Modes

To capture the large space of complex pose and shape deformations of humans, we experiment with four input modes:

- **RGB -** This setup is commonly used in rigid body reconstructions [21, 82]. However, we qualitatively and quantitatively show in Section 3.3.2 that this setup in inadequate for reconstructing non-rigid shapes.

- **D -** The premise behind this mode is that depth-maps give us information about the geometry of the object, which as seen in Section 3.3.2, help in significantly enhancing the reconstruction quality.

- **RGBD -** To exploit both the depth and color channels, we augment RGB with D in a 4 channel input setup.
- **RGB/D -** Lastly, we propose a unique training methodology that gives us superior performance in comparison with the above 3 modes. Here, at train time, for each mesh, we perform two feedforwards - one with input as RGB and another with input as Depth (replicated to all three channels). The outputs of both modalities are given equal weightage in the joint optimization as shown below (Equation 3.1), in which $L_{RGB}$ and $L_D$ are both given by Equation 3.2. Thus, this allows good reconstruction from just RGB at test time by leveraging upon Depth information at train time.

$$L_{final} = L_{RGB} + L_D \tag{3.1}$$

Intuitively, this strategy is equivalent to sharing weights between the RGB and D spaces, to exploit the coherence between RGB and D; thus combining the advantages from both the spaces.

## 3.3 Experiments & Results

In this section, we provide necessary implementation details and corroborate our method of single view non-rigid 3D reconstruction by means of qualitative and quantitative evaluation. For visual results, refer to our supplementary video [2, 4].

### 3.3.1 Datasets

(a) (b) (c) (d) (e)

Figure 3.5: **An Illustration of the diversity in poses, shapes and clothing in our 3D dataset**. While (a), (b) are relatively tight clothing, (c) salwar kameez, (d) dhoti (e) kurta are free-form clothing.

**Our Data:** As shown in Figure 3.6, we setup a calibrated multi-camera 3D Capture System with 5 v2 Microsoft Kinect [56] cameras vertically mounted on stands, all connected to a workstation grade server

to store and process the captured data in real time, from each of the devices. Our current motion capture system works at the rate of 30 fps, but, can be extended to high speed capture scenarios with the help of additional hardware (such as point-grey cameras) and the associated processing and synchronisation. Briefly, the setup and working of our system can be encapsulated in the 3 stages -

- *Calibration* - Each pair of Kinect cameras are first calibrated using a checkerboard pattern. The calibration process entails finding the corners of the checkerboard in each frame and solving the equation which gives the transformation between each pair of cameras.
- *Capture* - Each Kinect camera captures the RGB and Depth image. The Depth image provides the view-specific colored point clouds (2.5D) using the camera's intrinsic parameters. The transformation matrices between each pair of cameras obtained in the calibration process (step 1) are used to merge the view-specific 2.5D to get a consistent colored 3D point cloud.
- *Post-processing* - The 3D point clouds obtained in the previous step are often noisy due to errors in calibration, sensor noise, etc. and are therefore cleaned by thresholding (to remove noisy surface points). This cleaned 3D point cloud is converted to a mesh using surface reconstruction algorithms such as Poisson's Algorithm to obtained a coloured 3D mesh.

Figure 3.6: **Setup of our 5-Kinect 3D Motion Capture System.**

Using this system, we've captured the first of its kind 3D Models specific to the Indian population. Our continuously expanding dataset consists of male and female models of various shapes performing actions ranging from simple marching and punching sequences to complicated dance actions (as shown in Figure 3.5). In addition to the complexity in poses and shapes, we've captured data in a wide range of clothing attires from skin-tight clothing to traditional free-form clothing such as kurtas, salwar kameez, dhoti, etc.

As shown in Figure 3.5, this diverse data is very valuable as it can be used to solve several problems such as 3D Human Surface Reconstruction, Cloth Modeling, Texture Recovery, etc. A subset of this data

(with settings similar to Figure 3.5a) consisting of 5 mesh sequences, each containing 200 to 300 frames with significant pose and shape variation was used for training and testing our pipeline.

**MPI Datasets:** First, we use the parametric SMPL model [12, 85] to generate synthetic data as follows - 10 mesh sequences, each containing 300 frames, i.e., 3000 meshes, consisting of an equal number of male and female models. We use a virtual Kinect setup to obtain RGB and Depth from 8 locations for this setup. Secondly, we use data from FAUST [15] which consists of 300 high-resolution human meshes, each having approximately 250,000 vertices. Each scan is a high-resolution, triangulated, non-watertight mesh acquired with a 3D multi-stereo system. There are a total of 10 different subjects in 30 different poses. The 300 meshes come divided into 2 sequences, one having complete meshes and the other having broken/incomplete parts - the former used for training, and the latter for testing.

**MIT's Articulated Mesh Animation [90]:** This dataset consists of 5 mesh sequences (approx. 175 to 250 frames each). It provides RGB images from 8 views and the corresponding 3D meshes for each frame. The total number of meshes used from this dataset for training are 1,525.

### 3.3.2 Implementation Details

**Network's Training.** We used Nvidia's GTX 1080Ti, with 11GB of VRAM to train our models. A batch size of 5 with the ADAM optimizer having an initial learning rate of $10^{-4}$ and a weight decay of $10^{-5}$ is used to get optimal performance. Further, a standard $80:20$ split between the training and testing datasets is adhered to. In order to ensure that reconstruction is feasible from single as well as multiple views, we choose a random number of views from available views for training a mesh in each iteration. Using this randomization in training, we are providing sufficient view information to the network so that it can learn the space of body pose and shape variations and hence able to achieve single view reconstruction at test time.

**Loss Function.** We use Voxel-wise Cross Entropy to train the reconstruction models. It is the sum of the cross-entropies for each pair of predicted and target voxel values. Let '$p$' be the predicted value at voxel $(i, j, k)$ and '$y$' the corresponding expected value. Then, the loss is defined as :

$$L(p, y) = \sum_{i,j,k} y_{(i,j,k)} \log(p_{(i,j,k)}) + (1 - y_{(i,j,k)}) \log(1 - p_{(i,j,k)}) \tag{3.2}$$

**Evaluation Metric.** The primary metric used to evaluate our performance is the Intersection over Union (IoU), which is a comparison between the area of overlap and the total area encompassing both the objects. Larger its value, the better the quality of reconstruction.

| Dataset | Multi-View | RGB (Baseline) | D | RGBD | RGB/D (Ours) |
|---|---|---|---|---|---|
| MPI-SMPL [85] | 3D-GRU | 0.6903 | 0.7709 | 0.7541 | **0.8040** |
| | Max Pool | 0.7144 | 0.7633 | 0.7550 | **0.7816** |
| MIT [90] | 3D-GRU | 0.0103 | 0.7403 | - | **0.7547** |
| | Max Pool | 0.0081 | 0.7205 | - | **0.7480** |
| MPI-FAUST [16] | 3D-GRU | 0.8113 | 0.8629 | 0.8356 | **0.8644** |
| | Max Pool | 0.8150 | **0.8661** | 0.8366 | 0.8521 |
| OUR DATA | 3D-GRU | 0.6816 | 0.7963 | 0.8114 | **0.8241** |
| | Max Pool | 0.6883 | 0.7844 | 0.8017 | **0.8066** |

Table 3.1: A comparison of IoU values tested using a single view on datasets [90, 85, 16], under the various input modes, when trained with two different view modules.

Let '$p$' be the predicted value at voxel $(i, j, k)$ and '$y$' the corresponding expected value. '$I$' is an indicator function which gives a value of 1 if the expression it is evaluating is true, if not, it gives 0. '$t$' is an empirically decided threshold of 0.5 above which the cell is considered as filled.

$$IoU = \frac{\sum_{i,j,k}[I(p(i,j,k) > t)I(y(i,j,k))]}{\sum_{i,j,k}[I(p(i,j,k) > t) + I(y(i,j,k))]} \tag{3.3}$$

**Baseline.** As described in Section 2.2.2, there are several standard encoder-decoder networks that use single/multi-view RGB image(s) for voxelized reconstruction of rigid objects. Therefore, as a baseline, we interpret this setting of using only RGB image(s) for both training and testing the reconstruction network. Further, the qualitative and quantitative results are compared and evaluated on ground truth data generated using traditional Multi-View Geometry (MVG) setups (from up to 22 views), on four datasets of varying complexity. For rendered sample outputs, please refer to the supplementary video.

### 3.3.3 Results & Discussion.

Figure 3.7: **Clothing induced deformations captured by our proposed method, VolumeNet, on [90].**

Quantitative results (IOU metric) in Table 3.1 suggest that for a variety of datasets of varying complexity and irrespective of the method of combining multiple views, the depth information is very critical for accurate reconstruction of human models. It is interesting to notice that the difference in IoU values between RGB and RGB/D widens under two scenarios - a) when the dataset has very complicated poses (such as the handstand sequence in MIT) and b) when the background becomes more complicated. Figure 3.8 re-emphasizes this ineffectiveness of using RGB alone, in which the first and second rows show the reconstructions from MIT's handstand and our captured data, respectively. The intuition behind the working of this training paradigm is that the co-learning of the shared filter weights of the two modalities acts as a regularization for one another, thus enhancing the information seen by the network.

Figure 3.9 shows the robustness of the learned model performing a vast range of actions. This robustness while reconstructing from a single image can be attributed to the network's ability to exploit the symmetry associated with non-rigid human shapes. As a result of not imposing any body-model constraint, we were able to partially handle non-rigid deformations induced by free form clothing as shown in Figure 3.7. While the current pipeline has not been trained to explicitly capture the temporal information available in sequences, results on MIT's Samba dance sequence [90] shown in Figure 3.7 show great promise for reconstructing performance capture scenarios from a single image.

(a) Input RGB (b) Ground Truth (MVG) (c) Baseline (RGB) (d) Ours (RGB/D)

Figure 3.8: **Qualitative comparison of VolumeNet with the baseline.**

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3.9: **Qualitative Results of VolumeNet**. A comparison of 3D Shapes obtained using VolumeNet, our monocular reconstruction network (first row) with ground truth models (second row) obtained with through multi-view setup with 22 cameras.

## 3.4 PoShNet: Decoupling Pose and Shape

In this section, we try to build upon VolumeNet and simplify the regression problem by using domain-specific information about non-rigid human shapes. We decouple learning pose and shape as well as guide the shape estimation with a pose prior. It is to be noted that the ideas in this section are a work in progress and aren't peer-reviewed yet. Qualitative results are given at the end of the chapter.

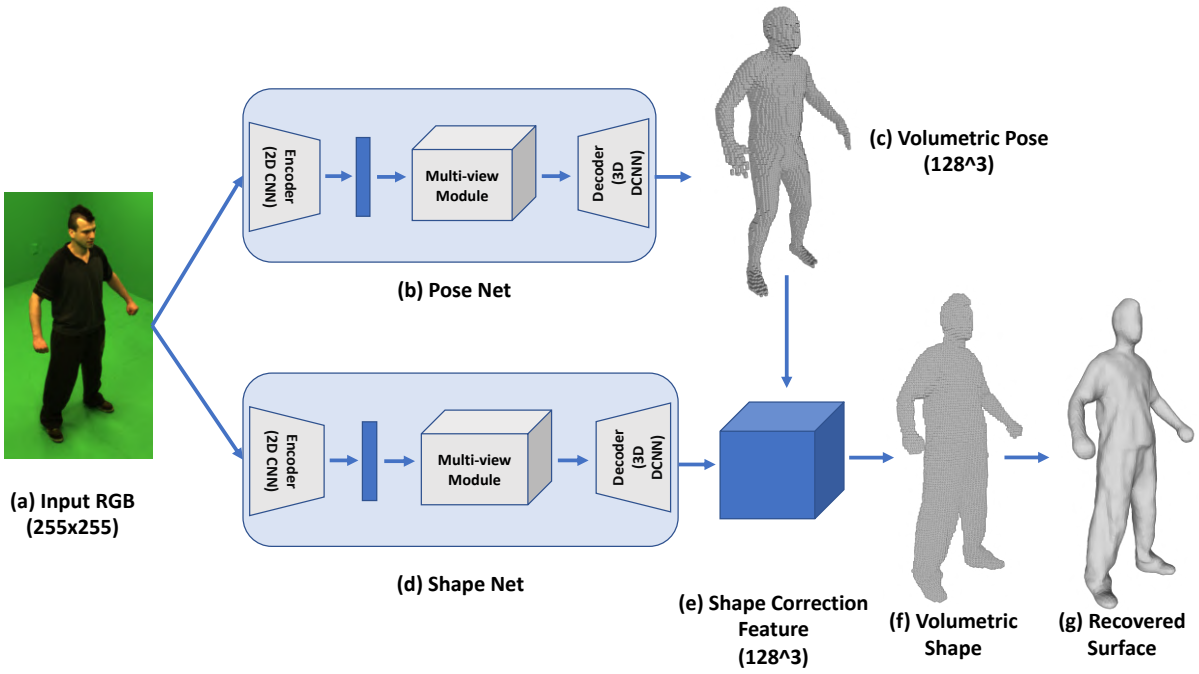



Figure 3.10: **Overview of the test-time flow of PoShNet - Decoupling of Volumetric Pose and Shape**. Given an input image (a), PoseNet (b) recovers a volumetric pose (c). Further, Shapenet (d) uses the same input image (a) to predict surface modifications to the volumetric pose (c) via the shape correction feature (e), to recover a modified volumetric shape (f) with accurate surface information. The final step involves Poisson's surface reconstruction algorithm [43] to recover a smooth mesh (g).

**Network Architecture.** We use domain-specific information by decomposing the learning into estimating the volumetric pose first, followed by volumetric shape completion/editing. As shown in Figure 3.10, we do so using PoseNet and ShapeNet, each having the same architecture as VolumeNet (see Figure 3.2). Hence, the training methodologies and input modes of VolumeNet extend here.

We follow a two-step procedure for training PoShNet -

- First, we train PoseNet to recover the volumetric pose of the human while ignoring shape information. This is facilitated by using the SMPL body model [51] to create ground truth 3D with correct pose information, but, template shape information.

| Dataset | VolumeNet | PoShNet |
|---|---|---|
| MPI-SMPL [85] | 0.804 | **0.921** |
| MIT [90] | 0.7547 | **0.844** |

Table 3.2: A quantitative comparison of IoU values between VolumeNet and PoShNet.

- Second, we train ShapeNet to use the same input image(s) as PoseNet and predict a shape correction feature (of the same size as volumetric pose grid), which when added to PoseNet's output enhances the surface information as well as corrects mistakes made by the PoseNet. Specifically, we add the final layer outputs produced by both PoseNet and ShapeNet, while back-propagating only through ShapeNet. Therefore, ShapeNet is forced to learn just surface information. The ground truth used in this step contains accurate surface geometry estimated with a calibrated multi-view setup.

**Loss Function.** While both ShapeNet and PoseNet use Voxel-wise Cross-Entropy (see Equation 3.2) for training, ShapeNet is trained by combining its output (shape completion feature) with that of PoseNet's prediction. This is elaborated as follows - at each voxel $(i, j, k)$, if $O^P$ is PoseNet's output and $O^S$ is ShapeNet's output, both are combined as follows to give '$p$', the final predicted value -

$$p_{(i,j,k)} = \quad O^P_{(i,j,k)} + O^S_{(i,j,k)} \forall (i, j, k) \tag{3.4}$$

This final predicted value, $p$, is used to calculate the loss against the ground truth's expected value using Equation 3.2.

**Results & Discussion.** As shown in Table 3.2, the decoupling strategy achieves superior surface reconstruction by breaking down the direct volumetric regression into two simpler sub-problems - pose and shape estimation, and guiding the shape estimation with a pose prior. Figure 3.12 shows the effectiveness of this approach, allowing ShapeNet to recover detailed deformations induced by free-form clothing. Such a modular approach opens up the possibility to extend this concept to add facial features, hair details, hand articulations, etc. in an end-to-end manner.

It is to be noted that our approach is superior to directly reconstructing the boundary voxels because we have eased the job of ShapeNet by - (a) already predicting the pose, and (b) providing an initial template volume in that pose, thereby allowing it to effectively utilize the volume prior via 3D convolutions.

**Shortcomings.** As a part of our experimentation, we explored different ways of combining and operating the pose and shape networks - (a) concatenation/additions at a multitude of intermediate feature levels, (b) directly operating on the volumetric pose's output with another 3D DCNN and (c) joint training of both networks, and (d) varying activation and loss function. Although the results of PoShNet prove to be superior, all of them (including PoShNet) suffer from instability. This can be perhaps be associated with the fact that two models control a single output space, causing multiple sources of error. Further, training

and testing PoShNet is extremely computationally expensive due to two pairs of 3D DCNNs, one each from PoseNet and ShapeNet.

## 3.5 Limitations

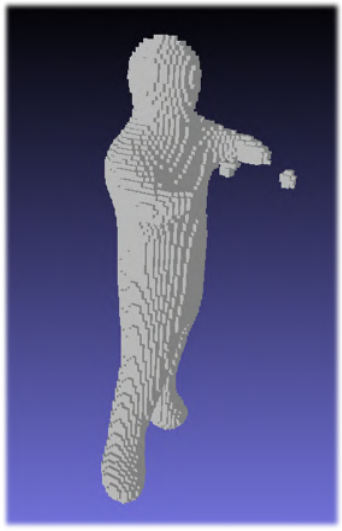


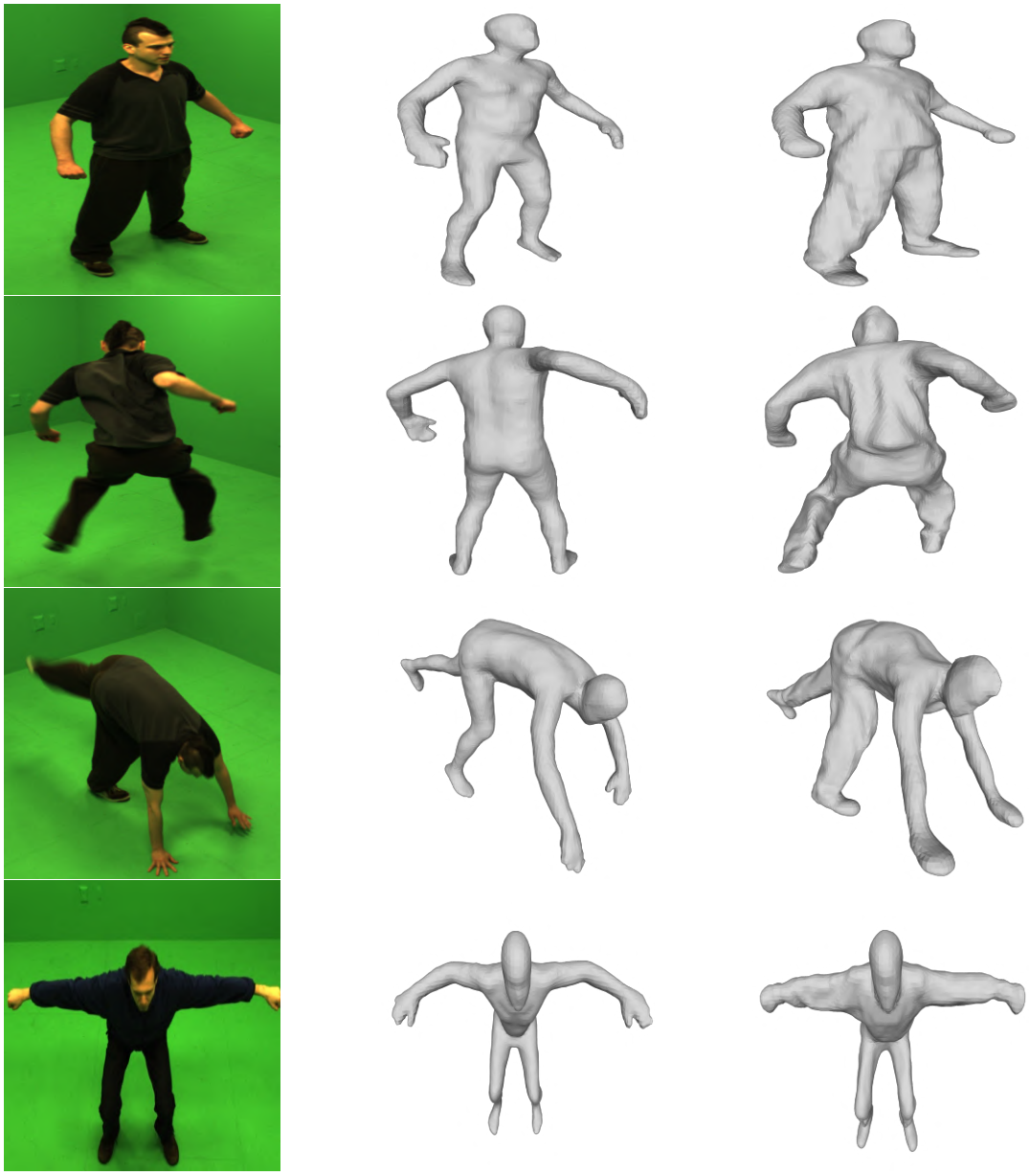
Figure 3.11: **Limitations with Volumetric Reconstruction** - 3D Models with Broken limbs reconstructed under certain difficult conditions - occlusions in input views and/or difficult poses.

- The human body has a natural symmetry and structure. Using an unbounded space (volumetric) to learn a structured object is sub-optimal. As seen in Figure 3.11, sometimes, the reconstruction can have broken hands/legs, i.e., exhibit the lack of a consistent topology. In other words, if we know that most humans have 2 hands, 2 legs, and one head, it would be more efficient to start the learning process from that point, rather than learning from scratch every-time. Hence, making use of a parametric template to learn the structure (pose), followed by a surface refinement for finer details might be a direction worth looking at.
- Learning a volumetric model is extremely expensive due to costly 3D convolutions. Further, although we're interested only in the surface boundary, a volumetric regression ends up filling the voxels within the surface as well, which adds additional strain on the learning algorithm.

## 3.6 Conclusion

Monocular 3D Human Reconstruction is a severely ill-posed problem due to self-occlusions caused by complex body poses and shapes, clothing obstructions, lack of surface texture, background clutter,

single view, etc. We proposed a novel deep learning pipeline that exploits the volumetric space to learn a more accurate surface and overcome these challenges with innovation in our training methodology by co-learning RGB with Depth cues as well as providing multi-view information so that single view reconstruction from only RGB is possible at test time. Further, we make the learning process easier by decoupling pose and shape estimation, while using the volumetric pose as a prior for better shape estimation. We show superior reconstruction performance using the proposed method in terms of quantitative and qualitative results on both publicly available datasets (by simulating the depth channel with virtual Kinect) as well as real RGBD data collected with a calibrated multi Kinect setup. As a part of future work, it will be practical to extend this to exploiting the temporal consistency for the task of reconstruction, for the case of continuous mesh sequences.



(a) Input RGB (b) PoseNet (c) ShapeNet

Figure 3.12: **Qualitative Results of PoShNet on MIT [90]**. Given an input image (a), the figure shows the volumetric pose predicted by PoseNet, followed by the modified surface predicted by ShapeNet (c). Note that the results depicted here are after smoothing [43].

# Chapter *4*

# Learning Statistical Body Models

Although volumetric regression enables recovering a more accurate surface reconstruction, it does so without an animatable skeleton. Such methods yield reconstructions of low resolution at higher computational cost (regression over the cubic voxel grid) and often suffer from an inconsistent topology via broken or partial body parts. Hence, statistical body models come as a natural choice because of the strong priors provided by the model offsets the ill-posed nature of the problem. Such models are created from thousands of scans of real people and inherently model anthropomorphic constraints such as limb/bone proportions. Further, their low dimensional space theoretically makes it easier to learn (in comparison with voxel-grids or point-clouds). An added advantage is that they easily integrate with existing graphics pipelines, thereby catering to a wide range of applications. Although they have often been critiqued with being incapable of capturing surface information, such models have the possibility of being integrated with physics-based clothing models in the future, and hence are an interesting representation to explore.

Recently, several end-to-end deep learning solutions for estimating the 3D parametric body model from a monocular image have been proposed [41, 59, 63, 80, 83, 98]. They all attempt to estimate the pose (relative axis-angles) and shape (PCA space) parameters of the SMPL [51] body model from a single image, which is a complex non-linear mapping. To get around this complex mapping, several methods transform them into rotation matrices [59, 63] or learn from 2D/3D keypoint and silhouettes projections of the predicted mesh [41, 59, 63]. Additionally, [41] proposes an alternate method for training (Iterative Error Feedback) as well as body-joint specific adversarial losses, which takes up to 5 days to train. Hence, learning the parametric body model hasn't been straightforward.

In this chapter, we explore a method to ease the learning process by providing a strong prior in the form of an approximate estimate. It is to be noted that the ideas in this chapter are a work in progress and are not yet peer-reviewed. We showcase some preliminary findings in this direction.

## 4.1 Contributions

To summarize, our primary contributions are as follows -

- Firstly, we propose a novel model - the CR framework to address the problem of complicated learning paradigms for parametric models.
- The proposed framework is the first of its kind for this problem, and aims at using classification to guide the body model regression. It acts as a strong prior by providing an approximate estimate from an image.

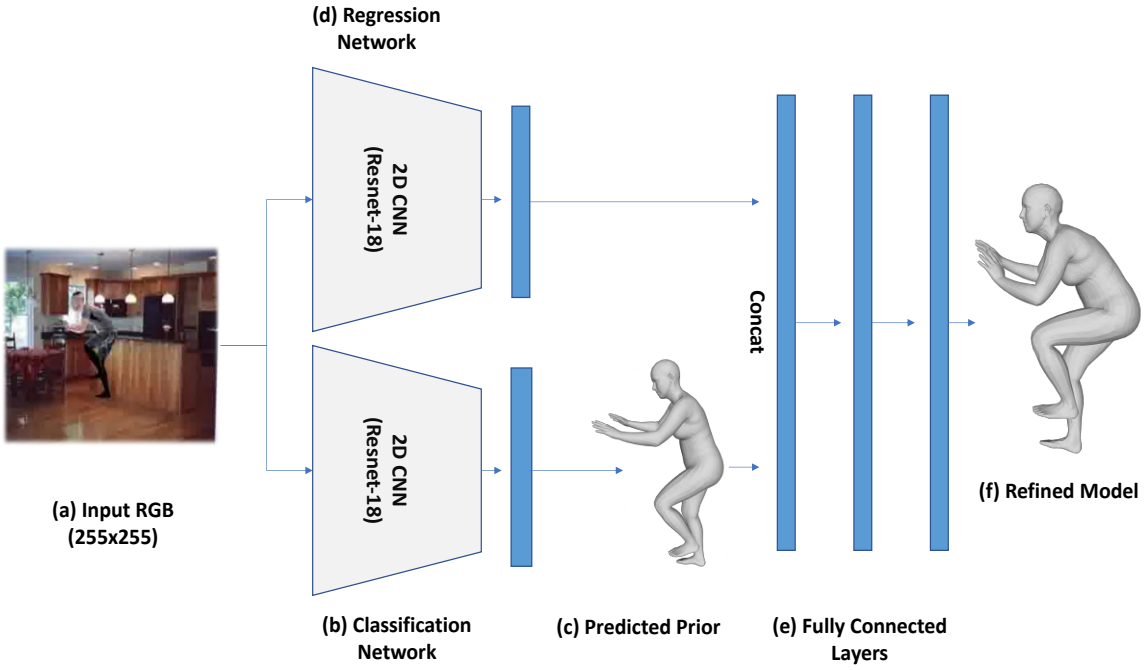



Figure 4.1: **Overview of CR Framework to predict parametric body models**. Given an input RGB image (a), the model first predicts an approximate prior (c) via the classification network (b). Then, from the same input RGB (a), the regression network (d), produces a CNN feature vector, which gets concatenated with the predicted prior (c). These are then passed through 3 fully connected layers (e) to predict the refined model (f).

## 4.2 Proposed Method: CR Framework

**Network Architecture.** In order to ease the learning process, we break down the problem into two parts - (a) to generate an approximate estimate and (b) to refine that estimate. As shown in Figure 4.1, we accomplish this with the help of (a) a 2D CNN (Resnet-18) that performs classification and (b) another 2D CNN (Resnet-18) that performs regression. A two-step procedure to train the CR Framework is outlined below -

- First, we solve a multi-label classification problem to map the input RGB image to 3 out of 500 possible priors (SMPL models). These 500 representative SMPL models have been learned from large-scale data to signify a wide range of poses and shapes, thereby acting as a smart initial estimate (refer to Section. 4.3.2 for more on data generation). The final predicted prior is then obtained from a linear combination of 3 candidate priors (weighed by the predicted probability).

- The second step involves a refinement of the prior. For this, the regression network's 1D feature vector is concatenated with the predicted prior and modified using fully-connected (FC) layers to obtain the refined model. The classification network's weights are frozen and loss is back-propagated only through the FC layers and regression network to guide the refining of the model with the RGB image.

**Loss Function.** In the first stage, we solve a multi-label classification problem and therefore use binary cross-entropy (Equation 4.1). If '$p$' is the predicted value, '$y$' is the ground truth and '$N$' is the total number of classes, then, the loss is given by -

$$L(p, y) = \frac{-1}{N} \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \tag{4.1}$$

Further, in the second stage, we follow NBF [59] and use the L2 norm (Equation 4.2) to capture differences between the predicted parameters and underlying joints with their respective ground truth. If '$P$', '$\hat{P}_i$' represent the predicted and ground truth model parameters and '$J$', '$\hat{J}_i$' represents the predicted and ground truth underlying 3D joints (recovered using the SMPL joint regressor [51]), then, the regression loss is defined as -

$$L = \sum_{\forall P_i, J_i} ||P_i - \hat{P}_i||_2 + ||J_i - \hat{J}_i||_2 \tag{4.2}$$

## 4.3 Experiment & Results

### 4.3.1 Datasets

**SURREAL.** This dataset provides synthetic image examples with 3D shape ground truth. The dataset draws poses from MoCap [38] and body shapes from body scans [68] to generate valid SMPL instances for each image. Although this dataset is synthetically generated, it emulates complex real poses and shapes, coupled with challenging input images that contain background clutter and are reflective with low resolution. It has a total of 1.6 million training and 15,000 test samples. Further, since [63, 86] show a good domain transfer to real data by training on SURREAL, the quality of SURREAL to correlate with real data is proven.

### 4.3.2 Implementation Details.

**Data Generation.** The ground truth for training the classification model is generated in a data-driven manner by first performing a Gaussian Mixture Model-based clustering [99] using a 1D concatenation of SMPL parameters [51] and 3D joint locations (flattened) as the feature, to obtain 500 cluster centers. The cluster centers obtained are representative of a wide variety of candidate models. This is followed by an assignment of the dataset's image-3D pairs to their respective top-3 cluster centers with a multi-label one-hot encoding. Due to numerical bottlenecks in clustering data points in the order of millions, we subsample the dataset and execute the above procedure.

Further, the input images are pre-processed by using ground truth bounding boxes given by the dataset to obtain a square crop of the human. This is a standard step performed by most comparative 3D human reconstruction models.

**Training Rubrics.** We use Nvidia's GTX 1080Ti, with 11GB of VRAM to train our models. A batch size of 96 is used for SURREAL. We use the ADAM optimizer with an initial learning rate of $10^{-4}$, to get optimal performance. Further, a standard $80:20$ split between the training and testing datasets is adhered to. Attaining convergence on SURREAL takes 10 hours for the entire framework.

**Evaluation Metric.** Since we're estimating a parametric model where the surface vertices are a linear combination of the transformations induced by the underlying joints (i.e., the surface is a function of the internal 3D skeleton), like NBF [59], we focus on the 3D joint error in estimating the quality of the fits, defined as the mean-per joint error between the ground truth and predicted joints in 3D.

**Baseline.** As a baseline, we use a Resnet-18 to directly predict parameters from the input image with an L2 loss on the parameters. This enables us to show the novelty introduced by our pipeline.

### 4.3.3 Results & Discussion.

| No. of Clusters | Clustering Feature | $JointError_{SegMask}$ | $JointError_{RGB}$ |
|---|---|---|---|
| 100 | Joint | 71.1 | 87.8 |
| 100 | Param | 72.3 | 89.3 |
| 500 | Joint | 69.3 | 86.2 |
| 500 | Param—Joint | **64.5** | **79** |

Table 4.1: Impact of different clustering parameters on the joint error with 2 input configurations - using Segmentation Mash and RGB Image.

**Ablation Study.** To provide an effective prior, it is essential for our clustering to effectively capture the diversity in the space and map it to the input images. Table 4.1 shows our experiments with different clustering parameters on two different input configurations - RGB and Segmentation Mask. We show

that as the number of clusters increase, the joint error reduces. This acts as a proof of concept that the CR Framework is effective. However, if the number of clusters becomes too high (above 500), it reduces the classification accuracy and induces noise into the estimated prior. As indicated in the table, the best feature for clustering was by flattening and concatenating the parameters and 3D joints. Also, much like shown in Chapter 5, the segmentation mask acts as a very powerful prior. Figure 4.2 gives the TSNE representation of 100 clusters in our dataset showing patterns/structures captured by the clustering.

Along similar lines, Table 4.2 shows a few different ways of providing the prior. 'Top 1' indicates providing only the parameters of the highest probability cluster center as prior. Similarly, 'Concat Top 3' indicates concatenating the top 3 predicted parameters and and 'L.C Top 3' indicates constructing the final prior by taking a linear combination of the top 3 predictions and weighing them by their predicted probabilities. Assuming that clustering captures the 3D Human pose and shape space, informally, the 500 clusters can be thought of as basis vectors and their linear combination can capture any motion. As indicated in the table, this configuration provides the best prior.

Table 4.3 shows the effectiveness of the CR framework with different input modalities. CR beats our baseline by a significant margin with RGB as its input. In configuration RGB/D, we further improve the learning by making use of our training methodology proposed in Chapter 3 of co-learning RGB with synthetically generated Depth to improve the test time reconstruction from only RGB. In this setting, the test time error of RGB reduces to 79, while that of Depth is 61mm. Knowing the effectiveness of Body Part Segmentation Masks from Chapter 5, we show that reconstruction from ground truth masks reduces the joint error to make it comparable to state-of-the-art RGB reconstruction methods. Further, with monocular depth, CR produces the best results, with a joint error of 57.2. Although these additional modalities work significantly better, we focus on reconstruction from RGB due to its ubiquitous nature.

| Method | Joint Error |
|---|---|
| Top 1 | 82.15 |
| Concat Top 3 | 80.72 |
| L.C Top 3 | 79 |

Table 4.2: Quantitative evaluation of the best method of constructing the final prior.

| Input | Network | Joint Error |
|---|---|---|
| RGB | Baseline | 104 |
| RGB | CR | 92 |
| RGB/D [87] | CR | 61(D), **79(RGB)** |
| Seg. Mask | CR | 64.5 |
| Depth | CR | 57.2 |

Table 4.3: An evaluation of the performance of CR with different input modalities.

| Method | Joint Error |
|---|---|
| Tung *et al.* [83] | 64.4 |
| SMPLR [52] | **55.8** |
| HMNet [88] | 71.9 |
| Baseline | 104 |
| CR | 79 |

Table 4.4: A comparison of CR with state-of-the-art methods.

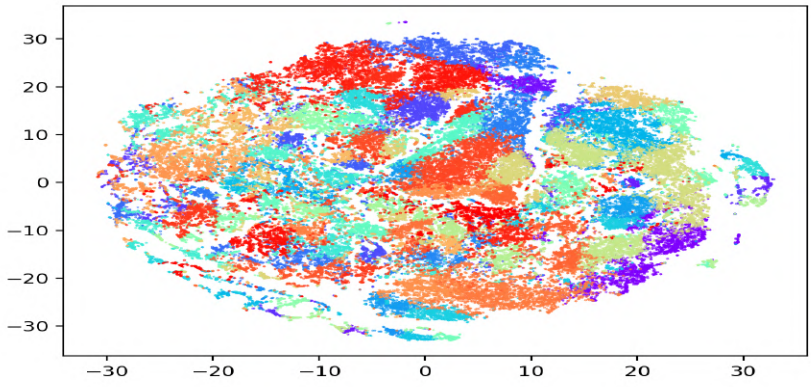


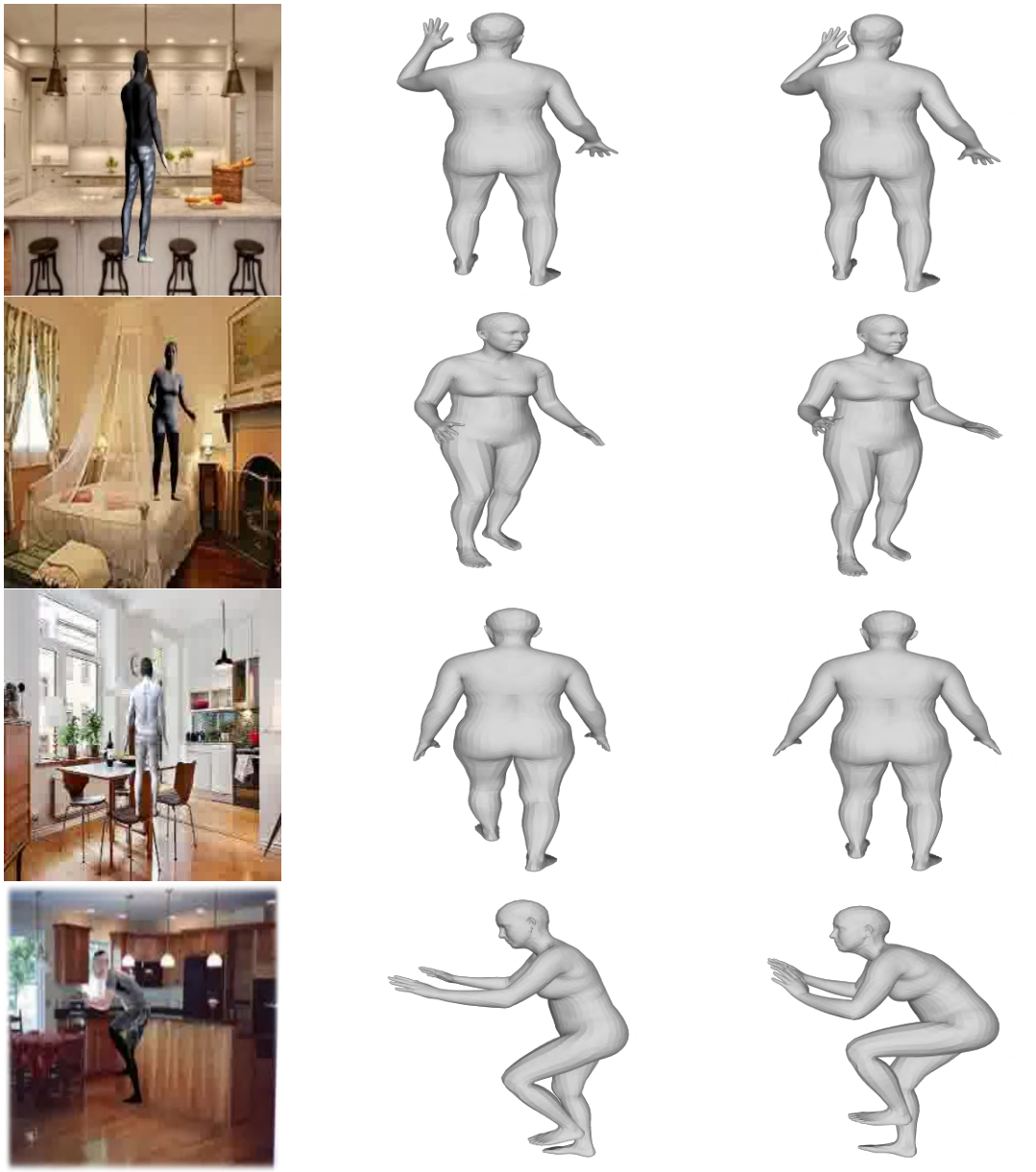

Figure 4.2: **TSNE representation of 100 clusters, each represented by a different colour.**

**Comparison with state-of-the-art.** With regard to monocular RGB reconstruction, CR falls behind state-of-the-art methods. This is indicated in Table 4.4. The qualitative results shown in Figure 4.3 provide some insight into this. As seen in the figure, the estimated prior is impressive, in terms of pose, shape, and global rotation. Hence, optimizing the classification network further isn't necessary. The regression network, on the other hand, isn't able to make use of this strong estimate and refine it. There might be two possible explanations for this -

- We found that although the combination of the two networks (C and R) happens at the later FC layers, the number of parameters is of the order of few millions, and increasing them reduces the performance. This could indicate that the difficulty of the problem is a bottleneck and it requires additional priors (possibly on the loss side) to learn better.
- Alternatively, the regression network's architecture is sub-optimal and requires better constructs and design to exploit the strong initial prediction.

## 4.4 Conclusion and Future Work

Monocular 3D Reconstruction is a severely ill-posed problem. Statistical Body Models are a natural choice since they are low-dimensional, yet constructed from thousands of real scans of humans. However, learning them hasn't been straightforward, with solutions relying on various projections of the predicted mesh. In this chapter, we presented the CR framework, as an attempt to simplify this learning procedure. The model provided an initial estimate via classification and attempted to refine it with a regression-based model. We showcased several results that acted as a proof of concept regarding the effectiveness of such a prior. However, the refinement of the prior requires more exploration. As a future direction, we would like to explore alternate architectures with increased 2D supervision and possibly joint learning.

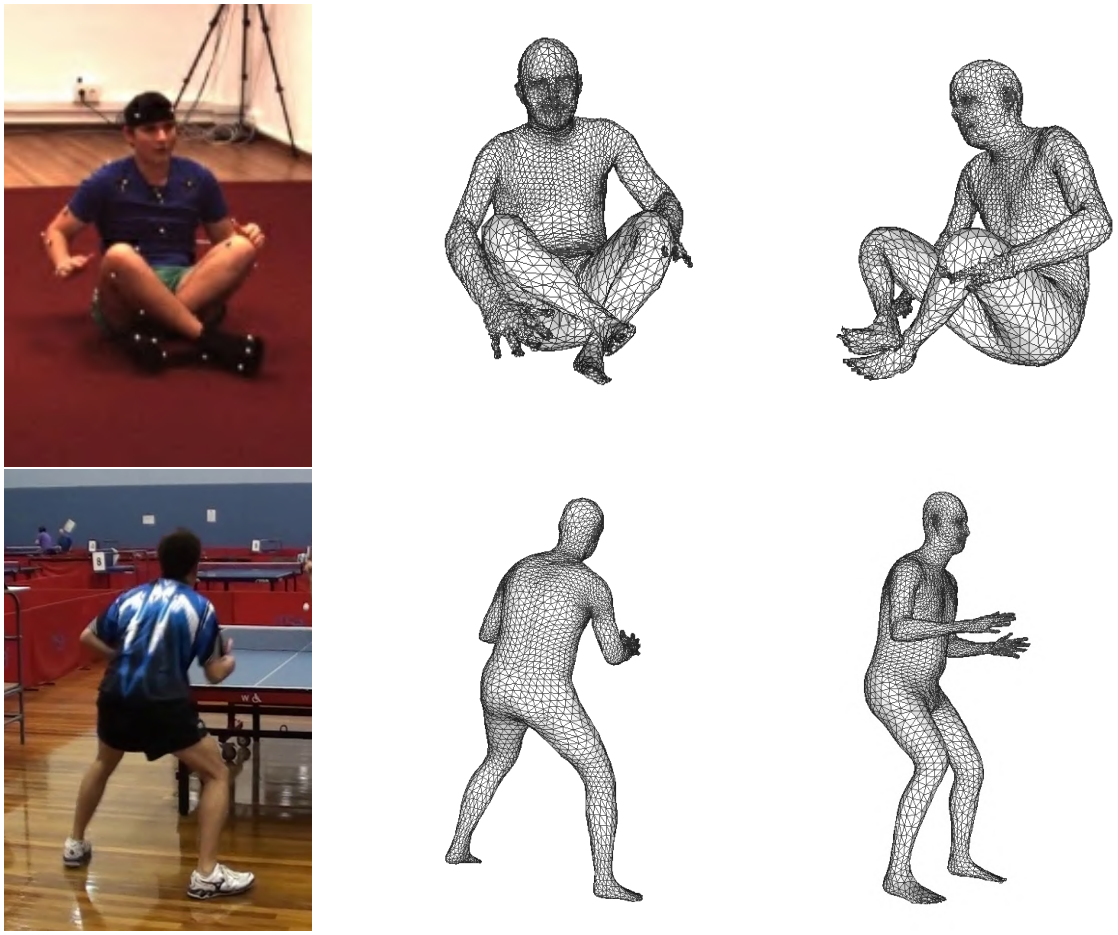


(a) Input RGB (b) Estimate from Classification (c) Refined Model

Figure 4.3: **Qualitative Results of the CR Framework on SURREAL [85]**.

# Chapter *5*

# Implicit Point Cloud Reconstruction



(a) (b) (c)

Figure 5.1: **We present an early method to integrate Deep Learning with the sparse mesh representation(b)**, to successfully reconstruct the 3D mesh of a human from a monocular image (a). (b) represents the reconstructed 3D mesh aligned to the input image, and (c) is a rotated version of the same, for aesthetic reasons.

Some of the recent deep learning methods employ volumetric regression to recover the voxel grid reconstruction of human body models from a monocular image [84, 87]. Although volumetric regression enables recovering a more accurate surface reconstruction, they do so without an animatable skeleton [87], which limits their applicability for some of the aforementioned applications. [84] attempted to overcome this limitation by fitting a parametric body model on the volumetric reconstruction using a silhouette reprojection loss. Nevertheless, in general, such methods yield reconstructions of low resolution at higher computational cost (regression over the cubic voxel grid) and often suffer from an inconsistent topology via broken or partial body parts.

Alternatively, the parametric body model [9, 51, 70] based techniques address some of the above issues, however, at the cost of accurate surface information [13, 34, 46, 74]. Recently, several end-to-end deep learning solutions for estimating the 3D parametric body model from a monocular image have been proposed [41, 59, 63, 80, 83, 98]. They all attempt to estimate the pose (relative axis-angles) and shape parameters of the SMPL [51] body model, which is a complex non-linear mapping. To get around this complex mapping, several methods transform them into rotation matrices [59, 63] or learn from the 2D/3D keypoint and silhouettes projections (a function of the parameters) [41, 59, 63]. Additionally, [41] proposes an alternate method for training (Iterative Error Feedback) as well as a body joint-specific adversarial loss, which takes up to 5 days to train. In other words, learning the parametric body model hasn't been straightforward and has resulted in complex and indirect solutions that in-fact rely on different projections of the underlying mesh.

Directly regressing to point cloud or mesh data from image(s) is a severely ill-posed problem and there are very few attempts in deep learning literature in this direction [54, 97]. With regard to point cloud regression, most of the attempts are focused on rigid objects, where learning is done in a class-specific manner. Apart from a very recent work [45], learning a mesh hasn't been explored much for reconstruction, primarily because of the lack of deep learning constructs to do so.

In this chapter, we attempt to work in between a generic point cloud and a mesh - i.e., we learn an "implicitly structured" point cloud. We hypothesize that in order to perform parametric body model-based reconstruction, instead of learning the highly non-linear SMPL parameters, learning its corresponding point cloud (although high dimensional) and enforcing the same parametric template topology on it is an easier task. This is because, in SMPL like body models, each of the surface vertices is a sparse linear combination of the transformations induced by the underlying joints i.e., implicitly learning the skinning function by which parametric models are constructed is easier than learning the non-linear axis-angle representation itself (parameters). Further, such models lack high-resolution local surface details as well. Therefore, there are far fewer "representative" points that we have to learn. Consequently, in comparison with generic point cloud regression, this is an easier task because of this implicit structure that exists between these points.

Going ahead, attempting to produce high-resolution meshes is a natural extension that is easier in 3D space than in the parametric one. Therefore, we believe that this is a direction worth exploring and we present an initial solution in that direction - HumanMeshNet that simultaneously performs shape

estimation by regressing to template mesh vertices (by minimizing surface loss) as well receives a body pose regularisation from a parallel branch in multi-task setup. The image to mesh vertex regression is further explicitly conditioned on the neighborhood constraint imposed by the mesh topology, thus ensuring a smooth surface reconstruction. Figure 5.2 outlines the architecture of HumanMeshNet.

Ours is a relatively simpler model as compared to the majority of the existing methods for volumetric and parametric model prediction (e.g., [84]). This makes it efficient in terms of network size as well as feed-forward time yielding significantly high frame-rate reconstructions. At the same time, our simpler network achieves comparable accuracy in terms of surface and joint error w.r.t. majority of state-of-the-art techniques on three publicly available datasets. The proposed paradigm can theoretically learn local surface deformations induced by body shape variations which the PCA space of parametric body models can't capture. In addition to predicting the body model, we also show the generalizability of our proposed idea for solving a similar task with different structure - non-rigid hand mesh reconstructions from a monocular image. *Note that the work in this chapter has been published in ICCV-W, 2019 [88].*

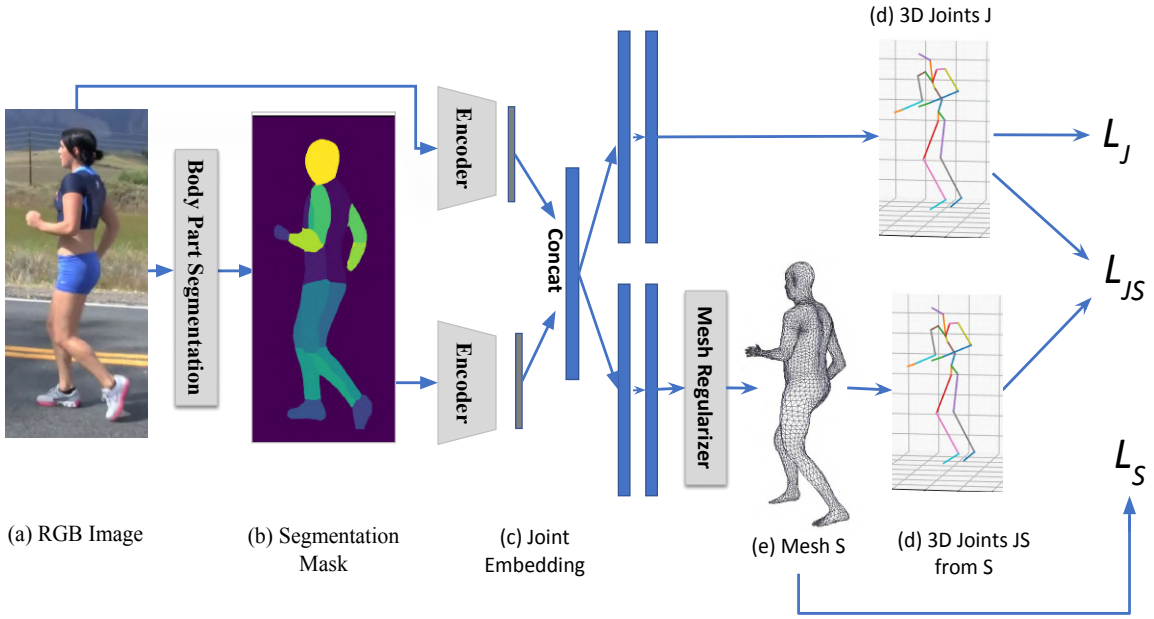



Figure 5.2: **Overview of HumanMeshNet [88] - A Multi-Task 3D Human Mesh Reconstruction Model**. Given a monocular RGB image (a), we first extract a body part-wise segmentation mask using [7] (b). Then, using a joint embedding of both the RGB and segmentation mask (c), we predict the 3D joint locations (d) and the 3D mesh (e), in a multi-task setup. The 3D mesh is predicted by first applying a mesh regularizer on the predicted point cloud. Finally, the loss is minimized on both the branches (d) and (e).

## 5.1 Contributions

To summarize, the key contributions of this work are:

- We propose a simple end-to-end multi-branch, multi-task deep network that exploits a "structured point cloud" to recover a smooth and fixed topology mesh model from a monocular image.
- The proposed paradigm can theoretically learn local surface deformations induced by body shape variations which the PCA space of parametric body models can't capture.
- The simplicity of the model makes it efficient in terms of network size as well as feed forward time yielding significantly high frame-rate reconstructions, while simultaneously achieving comparable accuracy in terms of surface and joint error, as shown on three publicly available datasets.
- We also show the generalizability of our proposed paradigm for a similar task of reconstructing the hand mesh models from a monocular image.

## 5.2 Proposed Method: HumanMeshNet

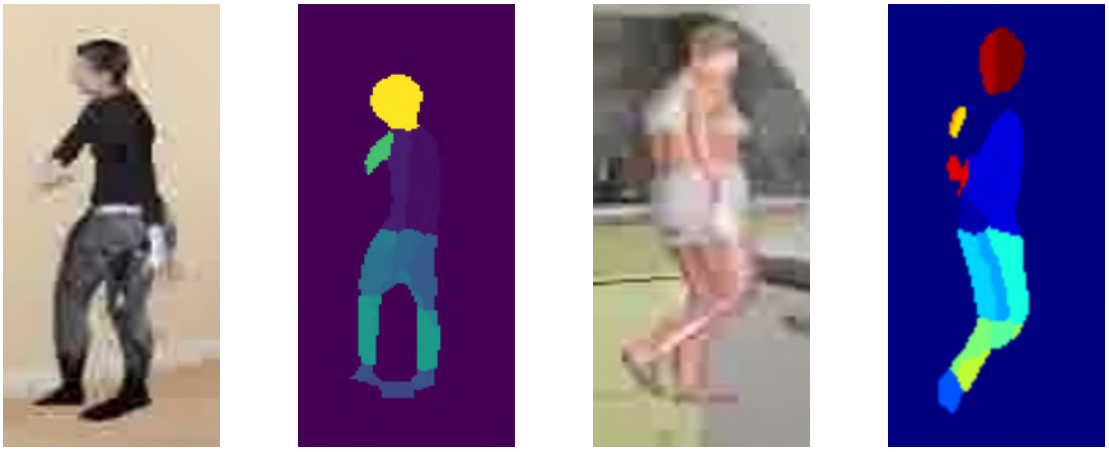
To learn this structured point cloud, we use an encoder- and multi-decoder model, which we describe in this section. Figure 5.2 gives an overview of our end-to-end pipeline. Our model consists of three primary phases:

**Phase 1 - RGB to Partwise Segmentation:** Given an input RGB image of size 224x224, we first predict a discrete body part label for each pixel in the input image (for a total of 24 body parts) using just the body part labeling network from [7]. A part-wise segmentation enables tracking of the human body in the image, making it easier for shape estimation.

**Phase 2 - Image Encoders and Joint Embedding:** Both the RGB image and segmentation mask are passed through separate encoders, each a Resnet-18, and their respective CNN feature vectors, each of dimension 1000 are concatenated together to obtain a joint embedding.

Such fusion of RGB and segmentation mask was employed to combine complementary information from each modality. This is important as a segmentation mask predictions can be very noisy in many scenarios (see Figure 5.3), e.g., low lighting, the distance of the person from the camera, sensing noise, etc., leading to failures like interchanged limbs or missing limbs.

**Phase 3 - Multi-branch Predictions:** From our concatenated feature embedding, we branch out into two complementary tasks via Fully Connected layers (FCs). Each branch consists of two FCs, each of dimension 1000 followed by the respective output dimensions for the 3D joints, and 3D surface respectively. It is to be noted that our predictions are in the camera frame.



(a) (b) (c) (d)

Figure 5.3: **Noisy Segmentation Masks predicted from images (a) and (c) in Phase 1**. The figure shows (b) missing body part masks (d) confusing between leg limbs.

**Loss Function.** We use a multi-branch loss functions to train our network i.e, $L_S$, $L_J$ and $L_{JS}$. We regularized the loss functions such that they contribute equally to the overall loss. This translates to Equation 5.1.

$$L = L_S + (\lambda_1 * L_J) + (\lambda_2 * L_{JS}) \tag{5.1}$$

The surface loss $L_S$ in Equation 5.2 gives the vertex-wise Euclidean distance between the predicted vertices $V_i$ and ground truth vertices $\hat{V}_i$ for the 3D mesh prediction branch in Figure 5.2 (e).

$$L_S = \sum_{\forall V_i} ||V_i - \hat{V}_i||_2 \tag{5.2}$$

However, this loss does not ensure prediction of smooth surfaces as each vertex is independently predicted.

Nevertheless, each mesh vertex has a neighborhood structure that can be used to further refine the estimate of an individual vertex. Here we make use of smoothing regularisation [77] (as shown in Equation 5.3), where the position of each vertex, $V_i$, is replaced by the average position, of its neighbours $N(V_j)$.

$$V_i = \frac{1}{|N(V_i)|} \sum_{V_j \in N(V_i)} V_j \qquad \forall V_i \tag{5.3}$$

This is achieved by first applying the smoothness mesh regularization given by Equation 5.3 and then calculating $L_S$. This helps in limiting the number of surface jitters or irregularities.

In order to enforce 3D joints consistency, we minimize joint loss $L_J$ defined in Equation 5.4, which gives the euclidean distance between the predicted joints $J_i$ and ground truth joints $\hat{J}_i$ in the 3D joint prediction branch as shown in Figure 5.2(d).

$$L_J = \sum_{\forall J_i} ||J_i - \hat{J}_i||_2 \tag{5.4}$$

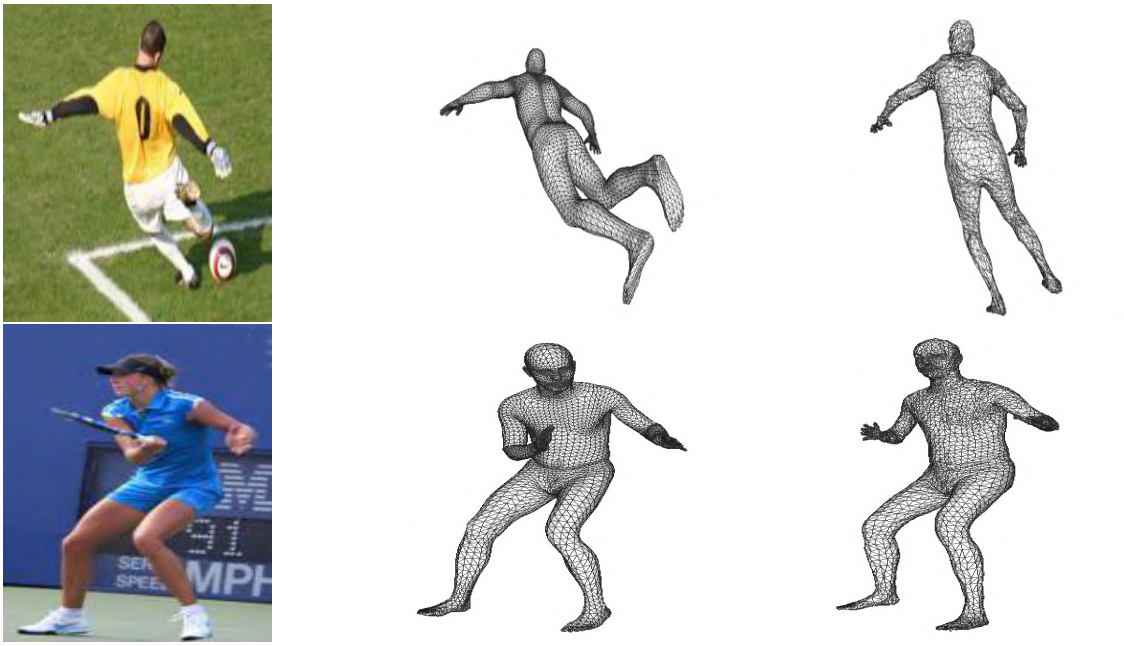
The 3D joints $JS_i$ under the surface are recovered using the SMPL joint regressor [51]. We also minimize the loss $L_{JS}$ defined in Equation 5.5 which gives the euclidean distance between the joints $J_i$ predicted from the joints branch and the joints $JS_i$ from the surface branch. It helps both the branches to learn consistently with each other.

$$L_{JS} = \sum_{\forall J_i} ||J_i - JS_i||_2 \tag{5.5}$$

**Network Variants:** We define two different variants of HumanMeshNet in order to perform an extensive analysis:

(a) HumanMeshNet (HMNet) - The base version which uses an "off-the-shelf" body part segmentation network ( [7]).

(b) HumanMeshNetOracle (HMNetOracle) - A refined version using a more accurate body part segmentation given by the dataset. However, in some datasets (e.g., UP-3D, [46]), these segmentation masks can be noisy due to manual annotations.



(a) Input Image (b) SMPLify [13] (c) HMNet (Ours)

Figure 5.4: **This figure depicts the poor quality of ground truth fits provided on UP-3D**. (a) The input RGB image is fit using SMPLify [13] to give (b) the ground truth. Our fit (c) makes use of more accurate markers or keypoints in a multi-branch setup, to account for noisy ground truth mesh data.

## 5.3 Experiments & Results

In this section, we show a comprehensive evaluation of the proposed model and benchmark against the state-of-the-art optimization and deep learning-based Parametric (P), Volumetric (V) and Surface-based (S) reconstruction algorithms. It is to be noted that we train on each dataset separately and report on its given test sets. All of the trained models and code shall be made publicly available, along with a working demo. Please view our supplementary video [3] for more results.

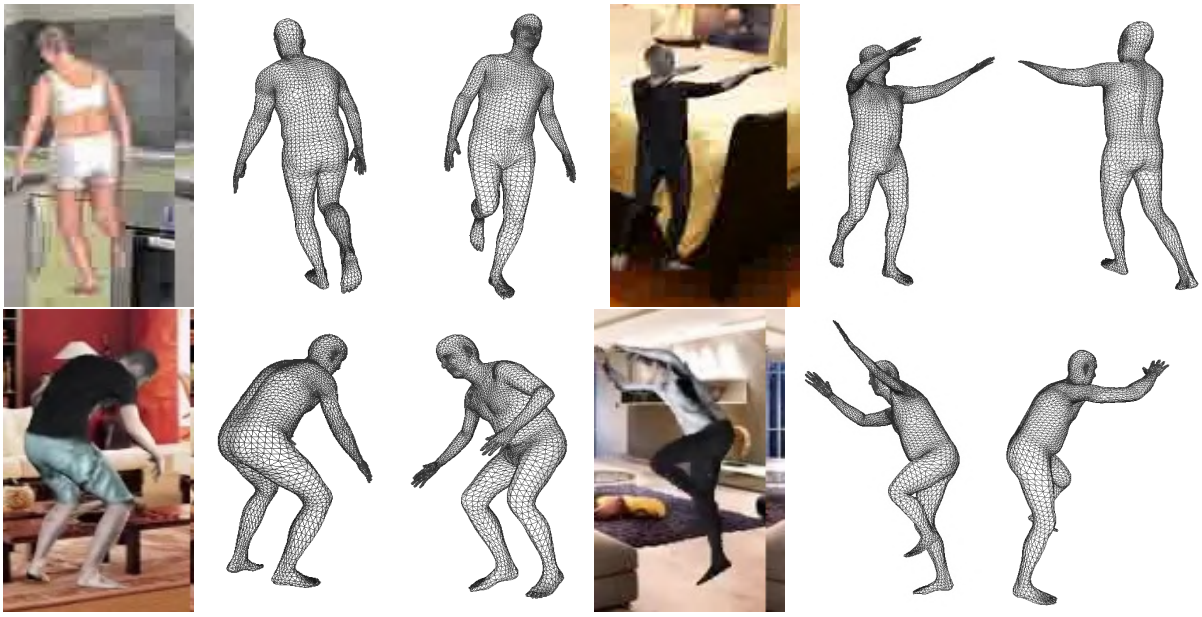


Figure 5.5: **Qualitative Results on SURREAL [86]** where (a) represents the input view, (b) our mesh reconstruction aligned to the input view, and (c) aligned to another arbitrary view.

### 5.3.1 Datasets

**SURREAL [86]**: This dataset provides synthetic image examples with 3D shape ground truth. The dataset draws poses from MoCap [38] and body shapes from body scans [68] to generate valid SMPL instances for each image. Although this dataset is synthetically generated, it emulates complex real poses and shapes, coupled with challenging input images that contain background clutter and are reflective with low resolution. It has a total of 1.6 million training and 15,000 test samples.

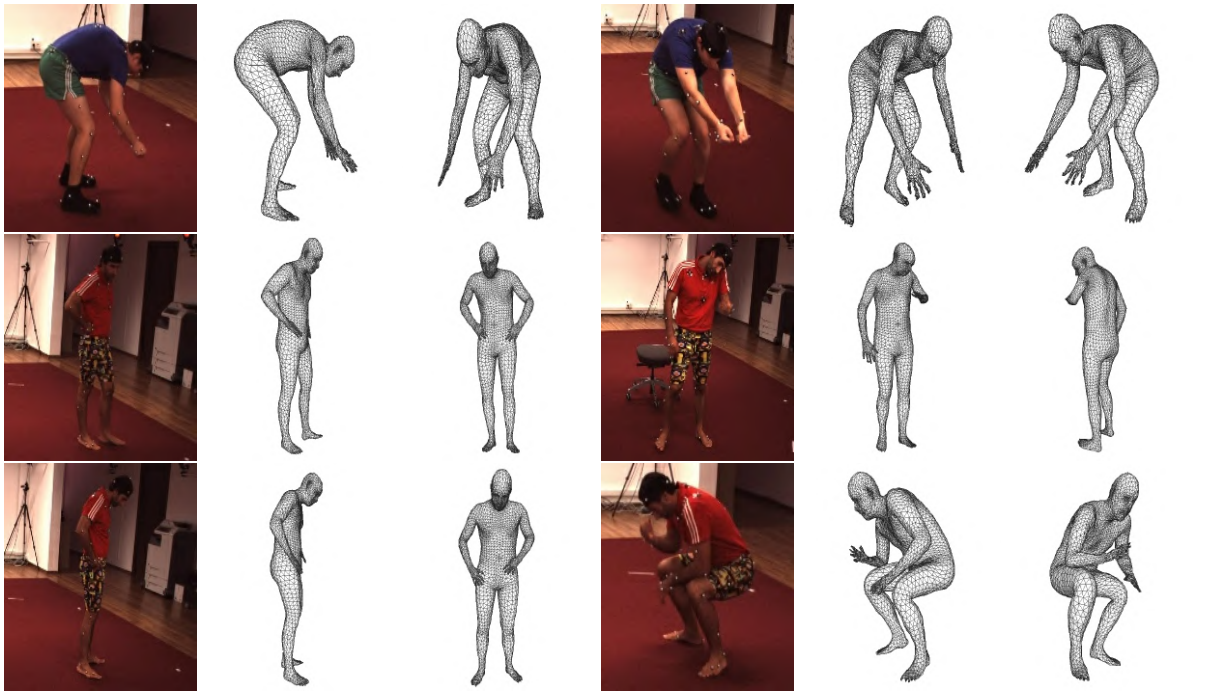
**UP-3D [46]**: It is a recent dataset that collects color images from 2D human pose benchmarks and uses an extended version of SMPLify [13] to provide 3D human shape candidates. The candidates were evaluated by human annotators to select only the images with good 3D shape fits. It comprises of 8,515 images, where 7,126 are used for training and 1,389 for testing. However, the ground truth meshes are sometimes inaccurately generated as shown in Figure 5.4. We separately train the network and report

results on the full test set of UP-3D.

(a) (b) (c) (a) (b) (c)

Figure 5.6: **Qualitative Results on UP-3D [46]** where (a) represents the input view, (b) our mesh reconstruction aligned to the input view, and (c) aligned to another arbitrary view.

**Human3.6M [38]**: It is a large-scale pose dataset that contains multiple subjects performing typical actions like "eating" and "walking" in a lab environment. It consists of a downsampled version of the original data with 300,000 image-3D joint pairs for training and 100,000 such for testing. Since ground truth 3D meshes for any of the commonly reported protocols [13] for evaluation aren't available anymore, we finetune SURREAL-pretrained network using joint loss only. We report the joint reconstruction error (trained as per Protocol 2 of [13]) and therefore compare with those methods that don't use mesh supervision for this dataset in Table 5.3.



(a) (b) (c) (a) (b) (c)

Figure 5.7: **Qualitative Results on Human3.6M**, where (a) represents the input view, (b) our mesh reconstruction aligned to the input view, and (c) aligned to another arbitrary view.

### 5.3.2 Implementation Details

**Data Pre-processing.** We use the ground truth bounding boxes from each of the datasets to obtain a square crop of the human. This is a standard step performed by most comparative 3D human reconstruction models.

**Network Training.** We use Nvidia's GTX 1080Ti, with 11GB of VRAM to train our models. A batch size of 64 is used for SURREAL and Human 3.6M datasets and a batch size of 16 for the UP3D dataset. We use the ADAM optimizer having an initial learning rate of $10^{-4}$, to get optimal performance. Attaining convergence on the SURREAL and Human3.6M takes 18 hours each, while on UP-3D takes 6 hours. We use the standard splits given by the datasets, for benchmarking, as indicated in Section 5.3.1.

**Procrustes Analysis (PA).** In order to evaluate the quality of the reconstructed mesh, we also report results after solving the Orthogonal Procrustes problem [32], in which we scale the output to the size of the ground truth and solve for rotation. Additionally, we also quantitatively evaluate without this alignment.

**Evaluation Metric.**

(a) Surface Error (mm): Gives the mean-per-vertex error between the ground truth and predicted mesh.

(b) Joint Error (mm): Gives the mean-per joint error between the ground truth and predicted joints. All reported results are obtained from the underlying joints of the mesh, rather than the alternate branch unless otherwise mentioned.

(c) PA. Surface/Joint Error (mm): It is the surface/joint error after Procrustes Analysis (PA).

| Output | Method | Surface Error | Joint Error |
|---|---|---|---|
| P | Tung *et al.* [83] | 74.5 | 64.4 |
| | Pavlakos *et al.* [63] | 151.5 | - |
| | SMPLR [52] | 75.4 | 55.8 |
| V | BodyNet [84] | 65.8 | - |
| S | Baseline | 101 | 85.7 |
| | HMNet[subsampled] | 86.9 | 72.4 |
| | HMNet | 86.6 | 71.9 |
| | HMNetOracle | **63.5** | **49.1** |

Table 5.1: Comparison with state-of-the-art methods on SURREAL's test set [86].

### 5.3.3 Comparison with State-of-the-art

**Baseline.** We define our baseline as the direct prediction of a point cloud from an RGB image, using a Resnet-50. This enables us to show the novelty introduced by our pipeline and the usefulness of learning in this output space.

**Results & Discussion.** For qualitative results on all of the three datasets refer to Figures 5.5- 5.7. A large amount of training data is required to learn a vast range of poses and shapes. However, [63, 86]

| Output | Method | Surface Error | Joints Error | PA. Surface Error | PA. Joint Error |
|---|---|---|---|---|---|
| P | Pavlakos *et al.* [63] | 117.7 | - | - | - |
| P | Lasner *et al.* [46] | 169.8 | - | - | - |
| P | NBF [59] | - | - | - | 82.3 |
| V | BodyNet [84] | 80.1 | - | - | - |
| S | Baseline | 151.4 | 130.8 | 93.8 | 83.7 |
| S | HMNet | 130.4 | 112.5 | 77.6 | 69.6 |
| S | HMNetOracle | **60.3** | **51.5** | **42.9** | **37.9** |

Table 5.2: Comparison with other methods on UP3D's full test set [46].

| 3D mesh Supervision | Method | PA. Joint Error |
|---|---|---|
| No | Ramakrishnan *et al.* [67] | 157.3 |
| | Zhou *et al.* [103] | 106.7 |
| | SMPLify [14] | 82.3 |
| | SMPLify 91 kps [46] | 80.7 |
| | Pavlakos *et al.* [63] | 75.9 |
| | HMR [41] | **56.8** |
| | HMNet(Ours) | 60.9 |
| Yes | NBF [59] | 59.9 |
| | SMPLR [52] | 56.4 |
| | CMR [45] | **50.1** |

Table 5.3: Joint Reconstruction error as per Protocol 2 of Bogo *et al.* [13] on Human 3.6M [38]. Refer to Section 5.3.1 for details on 3D mesh supervision.

show a good domain transfer to real data by training on the synthetic SURREAL dataset. Since our supervision is dominated by surface meshes, SURREAL plays an important role in benchmarking our method. We show comparable performance on it, as indicated by Table 5.1. In Table 5.1, we also show our results with a subsampled mesh (subsampled as per [45]) from 6890 to 1723 vertices with almost no change in reconstruction error. This is a good proof of our hypothesis that there are far fewer representative points to learn in this structured point cloud.

UP-3D is an "in the wild" dataset, however, has inaccurate ground truth mesh annotations, as shown in Figure 5.4. Most circumvent this issue, by avoiding 3D supervision altogether and projecting back to a silhouette or keypoints [41, 63]. Further, training on such a small dataset doesn't provide a good generalization. Therefore, we observe a higher error in HMNet. However, HMNetOracle produces a significant increase in accuracy with the increase in quality of the input image and segmentation mask (Table 5.4). Similar to state-of-the-art methods [45, 84, 87], we rely on 3D body supervision and providing more supervision like silhouette and 2D keypoint loss like [41, 84] can improve the performance further. For Human3.6m, we compare with those that don't use mesh supervision (since this data is currently unavailable) and achieve comparable performance.

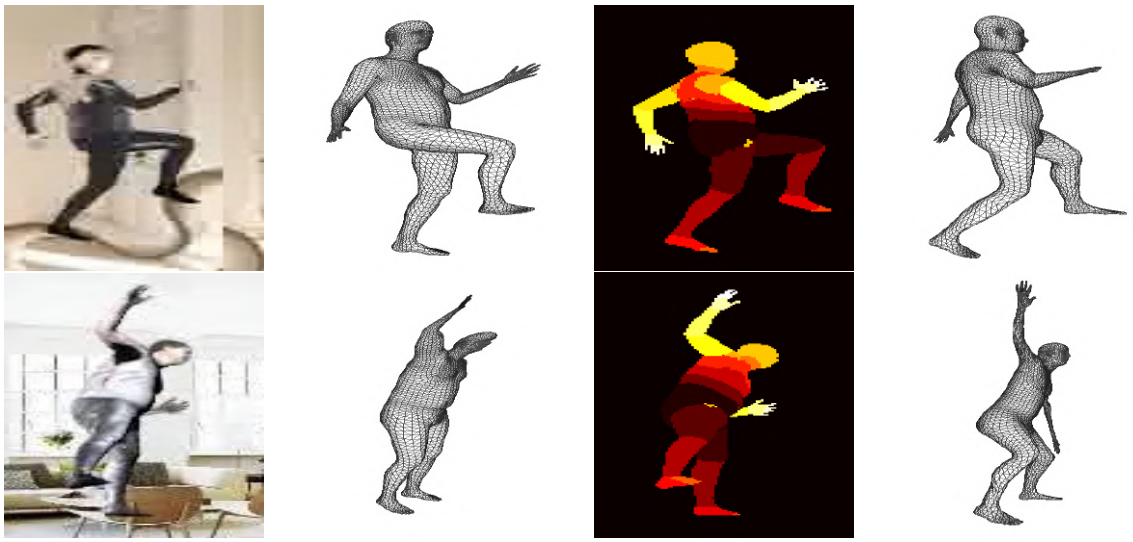
### 5.3.4 Discussion

**Ablation Study:** Directly regressing the mesh from RGB leads to sub-par performance. Limbs are typically the origin of maximum error in reconstruction, and the segmentation mask provides a better tracking in scenarios such as leg-swap shown in Figure 5.8. The first two rows of Table 5.4 quantitatively explain this behavior. Further, by having a more accurate segmentation mask, HMNetOracle achieves a significant reduction in surface error ($\downarrow 34.7mm$). In scenarios with inaccurate ground truth 3D (Figure 5.4), the reg-

| Config. | Input | PA. Surface Error | PA. Joint Error |
|---|---|---|---|
| Baseline | RGB | 93.8 | 83.7 |
| Single Task | $SM_{DP}$ | 82.9 | 74.6 |
| Single Task | RGB+$SM_{DP}$ | 79.2 | 71.04 |
| HMNet | RGB+$SM_{DP}$ | 77.6 | 69.6 |
| HMNetOracle | RGB+$SM_{GT}$ | **42.9** | **37.9** |

Table 5.4: Effect of each network module on the reconstruction error on UP-3D dataset. $SM_{DP}$ and $SM_{GT}$ denotes segmentation obtained from Densepose and groundtruth respectively.

ularisation 3D joint loss in our multi-branch setup helps us in recovering better fits (row 4 for UP3D). In datasets such as Human3.6m where accurate MoCap markers are given, this multi-branch loss provides a good boost - with and without joint loss, the joint reconstruction error is 60.9mm v/s 67.3mm respectively.



(a) Input RGB (b) Baseline (c) Predicted Mask (d) Our Output

Figure 5.8: **Shows the impact of the body-segmentation mask.** Given an input RGB image (a), this figure depicts a comparison of the baseline (b), against our output, HMNet (d). The predicted part-wise segmentation mask (c) assists HMNet to track the body parts and therefore solve the confusion between the legs as well as complex poses.

**Effect of Mesh Regularisation** Our mesh regularization module adds a smoothing effect while training, therefore ensuring that the entire local patch should move towards the ground truth for minimizing the error. Figure 5.9 shows the impact of this regularization. The error goes down from 83.7 to 63.5mm after the regularisation, in SURREAL.

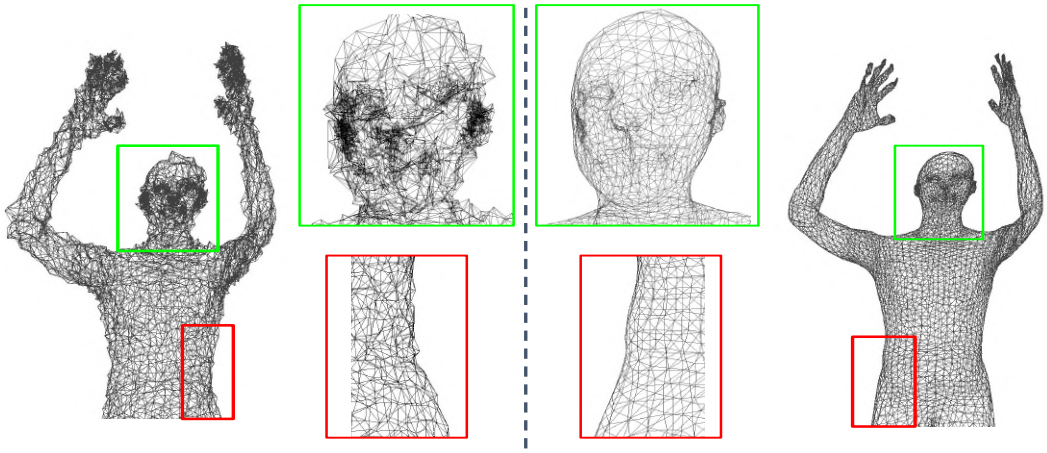
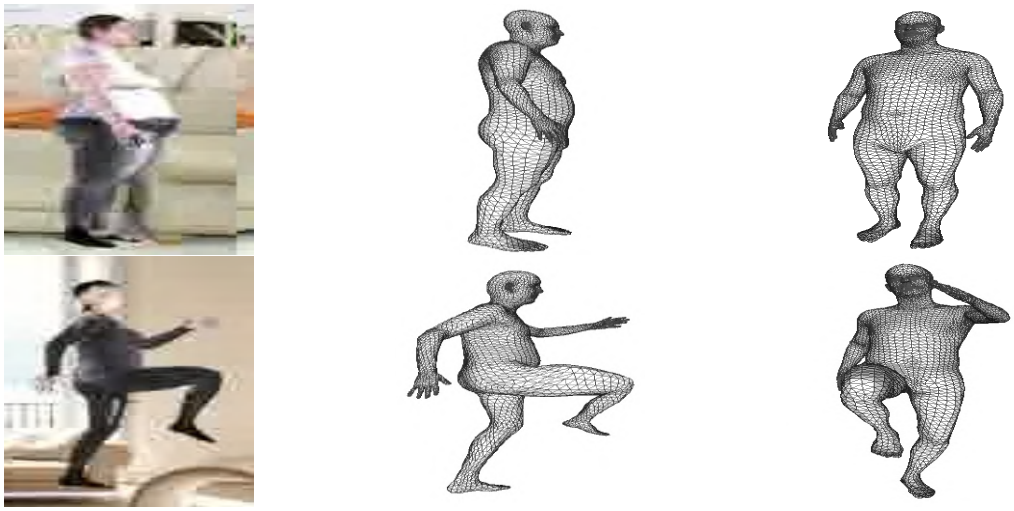


Figure 5.9: **Results showing the effect of our mesh regularization module while learning**. The figure on the left shows the irregularities in the mesh reconstructed, without our regularization, while the one on the right shows the smoothness induced by our regularizer.

**Recovering Shape Variations** Most parametric models prediction work with a neutral template model [41], and would have to learn the gender from the image. In our method, a direct mesh regression can learn the local shape variations (as long as training data has such variations) which extend to inherently learning gender invariant meshes. Two such samples are showing in Figure 5.10.



(a) Input RGB (b) Our Output (view 1) (c) Our output (view 2)

Figure 5.10: **Sample Shape Variations recovered by our model**, given a low resolution input image (a), rendered from the recovered view (b) and another arbitrary view (c)

**Generalizability to Hand Mesh Models.** We show the generalizability of our model to a similar task with a different structure. First, we populated a SURREAL like synthetic hand dataset using the MANO

hand model [70], similar to [27] with a total of 70,000 image-mesh pairs. We train our model on this dataset to predict hand surface and joints from an input RGB image using the same pipeline described in Figure 5.2. The training setting remains the same as earlier, and we obtain impressive qualitative results as shown in Figure 5.12. An interesting result is that of the mesh in the second block of Figure 5.12, in which the thumb and the index finger are merged. Due to using a surface representation directly, invalid configurations can be generated. The average surface error across the test dataset is 1mm, which acts as a proof of concept that polygonal mesh reconstruction of non-rigid hands (although in a simplistic scenario), is feasible.

**Network Runtime.** Table 5.5 list out run-time of various methods. Comparing this with HMNet with HMNetOracle, it is evident that a major part of HMNet's complexity arises from the multi-human pixel-wise class prediction, which runs at around 30 FPS for an image of size 224x224. [19] is an accurate real-time body part segmentation network which runs at 120 FPS, and can be incorporated into our system to produce accurate, real-time reconstructions.

| Method | Output | FPS |
|---|---|---|
| SMPLify [13] | P | 0.01 |
| SMPLify, 91 kps [46] | | 0.008 |
| Decision Forests [46] | | 7.69 |
| HMR [41] | P | 25 |
| Pavlakos [63] | | 20 |
| Direct Prediction [46] | | 2.65 |
| Baseline | S | 175.4 |
| HMNet | | **28.01** |
| HMNetOracle | | **173.17** |
| Fusion4D [24] | S | 31 |

Table 5.5: Overview of the run time (in Frames Per Second, FPS) of various algorithms. Numbers have been picked up from the respective papers. All methods have used 1080Ti or equivalent GPU.

**Limitations and Future Work.** Since we do not enforce any volume consistency, skewing/thinning artifacts (Figure 5.11)might be introduced in our meshes. We would like to account for these in a non-handcrafted anthropomorphically valid way by either learning the SMPL parameters on top of it using an MLP similar to [45] or by using a GAN to penalize fake/invalid human meshes. Further, we have made use of the mesh topology in two ways in this work - (a) implicitly, to make the learning easier, and (b) for smoothing. Going ahead, we would like to make use of the mesh topology and geometry details is a more explicit manner, by using intrinsic mesh/surface properties. We believe that this is a largely unexplored space and applying such a regularization can result in better exploitation of surface geometry for reconstruction.

Figure 5.11: **Failure Cases of Our Method.**

## 5.4 Conclusion

We proposed a multi-branch multi-task HumanMeshNet network that simultaneously regresses to the template mesh vertices as well as body joint locations from a single monocular image. The proposed method achieves comparable performance with significantly lower modeling and computational complexity on three publicly available datasets. We also show the generalizability of the proposed architecture for a similar task of predicting the mesh of the hand. Looking forward, we would like to exploit intrinsic mesh properties to recover a more accurate surface reconstruction.

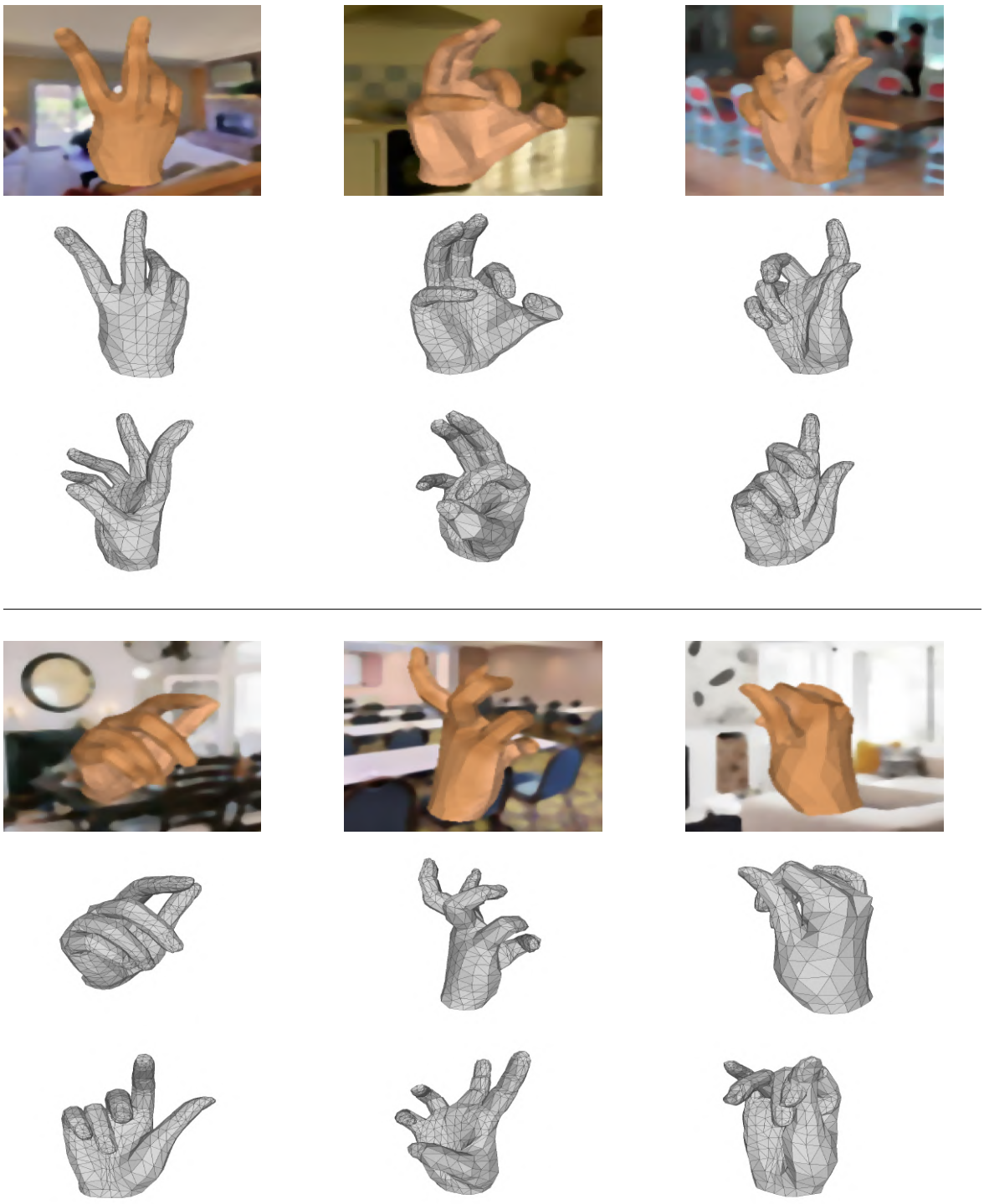


Figure 5.12: **Reconstruction Results on our Hand Mesh**. The first row denotes the input RGB image, the second the recovered mesh aligned to the same view and the final row aligned to an arbitary view.

# Chapter *6*

# Conclusion and Open Problems

Monocular 3D Human Reconstruction is a severely ill-posed problem due to self-occlusions caused by complex body poses and shapes, clothing obstructions, lack of surface texture, background clutter, single view, etc. Yet, the human sensory system can perceive, interpret, and predict unseen parts of know objects with ease. In this thesis, we attempt to empower machines with the capability to interpret 3D human poses and shapes from a single image in a manner that is non-intrusive, inexpensive, and has wide-scale applicability. In doing so, we progressively explore different 3D representations that are capable of producing accurate surface geometry, aimed at the long-term goal of recovering personalized 3D models.

Our first effort into learning a surface representation comes with VolumeNet, which predicts a voxel-grid from a monocular image. This was the first of its kind model for non-rigid human shapes at that time. To circumvent the ill-posed nature of this problem (aggravated by an unbounded 3D representation), we follow the ideology of providing maximal training priors with our unique training paradigms, to enable testing with minimal information. Specifically, co-learning of RGB and Depth, with randomized multi-view information while training enables superior reconstruction at test time from monocular RGB. Further, as we did not impose any body-model based constraint, we were able to recover deformations induced by free-form clothing. We extend VolumeNet to PoShNet by decoupling Pose and Shape, in which we learn the volumetric pose first, and use it as a prior for learning the volumetric shape. Although this proves to be far more accurate in recovering geometric details, the instability and complexity in learning 2 volumetric networks coupled with poor feed-forward time proved to be a disadvantage.

Although volumetric regression enables recovering a more accurate surface reconstruction, it does so without an animatable skeleton [87]. Such methods yield reconstructions of low resolution at higher computational cost (regression over the cubic voxel grid) and often suffer from an inconsistent topology via broken or partial body parts. Therefore, we explored learning parametric body models. Learning such models hasn't been straightforward due to the complex non-linear image to relative axis-angle mapping. It has resulted in complicated indirect solutions that rely on different projections of the underlying mesh (2D/3D keypoints, silhouettes, etc.). Therefore, to simplify the learning process, we proposed the CR framework which uses classification as a prior t the regression model. Although the CR framework

requires more exploration, it is evident that learning such templates can be very challenging without additional supervision. Moreover, recovering personalized models with high-resolution meshes isn't a direct possibility is this space.

Directly regressing to point cloud or mesh data from image(s) is a severely ill-posed problem and there are very few attempts in deep learning literature in this direction [54, 97]. With regard to point cloud regression, most of the attempts are focused on rigid objects, where learning is done in a class-specific manner. Apart from a very recent work [45], learning a mesh also hasn't been explored much for reconstruction, primarily because of the lack of deep learning constructs to do so.

As an alternative to directly learning parametric models, we focused on learning an implicitly structured point cloud (i.e., built from a statistical body model via a skinning function). In HumanMeshNet we predicted an implicitly structured point-cloud with the intuition that since each of the surface vertices is a sparse linear combination of the transformations induced by the underlying joints, learning this structured point-cloud becomes easier. We used the mesh topology as an implicit prior to easing the learning process for this highly ill-posed problem. This proposed paradigm can theoretically learn local surface deformations that body-model based PCA space can't capture. Further, learning high-resolution point-clouds/meshes are a natural extension in this space. The simplicity of the model makes it efficient in terms of network size as well as feed-forward time yielding significantly high frame-rate reconstructions, while simultaneously achieving comparable accuracy in terms of surface and joint error, as shown on three publicly available datasets.

Though there has been significant progress in this field in the past few years, there are still several important questions to answer and interesting solutions await us. A few directions to look out for -

- Extension to video-based 3D Human reconstruction is a natural step ahead. Human motion can act as a prior and frame-level reconstructions can be refined with trajectory information. Enforcing temporal consistency in such scenarios can turn out to be the key.

- Although the 3D shape and pose of hands, face, body, and clothing are learned separately today, it is evident that having a unified model that does all of the above in an end-to-end manner will be the gold-standard that several will aim at.

- Producing high resolution meshes with soft-tissue deformations from monocular images with deep-learning-based solutions.

- Reconstructing 3D models during human-human and human-object interactions will take us a step closer towards 3D scene understanding and manipulation, a very promising area with several applications in robotics.

- Multi-modal learning for reconstructing humans in dynamic outdoor scenes (such as skiing, biking, etc) or during everyday activities (such as praying, eating, etc). Due to limitations of optical-based systems for tracking in such scenarios, solutions will have to be built that make use of multi-sensory data.

- Reinforcement Learning based solutions for 3D Human Reconstruction, mimicking the natural human learning process.
- 3D Visual Question Answering, to enable robot/agent navigation.
- System Identification - Modeling and reconstructing real-world materials, textures, etc. to learn physical properties and how an agent can interact with them.

# Related Publications

- **Venkat, Abbhinav**, Sai Sagar Jinka, and Avinash Sharma. "Deep Textured 3D Reconstruction of Human Bodies." In British Machine Vision Conference (BMVC). 2018.
- **Venkat, Abbhinav**, Chaitanya Patel, Yudhik Agrawal, and Avinash Sharma. "HumanMeshNet: Polygonal Mesh Recovery of Humans." In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV-W). 2019.

## Other Publications

- Battan, Neeraj, **Abbhinav Venkat**, and Avinash Sharma. "DeepHuMS: Deep Human Motion Signature for 3D Skeletal Sequences." In Asian Conference on Pattern Recognition (ACPR), pp. 281-294. Springer, Cham, 2019.

## Fellowships

- Qualcomm's Innovation Fellowship (QIF), 2018-19 for proposal on "Textured 3D Reconstruction in a calibration-free environment".

# Bibliography


[1] 3d structured lighting scanner from artec 3d. https://www.artec3d.com.

[2] Qualitative result of the the dancing girl in volumenet. https://cvit.iiit.ac.in/research/projects/cvit-projects/3dcomputervision.

[3] Supplementary results of humanmeshnet. https://www.youtube.com/watch?v=zPMoNVfhlRs.

[4] Supplementary results of volumenet. https://youtu.be/7Z27I9XAuXI.

[5] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019.

[6] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. *arXiv preprint arXiv:1803.04758*, 2018.

[7] R. Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[8] S. Andrews, I. Huerta, T. Komura, L. Sigal, and K. Mitchell. Real-time physics-based motion capture with sparse sensors. In *Proceedings of the 13th European conference on visual media production (CVMP 2016)*, pages 1–10, 2016.

[9] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Transaction on Graphics*, 24:408–416, 2005.

[10] C. Barron and I. A. Kakadiaris. Estimating anthropometry and pose from a single image. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 669–676. IEEE, 2000.

[11] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5420–5430, 2019.

[12] 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 (ECCV)*, 2016.

[13] 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 (ECCV)*, 2016.

[14] 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*, pages 561–578. Springer, 2016.

[15] F. Bogo, J. Romero, M. Loper, and M. J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ, USA, June 2014. IEEE.

[16] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] A. Boukhayma, R. de Bem, and P. H. Torr. 3d hand shape and pose from images in the wild. *arXiv preprint arXiv:1902.03451*, 2019.

[18] C. Canton-Ferrer, J. R. Casas, and M. Pardas. Marker-based human motion capture in multiview sequences. *EURASIP Journal on Advances in Signal Processing*, 2010(1):105476, 2010.

[19] J. J. Charles, I. Budvytis, and R. Cipolla. Real-time factored convnets: Extracting the x factor in human parsing. 2018.

[20] J.-H. Cho, S.-Y. Kim, Y.-S. Ho, and K. H. Lee. Dynamic 3d human actor generation method using a time-of-flight depth camera. *IEEE Transactions on Consumer Electronics*, 54(4):1514–1521, 2008.

[21] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *European Conference on Computer Vision (ECCV)*, pages 628–644, 2016.

[22] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015.

[23] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *2016 fourth international conference on 3D vision (3DV)*, pages 108–117. IEEE, 2016.

[24] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transaction on Graphics*, 35(4):114:1–114:13, July 2016.

[25] J. G. D. França, M. A. Gazziro, A. N. Ide, and J. H. Saito. A 3d scanning system based on laser triangulation and variable field of view. In *IEEE International Conference on Image Processing 2005*, volume 1, pages I–425. IEEE, 2005.

[26] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Proceedings cvpr ieee computer society conference on computer vision and pattern recognition*, pages 73–80. IEEE, 1996.

[27] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[28] S. Ge and G. Fan. Non-rigid articulated point set registration with local structure preservation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–133, 2015.

[29] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019.

[30] V. Golyanik, G. Reis, B. Taetz, and D. Strieker. A framework for an accurate point cloud based registration of full 3d human body scans. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 67–72. IEEE, 2017.

[31] V. Golyanik, B. Taetz, G. Reis, and D. Stricker. Extended coherent point drift algorithm with correspondence priors and optimal subsampling. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[32] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, Mar 1975.

[33] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018.

[34] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009.

[35] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3083–3091, 2015.

[36] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[37] L. Herda, P. Fua, R. Plankers, R. Boulic, and D. Thalmann. Skeleton-based motion capture for robust reconstruction of human motion. In *Proceedings Computer Animation 2000*, pages 77–83. IEEE, 2000.

[38] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2014.

[39] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039, 2017.

[40] D. Joseph Tan, T. Cashman, J. Taylor, A. Fitzgibbon, D. Tarlow, S. Khamis, S. Izadi, and J. Shotton. Fits like a glove: Rapid and reliable hand shape personalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2016.

[41] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose.

[42] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[43] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):29, 2013.

[44] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[45] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

[46] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations.

[47] H. Liu, X. Wei, J. Chai, I. Ha, and T. Rhee. Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games*, pages 133–140, 2011.

[48] K. Liu, Y. Wang, D. L. Lau, Q. Hao, and L. G. Hassebrook. Dual-frequency pattern scheme for high-speed 3-d shape measurement. *Optics express*, 18(5):5229–5244, 2010.

[49] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2720–2735, 2013.

[50] M. Loper, N. Mahmood, and M. J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014.

[51] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[52] M. Madadi, H. Bertiche, and S. Escalera. Smplr: Deep smpl reverse for 3d human pose and shape recovery. *arXiv preprint arXiv:1812.10766*, 2018.

[53] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Heloir, and D. Stricker. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. *arXiv preprint arXiv:1808.09208*, 2018.

[54] P. Mandikal, N. K. L., M. Agarwal, and V. B. Radhakrishnan. 3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. In *British Machine Vision Conference*, page 55, 2018.

[55] R. A. Morano, C. Ozturk, R. Conn, S. Dubin, S. Zietz, and J. Nissano. Structured light using pseudorandom codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):322–327, 1998.

[56] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011.

[57] T.-N. Nguyen, H.-H. Huynh, and J. Meunier. 3d reconstruction with time-of-flight depth camera and multiple mirrors. *IEEE Access*, 6:38106–38114, 2018.

[58] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011.

[59] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. *arXiv preprint arXiv:1808.05942*, 2018.

[60] S. I. Park and J. K. Hodgins. Capturing and animating skin deformation in human motion. *ACM Transactions on Graphics (TOG)*, 25(3):881–889, 2006.

[61] S. I. Park and J. K. Hodgins. Data-driven modeling of skin and muscle deformation. In *ACM SIGGRAPH 2008 papers*, pages 1–6. 2008.

[62] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[63] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *arXiv preprint arXiv:1805.04092*, 2018.

[64] J. Piao, C. Qian, and H. Li. Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9398–9407, 2019.

[65] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taixe, M. Mueller, H.-P. Seidel, and B. Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *2011 International Conference on Computer Vision*, pages 1243–1250. IEEE, 2011.

[66] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[67] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*, pages 573–586. Springer, 2012.

[68] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming. Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Technical report, SYTRONICS INC DAYTON OH, 2002.

[69] D. Roetenberg, H. Luinge, and P. Slycke. Moven: Full 6dof human motion tracking using miniature inertial sensors. *Xsen Technologies, December*, 2(3):8, 2007.

[70] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245, 2017.

[71] L. A. Schwarz, D. Mateus, and N. Navab. Discriminative human full-body pose estimation from wearable inertial sensor data. In *3D physiological human workshop*, pages 159–172. Springer, 2009.

[72] A. Sharma, R. Horaud, J. Cech, and E. Boyer. Topologically-robust 3d shape matching based on diffusion geometry and seed growing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2481–2488, 2011.

[73] R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

[74] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in neural information processing systems*, pages 1337–1344, 2008.

[75] A. Sinha, J. Bai, and K. Ramani. Deep learning 3d shape surfaces using geometry images. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *European Conference on Computer Vision (ECCV)*, pages 223–240, 2016.

[76] A. Sinha, A. Unmesh, Q. Huang, and K. Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. *CoRR*, abs/1703.04079, 2017.

[77] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184. ACM, 2004.

[78] K. H. Strobl, E. Mair, and G. Hirzinger. Image-based pose estimation for 3-d modeling in rapid, hand-held motion. In *2011 IEEE International Conference on Robotics and Automation*, pages 2593–2600. IEEE, 2011.

[79] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *Proceedings of the IEEE international conference on computer vision*, pages 1868–1876, 2015.

[80] V. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction.

[81] A. Tkach, M. Pauly, and A. Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (TOG)*, 35(6):222, 2016.

[82] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Computer Vision and Pattern Recognition (CVPR)*, pages 209–217, 2017.

[83] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017.

[84] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. *arXiv preprint arXiv:1804.04875*, 2018.

[85] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *CVPR*, 2017.

[86] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[87] A. Venkat, S. S. Jinka, and A. Sharma. Deep textured 3d reconstruction of human bodies. *arXiv preprint arXiv:1809.06547*, 2018.

[88] A. Venkat, C. Patel, Y. Agrawal, and A. Sharma. Humanmeshnet: Polygonal mesh recovery of humans. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[89] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popović. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)*, 26(3):35–es, 2007.

[90] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH*, pages 97:1–97:9, 2008.

[91] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, volume 36, pages 349–360. Wiley Online Library, 2017.

[92] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

[93] T. Weber, R. Hänsch, and O. Hellwich. Automatic registration of unordered point clouds acquired by kinect sensors using an overlap heuristic. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102:96–109, 2015.

[94] O. Wiles and A. Zisserman. Silnet : Single- and multi-view reconstruction by learning from silhouettes. In *British Machine Vision Conference*, 2017.

[95] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Neural Information Processing Systems (NIPS)*, pages 82–90, 2016.

[96] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015.

[97] Y. Xia, Y. Zhang, D. Zhou, X. Huang, C. Wang, and R. Yang. Realpoint3d: Point cloud generation from a single image with complex background. *CoRR*, abs/1809.02743, 2018.

[98] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *arXiv preprint arXiv:1812.01598*, 2018.

[99] D. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[100] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Neural Information Processing Systems (NIPS)*, pages 1696–1704, 2016.

[101] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Yong Chang, K. Mu Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[102] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, volume 2, page 3, 2017.

[103] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4447–4455, 2015.

[104] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33(4):1–12, 2014.

[105] M. Zollhofer, M. Niessner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transaction on Graphics*, 33(4):156:1–156:12, 2014.

[106] S. Zuffi, A. Kanazawa, and M. J. Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3955–3963, 2018.

[107] S. Zuffi, A. Kanazawa, D. W. Jacobs, and M. J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017.