Estimating 3D Human Pose, Shape, and Correspondences from Monocular Input

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

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

by

Amogh Tiwari
2018111003
amogh.tiwari@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500032, INDIA
June, 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled ***"3D Human Pose, Shape and Correspondence Estimation from Monocular Input"*** by Amogh Tiwari, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Avinash Sharma

To all the known and the unknown people who have contributed to providing me with the extraordinary life that I have today.

Acknowledgments

I would like to start by thanking my advisor, Dr. Avinash Sharma, who is not only a great advisor but also a remarkable human being. My journey from being a complete beginner in research to becoming a 'somewhat' decent researcher would not have been possible without his constant support, guidance, and encouragement. The freedom and support he provides to all his students to explore different research problems are truly commendable. He is immensely supportive and patient whenever we encounter roadblocks, always ensuring that we remain confident. Beyond his technical guidance, he also deeply cares for his students on a personal level. He constantly checks in on us, inquires about our personal goals, and does his best to help us achieve them in any way he can. Having him as my advisor has taught me a great deal, both professionally and personally, and I am confident that many of these lessons will benefit me for a long time, both in my research and in my life.

I would also like to thank my collaborators from different projects, who provided me the opportunity to work with and learn from them. In particular, I would like to thank Dr. Vineet Gandhi, Dr. Mahdava Krishna, Kanishk, and Varun, with whom I collaborated on a project on visual grounding, leading to my first accepted publication. I also extend my gratitude to my collaborators at TCS: Sushovan, Dr. Lokender Tiwari, Dr. Brojeshwar Bhowmick, and Hrishav. I would like to give a special shoutout to Kanishk and Sushovan, who have been exceptional teammates, often taking on significant responsibilities and doing the 'heavy lifting' in the projects whenever required.

Next, I would like to thank my family, who have positively influenced me in a myriad of ways. While many close family members have contributed significantly to who I am today, a few have played huge roles. Firstly, my parents, who have worked extremely hard and sacrificed a lot to provide me with an environment where I could pursue my goals without any other having to worry about anything else. They also ensured that I have some of the best role models available in-house! When I think of the most hardworking people I know, my mother, Dr. Chetna, is one of the first that comes to mind. She is also one of the most caring individuals I know, always going out of her way to help anyone in need, even those she barely knows. She raised me single-handedly when my father was away on his military duties and has made significant decisions, like relocating cities at a moment's notice (sometimes at a great personal loss), just to ensure my comfort and benefit. Much of my success today would not be possible without her constant effort and sacrifice. My father, Dr. Kailash, has likely influenced me in more ways than either of us realizes. While he never explicitly asked me to follow him, and neither did I try to consciously copy him, many of my 'unique' traits stem from him. My interest in fitness and

sports, my (relatively) healthy diet choices, my organization skills, my independent living style, a strong inclination towards honesty, and perhaps even my interest in research have all been shaped significantly by his actions and example. In fact, if some of my friends are to be believed, our speaking styles are also quite similar! Apart from my parents, I would like to explicitly thank my *Nani* (maternal grandmother) and *Mama* (maternal uncle) for taking care of me from my childhood to today. I also want to give a shoutout to my pet dogs, Bunny and Brad, who are like my younger siblings and have been a great source of love and happiness in my life.

Finding intelligent and hardworking colleagues is hard. Finding nice and helpful colleagues who are also great friends is harder. Finding colleagues who are not just intelligent and hardworking but also great friends willing to always help you out is a dream. While at CVIT, this dream came true for me. All my colleagues at the lab - Aakash, Anjali (not exactly a labmate, but close enough), Aryamaan, Astitva, Dhawal, Abhinaba, Adhiraj, Aparna, Ayan, BVK, Chandradeep, Chirag (both Jain and Parikh!), Deepti, Ishaan, Nakul, Naren, Neha, Neil, Shanthika, Pranav, Rahul, Ritam, and Vishal - are all some of the most hardworking and ambitious people I have met. Yet, they are also some of the most helpful and fun people to be around. Despite everyone's extremely busy schedules, you can always rely on them to help debug a codebase, discuss a research problem, or simply fool around at the canteens by discussing an out-of-the-world theory and acting as if it's the most normal thing in the world.

In particular, I would like to thank Chandradeep (who, unlike Astitva, is from the same state as me), for helping me with many 'simple' queries related to my projects and for inspiring me to be more disciplined and muscular! Shanthika, who has also helped me with many 'simple' doubts, especially when I was a beginner, and has been a great example of perseverance in the face of adversity. She has also shown me that one can be a great researcher without sacrificing sleep! Ishaan, who, among many other things, has been our tech support guy. Rahul, who, has been my partner-in-crime in running (along with Deepti, Chirag J and P, and occasionally Shanthika) as well as in my 'preachings' about fitness. Pranav, Deepti, and Aparna, who have made our late-night JC conversations extremely fun by being good sports about all the (many times unfair, though not always...) taunts and insults they receive. Nakul and Vishal, for our passionate discussions on sports. Naren, for our philosophical discussions. BVK, for the scooty rides and shoulder massages! And finally, Aakash and Neil, who have been great examples of maintaining a good work-life balance, have provided a sane and mature voice among a bunch of clownish individuals, and also being the group's cab drivers whenever necessary! These people have transformed what was supposed to be a challenging experience into not just a manageable but an extremely enjoyable experience. Moving forward, their company is undoubtedly the aspect that I will miss the most about not being at CVIT.

One person from the lab who deserves a special mention, is Astitva. Astitva was my first 'proper' (student) mentor at CVIT, and some of the works in this thesis wouldn't have been possible without his help. He has taught me a great deal about research and has always been available to help me with any doubts. No matter what doubt you have (because he knows pretty much everything about everything!), at what time of the day (because he hardly sleeps!), you can always rely on Astitva to help you resolve

your query. It is beyond my comprehension how someone can work so hard, sleep so little (though I definitely do not encourage this!), handle so many different projects together, and still be so helpful and calm and composed. I have never seen him lose his patience even when junior students (including me in my early days) (or sometimes even senior students!) ask 'silly' questions repeatedly or make serious mistakes. He has made an invaluable contribution (and continues to do so) to my research journey and to those of many others in the lab. I wish I would have met him earlier!

I would also like to thank my 'basketball gang,' with a special shoutout to Aadarsh, Janardan, and Revanth. During my time in college, I have had two ankle fractures, one thumb fracture, one finger that wouldn't fully straighten due to an injury, and god knows how many other 'jammed' fingers, all from playing basketball (I am quite injury resilient outside the court !). Despite all these injuries, I have constantly gone back to playing basketball, and a large reason for this has been these people. Playing basketball has not only helped me maintain my physical fitness (Contrary to popular opinion, physical fitness is among the last thing in the minds of most serious athletes. Our primary motivation is to just play. Fitness is a by-product.), but it has also helped me maintain a good daily routine, have more discipline, and have loads of fun. Playing basketball allowed me to disconnect from all the stresses of research and helped me clear my mind so that I could go back to research with renewed enthusiasm. I will miss all the fun moments we have had on (and outside) the court and will forever rue the fact that I couldn't be the dominating center they deserved :(. Also, Jana - you ARE short.

Another special person in my life who deserves a big shoutout is Aastha. While (unfortunately) she didn't get to directly be a part of this thesis (how good would it have been to have her in IIIT, though!), she has still had an invaluable role in my life and, consequently, in this work. Very few people have the patience to listen to my long monologues about fitness, and everything else that is wrong with the world. But she is someone who not just patiently listens to them (in most cases) but also comes up with follow-up questions. Imagine you come from a non-engineering background and are forcibly given a lesson on perspective projection on your first date. What are the odds that you would go back and come up with a bunch of follow up questions later? Well, she did. She has been extremely supportive and understanding when I couldn't give her sufficient time due to my research commitments (or sometimes, just poor time management). Her presence and opinions have encouraged me to 'slow down' a little and appreciate the more abstract things in life, like love! Our 6:30 PM calls, where we discussed our daily lives, ranted about different things, and many times, just had a constant back-and-forth of "meri life to boring hai, tu kuch bata na ... ", have been a constant source of joy for me. I wish one day I can use my research experience to make a device that would allow her to read my thoughts and make her realize that I appreciate her presence much more than what I can ever express or she can ever fathom.

Finally, I would like to thank many other people I could not explicitly name above but who have played an invaluable role in my development as an individual. Without their help at different stages of my life, I wouldn't be what I am today, and this thesis wouldn't be what it is now. This list includes a lot of people like my school teachers, my friends - from primary school to now (I have always been incredibly lucky when it comes to getting a good friend group), my cousins, a few other close relatives,

and many more. I would also like to extend my thanks to the many unseen and unsung heroes who work tirelessly to ensure the smooth functioning of our institutions and allow people like me to focus on tasks like research without having to worry about mundane things. In this list, I would like to particularly give a shoutout to the armed forces members and their families. These people silently sacrifice much more than any of us can ever imagine or do, just to ensure a safer life for us.

I would like to once again thank all of the above people for their unwavering support and encouragement from the depth of my heart. I hope to repay their kindness by providing them with similar support in the future and also pay forward their kindness to my other juniors, colleagues, and family members.

Abstract

In recent years, advances in computer vision have opened up multiple applications in virtual reality, healthcare, robotics, and many other domains. One crucial problem domain in computer vision, which has been a key research focus lately, is estimating the 3D human pose, shape, and correspondences from monocular input. This problem domain has applications in various industries like fashion, entertainment, healthcare, etc. However, it is also highly challenging due to various reasons like large variations in the pose, shape, and appearance of humans and clothing details, external and self-occlusions, challenges with ensuring consistency etc.

As part of this thesis, we tackle two key problems related to 3D human pose, shape, and correspondence estimation. First, we focus on the problem of temporally consistent 3D human pose and shape estimation from monocular videos. Next, we focus on dense correspondence estimation across images of different (or the same) humans. We show that despite receiving a lot of research attention lately, existing methods for these tasks still perform sub-optimally in many challenging scenarios and have significant scope for improvement. We aim to overcome some of the limitations of existing methods and advance state-of-the-art (SOTA) solutions to these problems.

First, we propose a novel method for temporally consistent 3D human pose and shape estimation from a monocular video. Instead of using the traditionally used, generic ResNet-like features, our method uses a body-aware feature representation and an independent per-frame pose and camera initialization over a temporal window followed by a novel spatio-temporal feature aggregation by using a combination of self-similarity and self-attention over the body-aware features and the per-frame initialization. Together, they yield enhanced spatio-temporal context for every frame by considering the remaining past and future frames. These features are used to predict the pose and shape parameters of the human body model, which are further refined using an LSTM.

Next, we expand our focus to the task of dense correspondence estimation between humans, which requires understanding the relations between different body regions (represented using dense correspondences), including the clothing details, of the same or different human(s). We present Continuous Volumetric Embeddings (ConVol-E), a novel robust representation for dense correspondence-matching across RGB images of different human subjects in arbitrary poses and appearances under non-rigid deformation scenarios. Unlike existing representations, ConVol-E captures the deviation from the underlying parametric body model by choosing suitable anchor/key points on the underlying parametric body surface and then representing any point in the volume based on its Euclidean relationship with

the anchor points. This allows us to represent any arbitrary point around the parametric body (clothing details, hair, etc.) by an embedding vector. Subsequently, given a monocular RGB image of a person, we learn to predict per-pixel ConVol-E embedding, which carries a similar meaning across different subjects and is invariant to pose and appearance, thereby acting as a descriptor to establish robust, dense correspondences across different images of humans.

We thoroughly evaluate our methods on publicly available benchmark datasets and show that our methods outperform existing SOTA. Finally, we provide a summary of our contributions and discuss the potential future research directions in this problem domain. We believe that this thesis improves the research landscape for the domain of the human body, pose, shape, and correspondence estimation and helps accelerate progress in this direction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, advances in computer vision have opened up new avenues for applications in virtual reality, healthcare, robotics, and many other domains. One key problem domain in this area is estimating 3D human pose, shape, and correspondences from monocular input. This problem domain has applications in various industries like fashion, entertainment, healthcare, etc. However, it is also a highly challenging problem for various reasons, like large variations in the pose, shape, and appearance of humans and clothing details, external and self-occlusions, challenges with ensuring consistency across, etc. In this chapter, we discuss our motivation for exploring this problem domain and describe our specific problem statements and the related research challenges. Subsequently, we briefly discuss the existing literature and its limitations, followed by providing an outline of our major contributions.

1.1 Motivation

Vision is the most dominant human sense by a large margin. According to a study [53], it is responsible for up to 80% of our perception, cognition, and learning abilities. Consequently, significant research efforts have focused on mimicking (and potentially outperforming) human visual capabilities using machines. While this field of study - Computer Vision - has made remarkable strides in recent years, many important problems remain to be solved. One such problem domain, which still has significant scope for improvement, is estimating human body pose, shape, and the correspondences across humans from a given visual input (videos or images).

As illustrated in Figure 1.1, estimating human pose, shape, and the correspondences between humans is an important problem with diverse applications across various fields. For instance, estimating human shape aids in creating personalized virtual avatars, while pose estimation helps in accurately posing these avatars. This is useful in AR/VR, gaming, and the movie industries, where the need to animate virtual characters for specific actions is common. Pose estimation is also beneficial for sports analytics and healthcare for analyzing and identifying any potential issues in the movement patterns of a player or a patient. Another application for pose estimation is in robotics, where understanding a person's motion and predicting their future motion is helpful for tasks like collision avoidance.

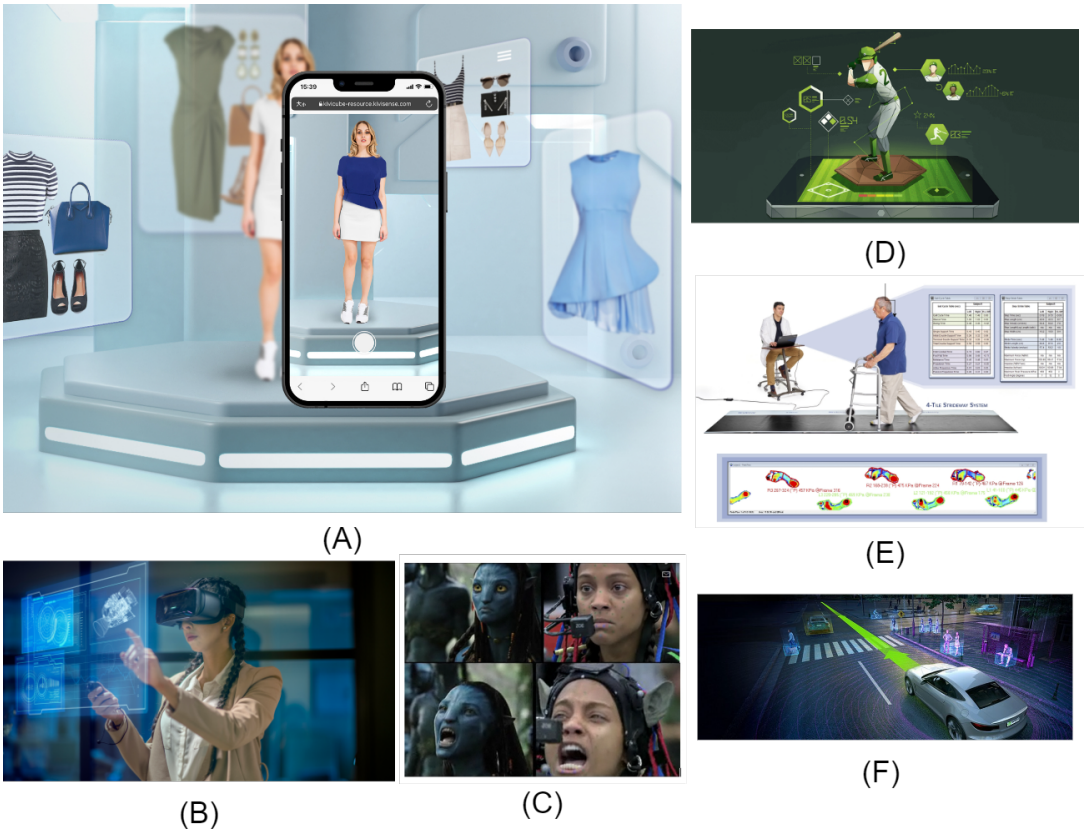


Figure 1.1 Examples of diverse domains where estimating human pose, shape and correspondences is beneficial. (A) A virtual cloth try-on platform. (B) An augmented reality platform. (C) Motion transfer example. (D) A sports analytics platform. (E) Healthcare application - Gait analysis. (F) Robotics application - Avoiding collisions with humans. (Image Credits: (A)-Kivisense; (B)-UK Gov Blog; (C)-AvatarBlog @ Typepad; (D)-AIWS; (E)-Tekscan; (F)-IEEE Spectrum).

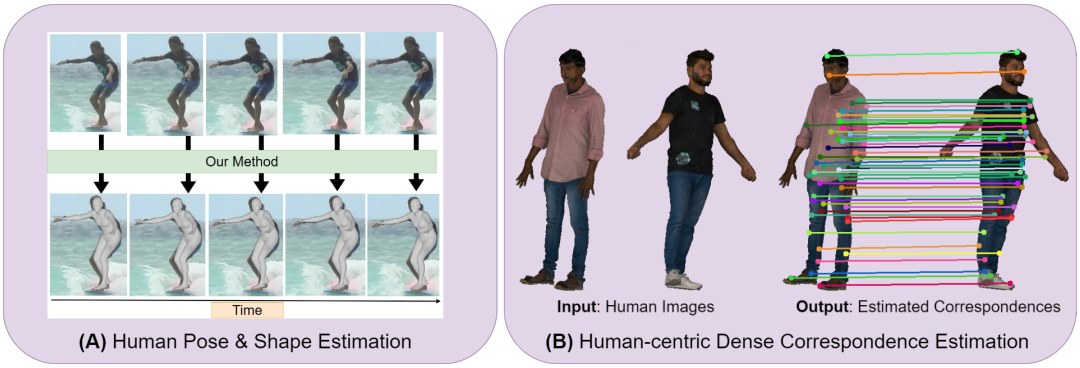


Figure 1.2 Our problem statements. (A) depicts the problem of temporally consistent 3D human pose and shape estimation. The input is a RGB sequence, and the output is the estimated parametric body model, visualized by overlaying it on the RGB frames. (B) depicts the problem of dense human-centric correspondence estimation. The input is images of humans, and the output is the estimated dense correspondences.

In addition to estimating the pose and shape of humans, certain problems also require understanding the relations between different body regions, including the clothing details, of the same or different human(s). For instance, 3D reconstruction requires feature matching across images. Similarly, tasks like virtual try-on and animation/motion transfer, where we want to transfer attributes (such as clothing details or motion) from a source human to a target can also benefit from understanding the correspondences between different body regions. Thus, in addition to estimating the pose and shape, we also focus on the task of dense correspondence estimation across humans.

1.2 Problem Statement

We focus on two important problems in the domain of human pose, shape, and correspondence estimation from monocular input, which are described below.

- **Temporally Consistent 3D Human Pose and Shape Estimation from Monocular Videos:** Given a monocular input video, our aim in this problem is to estimate the underlying human body pose and shape dynamics by fitting a parametric body model to the individual frames in a temporally consistent fashion. This is illustrated in Figure 1.2-(A).
- **Dense Human-Centric Correspondence Estimation:** In this problem, given two or more images containing the same or different human(s), we aim to establish dense (pixel-to-pixel) 2D correspondences across the humans, including their clothing details, present in those images. This must be done along with dealing with variations in identity, background information, appearance, and occlusions by garments - especially loose garments. This is illustrated in Figure 1.2-(B).

1.3 Research Challenges

Both of the aforementioned problem statements present a wide range of research challenges, which are illustrated in Figure 1.3, and described below:

- **Large variations in pose, shape, and appearance:** Both humans and clothing details can exhibit significant variations in their pose, shape, and appearance (see Figure 1.3-(A, E)), making it difficult to capture these diverse characteristics accurately.

- **Occlusions:** Occlusions (Figure 1.3-(B)) caused by external objects in the scene (external occlusion) or by the human's own body parts like crossed limbs (self-occlusions) introduce visibility obstruction and ambiguity, thus complicating the estimation process.

- **Poor Lighting Conditions:** Diminished input quality due to poor lighting conditions (Figure 1.3-(C)) makes it challenging to localize humans (and other objects) in the scene or reason about them effectively.

- **Camera viewpoint variations:** Variations in camera viewpoints (Figure 1.3-(D)) add additional complexity, as certain viewpoints (like side-views or back-views) provide much lesser information than other views, forcing our methods to infer about the non-visible parts of the human body.

- **Clothing details:** Clothing, like human bodies, exhibits a wide range of complex poses, shapes, and appearances (Figure 1.3-(E)). Further, clothing details, particularly loose clothing, can make it difficult to localize the underlying human body, making the task of pose, shape, and correspondence estimation more challenging.

- **Ensuring Temporal Consistency**: Temporal data, such as videos, require maintaining temporal consistency across frames to ensure the plausibility of estimated results (refer to Figure 1.3). However, achieving temporal consistency in estimates is not straightforward due to the complex nature of human motion, which makes effective integration of information across frames challenging.

- **Inherent ambiguity of recovering 3D information from monocular 2D input:** Recovering 3D information from monocular images is inherently challenging due to the loss of depth information during imaging. For example, in Figure 1.3-(G), the elbow joint appears incorrectly positioned above the head despite being behind in 3D.

- **Limited reliable real-world data:** Obtaining reliable ground-truth data for human pose, shape, and correspondences is challenging. Establishing ground truth requires fitting a parametric model to raw 3D human scans using some fitting method (usually multi-view optimization-based). However, this restricts the accuracy of ground truth to the accuracy of the fitting method used. Figure 1.3-(G) illustrates an example of inaccurate annotations from a popular dataset, where the

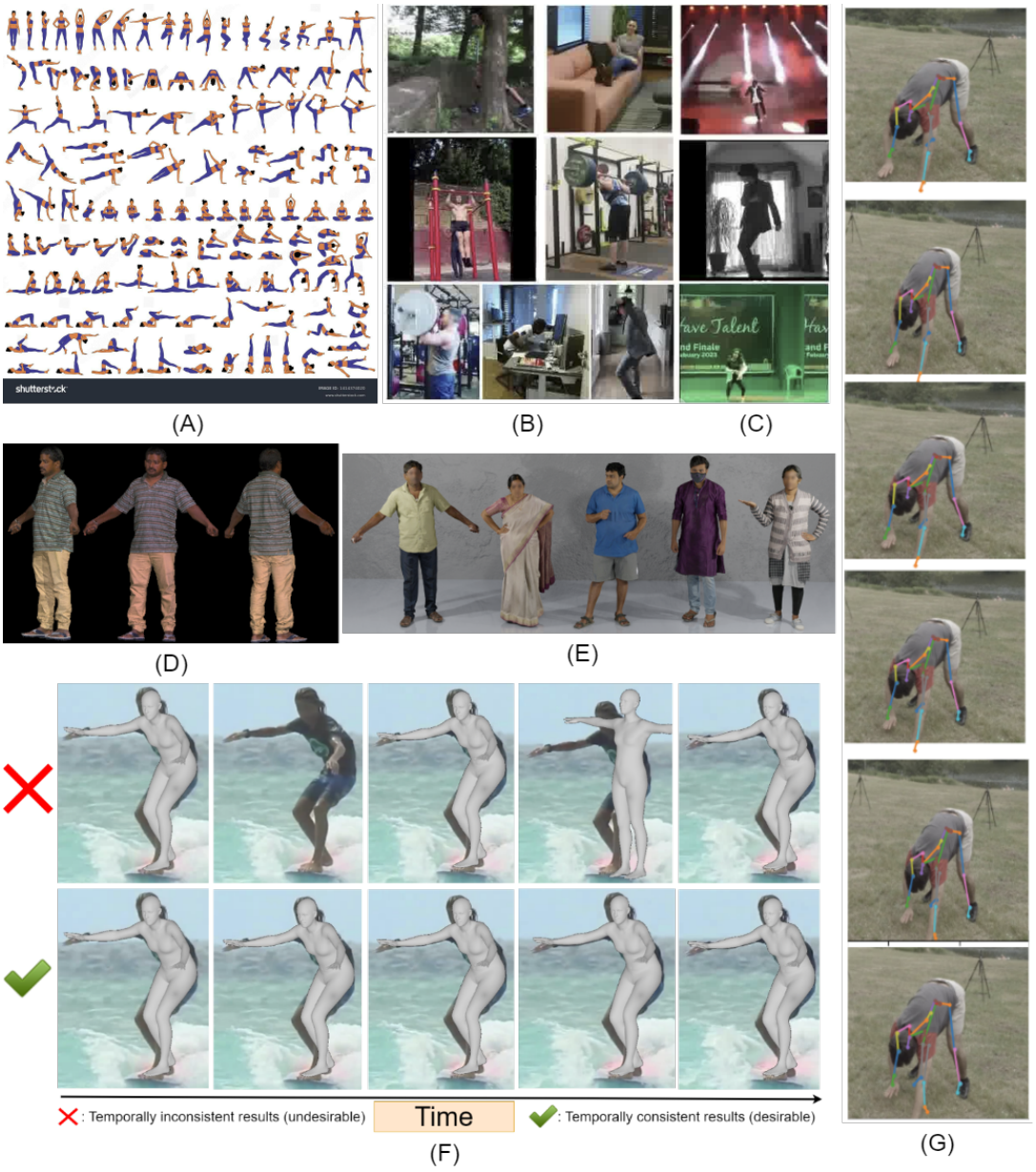
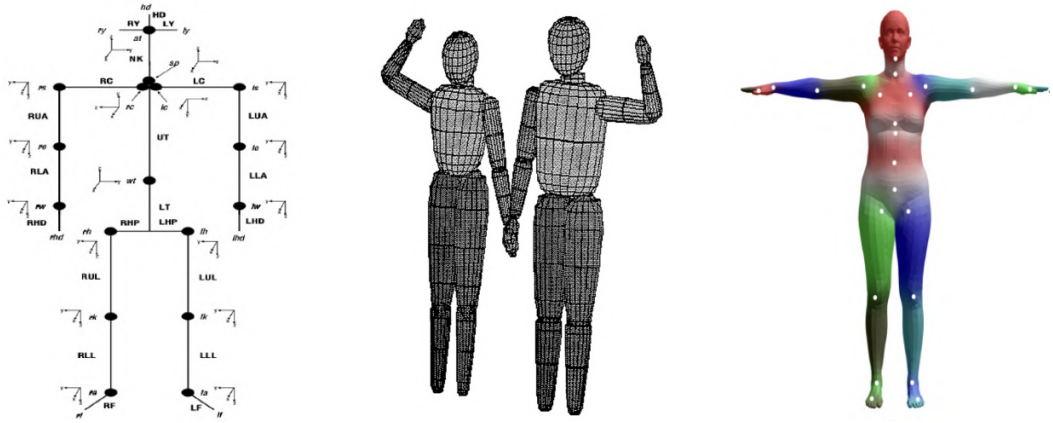


Figure 1.3 Various research challenges in our target problem domain. (A) Large pose, shape and appearance variation in humans. (B) Occlusions (external and self). (C) Poor lighting conditions. (D) Camera viewpoint variations. (E) Clothing variations. (F) Temporally consistent (and inconsistent) results example. (G) Ambiguity of recovering 3D information from monocular 2D input; And noisy ground truth annotations. (Image credits: (A)-Shutter Stock; (B,C)-YouTube Videos; (D,E)-3DHumans Dataset [29]; (F) - Our results on a YouTube video; (G) MPI-INF-3DHP Dataset [45]).



(a) (b) (c)

Figure 1.4 Evolution of human body representations over the years. (A) Skeleton based representation, e.g., [3] (B) Geometric primitive based representation, e.g., [12], (C) Statistical body model, e.g., SMPL [40]. (Image Credit: [61]).

annotation erroneously depicts the right elbow as bent towards the person's head, even though it is bent away.

1.4 Research Landscape

Humans can exhibit a wide variety of body poses and shapes. However, the human body has a well-defined structure, and thus can be represented using parametric body models. Figure 1.4 illustrates the evolution of human body models over the years. The earliest body models were stick-based models [12], where labeled landmarks were used to represent the joints, and the edges between those landmarks indicated the connectivity. These models evolved into geometric primitive-based models [12], where geometric primitives like cubes, cylinders, and spheres were used to provide a more human-like representation of the body. More recently, state-of-the-art parametric models rely on thousands of 3D scans of people to learn the shape and structural statistics of the human body and can provide a very realistic representation of humans [1, 40]. These parametric body models can model the human body in much lower dimensions than voxel-grid or point clouds. Since a lower-dimensional representation is easier to learn, problems related to reasoning about the human body, especially pose & shape estimation, can, thus, be modeled as a problem of estimating the pose and shape parameters of a parametric body model.

The optimal setting for recovering the 3D information about a human would involve a multi-view calibrated camera setup. However, constructing and managing such a setup is challenging and incurs significant costs. Dependence on such a setup would limit the applicability of related technologies to specialized labs capable of affording such intricate setups. Therefore, to ensure broader accessibility,

there is a need to develop methods that can work with simple commodity input sensors like a single mobile phone camera. Consequently, a significant fraction of research in this problem domain operates under the challenging constraint of a single-view (monocular) input.

Initial efforts in the problem domain [13, 30, 35, 49, 50] focused on the task of 3D human pose and shape estimation from monocular images. These methods typically employ a CNN-based image feature extractor, followed by an MLP-based regressor to regress the SMPL parameters. This may be followed by additional (learning or non-learning based) constraints to ensure physically plausible pose and shape estimates. However, many applications critically depend on the temporal consistency of the estimated human motion, where single-image-based methods give jittery and implausible results due to the lack of any temporal consistency constraints. Further, when used on videos, single image-based methods, by their very design, fail to effectively capture the temporal information available across frames and thus perform sub-optimally.

To overcome the above-mentioned limitations of image-based methods, specialized video-based motion estimation methods have been proposed [33, 7, 65, 57, 68]. These methods typically introduce a CNN module similar to image-based methods, followed by an RNN module to perform spatio-temporal feature aggregation from neighboring frames before estimating the human pose and shape. However, despite recent research efforts in this problem domain, many methods fail to capture long-term temporal dynamics and show poor performance when the body is under partial occlusion. Thus, there is a need to develop more robust video-based methods.

For the task of dense correspondence estimation, one approach would be to use one of the above (or similar) methods to establish dense correspondences between the human image and the parametric body model, followed by establishing correspondences between the humans in the parametric body model space. DensePose [17] and Continuous Surface Embeddings (CSE) [48] follow a similiar approach. However, since parametric body models are designed to model only the underlying human body, this approach will not account for points lying away from the underlying human body, like clothing or hair details. BodyMap [24] proposes to overcome this limitation by doing a geodesic distance-based extrapolation of the CSE embeddings [48] in the SMPL's UV space. However, such extrapolation is insufficient to prevent distant vertices from having similar embedding values, resulting in false matching across different regions of the human body. Additionally, BodyMap's approach does not guarantee consistent pixel-wise embeddings for loose clothing scenarios, as the effect of geodesic distance will diminish in the far-apart regions of the UV space. Thus, there is a need to develop a representation that is unique across different body parts but consistent across the same body parts of different humans, even for points at arbitrary distances from the underlying human body.

1.5 Thesis Contributions

As part of this thesis, we propose solutions for the problems of (1) Temporally Consistent 3D Human Pose and Shape Estimation from Monocular Videos and (2) Dense Human-Centric Correspondence Estimation. Our major contributions as part of these works are outlined below.

1. **Enhanced Spatio-Temporal Context for Temporally Consistent Robust 3D Human Motion Recovery from Monocular Videos:** In this work, we propose a method for temporally consistent 3D human body pose and shape estimation, where our key contributions are:

   (a) Using a body-aware feature representation instead of generic ResNet-like features, which allows us to exploit prior knowledge about the human body and leads to more robust performance.

   (b) A self-attention and self-similarity based improved spatio-temporal feature aggregation scheme which yields enhanced per-frame spatio-temporal context for our task, leading to more accurate and temporally consistent results.

   (c) An LSTM-based motion refinement module for fine-grained refinement of the initial coarse pose and shape estimation.

2. **ConVol-E: Continuous Volumetric Embeddings for Human-Centric Dense Correspondence Estimation:** In this work, we propose a method for dense correspondence estimation across humans, where our key contributions are:

   (a) A novel volumetric representation that can be used to establish dense correspondences across images of humans, including correspondences for points lying significantly far away from the underlying human body. This allows us to model details like loose clothing or hair.

   (b) A novel evaluation metric to better evaluate the richness of a given representation.

1.6 Organization of the Thesis

In this chapter, we have introduced our target problem domain, our motivation for exploring it, the related research challenges, and provided a brief overview of the existing literature and its limitations. The remainder of this thesis is structured as follows: In Chapter 2, we describe some of the key methods we build upon, providing the necessary background for the remainder of the thesis. In Chapters 3 and 4, we discuss our proposed solutions for the problem of temporally consistent 3D human pose and shape estimation from monocular videos and the problem of dense human-centric correspondence estimation in detail. Finally, in Chapter 5, we summarize our key contributions and discuss the potential future research directions in this problem domain.

Chapter 2

Background

In this chapter, we discuss the relevant prior literature, useful for the rest of the thesis. Specifically, we talk about the Skinned Multi-Person Linear Model (SMPL) [40], Continuous Surface Embeddings (CSE) [48], and geodesic distance calculation over meshes.

2.1 Skinned Multi-Person Linear Model (SMPL)

The Skinned Multi-Person Linear Model (SMPL) [40] is a learned statistical body model, compatible with existing rendering engines, designed to represent the human body in various poses and shapes, including the pose-dependent shape variations, using a set of pose and shape parameters. Each SMPL mesh comprises $N = 6890$ vertices and $K = 23$ (body) joints, along with an additional root joint. The joints and vertices are arranged in a kinematic tree structure. SMPL is learnt and made available in three mesh representations - *Male*, *Female*, and *(Gender) Neutral*.

SMPL's body pose is parameterized using $\theta = [\theta_0, \theta_1, \ldots, \theta_K]$, where, $\forall i \in [1, 23]$, θ_i denotes the relative angle formed by the i^{th} joint with its parent joint in the kinematic tree, while, θ_0 denotes the orientation of the root joint. The body shape is parametrized by a shape vector, $\beta = [\beta_1, \beta_2, \ldots, \beta_{nc}]$. Here, nc refers to the number of shape coefficients, typically chosen to be 10 (but can be up to 300). Overall, the model can be expressed mathematically as: $M(\vec{\beta}, \vec{\theta}; \Phi) : \mathbb{R}^{|\theta| \times |\beta|} \rightarrow \mathbb{R}^{3N}$. It should be noted, that here, Φ denotes the learned model parameters, which remain fixed once the model has been learnt, while θ and β, respectively, refer to the pose and shape parameters as described above, and can take different values during inference.

Figure 2.1 illustrates the working of the SMPL body model. We begin with a template mesh $\tilde{\mathbf{T}}$ and a set of learned blend weights, $\mathcal{W}$. Next, an identity-driven blend shape function ('identity-driven' as it is just based on the shape parameters and does not account for the pose-dependent shape deformations) is used to deform the mesh based on the shape parameters, as $B_s(\vec{\beta}) : \mathbb{R}^{|\vec{\beta}|} \rightarrowtail \mathbb{R}^{|3N|}$. Following this, the pose-driven blend shape function, $B_p(\vec{\theta}) : \mathbb{R}^{|\vec{\theta}|} \rightarrowtail \mathbb{R}^{|3N|}$ is used to add the pose-dependent shape deformations. Finally, a standard blend skinning function $W(\cdot)$ is used to obtain the final mesh by rotating the vertices around the joint centers as per the given body pose.

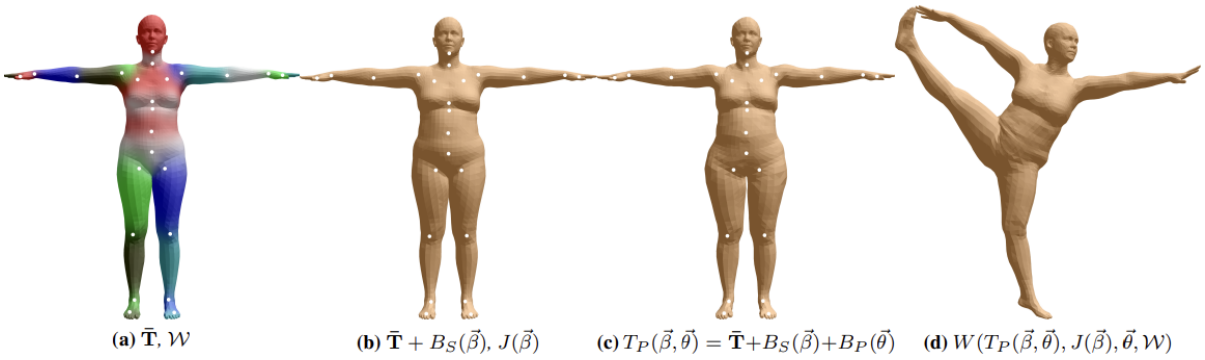


Figure 2.1 Working of the SMPL body model. (a) shows the template mesh, where the different colors represent the blend weights, while the joints are highlighted in white. (b) shows the identity-driven (driven by the shape parameters only) shape deformation of the template mesh. (c) shows the pose-dependent shape deformation of the mesh. It can be seen that the hips in (c) are expanded as compared to (b). (d) shows the final re-posed body in the target shape and pose. (Image Credit: [40].)

Mathematically, this is formulated as:

$$M(\vec{\beta}, \vec{\theta}; \Phi) = W(T_P(\vec{\beta}, \vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W}) \tag{2.1}$$

Here, $M(\vec{\beta}, \vec{\theta}; \Phi) : \mathbb{R}^{|\theta| \times |\beta|} \rightarrowtail \mathbb{R}^{3N}$. $W(\cdot)$ is a standard blend skinning function. T_P, (as defined in Equation 2.2) is the mesh obtained after deforming the template mesh with the identity-driven, $B_S(\vec{\beta})$ (see Equation 2.3), and the pose-driven $B_S(\vec{\theta})$ (see Equation 2.4) shape deformations. $J(\vec{\beta})$ (see Equation 2.5) are the joint locations in the deformed mesh. While, θ and $\mathcal{W}$ are the joint angles, and the blend skinning weights, respectively.

$$T_P(\vec{\beta}, \vec{\theta}) = \bar{\mathbf{T}} + B_S(\vec{\beta}) + B_P(\vec{\theta}) \tag{2.2}$$

$$B_S(\vec{\beta}; \mathcal{S}) = \sum_{n=1}^{|\vec{\beta}|} \beta_n \mathbf{S}_n \tag{2.3}$$

$$B_P(\vec{\theta}; \mathcal{P}) = \sum_{n=1}^{9K} (R_n(\vec{\theta}) - R_n(\vec{\theta^*}))\mathbf{P}_n \tag{2.4}$$

$$J(\vec{\beta}; \mathcal{J}, \bar{\mathbf{T}}, \mathcal{S}) = \mathcal{J}(\bar{\mathbf{T}} + B_S(\vec{\beta}; \mathcal{S})) \tag{2.5}$$

It should be noted that θ^* in Equation 2.3 above denotes the SMPL rest pose. Similarly, R_n in Equation 2.4 denotes the n^{th} element of $R(\vec{\theta})$, the pose rotation vector for all joints, as per the rotation

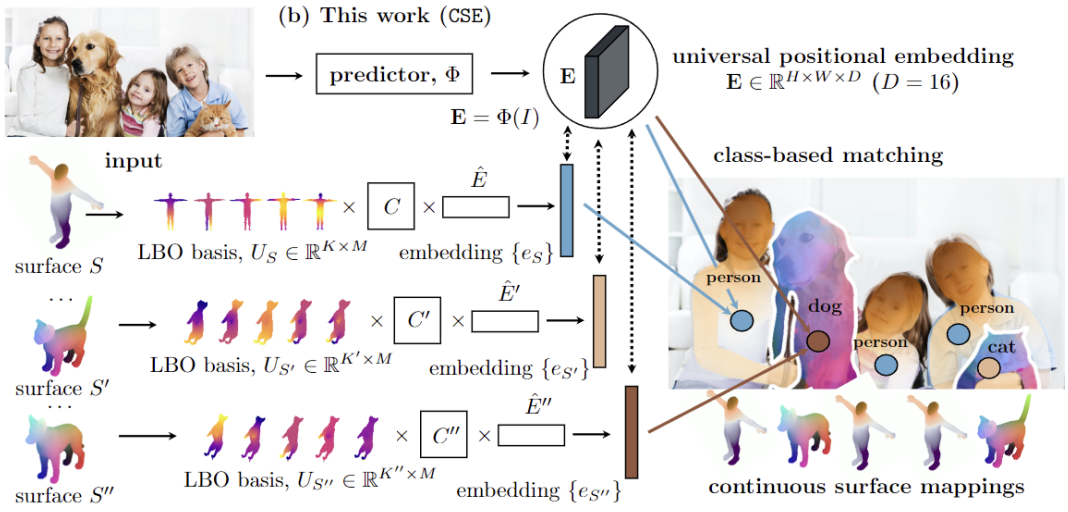


Figure 2.2 Overview of CSE. Given an input image containing an object and a canonical mesh for that object class, CSE helps us establish dense correspondences by learning a common embedding space for the image and the canonical mesh (Image Credits: [48]).

matrix representation. Further, as per the notational convention, the function parameters learnt during training but kept fixed during inference are written after the ';' (semi-colon), while the parameters which can be varied during inference are specified before the semi-colon.

2.2 Continuous Surface Embeddings

Continuous Surface Embeddings (CSE) [48] focuses on the task of learning dense correspondences between deformable object categories. Methods prior to CSE provided ad-hoc solutions for specific object types, often with significant manual work involved. The key idea of CSE is to learn an image-based representation of dense correspondences. CSE learns to predict an embedding vector for each pixel in a 2D image, such that the embedding vector corresponds to the vertex corresponding to the pixel in a canonical object mesh. The authors demonstrate that compared to existing methods, CSE achieves competitive performance on the task of dense pose estimation for humans while also being able to generalize to new object categories by sharing a single dense pose predictor across categories.

Figure 2.2 provides an overview of the method of CSE. Given an input image I and a trained predictor network Φ, the per-pixel embeddings for the image are obtained as $E = \Phi(I)$. Now, the per-pixel embeddings for the image, E, and the per-vertex embeddings for the canonical mesh, E', can be used to establish dense correspondences as follows:

$$p(k|x, I, E', \Phi) = \frac{\exp(-\langle e_k, \Phi_x(I)\rangle)}{\sum_{k=1}^{K} \exp(-\langle e_k, \Phi_x(I)\rangle)} \tag{2.6}$$

here, $p(k|x, I, E', \Phi)$ is the probability, represented using a softmax-like function, that the k^{th} vector of the canonical mesh corresponds to x^{th} pixel of the image I. While $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors.

The model is learnt by using a modified version of the cross-entropy loss. Instead of using the vanilla cross-entropy loss, the cross-entropy between a gaussian-like distribution centered on the point k and the predicted posterior is minimized, as follows:

$$\begin{aligned} \mathcal{L}_\sigma &= \sum_{q=1}^{K} g_s(q;k)\mathcal{L}(E,\Phi) \\ \mathcal{L}(E,\Phi) &= - \underset{(I,x,k)\in\tau}{avg} \log p(k|x,I,E,\Phi) \\ g_s(q;k) &\propto \exp\left(-\frac{1}{2\sigma^2} d_s(X_q, X_k)\right) \end{aligned} \tag{2.7}$$

Additionally, CSE authors use spectral analysis to inject geometrical knowledge into their embeddings. This helps them to ensure that their embeddings are (1) of (relatively) low dimensions, (2) agnostic to mesh discretization, and (3) easy to relate across two different object categories (like humans and chimps). We refer readers to the CSE [48] paper for more details.

For our problem statements, the pre-trained CSE network is beneficial as it can help us obtain rich information about the human body from the provided input. For the task of 3D human pose and shape estimation, we use CSE as an initializer to provide a strong prior about the per-pixel relations with the SMPL mesh. This leads to more robust and accurate pose and shape estimation. For the task of dense correspondence estimation, while CSE does not solve our problem, as it can not handle points lying away from the body model (like clothing details), it can still provide us with a strong prior about the underlying human body, which can be used as a foundation for further reasoning about the clothing details and other points lying away from the body model.

2.3 Geodesic Distance Calculation over Meshes

Geodesic distance represents the shortest distance between two points along a given surface, such as a sphere or mesh surface. Unlike Euclidean distance, which measures the direct, 'straight-line' distance between points in a flat, Cartesian space, geodesic distance considers the curvature and topology of the underlying space. This concept is illustrated in Figure 2.3, where the straight line between points A and B depicts the Euclidean distance, while the curved line along the surface depicts the geodesic distance between the two points. Calculating geodesic distance becomes crucial when dealing with non-flat surfaces, where a straight-line path may not accurately reflect the actual distance between two points. For instance, using Euclidean distance to measure distances on the Earth's surface would lead to inaccuracies due to the Earth's curvature.

A few key application areas where geodesic distances are commonly used are:

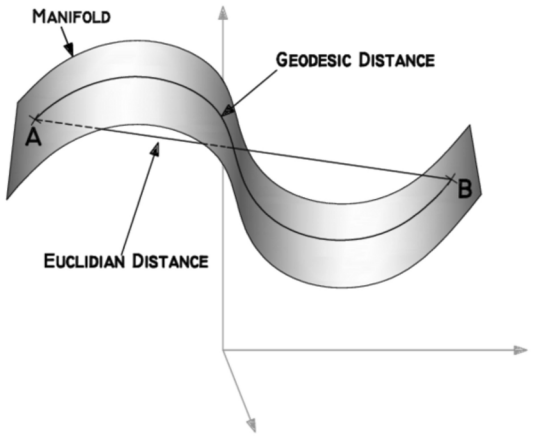


Figure 2.3 Geodesic Distance vs Euclidean distance: The straight line between points A and B shows the Euclidean distance between the two points, while the curved line, along the surface, shows the geodesic distance between the points. (Image Credits: [23])

- **Geographic Information Systems (GIS):** Geodesic distances are widely employed in Geographic Information Systems for tasks like computing distances over topographical maps while accounting for roads or other permissible paths for travelers.
- **Manifold Learning:** Manifold learning refers to a class of Machine Learning problems related to projecting high-dimensional data onto lower-dimensional manifolds. These techniques commonly use geodesic distances between points to compute the similarity or proximity between points in a dataset.
- **3D Computer Vision:** Another common application area of geodesic distances, and the one most relevant to us, is 3D Computer Vision. In 3D computer vision, geodesic distances are used to establish relations between features on a surface. For example, after we detect the eyes and mouth of a face, the inter-feature distances can be used to help characterize a surface for recognition or correspondence.

$$|\nabla_\phi| = 1 \tag{2.8}$$

A popular class of algorithms for computing geodesic distances involves solving the eikonal equation (see Equation 2.8), subject to boundary conditions $\phi|_\gamma = 0$ over some subset γ of the domain (like a point or curve). Typically, the equation is solved using *fast marching* [56] and *fast sweeping*. Fast marching methods involve propagating distance information in wavefront order using a priority queue and provide a fast but approximate solution for computing geodesic distances over the meshes. On the other hand, heat-diffusion-based methods [10] are another class of methods that provide fast but approximate solutions for computing geodesic distances. Heat-diffusion-based methods offer even higher

computational efficiency than fast-marching methods and are generally more robust for different data types. However, both these classes of algorithms provide approximate solutions. For computing the exact distances, the MMP algorithm is one popular algorithm [46] for computing exact geodesic distances is the MMP algorithm. This method provides an exact solution for the geodesic distance. However, it offers limited computational and memory efficiency. We refer the readers to [9] for a more detailed discussion on this topic.

Chapter 3

Enhanced Spatio-Temporal Context for Temporally Consistent Robust 3D Human Motion Recovery from Monocular Videos

As introduced in chapter 1, recovering temporally consistent 3D human body pose, shape and motion from a monocular video is a challenging task due to (self-)occlusions, poor lighting conditions, complex articulated body poses, depth ambiguity, and limited availability of annotated data. Further, doing a simple per-frame estimation is insufficient as it leads to jittery and implausible results. In this chapter, we propose a novel method for temporally consistent motion estimation from a monocular video. Instead of using generic ResNet-like features, our method uses a body-aware feature representation and an independent per-frame pose and camera initialization over a temporal window followed by a novel spatio-temporal feature aggregation by using a combination of self-similarity and self-attention over the body-aware features and the per-frame initialization. Together, they yield enhanced spatio-temporal context for every frame by considering remaining past and future frames. These features are used to predict the pose and shape parameters of the human body model, which are further refined using an LSTM. Experimental results on the publicly available benchmark data show that our method attains significantly lower acceleration error and outperforms the existing state-of-the-art methods over all key quantitative evaluation metrics, including complex scenarios like partial occlusion, complex poses and even relatively low illumination.

3.1 Introduction

Recovering 3D human body pose, shape and motion from a monocular video is an important task that has tremendous applications in augmented/virtual reality, healthcare, gaming, sports analysis, human-robot interaction in virtual environments, virtual try-on, etc. A lot of work has been done in estimating 3D body pose and shape from a single-image [13, 30, 35, 49, 50] by learning to regress the explicit 3D skeleton or parametric 3D body model like SMPL [40] (Please see section 2.1 for a detailed discussion about SMPL). However, many applications such as human motion analysis, sports analytics, behavior analysis, etc., critically depend on the temporal consistency of human motion where single-image-based

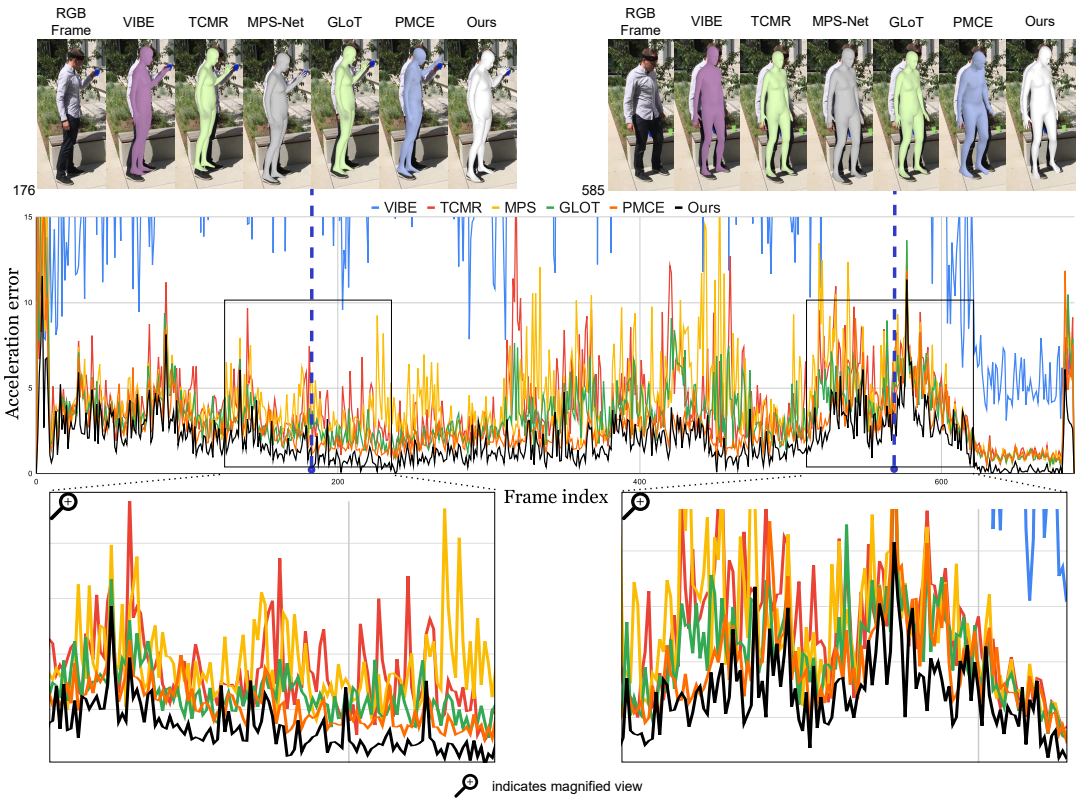


Figure 3.1 Acceleration error plot on unseen test video from 3DPW dataset [62]

.

methods seem to fail frequently. Temporally consistent 3D human pose, shape and motion estimation from a monocular video is a challenging task due to (self-) occlusions, poor lighting conditions, complex articulated body poses, depth ambiguity, and limited availability of annotated data. Efforts on monocular video-based motion estimation [33, 41, 7, 36, 51, 60] typically introduce a CNN or RNN module to perform spatio-temporal feature aggregation from neighboring frames followed by SMPL [40] parameters regression, thus modeling relatively local temporal coherence. However, these methods tend to fail while capturing long-term temporal dynamics and show poor performance when the body is under partial occlusion. Another class of works [52, 20, 44, 70] attempt to model the generative space of motion modeling using Conditional VAEs, often followed by a global, non-learning-based optimization at inference time using the entire video. Such global optimization is also used in [71] with a plug-and-play post-processing step for improving the existing methods by exploiting long-term temporal dependencies for human motion estimation. However, due to the post-processing over the entire sequence, such methods find limited applicability to real-world scenarios.

A relevant work, MPS-Net [65], explicitly models the visual feature similarity across RGB frames and uses it to guide the learning of the self-attention module for spatio-temporal feature learning followed by a local to global temporal feature aggregation for per-frame SMPL prediction. A recent work GLoT [57] uses random masking along with a global (at a window-level) encoder to regress a coarse global mesh sequence followed by per-frame local parameter correction using a local transformer-based encoder and a Hierarchical Spatial Correlation Regressor (HSCR). Another recent work, PMCE [68], attempts to directly regress the SMPL mesh vertices by jointly updating off-the-shelf temporal image features and 2D pose estimates using an attention based co-evolution decoder. Nevertheless, similar to the majority of the existing methods, these methods neglects the body-aware spatial context in the images, while using generic ResNet [21] features extracted from the RGB frames.

In this chapter, we propose a holistic method that exploits enhanced spatio-temporal context and recovers temporally consistent 3D human pose/shape from monocular video. At first, we select a set of continuous frames in a temporal window and pass it to the *Initialization module* which extracts the body-aware deep features from individual frames and in-parallel predict initial per-frame estimates of body pose/shape and camera pose using an off-the-shelf method. Subsequently, we pass these initial estimates and features to novel *Spatio-Temporal feature Aggregation (STA) module* for recovering enhanced spatio-temporal features. Finally, we employ our novel *Motion estimation and Refinement module* to obtain temporally consistent pose/shape estimation using these enhanced features. Figure 3.2 provides outline of our method.

In regard to functionality/relevance of these modules, the initialization module extracts a body-aware feature representation [48] for each frame of the local non-overlapping temporal frame window, instead of the generic ResNet feature used by existing methods and the independent per-frame pose and camera initialization estimated using [14]. This provides a strong spatial prior to our method. Further, our proposed novel STA module computes the self-similarity and the self-attention on initial spatial priors provided by the previous module. In particular, the self-similarity between the body-aware features in a

temporal window helps us to correlate the body parts across frames even in the presence of occlusion. Similarly, the self-similarity among the pose parameters and the camera reveals the continuity of the human motion along with the camera consistency. We also use self-attention on the camera parameters and the body-aware features. Together, they yield spatio-temporal aggregated features for every frame by considering the remaining past and future frames inside the window. Here, the joint characteristics of the self-similarity and the attention map find the more appropriate range in the input video to reveal the long-horizon context. Finally, our novel motion estimation and refinement module first predicts the per-frame coarse estimation of pose/shape using the spatio-temporally aggregated features from the STA module and subsequently passes it to an LSTM-based joint-temporal refinement network to recover the temporally consistent robust prediction of pose/shape estimates. In order to generate continuous predictions near the temporal window boundaries, we average the pose/shape parameters for consecutive border frames across neighboring windows. We empirically observed that applying LSTM-based joint refinement on pose/shape yields superior performance instead of applying it on STA features and then predicting pose/shape parameters (see subsection 4.4.6).

As a cumulative effect, our method produces significantly lower acceleration errors in comparison to SOTA methods (see subsection 3.4.2). Figure 4.1 shows a plot of acceleration where our method (in black) yields significantly lower acceleration errors compared to other methods. Moreover, owing to our enhanced spatio-temporal context and motion refinement, our method significantly outperforms the state-of-the-art (SOTA) methods even in relatively poor illumination and occlusions (see subsection 3.4.3).

3.2 Related Work

Image based 3D human pose, shape and motion estimation: Existing methods either solve for the parameters of SMPL [40] from the images or directly regress the coordinates of a 3D human mesh [18]. [30, 35, 49] are some of the early succesful works for human pose and shape estimation from monocular images.

HyBrik [38] and KAMA [27] leverage 3D key points for the 3D mesh reconstruction. In particular, HyBrik uses twist and swing decomposition for transforming the 3D joints to relative body-part rotations. Instead of full body, methods like HoloPose [16] and PARE [34] have introduced parts parts-based model. While HoloPose does part-based parameter regression, PARE uses a part-guided attention mechanism for exploiting the visibility of individual body parts and predicting the occluded parts using neighboring body-part information. While these methods are quite effective for estimating the 3D pose and shape from images, they are not capable of producing temporally consistent 3D human motion from video by frame-based processing.

Video based 3D human and pose estimation: Recently, a considerable amount of work has been carried out to address the challenge of temporally consistent 3D human pose and shape estimation from

video. For instance, HMMR [31] trains a temporal encoder that learns a representation of 3D human dynamics from a temporal context of image features. Along with 3D human pose and shape, such representation is also used for capturing the changes in the pose in the nearby past and future frames. Similarly, VIBE [33] proposes a temporal encoder that encodes static features into a series of temporally correlated latent features and feeds them to a regressor to estimate the SMPL parameters. MEVA [41] uses a two-stage model that first captures the coarse overall 3D human motion followed by a residual estimation that adds back person-specific motion details. However, these methods fail to reconstruct the humans under partial occlusions. TCMR [7] uses GRU-based temporal encoders with different encoding strategies to learn better temporal features from images. They also propose a feature integration from the three encoders for the SMPL parameter regressor. GRU-based techniques can only deal with local neighborhoods which makes it difficult for them to learn long-range dependencies. Hence, [2, 75] use a transformer to learn long-range temporal dependencies. However, such methods require a large number of consecutive frames (around 250), making them slower. Another class of methods like HuMoR [52] and GLAMR [70] use a variational autoencoder which takes single frame-based human pose estimates to predict the human motion sequence in an auto-regressive way followed by a non-learning based global optimization on the human pose and trajectory obtained from the entire video for temporal refinement. Similarly, SmoothNet [71] also does a global optimization on the estimated trajectory of any human pose estimation method to improve their temporal continuity. The global optimization in the test time limits the applicability of such methods. MPS-Net [65] tries to produce locally global temporal coherence using a MOtion Continuity Attention (MOCA) module. More specifically, their method explicitly models the visual feature similarity across RGB frames and uses it to guide the learning of the self-attention module for spatio-temporal feature learning. MOCA enables focusing on an adaptive neighborhood range for identifying the motion continuity dependencies. This is followed by a Hierarchical Attentive Feature Integration (HAFI) module to achieve local to global temporal feature aggregation. GLoT [57] uses a combination of global and local encoder-decoder networks, where the global branch helps perform feature aggregation across a larger temporal window, while the local branch operates over a smaller temporal window and learns the local pose corrections using a Hierarchical Spatial Correlation Regressor (HSCR). Unlike most other methods, which regress the joint poses and shapes together, HSCR regresses the body pose and shape in a sequential fashion, following the kinematic structure of the parametric body. PMCE [68] uses a GRU over per-frame, off-the-shelf image features to obtain temporal image features, and a spatial-temporal transformer over per-frame, off-the-shelf 2D pose estimates to obtain the mid-frame 3D pose. The temporal image features and mid-frame 3D pose are used to jointly update the vertex positions of a coarse template SMPL mesh using and the mid-frame 3D pose, using an attention based architecture such that they agree with the visual cues available in the image features. Unlike other methods discussed above, PMCE does not predict the SMPL parameters. Instead, it directly regresses the vertices of the body model.

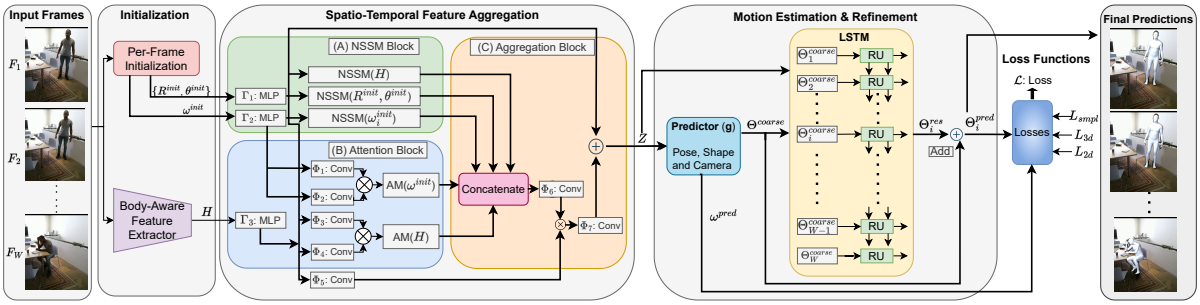


Figure 3.2 Architecture overview of our proposed method.

3.3 Method

In this section, we provide a detailed overview of the key modules of our proposed method. As discussed in section 3.1 (and outlined in Figure 3.2), our method takes a set of consecutive frames as input and feeds it to the three key modules, namely, initialization, spatio-temporal feature aggregation and motion prediction & refinement, to predict temporally consistent body pose and shape parameters of SMPL [40], a statistical body model.

More specifically, given an input video $V = \{F_i\}_{i=1}^{N}$ composed of N frames, with F_i representing the i^{th} frame, we aim to recover SMPL-based human body pose and shape parameters for each frame, i.e., $\Theta_i^{pred} = \{T_i, R_i, \theta_i, \beta_i\}$. Here, $T_i \in \mathbb{R}^3$ and $R_i \in \mathbb{R}^3$ represents the translation and rotation (in axis-angle format) of the root joint, $\theta_i \in \mathbb{R}^{23\times3}$ represents the relative rotations of the remaining 23 joints while $\beta_i \in \mathbb{R}^{10}$ represents body shape parameters. Please note that we sample a temporal window (a subset of continuous frames) of size W (we choose $W = 16$) from the input video and learn/infer over it instead of doing inference on all frames in the video sequence.

3.3.1 Initialization

Per-frame Body Pose and Camera Estimation: We perform independent estimation of per-frame body pose (θ^{init}) and camera (ω^{init}) parameters using a SOTA method (HMR2.0 [14]) and feed it as initialization to our STA module.

Body-aware Spatial Feature Extraction: Continuous Surface Embedding (CSE) [48], as described in section 2.2, proposed to learn a body-aware feature representation for obtaining dense correspondences across images of humans. CSE predicts, per pixel 16-dimensional embedding vector (associated with the corresponding vertex in the parametric human mesh), thereby establishing dense correspondences between image pixels and 3D mesh surface, even in the presence of severe illumination conditions and (self-) occlusions. Figure 3.3 shows the color-coded visualization of CSE embeddings demonstrating

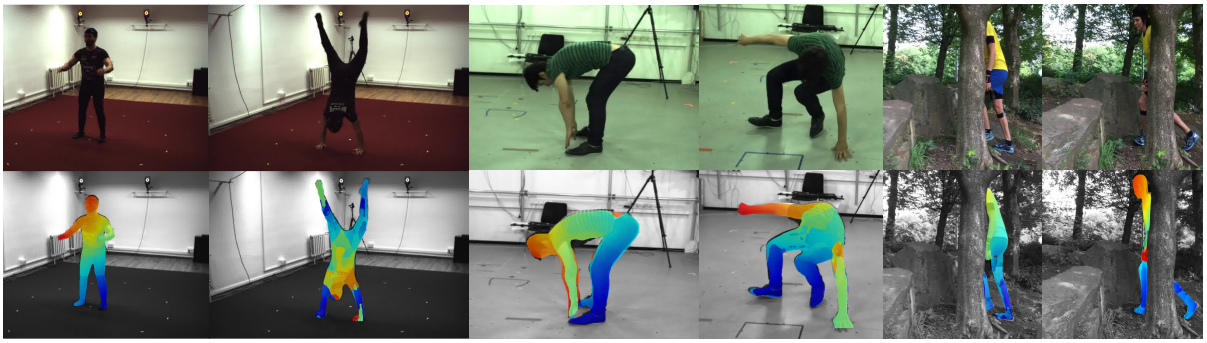


Figure 3.3 Sample 3-channel visualization of CSE embedding (row-1 : RGB frame & row-2: embedding plot).

its robustness to severe illumination and/or occlusion scenarios. Thus, we propose to extract and use the 16-dimensional body-aware spatial features $H = \{H_i\}_{i=1}^{N}$ using a pre-trained CSE encoder for each frame F_i, such that:

$$H_i = \Psi(F_i) \tag{3.1}$$

3.3.2 Spatio-Temporal Feature Aggregation (STA)

The spatial features H_i extracted from each frame can be directly regressed to estimate per-frame motion and shape parameters. However, this typically leads to jittery and implausible motion estimates as the predictions are not temporally consistent. One possible remedy to this is to use self-attention across frames in a temporal window [63]. Interestingly, [65] showed that a regular attention network is unreliable and can give high attention scores between temporally distant frames which would lead to inaccurate results. They address this problem by using a **Normalized Self-Similarity Matrix (NSSM)** in their MOCA module. Nevertheless, their method only exploited the spatial features for such self-attention guidance. Instead, as per the recent trend of exploiting per-frame pose initialization [70, 37], we propose to encode additional information to our temporal features in terms of initial estimates of body pose and camera parameters. More specifically, we obtain for each i^{th} frame the initial pose/shape and camera parameters using [14] as: $\Theta_i^{init} = \{T_i, R_i, \theta_i, \beta_i\}$ and camera parameters $\omega_i^{init} \in \mathbb{R}^3$ (assume a weak perspective camera model). It is important to note that we represent rotation using the 6-dimensional vector representation [73] and then flatten them into a single 144-dimensional vector to recover body pose as: $[R_i, \theta_i] \in \mathbb{R}^{144}$.

Our STA module has three key blocks: (1) Frame-wise Similarity Computation, (2) Frame-wise self-attention, and (3) Feature Aggregation.

The first block deals with the computation of the three $\{W \times W\}$ self-similarity matrices, namely,

NSSM (H) for the Body-aware spatial features, NSSM ($[R, \theta]$) for initial body pose estimates and NSSM (ω^{init}) for initial camera estimates. More specifically, we uplift $[R_i, \theta_i]$ and $\omega_i^{init} \in R^3$ to 512 dimensions using linear layers Γ_1 and Γ_2 and similarly transform the spatial feature H_i to 2048 dimensions using Γ_3. These multiple NSSMs help us to correlate the frames based upon body parts appearance, body pose, and cameras thereby giving robustness to occlusions as well as revealing the continuity of the human motion along with the camera consistency.

The second block obtains a self-attention map on our spatial features i.e., AM (H) and initial camera estimates i.e., AM (ω^{init}), respectively. When applying self-attention on our spatial features H_i , first we transform H_i to 2048 dimension using a linear layer Γ_3 , and later down-sample them to 1024 by learning two different 1×1 convolution layers Φ_3 and Φ_4. Similarly, when applying self-attention on the initial camera estimates, we first uplift this vector to 512 dimension vector using an MLP Γ_2 and subsequently learn two different 1×1 convolution layers Φ_1 and Φ_2. This self-attention on the camera parameters and the body-aware features help us adaptively find the range which is important to capture the temporal smoothness.

Finally, the feature aggregation block first concatenates all the attention and NSSM maps to get a $W \times W \times 5$ tensor and later resize it to $W \times W$ matrix using a 1×1 convolution layer (Φ_6). This $W \times W$ matrix represents the consolidated similarity between frames across the window. This feature is subsequently multiplied with the down-sampled spatial features (of 1024 dimension obtained by Φ_5) and the result is then uplifted (using convolution layer Φ_7) to get $Y \in \mathbb{R}^{W \times 2048}$. Thus, together, they yield spatio-temporal aggregated features for every frame by considering the remaining past and future frames inside the window. The per-frame temporally aggregated feature Y_i is finally added to the spatial features to get the spatio-temporally aggregated features Z_i for i^{th} frame as:

$$Z_i = H_i + Y_i. \tag{3.2}$$

3.3.3 Motion Estimation & Refinement

Once we have the spatio-temporal features Z_i, we obtain an independent coarse pose/shape, and camera estimation for each frame using predictor network (g)

$$\Theta_i^{coarse}, \omega_i^{pred} = g(Z_i) \tag{3.3}$$

where g predicts the SMPL parameters i.e., $\Theta_i^{coarse} \in \mathbb{R}^{85}$ and the camera parameters i.e., $\omega_i^{pred} \in \mathbb{R}^3$ for frame F_i.

We propose to further refine these estimated independent coarse poses and shapes (obtained using spatiotemporally aggregated features) using an LSTM [22] based joint *residual* prediction. The LSTM ζ takes as input the features Z_i and coarse SMPL pose estimates Θ_i^{coarse} and predicts the residual $\Theta_i^{res} \in \mathbb{R}^{85}$, which is subsequently added to Θ_i^{coarse} in order to recover the refined pose and shape

parameters Θ^{pred}.

$$\Theta_i^{res} = \zeta(Z_i, \Theta_i^{coarse}) \tag{3.4}$$

$$\Theta_i^{pred} = \Theta_i^{coarse} + \Theta_i^{res} \tag{3.5}$$

To ensure temporally consistent predictions at the window boundaries, we average the pose and shape parameter estimates of bordering frames across the neighboring windows.

3.3.4 Loss Functions

Similar to existing literature [65, 30, 33], we adopt loss functions on body pose and shape (L_{SMPL}), 3D joint coordinates (L_{3D}), and 2D joint coordinates (L_{2D}) obtained with predicted weak-perspective camera parameters (ω^{pose}). These loss functions are briefly explained below.

$$\begin{aligned} L_{SMPL} = \lambda_{shape}||\hat{\beta}_i - \beta_i||_2 \\ + \lambda_{pose}||\{\hat{R}_i, \hat{\theta}_i\} - \{R_i, \theta_i\}||_2 \end{aligned} \tag{3.6}$$

where β_i and $\{R_i, \theta_i\}$ respectively are the predicted pose and shape parameters for the i^{th} frame, and $\hat{\beta}_i$, $\{\hat{R}_i, \hat{\theta}_i\}$ are the corresponding ground truths.

$$L_{3D} = ||\hat{J}_i^c - J_i^c||_2 \tag{3.7}$$

where J_i^c represents predicted the 3D joint coordinates for the i^{th} frame and $\hat{J}_i^c$ are the corresponding ground truth 3D joint coordinates.

$$L_{2D} = ||\hat{x}_i - \Pi(J_i^c)||_2 \tag{3.8}$$

where $\hat{x}_i$ represents the ground truth 2D keypoints for the i^{th} frame and Π represents the 3D-2D projection obtained from the predicted camera parameters ω^{pred}.

The final loss function is a linear combination of these losses defined as:

$$L_{final} = \lambda_1 L_{SMPL} + \lambda_2 L_{3D} + \lambda_3 L_{2D} \tag{3.9}$$

It is important to note that our model is trained in an end-to-end trainable fashion where L_{final} is applied on the final predicted pose and shape parameters obtained from LSTM ζ. There is no separate training performed for the coarse estimation predictor g.

3.4 Experiments & Results

In this section, we evaluate our method on different publicly-available datasets and report superior qualitative as well as quantitative results in comparison with SOTA methods.

3.4.1 Experimental Setup

Datasets Details: We evaluate our method on the standard test split of Human3.6M [26] and 3DPW [62] used in existing literature [33, 7, 65, 57, 68]. We separately discuss the challenges associated with the quality of ground truth annotations in another popularly reported dataset MPI-INF-3DHP [45] in subsection 3.5.1. Nevertheless, for a fair comparison with existing methods, we still use MPI-INF-3DHP [45] for training, along with Human3.6M [26] and 3DPW [62] datasets. A brief description of the datasets is given below.

Human3.6M is a large scale dataset containing video sequences with corresponding 3D pose and shape annotations of various subjects performing different actions like discussion, smoking, talking on the phone etc. It is collected using a multi-view calibrated system and a high-speed motion capture system. Similar to existing work [33, 7, 65, 57, 68], we use the sub-sampled dataset (25 FPS) for our experiments.

3DPW is an in-the-wild dataset, captured with a moving cell-phone camera. It uses inertial measurement unit (IMU) sensors patched to the human body parts to calculate the ground truth SMPL [40] parameters. It contains 60 video sequences with 18 3D models in different clothing, performing daily-life activities like walking, buying vegetables etc.

MPI-INF-3DHP is a dataset which provides the pose annotations of 8 subjects with 16 videos per subject. It is captured in a combination of indoor and outdoor settings, with actions ranging from simple actions like walking and sitting, to complex dynamic actions like exercising. It is captured using a markerless motion capture system using a multi-view camera setup.

Further, in order to evaluate the generalization ability of our method to unseen data, we use two additional datasets: i3DB [47] and PROX [19]. These datasets contain action sequences of humans interacting with objects in an indoor setting like a room/office. These sequences contain occlusions and are fairly different from our training datasets. Thus, serving as a good measure to evaluate generalization ability.

Evaluation Metrics: We use the standard evaluation metrics used in existing literature [33, 41, 7, 65, 57, 68] to evaluate our method's performance. These are discussed in more details below.

Mean Per Joint Position Error (MPJPE) is defined as the mean of the Euclidean distances between the ground truth and the predicted joint positions. It is measured in millimeters (mm).

Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) is defined as the MPJPE computed after using Procrustes alignment (PA) to solve for translation, scale and rotation between the estimated body and the ground truth. Similar to MPJPE, PA-MPJPE is also measured in millimeters (mm).

Acceleration Error (ACC-Err) is defined as the mean difference between the accelerations of the ground truth and predicted 3D joints. Specifically, the change in position of the 3D joints in unit time (i.e. across two consecutive frames) gives us the velocity of the joints, and the change in velocity in unit time gives us the acceleration. Acceleration error is then measured by finding the difference between the groundtruth and predicted accelerations. It is measured in mm/t^2 (where t denotes unit time - the time interval between two consecutive frames).

Mean Per Vertex Position Error (MPVPE) is given by the mean of the Euclidean distances between the ground truth and the predicted vertex positions of each vertex in the SMPL [40] body model constructed using the predicted pose/shape parameters. It is also measured in milimeters (mm).

Comparison with SOTA: We compare our method with existing SOTA methods for monocular video-based pose and shape estimation. Speicifically, we provide comparison with VIBE [33], MEVA [41], Uncertainity Aware [36], TCMR [7], HuMoR [52], MPS-Net [65], GLAMR [70], GLoT [57], PMCE [68]. A brief discussion about these methods is provided in section 3.2.

Evaluation Protocol: Existing methods report best performance by learning different models for specific datasets with hyper-parameter tuning optimal for a given dataset. For example, TCMR, MPS-Net and GLoT use the same model for evaluation on Human3.6M and MPI-INF-3DHP, but use a model trained with different hyperparameters for evaluation on 3DPW. A few other methods report results by training and evaluating on a single dataset. For example, D&D reports results by performing individual training and evaluation on Human3.6M and 3DPW. While, PMCE uses a common model (trained on Human3.6M, 3DPW and MPI-INF-3DHP) for reporting results on the 3DPW and MPI-INF-3DHP datasets, but reports results on Human3.6M by training and evaluating only on Human3.6M. Additionally, some methods also train their common model on additional large datasets on which they do not evaluate on. For example, PMCE trains on the COCO dataset.

For consistency, we adopt two protocols of reporting quantitative results across all methods. Namely, *prot-1* : training a common model on the three datasets (Human3.6, 3DPW, MPI-INF) and reporting inference results on individual datasets; *prot-2* : training an optimal model for specific dataset with hyper-parameter tuning using the single or multiple datasets (only for Human3.6, 3DPW, MPI-INF).

Implementation Details: We obtain the body aware features and per-frame pose/camera initializations using the pre-trained CSE [48] and HMR2.0 [14] models, respectively. Similar to existing work [33, 7, 65, 57], we initialize our pose, shape, and camera predictor in the motion estimation and refinement module with the pre-trained SPIN [35] checkpoint. In the same module, the LSTM has 3 layers and uses 2048 as the hidden feature size. We use a mini-batch size of 32 and an initial learning rate of 5×10^{-5}. The learning rate is reduced by a factor of 10 every time the 3D pose accuracy does not improve for the 5 consecutive epochs. Adam Solver [32] is used for optimization. For our experiments, we use a

Method	Human3.6M [26]			3DPW [62]			
	PA-MPJPE↓	MPJPE ↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	MPVPE↓	ACC-ERR↓
VIBE [33]	41.4	65.6	-	51.9	82.9	99.1	23.4
MEVA [41]	53.2	76.0	15.3	54.7	86.9	-	11.6
Uncertainty-Aware [36]	38.4	58.4	6.1	52.2	92.8	106.1	6.8
TCMR [7]	52.0	73.6	3.9	52.7	86.5	102.9	7.1
HUMOR [52]	47.3	69.3	4.2	51.9	74.8	81.4	6.3
MPS-Net [65]	47.4	69.4	3.6	52.1	84.3	99.7	7.4
GLAMR [70]∗	48.3	72.8	6.0	51.7	72.9	86.6	8.9
D&D [37]	35.5	52.5	6.1	42.7	73.7	88.6	7.0
GLoT [57]	46.3	67.0	3.6	50.6	80.7	96.3	6.6
PMCE [68]	37.7	53.5	3.1	46.7	69.5	84.8	6.5
Our Method	**30.1**	**41.1**	**3.0**	**39.2**	**63.5**	**61.8**	**5.3**

Table 3.1 Quantitative comparison of mean error values of our methods with other monocular video-based methods as per *prot-2*. Best results are in **bold** and second best are underlined. (∗: GLAMR uses Human3.6M, 3DPW and AMASS [43] as 3D datasets.)

Method	Human3.6M [26]			3DPW [62]			
	PA-MPJPE↓	MPJPE ↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	MPVPE↓	ACC-ERR↓
MPS-Net [65]	53.2	72.5	3.8	52.1	84.3	99.7	7.4
GLoT [57]	51.4	72.0	**3.3**	52.3	82.8	98.5	6.5
PMCE [68]	34.4	48.6	3.8	46.7	70.4	85.3	6.7
Our Method	**31.0**	**41.3**	**3.3**	**39.2**	**63.5**	**61.8**	**5.3**

Table 3.2 Quantitative comparison of our methods with SOTA methods as per *prot-1*. Best results are in **bold** and second best are underlined.

Method	Human3.6M [26]			3DPW [62]			
	PA-MPJPE↓	MPJPE ↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	MPVPE↓	ACC-ERR↓
MPS-Net [65]	17.7	22.3	<u>3.2</u>	23.2	31.4	37.2	14.8
GLoT [57]	18.3	23.1	3.4	22.7	31.0	36.7	13.4
PMCE [68]	<u>11.0</u>	<u>15.4</u>	4.0	<u>20.0</u>	<u>27.2</u>	<u>27.2</u>	<u>12.7</u>
Our Method	**6.0**	**7.4**	**0.6**	**12.4**	**5.2**	**16.3**	**3.3**

Table 3.3 Quantitative comparison of our method with other SOTA methods using standard deviations of evaluation metrics. Best results are in **bold** and second best are <u>underlined</u>.

window size of 16 (see Table 3.7 for discussion on choice of window size). Training is done for 40 epochs and takes about 8 hours using 3 NVIDIA RTX A-6000 GPUs.

3.4.2 Quantitative Results

Table 3.1 provides a quantitative comparison between our method and existing SOTA methods following the evaluation protocol *prot-2* and standard evaluation metrics, as introduced in section 3.4. More specifically, *prot-2* evaluation protocol entails training an optimal model for specific dataset with hyper-parameter tuning using the single or multiple datasets. However, some methods perform different hyper-parameter tuning for different datasets. Nevertheless, all methods use the same combination of 3D datasets for training, except GLAMR, which uses Human3.6M, 3DPW and AMASS [43].

It can be observed from Table 3.1 that our method significantly outperforms all other methods across different evaluation metrics on both the Human3.6M and 3DPW datasets. Additionally, there is no consistently second best performing method across all evaluation metrics. On Human3.6M dataset, D&D is the second-best performer in terms of the PA-MPJPE and MPJPE metrics where we outperform it by a margin of 5mm and 10mm, respectively. However, it has significantly higher Acc-Err values in comparison with other competing methods. On the other hand, PMCE, which is the second-best performer on Acc-Err on Human3.6M dataset has higher PA-MPJPE and MPJPE values than D&D. Nevertheless, our method outperforms PMCE in terms of PA-MPJPE and MPJPE by a margin of 7mm and 12mm, respectively. Similarly, on 3DPW dataset, we observe that D&D is the second-best by a margin of 3mm on PA-MPJPE, PMCE is the second-best by a margin of 6mm on MPJPE while HuMoR is the second-best on MPVPE and Acc-Err metrics. It is important to note that we have a huge improvement of 20mm on the MPVPE metric and a considerable improvement of 1mm/(unit time) on the Acc-Err metric. Thus, we outperform these SOTA methods while there seems to be no single second-best performing method across all evaluation metrics.

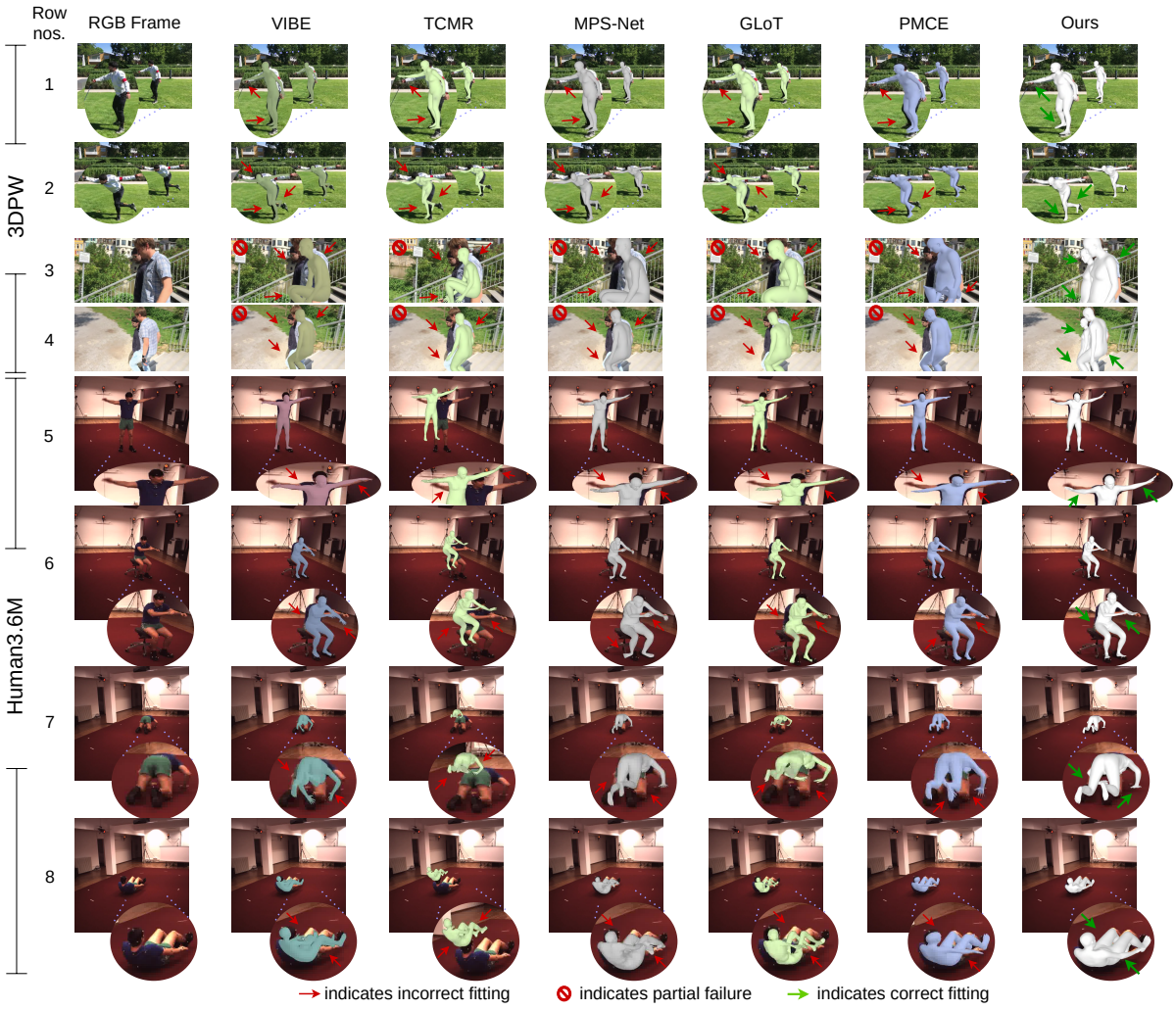


Figure 3.4 Qualitative results showing the estimated pose overlaid on the frames of videos from the test sets of Human3.6M [26] and 3DPW [62] datasets.

Method	i3db [47]			PROX [19]		
	PA-MPJPE ↓	MPJPE ↓	ACC-ERR↓	PA-MPJPE ↓	MPJPE↓	ACC-ERR↓
MPS-Net [65]	29.8	39.9	2.6	22.7	31.1	2.5
GLOT [57]	26.4	36.8	2.4	19.1	28.0	2.1
PMCE [68]	24.7	32.0	2.4	17.0	23.5	2.1
Ours	**20.2**	**29.5**	**1.8**	**12.1**	**18.3**	**1.6**

Table 3.4 Generalization results on unseen datasets.

Further, we also evaluate and report the comparison with most recent/relevant methods, namely, MPS-Net, GLoT and PMCE in Table 3.2 where we follow the evaluation protocol *prot-1* (as described in section 3.4). In contrast to *prot-2*, *prot-1* protocol use a common model trained with the same dataset combination. Once again, it can be seen that our method significantly outperforms other methods across all metrics on Human3.6M and 3DPW datasets.

It is important to note that the aforementioned results are only reporting the mean values of the evaluation metrics computed over the test split. However, the mean value of evaluation metric can be significantly affected by few (outlier) sequences. Thus, for more robust evaluation, we report the standard deviations of the evaluation metric for most relevant methods (results computed using *prot-1*) in Table 3.3. Once again, it can be observed that our method significantly outperforms other methods in terms of standard deviation of evaluation metrics across Human3.6M and 3DPW datasets. The tighter standard deviation bounds indicate that our method yields superior as well as more stable/robust performance over the SOTA methods.

3.4.3 Qualitative Results

Figure 3.4 visualizes qualitative comparison with SOTA methods where red and green arrows indicates regions with inaccurate and accurate SMPL fitting, respectively. More specifically, rows 1 & 2 show selected frames from a 3DPW dataset sequence where it can be observed that our method provides more accurate SMPL fitting in comparison to other methods. Similarly, rows 3 & 4 show selected frames from another 3DPW sequence where it can be observed that our method is able to accurately detect both the humans in the image (and get better SMPL fitting) while other methods only detect one human. Further, we can also observe that in row 3, where a large portion of the body is occluded, our method provides a much more plausible pose prediction. In regard to performance on Human3.6M dataset, rows 5-8 show results of SMPL fitting on selected frames from different sequences that include diverse poses like sitting on a chair and lying on the ground. It can be observed from these qualitative results that our method consistently provides more accurate SMPL fitting across all these sequences.

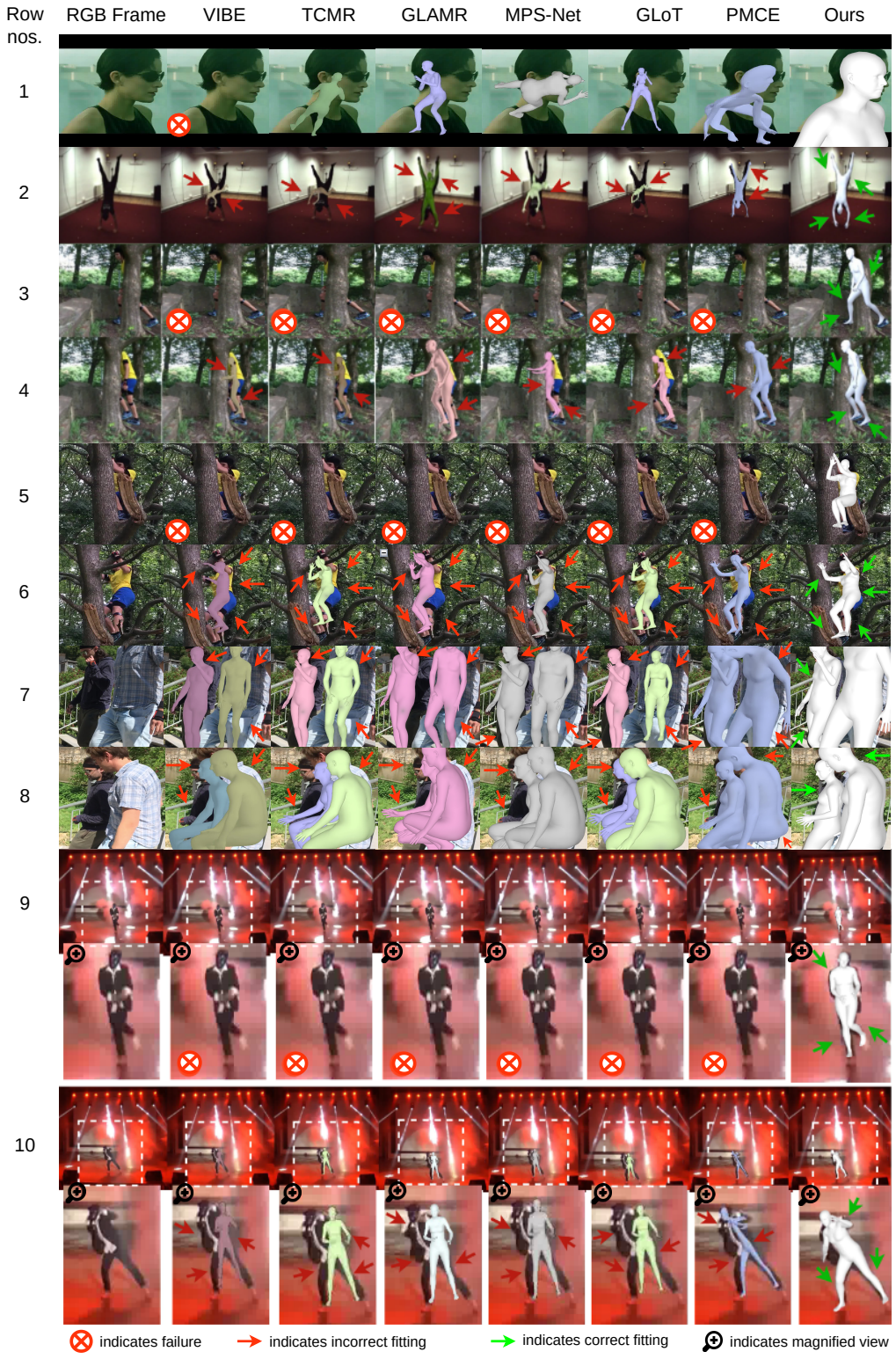


Figure 3.5 Qualitative comparison on challenging sequences on selected frames from dataset as well as in-the-wild internet images .

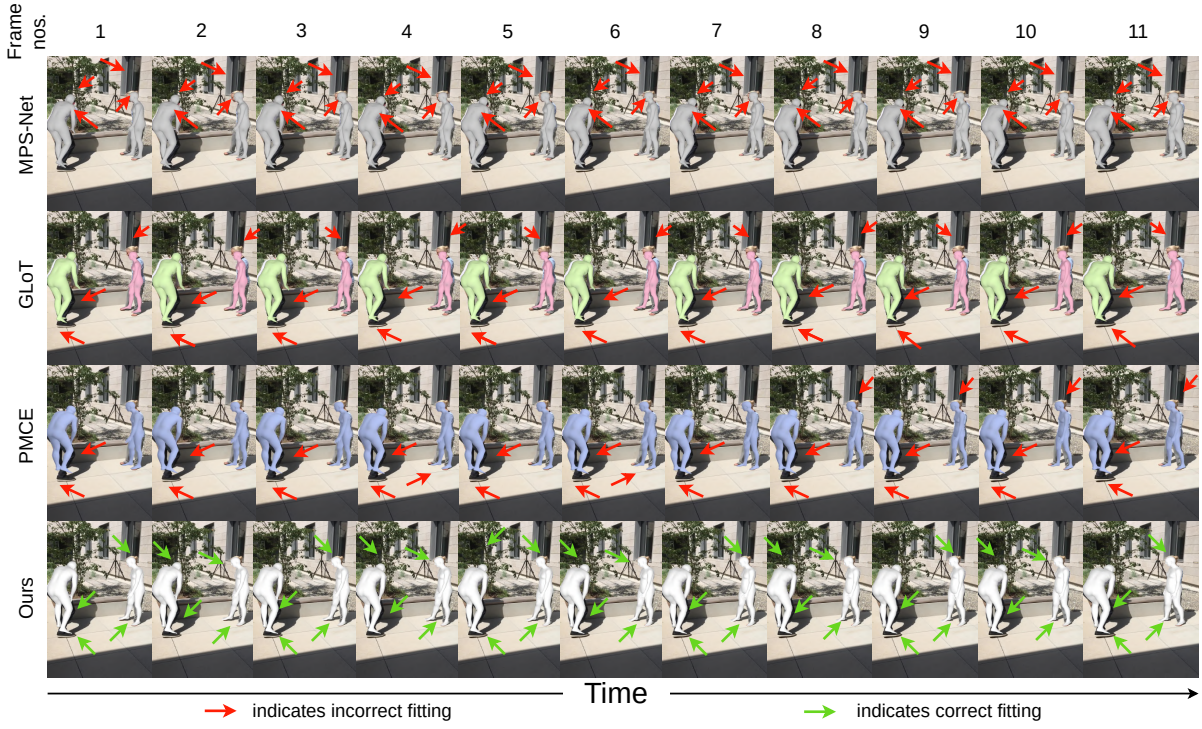


Figure 3.6 Qualitative comparison across frames on a sequence of 3DPW dataset. The green arrows show the regions with improved SMPL fitting compared to the red ones .

Further, in Figure 3.5 we also depict results on selected frames from a few challenging sequences selected from the aforementioned standard datasets, as well as in-the-wild images taken from the internet. The results demonstrate our method's ability to handle significantly challenging cases, where other methods often fail.

Additionally, in Figure 3.6, we demonstrate the results of the most relevant/recent methods and our method across multiple successive frames of a video sequence from 3DPW where the person is trying to lift a bag. It can be observed that our method provides more accurate SMPL fitting across the frames in comparison to other SOTA methods, demonstrating the superior temporal consistency ability of our method.

3.4.4 Generalization to Unseen Datasets

We also test the generalization ability of our method by evaluating its performance on completely unseen i3DB [47] and PROX [19] datasets. These datasets contain diverse scenarios and were not part of training data. As reported in Table 3.4, our method significantly outperforms existing SOTA on unseen datasets, demonstrating superior generalization ability of our method. We also show qualitative

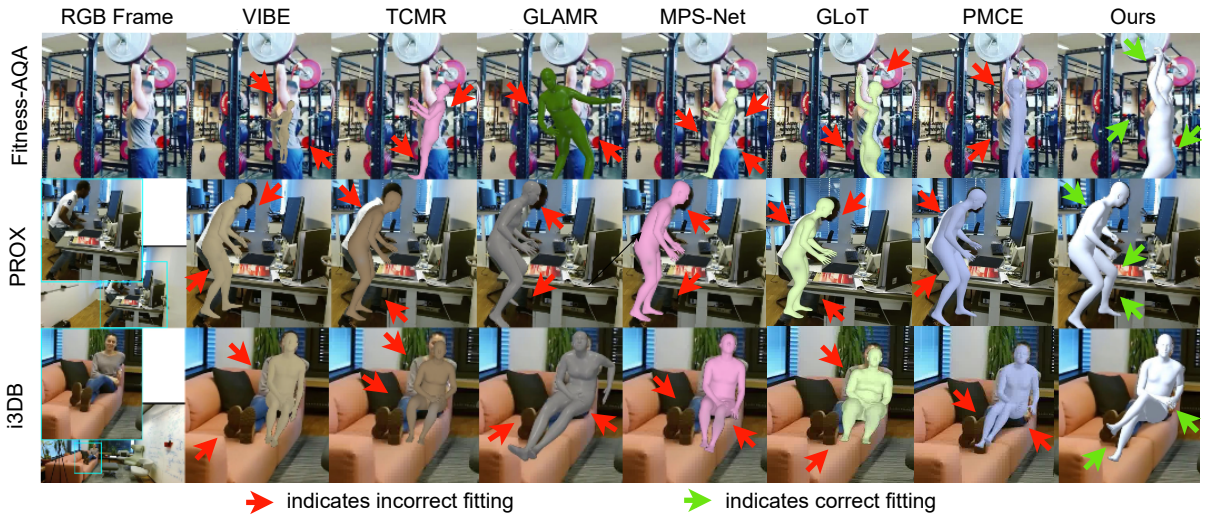


Figure 3.7 Qualitative comparison demonstrating generalization ability of our method on unseen datasets.

comparison for the same in Figure 3.7 where we can observe that our method yields more accurate pose and shape estimates in comparison to SOTA methods.

3.4.5 Ablation Study

We perform a detailed ablative study to analyze the contributions of different components of our method. Table 3.5 provides the quantitative ablative results, which list our final method's performance in row 8. We sequentially removed each component of our method and reported the performance drop in rows 1-7. More specifically, row 1 reports the results where we replace our body-aware feature encoder with generic ResNet. This leads to a drop in performance, demonstrating the contribution of the body aware features to our overall performance. In row 2, we train our network without using the per-frame pose and camera initialization. This too leads to a drop in the model performance. In row 3, we report the performance after removing the NSSM and Attention blocks. And we see a considerable drop in performance, demonstrating the importance of a well-designed feature aggregation strategy. In row 4, we further remove the 2D loss L_{2D} during training, and find a further drop in performance, demonstrating the contribution of L_{2D} (please see Table 3.8 for a more detailed analysis of the contribution of different loss terms to the model's performance). In row 5 & row 6, we report the performance of the model by individually removing the pose initialization and camera initialization. The results demonstrate that both pose initialization and camera initialization contribute individually to our method's performance. In row 7, we report the performance by removing the LSTM-based motion refinement component, and

Configuration	Human3.6M [26]			3DPW [62]		
	PA-MPJPE ↓	MPJPE ↓	ACC-ERR↓	PA-MPJPE ↓	MPJPE↓ ↓	ACC-ERR↓
1. Ours *w/o* Body-Aware Features (i.e., *w/o* H)	34.8	45.0	3.5	41.6	67.3	5.3
2. Ours *w/o* Per-Frame initialization (i.e., *w/o* $\{R^{init}, \theta^{init}\}$ and *w/o* ω^{init})	43.8	71.9	4.2	49.8	78.5	5.5
3. Ours *w/o* NSSM and *w/o* Attention (simply concatenation)	55.1	67.7	6.0	63.8	88.6	6.5
4. Ours *w/o* NSSM, *w/o* Attention and *w/o* L_{2D}	57.8	68.3	6.2	65.5	89.0	6.6
5. Ours *w/o* pose initialization (i.e., *w/o* $\{R^{init}, \theta^{init}\}$)	41.5	53.2	3.6	48.1	74.7	5.5
6. Ours *w/o* camera initialization (i.e., *w/o* ω^{init})	37.3	49.1	3.5	47.3	73.8	5.4
7. Ours *w/o* LSTM based refinement on coarse estimates (i.e., *w/o* ζ)	32.7	42.8	3.5	40.3	69.3	5.6
8. Ours - Final	**31.0**	**41.3**	**3.3**	**39.2**	**63.5**	**5.3**
9. Ours + AM on pose (i.e., AM on $\{R^{init}, \theta^{init}\}$)	33.8	44.8	3.4	42.7	68.0	5.3
10. Ours *w.* LSTM on Feature Space (followed by motion estimation)	39.2	47.2	4.0	44.3	72.9	5.8
11. Ours *w.* Transformer in place of LSTM (8 head transformer)	41.9	51.4	3.8	42.7	67.8	5.4
12. Ours *w.* Transformer in place of LSTM (16 head transformer)	43.3	53.7	3.9	49.1	73.4	5.5

Table 3.5 Ablation study on our method's performance while considering different architectural configurations. (Best results are in **bold**.)

once again find a drop in performance, especially in the ACC-Err metric, demonstrating the contribution of the motion refinement module.

We also report three additional ablative results in the last three rows of Table 3.5 as modifications to our proposed method. Specifically, row 9 reports the performance of the modified method by adding the self-attention on the body pose initialization to our method. However, unlike self-attention on body-aware features and camera pose, we empirically find that self-attention on body pose leads to a degradation in performance. One possible explanation for this degradation is that self-attention to body poses can sometimes be misleading due to the frequently repeating body poses in a temporal window (e.g. walking involves very similar body poses). Nevertheless, we observed that using self-similarity (NSSM) on body pose helps as it exploits the spatio-temporal ordering (see row 3 & row 4). In row 10, we report the performance of an alternate setup for temporal refinement where we use the LSTM to aggregate temporal features before passing them to the pose/shape predictor, thereby eliminating the coarse prediction step. However, this leads to a drop in performance. As an explanation to this, we hypothesize that learning pose/shape corrections is more conducive to LSTM and hence our method provides a better estimate of body pose. Finally, in row 11 and row 12, we report the performance of our method when using a transformer in place of LSTM for motion refinement. We report this in two settings - transformer with 8 heads (row 11), and transformer with 16 heads (row 12). We observe that simpler architectures perform better in this setting, as results obtained with the 8 head transformer, outperform the results obtained with the 16 head transformer. While, the results obtained with LSTM (our final method; row 8) outperform the results obtained with both the 8 head and 16 head transformers. This observation seems consistent with results also reported in [70] (section 4.2), where the authors report that they obtain better results with LSTM compared to a transformer.

Method	Human3.6M [26]			3DPW [62]		
	PA-MPJPE↓	MPJPE↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	ACC-ERR↓
PARE [34]	53.8	72.8	6.9	46.5	74.5	7.1
Ours w. PARE	42.6	66.4	4.1	49.3	67.3	5.4
CLIFF [39]	32.7	47.1	6.7	43.0	69.0	7.3
Ours w. CLIFF	31.2	42.7	5.7	39.3	65.1	6.7
HMR 2.0 [14]	33.8	45.3	3.8	44.4	69.8	5.6
Our Method	**31.0**	**41.3**	**3.3**	**39.2**	**63.5**	**5.3**

Table 3.6 Evaluation of our method with different per-frame initializers. (Best results are in **bold**.)

Window size	Human3.6M [26]			3DPW [62]			
	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	ACC-ERR ↓
8 frames	31.3	41.7	3.4	40.1	64.4	62.3	5.4
16 frames	**31.0**	**41.3**	**3.3**	**39.2**	**63.5**	**61.8**	**5.3**
32 frames	32.0	42.0	3.8	40.7	65.8	63.1	5.4

Table 3.7 Ablation study on performance of our method with different temporal window sizes. (Best is in **bold**.)

Configuration	Human3.6M [26]			3DPW [62]			
	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	ACC-ERR ↓
Our Method	**31.0**	**41.3**	**3.3**	**39.2**	**63.5**	**61.8**	**5.3**
w/o L_{SMPL}	43.5	56.4	3.7	47.8	76.7	79.2	5.8
w/o L_{SMPL} & L_{2D}	50.2	64.6	4.1	56.8	81.4	87.2	6.0
w/o L_{3D} & L_{2D}	37.8	49.2	3.5	43.6	70.4	72.9	5.5

Table 3.8 Ablation study on the effect of different loss terms on training. (Best is in **bold.**)

Method	MPI-INF-3DHP [45]		
	PA-MPJPE↓	MPJPE ↓	ACC-ERR↓
MPS-Net [65]	65.9	100.0	8.6
GLoT [57]	62.6	95.5	<u>7.7</u>
PMCE [68]	<u>55.0</u>	**80.4**	**7.4**
Our Method	**53.2**	<u>88.7</u>	8.1

Table 3.9 Quantitative comparison of our method with other SOTA methods on MPI-INF-3DHP dataset (as per *prot-1* described in section 3.4). Best results are in **bold** and second best are <u>underlined</u>.

Next, we evaluate the performance of our method with different per-frame initialization methods and report results in Table 3.6. It can be observed that our method consistently improves over the per-frame initialization methods (especially in terms of acceleration errors).

In Table 3.7, we report the performance of our method when using different temporal window sizes. Similar to existing works [33, 7, 65], we find that a temporal window of size 16 provides optimal performance.

In Table 3.8, we report the results of ablative experiment evaluating the impact of different loss terms (or a combination thereof) on model performance. In row 2, we report results of our model trained without using L_{SMPL} (i.e. only using L_{2D} and L_{3D}) during training. It can be observed that this leads to a drop in performance, indicating the contribution of L_{SMPL} loss term to the learning process. In row 3, we report results with further removing L_{2D} (i.e. training our model with only L_{3D}) and observe that this leads to a further degradation in performance, indicating the contribution of the L_{2D} loss term. In row 4, we report the results of removing L_{3D} and L_{2D} (i.e. using only L_{SMPL} for training), and observe that this too, leads to a drop in performance compared to our final method (row 1). However, this drop in performance is not as severe as that of observed in rows 2 and 3 (i.e., when L_{SMPL} is removed). This suggest that L_{SMPL} seems to be contributing more to the learning process than L_{3D} and L_{3D}.

3.5 Discussion

3.5.1 Noisy Annotations

Accurate ground truth annotations are essential for training supervised learning methods, and also for reliably evaluating different methods. However, ensuring availability of reliably annotated ground truth data is a major challenge in this problem domain due to the large amount of data that needs to be annotated. One popular dataset which seems to suffer from this issue, is the MPI-INF-3DHP [45]

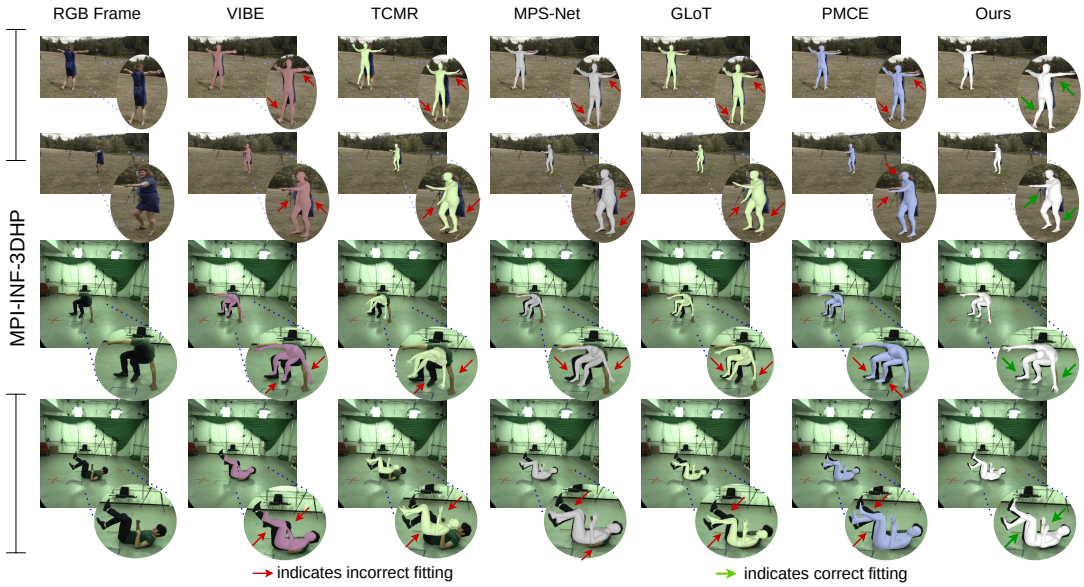


Figure 3.8 Qualitative results showing the estimated pose overlaid on the frames of videos from the test set of MPI-INF-3DHP.

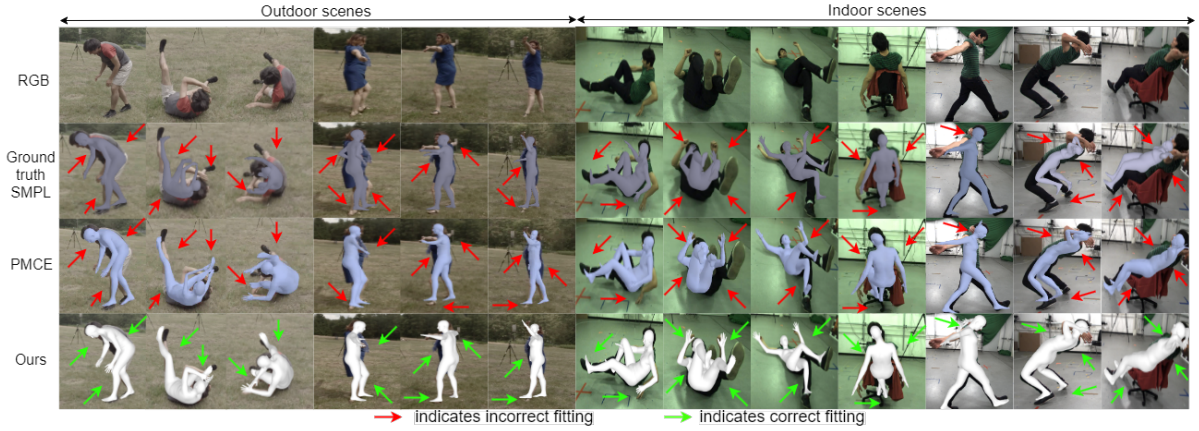


Figure 3.9 Visualization of noisy ground truth annotations on few sample sequence frames from MPI-INF-3DHP.

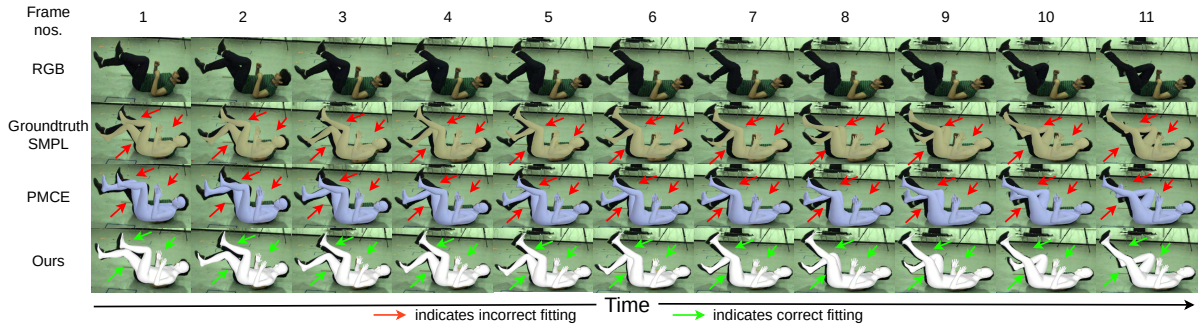


Figure 3.10 Additional examples of a sequence of frames from MPI-INF-3DHP showing noisy ground truth annotations for a sequence of consecutive frames.

dataset. Few such examples of noisy ground truth annotations are shown in Figure 3.9 (selected random frames) and Figure 3.10 (selected successive frames). We can attribute this to the fact that unlike other datasets, the capture setup used by MPI-INF-3DHP had only RGB cameras and lacked any additional sensing modality like IR sensors, depth sensors or inertial measurement units (IMU) during capture to obtain improved fidelity. It may be noted that since the MPI-INF-3DHP dataset only provides the 3D joint locations, we have estimated the SMPL body in Figure 3.9 by using SMPL shape parameter estimates from an off-the-shelf method, along with the provided ground truth 3D joint locations to obtain the posed SMPL body. This SMPL body is then overlaid on the RGB frames to get the visualization in the figure.

As a consequence, although our method yields superior qualitative results on MPI-INF-3DHP dataset, in comparison to other methods (as shown in Figure 3.8), our quantitative results on this dataset (as reported in Table 3.9) are not consistently best performing across all metrics. Thus, owing to the noisy ground truth annotations, PMCE seems to obtain better quantitative results as reported in Table 3.9, however, our qualitative results (shown in Figure 3.9 Figure 3.10) indicate superior SMPL fitting with our method. We argue that the use of the pre-trained body-aware feature network in our method regularizes the learning of our method, and prevents it from learning the noise in the dataset.

3.5.2 Recovering from Inaccurate Initialization

The spatio-temporal feature aggregation (STA) module incorporated in our method plays a crucial role in providing temporal context by analyzing both preceding and subsequent frames. This module enables our method to precisely recover the pose and shape, even in scenarios where CSE [48] and HMR2.0 [14] fail to provide adequate initialization. Figure 3.11 demonstrates qualitative examples where our method is able to estimate accurate pose and shape despite inaccurate initialization on selected frames from two different sequences in challenging scenarios (such as low light and occlusion). We can

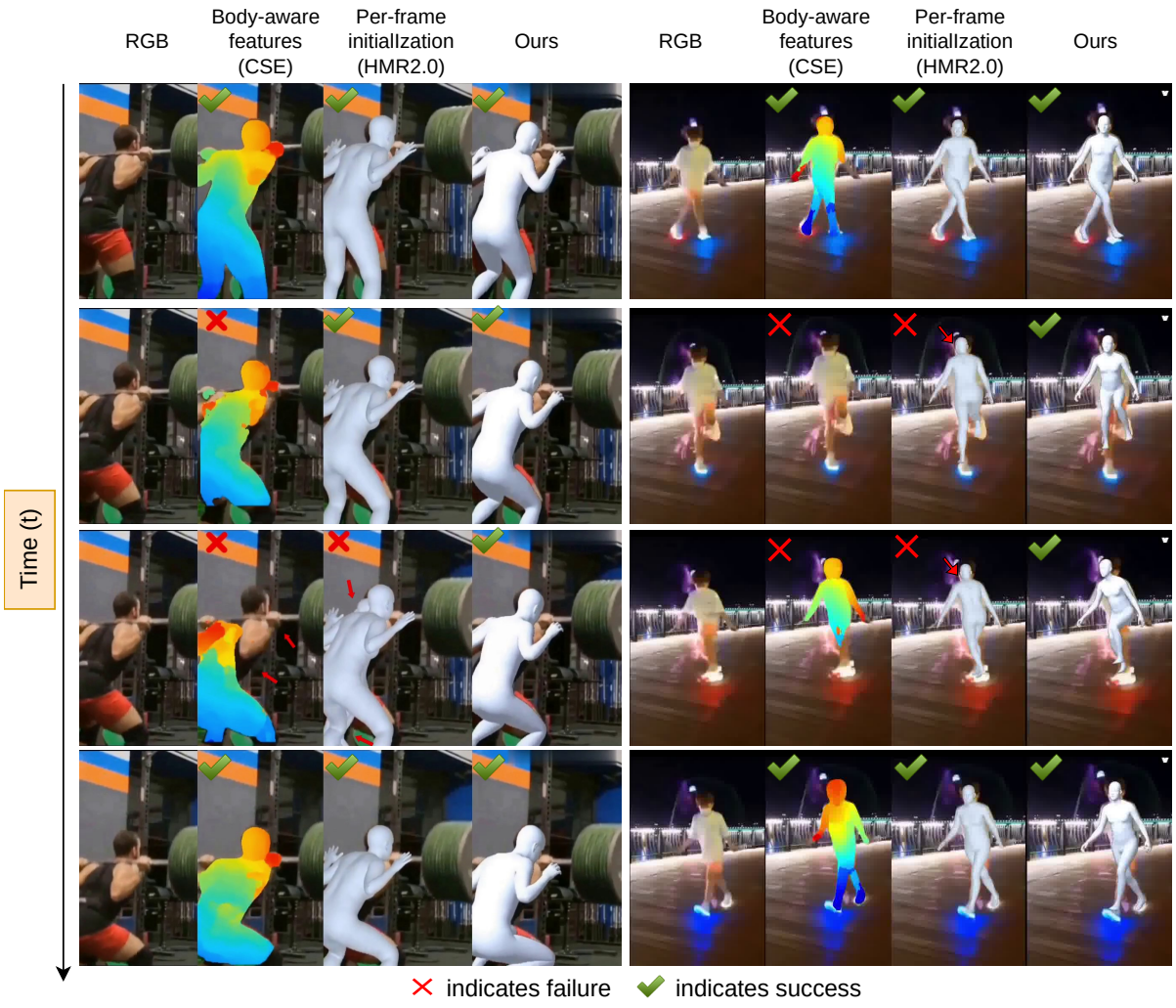


Figure 3.11 Qualitative results showing efficacy of our method in case of inaccurate initialization of body-aware features and per-frame SMPL fitting.

Figure 3.12 Failure cases involving extremely loose clothing or occlusion on in-the-wild sequences taken from the internet.

observe in row 2 that the CSE initialization misses the leg region in the left sequence, and completely misses the body in the right sequence. Similarly, in row 3, CSE misses the head region in the left sequence and misses the leg region in the right sequence. Likewise, it can be observed that in row 3, HMR2.0 is estimating two humans for the left sequence, while the orientation of the estimated human is wrong in both rows 2 and 3 for the right sequence. Nevertheless, we can observe that our method is able to accurately estimate the body pose and shape for these cases, as it is able to exploit the temporal context from the nearby frames with accurate initialization (rows 1 and 4).

3.5.3 Limitations and Future Work

Our method excels in estimating human pose and shape under diverse conditions including occlusions, low lighting, and intricate poses, yielding improved temporally consistent outcomes. Nevertheless, it still has certain limitations that future research needs to address. Similar to various other studies in human mesh recovery, our approach is restricted to retrieving SMPL parameters, thereby overlooking intricate details like clothing in human meshes. Moreover, as shown in Figure 3.12, our method can fail in certain scenarios with humans with extremely loose clothing as it is difficult to localize the underlying body in such scenarios. Similarly, in scenarios involving very severe occlusions of the human body, our method is prone to failure as the missing information makes it difficult to understand the pose of the occlulded body regions. We plan to explore an extension of our work to very loose clothing (e.g., robes/abaya) in the near future.

3.6 Conclusion

In this chapter, we described our novel method for recovering temporally consistent 3D human pose and shape from monocular video. Our method utilizes body-aware spatial features along with initial per-frame SMPL pose parameters to learn spatio-temporally aggregated features over a window. These features are then used to predict the coarse SMPL and camera parameters which are then further refined using a joint prediction of motion with LSTM. We demonstrate that our method consistently outperforms the SOTA methods both qualitatively and quantitatively. We also reported detailed ablative studies to establish relevance of key components of proposed method. As part of future work, it will be interesting to see extension of this work for humans with very loose garments.

Chapter 4

ConVol-E: Continuous Volumetric Embeddings for Human-Centric Dense Correspondence Estimation

In the previous chapter, we have described our method for temporally consistent 3D human pose and shape estimation. In this chapter, we describe our next problem statement - dense human-centric correspondence estimation. As introduced in section 1.1, dense human centric correspondence estimation is another important yet highly challenging problem in this domain. To this end, we present Continuous Volumetric Embeddings **(ConVol-E)**, a novel robust representation for dense correspondence-matching across RGB images of different human subjects in arbitrary poses and appearances under non-rigid deformation scenarios. Unlike existing representations [48, 24], ConVol-E captures the deviation from the underlying parametric body model by choosing suitable anchor/key points on the underlying parametric body surface and then representing any point in the volume based on its euclidean relationship with the anchor points. It allows us to represent any arbitrary point around the parametric body (clothing details, hair, etc.) by an embedding vector. Subsequently, given a monocular RGB image of a person, we learn to predict per-pixel ConVol-E embedding, which carries a similar meaning across different subjects and is invariant to pose and appearance, thereby acting as a descriptor to establish robust dense correspondences across different images of humans. We empirically evaluate our proposed embedding using a

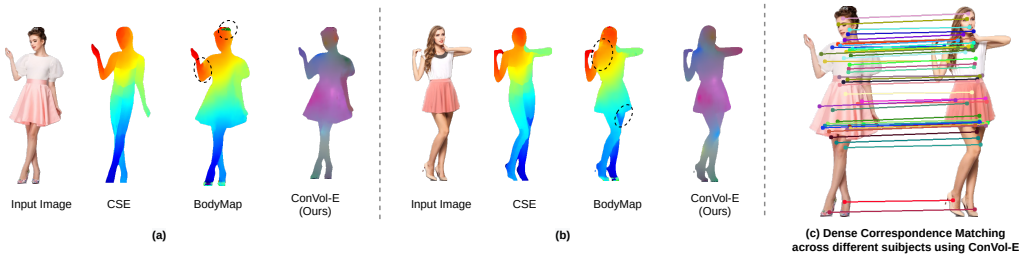


Figure 4.1 Comparing correspondences on 3D meshes when encoded with BodyMap [24] (left) and ConVol-E (right). Multiple false matching can be seen in the representation of BodyMap whereas, ConVol-E provides robust matching even in presence of loose clothing scenario.

novel metric and show superior performance compared to the state-of-the-art for the task of in-the-wild dense correspondence matching across different subjects, camera views, and appearance.

4.1 Introduction

Dense pixel-level understanding and labelling of humans in images is a well-attempted, yet challenging research problem in computer vision. Traditionally, it helps estimate body pose & shape, part semantics, dense correspondence/flow with key applications, including instance level segmentation, human tracking, gait analysis, 3D/4D human body reconstruction, virtual try-on, etc. In particular, the dense correspondence estimation can immensely benefit by associating each pixel with body pose/shape/appearance agnostic characterization. The key idea for establishing dense correspondences is to identify a per-pixel feature-based representation that can explain the relationship across different images of humans. The representation should be agnostic to appearance, i.e., it should carry the same meaning across images of different individuals. The formulation of such a representation is non-trivial owing to challenges such as large space of complex pose articulations, significant variations in body shape & size and large camera viewpoint variations. Moreover, the arbitrary and non-rigid nature of the garments causes deformations in the topology, which are extremely hard to model just from an image. The underlying representation should also understand the relationship between the garments and the body, especially under the loose clothing setup, which is a highly ill-posed problem.

Continuous Surface Embeddings (CSE) [48] is one such pixel-level representation that leverages the parametric human body model by estimating common embedding space between vertices of the SMPL [40] mesh and the pixels occupied by the humans in RGB image. However, SMPL doesn't capture high-frequency details such as clothing and hair, as shown in Figure 4.1(a,b). Recently, BodyMap [24] proposed to extend this representation to include these high-frequency details by assigning a three-dimensional embedding to the vertices of a human scan by extrapolating the CSE embedding, represented as simple RGB values, in the UV space based on the geodesic distance. This allows them to establish a relationship between different human scans by estimating similar extrapolated pixel-level embeddings. Subsequently, a network is trained to predict these dense three-dimensional embeddings in the form of color-coded RGB maps from the rendered images of the ground truth scans. However, such extrapolation of the RGB colors in UV space can not prevent distant vertices from having similar colors (as stated by the authors in their original paper [24]), thereby resulting in false matching across different regions of the human body, as shown in Figure 4.2. Additionally, it doesn't guarantee to produce consistent pixel-wise embedding for loose clothing scenarios as the effect of the geodesic distance will diminish in the far-apart regions of the UV space.

In this chapter, we propose *Continuous Volumetric Embeddings* (ConVol-E), a novel representation for establishing dense correspondence across humans in arbitrary poses and appearances. Our representation can handle any arbitrary point in the volume occupied by the human subject, i.e., each point in the 3D space is associated with a continuous value representing its relationship with the underlying

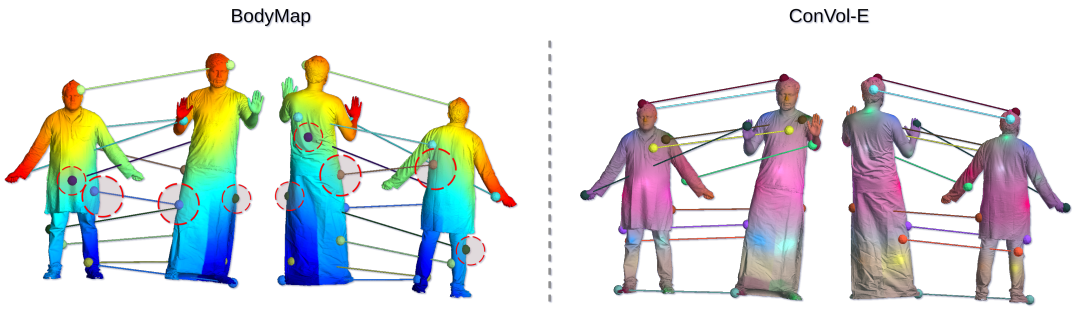


Figure 4.2 Comparing correspondences on 3D meshes when encoded with BodyMap [24] (left) and ConVol-E (right). Multiple false matching can be seen in the representation of BodyMap whereas, ConVol-E provides robust matching even in presence of loose clothing scenario.

parametric body model (SMPL to be specific). To ensure uniqueness and avoiding repetitions of the embeddings, we carefully designate *anchor nodes* on the SMPL surface. The embedding values for the *anchor nodes* are assigned such that the extrapolated embeddings vary significantly across different regions of the body and the volume, thereby ensuring a minimal chance of repeated values for far-away points in different directions (refer to supplementary for a detailed analysis). The anchor nodes are designated for a gender-neutral SMPL model, and their embedding values are extrapolated to all the vertices of a 3D human scan by registering the shape and pose parameters of SMPL with the scan. It is important to note that unlike BodyMap [24], our approach of volumetric extrapolation inherently addresses the challenge of far-away surface deformations (typically caused by loose clothing). Subsequently, we propose to learn dense pixel-wise ConVol-E values using a U-Net [55] encoder-decoder network given an input image of a human with a corresponding ground truth scan, which can later be inferred on in-the-wild images with high accuracy. Finally, the predicted embeddings are used for dense correspondence matching across different viewpoints and subjects, as shown in Figure 4.1(c). We perform a thorough evaluation of ConVol-E and compare with existing state-of-the-art methods and demonstrate applications like segmentation label transfer and appearance transfer.

4.2 Related Works

Estimating dense correspondence embeddings across different images of humans is an active area of research, with tons of potentially useful human-centric applications. The problem is well-attempted for general objects, and many solutions exist [8, 67, 54]. However, the complexity and difficulty increase drastically for humans, due to articulation, non-rigid deformation and clothing. Initial attempts were made to first solve the problem in a sparse way using pose estimation [6, 5, 64, 25], which mostly involves fitting either a human joint-skeleton or a parametric model such as SMPL [40], as introduced in chapter 2. While such solutions are widely used in the case of 3D human body re-

construction [72, 29, 66] and garment reconstruction from images [4, 74, 58], they can not provide dense universal correspondences across humans. DensePose [17] aims at establishing dense correspondence from 2D images to a surface-based representation of human body (SMPL), but requires a lot of hand-annotated data. Moreover, it either provides sparse image-to-surface correspondence landmarks (DensePose-COCO) or part-specific UV coordinates on top of the input image (DensePose-RCNN). It doesn't provide pixel-wise unique embedding, which is essential for dense correspondence matching.

Continuous Surface Embeddings (CSE) [48], as introduced in section 2.2 propose a drop-in replacement of DensePose by introducing a better and more flexible representation of correspondences using learnable positional embeddings. Given a canonical surface model of humans (SMPL), the idea is to estimate the deformation variant identity of any point on the canonical surface, and additionally, train a neural network to predict a per-pixel color-coded embedding corresponding to one such surface points, visible in the given image. Since these embeddings vary smoothly over a 3D manifold, they are continuous in nature. CSE provides a reliable way to match pixel-wise color-coded embeddings across different images of humans, however, it is not guaranteed that every pixel belonging to the human in the image is assigned some unique embedding. However, many pixels are left out, as SMPL does not cover all the intricate details, e.g. hair, clothing, skin deformations, etc. Nevertheless, it provides a universal intrinsic representation applicable to any human body, agnostic to appearance, body shape & pose. HumanGPS [59] tries to circumvent the issues in CSE prediction by proposing to use geodesic distances between corresponding points on the surface of a human scan, but it does not produce an explicit per-pixel mapping from image to scan, and additionally does not generalize well to loose clothing as reported in [24].

Recently, BodyMap [24] proposes to build on top of CSE to include the aforementioned intricate details. The authors propose to extrapolate CSE representation to human scans, by first registering a canonical SMPL mesh to the scan, and then extrapolating the embeddings from SMPL vertices to Scan vertices in the UV space based on the geodesic distance between them. This approach is reasonable as it becomes easier to render images for both the RGB and dense per-pixel correspondences to generate the training data. However, it does not handle loose clothing deformations very well. Modelling extreme deformations that lie far apart from the underlying body can not be achieved in UV space while preserving the uniqueness of the embedding values. Far apart values can have repeated values as they are only influenced by geodesically closer vertices. Although, the authors said that this can be mitigated partially by putting additional constraints on the learning side, however, this is still an inherent flaw in the representation that needs to be addressed.

Another relevant work, Virtual Correspondence [42], aims at establishing correspondences across different views of a human subject in a fixed pose, by fitting a common SMPL model to multi-view images of the subject. However, the method does not establish correspondences across different subjects or even the same subject in a different pose. Hence, we propose our appearance agnostic representation that can be used to establish dense correspondences across the different subjects, viewpoints, and mainly to handle loose clothing deformations during the matching.

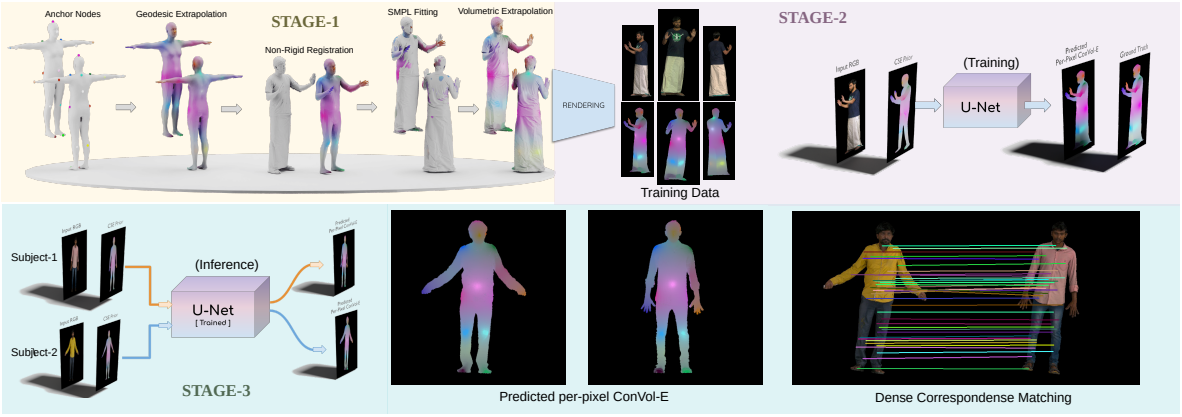


Figure 4.3 Overview of the three-stage method-pipeline for learning ConVol-E representation on the human images.

4.3 Our Method

We aim to find a novel pixel-wise unique characterization of in-the-wild clothed humans with the goal of establishing appearance-agnostic dense correspondences across multiple images. Our method consists of three key stages as shown in Figure 4.3. In Stage-1, we prepare the training data using high-quality human scans registered with corresponding SMPL meshes, which are rendered to generate RGB images and ConVol-E encoded maps. As part of Stage-2, we train a U-Net based encoder-decoder to predict the ConVol-E maps given RGB images as input. Finally, Stage-3 use the trained U-Net to predict ConVol-E maps for unseen images and then perform dense correspondence matching based on predicted embeddings.

4.3.1 Continuous Volumetric Embeddings

The proposed ConVol-E is the representation of an arbitrary 3D point embedded in the volume of an underlying parametric model. Unlike BodyMap [24], which defines these embedding in 2D UV space, ConVol-E encodes the volume around a parametric model in 3D Euclidean space. **ConVol-E** can be formally defined as a mapping $\mathcal{F} : \mathbb{R}^3 \rightarrow \mathbb{R}^k$ which takes a 3D point $x \in \mathbb{R}^3$ and assigns it an embedding vector $e \in \mathbb{R}^k$ (we choose k=3 in our experiments). The embedding vector precisely captures the information about where a given point x lies in the vicinity of a given parametric human model. More specifically, let $\mathcal{M} = \{\mathcal{V}, \mathcal{F}\}$ be a 3D human mesh scan (obtained either from an off-the-shelf 3D reconstruction solution or using a 3D scanning methods) and $\mathcal{M}^c = \{\mathcal{V}^c, \mathcal{F}^c\}$ be the parametric human mesh (SMPL [40] in our case) in neutral pose and shape. Here, $\mathcal{V}^c$ and $\mathcal{F}^c$ are the fixed number of vertices and faces of the canonical mesh in canonical pose and shape.

First, we select a set of anchor vertices $\mathcal{V}^c_{anchor} \subset \mathcal{V}^c$. These anchor vertices are distributed across the SMPL mesh at 19 key locations like pelvis, shoulder, feet etc. (see supplementary for visualization), and each anchor vertex is assigned a unique value denoted by $\hat{e}_{v^c} \in \mathbb{R}^k$. The embeddings for the remaining vertices $\mathcal{V}^c_{non-anchor} \in \mathcal{V}^c$ (such that $\mathcal{V}^c_{non-anchor} \cap \mathcal{V}^c_{anchor} = \phi$) are computed using the weighted geodesic distance from all the anchor vertices over the canonical mesh. Thus, we can compute embedding for every canonical mesh vertex $v^c_j \in \mathcal{V}_{non-anchor}$ as

$$e_{v^c_j} = \frac{\sum_{i=1}^{|\mathcal{V}_{anchor}|} w_i * \hat{e}_{v^c_i}}{\sum_{i=1}^{|\mathcal{V}_{anchor}|} w_i} \tag{4.1}$$

$$w_i = \frac{1}{g(v^c_j, v_i)} \tag{4.2}$$

where, $g(v^c_j, v^c_i)$ is the geodesic distance between $v^c_j \in \mathcal{V}^c_{non-anchor}$ and $v^c_i \in \mathcal{V}^c_{anchor}$.

It is important to note that, such embedding is only defined for vertices of canonical SMPL mesh $\mathcal{M}^c$. To obtain ConVol-E ebemdding for every vertex of a 3D human mesh $\mathcal{M}$ (in arbitrary pose and shape), we first perform a non-rigid registration with canonical SMPL mesh. This yields aligned SMPL mesh surface close to input 3D human mesh scan.

Subsequently, for each vertex $v_j \in \mathcal{V}$ of $\mathcal{M}$, we compute the nearest neighbor set $\mathcal{N}_j \in \mathcal{V}^c$ consisting of $p = 32$ closest vertices of the registered SMPL mesh, and assign the vertex an embedding value using the following equation:

$$e_{v_j} = \frac{\sum_{i=1}^{|\mathcal{N}_j|} w_i * e_{v^c_i}}{\sum_{i=1}^{|\mathcal{N}_k|} w_i} \tag{4.3}$$

$$w_i = \frac{1}{d(v_j, v^c_i)}, \forall v^c_i \in \mathcal{N}_j \tag{4.4}$$

where, $d(\cdot, \cdot)$ is the Euclidean distance. The choice of anchor vertices and the embedding values assigned to them is important and empirically chosen to allow highly diverse values during the extrapolation, so that each vertex is assigned a sufficiently unique embedding value.

Neighborhood Consistency Score : The underlying representation for dense correspondence estimation should be rich and varied enough to avoid repetitions in the feature space when extrapolated, otherwise different body parts would map nearby in the embedding space. More specifically, geodesically far-apart vertices should map far apart in the embedding space and vice-versa. Keeping this idea in mind, in order to quantify the efficacy of the proposed ConVol-E embeddings with other representations,

we design a novel metric named Neighborhood Consistency Score(NCS), for each vertex v_i of the scan mesh, and is calculated as follows:

$$NCS_i = (NCS_{near_i} + NCS_{far_i})/2 \tag{4.5}$$

$$NCS_{near_i} = \frac{1}{q^2}\sum_{i=1}^{q} min(|\mathcal{N}_{geo}^{rank} - \mathcal{N}_{emb}^{rank}|, q) \tag{4.6}$$

$$NCS_{far_i} = \frac{1}{q^2}\sum_{i=1}^{q} min(|\mathcal{F}_{geo}^{rank} - \mathcal{F}_{emb}^{rank}|, q) \tag{4.7}$$

where, $\mathcal{N}_{geo}^{rank}$ & $\mathcal{N}_{emb}^{rank}$ denotes the ranks (relative orders) of q-nearest neighbors of v_i in both geodesic and embedding space, and similarly, $\mathcal{F}_{geo}^{rank}$ & $\mathcal{F}_{emb}^{rank}$ denotes the ranks of q-farthest neighbors of v_i in both geodesic and embedding space (q is emprically set as 32). Thus, NCS penalizes the representation if the rank of these nearest/farthest neighbours in geodesic and embedding space doesn't match, i.e. j-th nearest-neighbor in geodesic space should be j-th nearest-neighbor in embedding space as well. Any neighbor among the q-neighbors of v_i can take maximum rank as q, so we divide by q^2 for normalization. Hence, NCS takes values between 0 and 1 where lower values are preferred. We compare the efficacy of ConVol-E with BodyMap[24] in subsection 4.4.4.

4.3.2 Learning Embeddings in Image Space

Given an input image $\mathcal{I}_{rgb}$ and the prior $\mathcal{I}_{cse}$, of size $\mathcal{W}$x$\mathcal{H}$ we train a U-Net [55] style encoder-decoder network to predict the per-pixel embeddings, represented as a three-channel feature map $\mathcal{I}_{\mathcal{E}}$. We train the U-Net by minimizing the $L1$ loss between the foreground pixels of predicted feature map $\mathcal{I}_{\mathcal{E}}$ and the corresponding ground-truth $\hat{\mathcal{I}}_{\mathcal{E}}$ generated in the previous stage.

It should be noted that we estimate the prior $\mathcal{I}_{cse}$ using a pre-trained Densepose-CSE [48] network, however with a key difference that instead of their default per-vertex embedding, we replace it with our proposed ConVol-E embedding. Additionally, BodyMap uses a Vision Transformer (ViT) [11] architecture instead of U-Net for predicting dense pixel-wise embeddings. However, the authors treat U-Net as a baseline and show that the performance of U-Net is on par with the ViT and convergence of ViT is slow and challenging in general. Therefore to avoid overkill, we decide to go with U-Net style encoder-decoder network.

4.3.3 Dense Correspondence Matching

Let $\mathcal{I}_{rgb_1}$ and $\mathcal{I}_{rgb_2}$ be two input RGB images with the known foreground, where we aim to establish dense correspondences between them. These images can have the same or different human subject, viewpoint, pose, clothing, etc. We predict the respective per-pixel ConVol-E embedding $I_{\mathcal{E}_1}$ and $I_{\mathcal{E}_2}$ using the U-Net trained in the previous stage. Following this, we can establish correspondences by

finding, for each pixel $p_1 \in \mathcal{I}_{rgb_1}$, the closest matching pixel $p_2 \in \mathcal{I}_{rgb_2}$, where the matching is established if the absolute difference in embedding values of pixels p_1 & p_2 is below a threshold, i.e. $|I_{\mathcal{E}_1}(p_1) - I_{\mathcal{E}_2}(p_2)| \leq t_{match}$. Further, to ensure more robustness in matching, we provide an additional bi-directional constraint that the matching pixels should mutually be the best matches of each other. More specifically, we consider a correspondence match between pixels p_1 and p_2 to be a valid correspondence if and only if p_1 is the best match of p_2 and p_2 is the best match of p_1. Figure 4.3 shows the obtained dense correspondences across different subjects and different viewpoints.

4.4 Experiments and Results

4.4.1 Dataset Details

We perform quantitative and qualitative evaluation of our method on two publicly available datasets - **3DHumans** [29] and **THUman2.0** [69]. 3DHumans, contains around 180 meshes of people in diverse body shapes in various garments styles and sizes, including a wide variety of clothing styles ranging from loose robed clothing to relatively tight fit clothing, like shirts and trousers. THUman2.0 contains 500 high-quality scans of multiple human subjects in arbitrary clothing and poses. We perform a random 80:20 split for training and testing for both datasets. We render RGB images and corresponding embeddings for each textured scan from 70 viewpoints using a Pre-computed Radiance Transfer (PRT)-based renderer.

4.4.2 Implementation Details

We adopt Pix2Pix [28] architecture to build our U-Net encoder-decoder network. Since, the final task involves regression and not synthesis, we do not require adversarial training and hence we remove the discriminator from the original Pix2Pix [28] architecture, retaining only the generator network. The generator is a U-Net style encoder-decoder, with 5 convolution and 5 transposed-convolution layers. We train the network to minimize L1 loss, with an initial learning rate of 0.0002 and the standard LR-decay. An input images resized to 512 × 512 resolution before passing through the encoder. For all the experiments, the network is trained with a batch size of 4 for 200 epochs.

4.4.3 Quantitative Evaluation Metric

Efficacy of ConVol-E: In terms of quantitative evaluation, we first intend to compare the efficacy of ConVol-E representation (i.e., ability to preserve the geodesic neighborhood in the embedding space) in comparison with BodyMap[24] representation. This would indicate the robustness of the underlying representations for the task of correspondence matching in 3D space itself (i.e., on the mesh surface). To this end, we compute Neighborhood Consistency Score (NCS) using Equation 4.5 for both ConVol-E

and BodyMap.

Representation	NCS (3DHumans [29]) ↓	NCS (THUman2.0 [69]) ↓
BodyMap [24]	0.955	0.957
ConVol-E (Ours)	**0.838**	**0.835**

Table 4.1 Comparison between BodyMap [24] and ConVol-E using the proposed Neighborhood Consistency Score.

Evaluation of Predicted 2D Embedding Maps: Inspired from [59], we develop another metric ***Geodesic Distance Error (GDE)*** to quantitatively evaluate the predicted pixel-level embedding against the ground-truth. Specifically, we find the geodesic distance between the corresponding ground truth and predicted vertices for each pixel, followed by computing the percentage of pixels having geodesic error less than a particular distance threshold. GDE is computed as:

$$GDE^t = \frac{1}{N}\sum_{i=1}^{N} g(v_i, v_i') < t \tag{4.8}$$

where t represents the distance threshold, $i \in \{1...N\}$ represents the indices of all foreground pixels and v_i, v_i' represent the corresponding ground-truth and predicted vertices. We compute threshold-specific numbers as that gives us additional information about performance of a method for different thresholds.

4.4.4 Quantitative Results

We compare our method with the current state of the art method BodyMap [24]. Firstly, we report *NCS* in Table 4.1 where our ConVol-E representation outperform BodyMap [24] by attaining lower *NCS* score on two datasets.

Table 4.2 report GDE values where our method significantly outperforms BodyMap across thresholds. The observed improvement in performance is even higher for smaller distance thresholds, indicating that our method is better than BodyMap [24] for correspondence estimation, both overall, and specially, for fine-grain correspondence estimation.

Further, in Table 4.3 we also report the $L1$ and $L2$ loss between the ground-truth per-pixel embeddings and the predicted per-pixel embeddings for both ConVol-E and BodyMap representations. It can be seen that the $L1$ and $L2$ loss values for ConVol-E are lower than the values for BodyMap for both the cases - RGB only and RGB with CSE prior, respectively. This shows that along with being better than BodyMap in terms of the richness of the representation, our ConVol-E representation is also more easily learnable by a U-Net style encoder-decoder network.

4.4.5 Qualitative Results

Figure 4.5 shows a qualitative comparison with BodyMap on test samples from 3DHumans and THUman2.0 datasets, and internet images. Figure 4.6 and Figure 4.7 show additional qualitative results

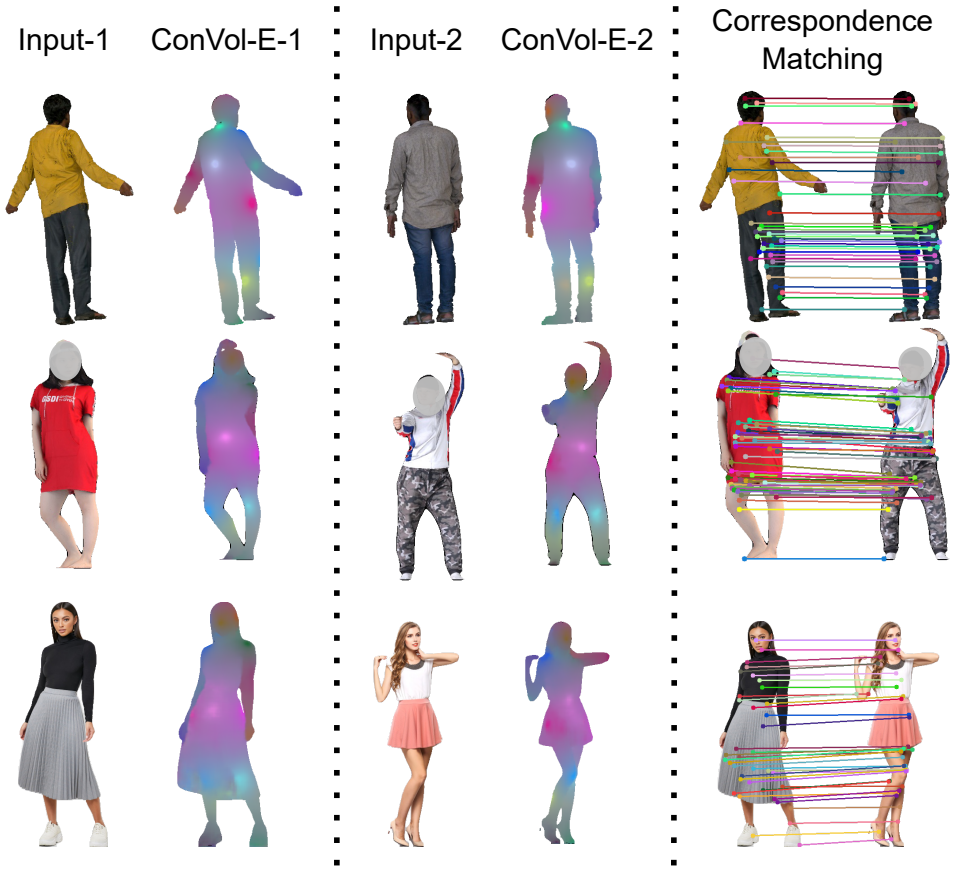


Figure 4.4 Predicted ConVol-E maps and dense correspondence matching on samples from 3DHumans[29] (first row), THUman2.0[69] (second row) & internet images (third row) [Some faces have been blurred according to the dataset T&C].

Method	GDE (3DHumans [29])			GDE (THUman2.0 [69])		
	5cm ↑	10cm ↑	15cm ↑	5cm ↑	10cm ↑	15cm ↑
BodyMap [24]: RGB-only	24.85	44.62	58.33	16.37	33.60	48.47
BodyMap [24]: RGB+CSE	25.45	45.02	58.65	21.77	40.76	55.44
ConVol-E (Ours): RGB-only	58.89	68.41	73.30	41.60	53.85	61.72
ConVol-E (Ours): RGB+CSE	63.98	72.40	76.17	51.94	62.35	68.72

Table 4.2 Comparison between BodyMap [24] and ConVol-E using **GDE** (eq. 4.8) for varying values of threshold t={5cm,10cm,15cm}.

Method	3DHumans [29]		THUman2.0 [69]	
	L1 ↓	L2 ↓	L1 ↓	L2 ↓
BodyMap [24]: RGB-only	0.064663	0.000713	0.178216	0.001234
BodyMap [24]	0.060392	0.000699	0.088864	0.000847
ConVol-E (Ours): RGB-only	0.046234	0.000212	0.090537	0.000412
ConVol-E (Ours)	0.038513	0.000191	0.061203	0.000280

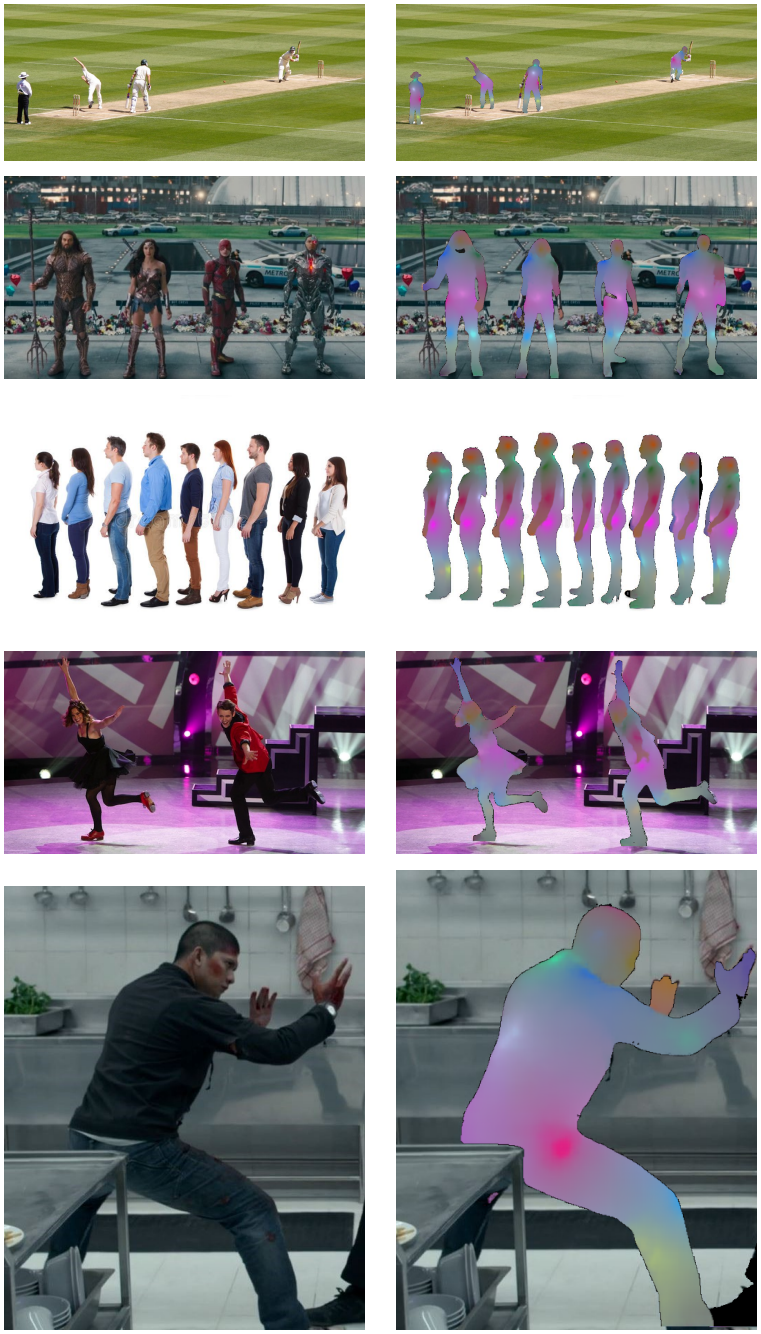
Table 4.3 Comparison of L1 and L2 loss between predictions and ground truth across datasets for BodyMap [24] and ConVol-E.

(a) RGB (b) CSE (c) BodyMap (d) Ours

Figure 4.5 Qualitative comparison between CSE [48], BodyMap [24] and the proposed ConVol-E representation on internet images.

(a) Input RGB Image (b) Output ConVol-E (c) Input RGB Image (d) Output ConVol-E

Figure 4.6 Visualization of ConVolE embeddings of our method on internet images.



(a) Input RGB Image (b) Output ConVol-E

Figure 4.7 Results of our method on internet images with multiple humans and occlusions.

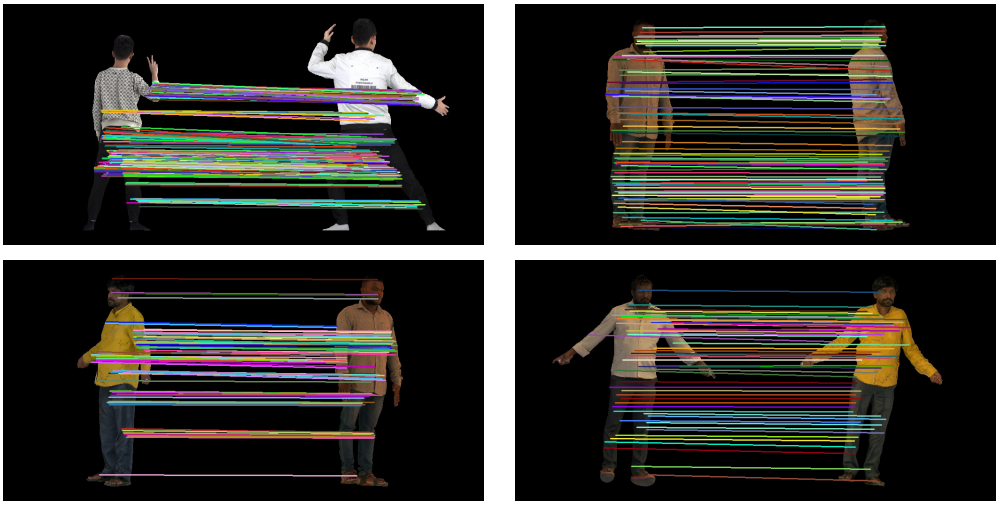


Figure 4.8 Additional results for correspondence matching across images (number of correspondences have been sampled for visualization.)

Method	THUman2.0 [69]		3DHumans [29]	
	L1	**L2**	**L1**	**L2**
Ours: RGB-only	0.090537	0.000412	0.046234	0.000212
Ours: RGB + PGN	0.066555	0.000301	0.043039	0.000192
Ours: RGB + CSE	0.061203	0.000280	0.038513	0.000191

Table 4.4 Effect of different input priors on $L1$ and $L2$ errors between predictions and ground truth of our method.

with the predicted ConVol-E embeddings. Figure 4.4 and Figure 4.8 show qualitative results for dense correspondence matching obtained using the ConVol-E embeddings. Overall, the results demonstrate ConVol-E's ability to generalize on challenging scenarios involving loose clothing deformations, where CSE and BodyMap fail drastically.

4.4.6 Ablation Study

4.4.6.1 Choice of Input Prior

We perform an ablative study by providing different inputs to our network and compare its performance. Specifically, we use following input setup : (1) Using only RGB images as the input, (2) Using output of Part Grouping Network (PGN) [15], which provides a semantic prior for different human body parts to the network and (3) Using *ConVol-E encoded CSE prior*. The results are reported in Table 4.4 where we can conclude that RGB + CSE prior outperforms any other setting, and yields lower $L1$ and $L2$ errors. This observation is supported by GDE values reported in Table 4.5 where RGB + CSE prior input setup outperform other two setups. This is due to the fact that CSE prior (encoded with proposed

Method	GDE (3DHumans [29])			GDE (THUman2.0 [69])		
	5cm ↑	10cm ↑	15cm ↑	5cm ↑	10cm ↑	15cm ↑
Ours: RGB-only	58.89	68.41	73.30	41.60	53.85	61.72
Ours: RGB + PGN	61.87	71.06	75.81	49.99	60.83	67.56
Ours: RGB + CSE	63.98	72.4	76.17	51.94	62.35	68.71

Table 4.5 Effect of different input priors for our method shown using GDE for varying values of threshold t=$\{5,10,15\}$.

ConVol-E embedding) provides a good initialization for the network to further refine the intricate details covering hair, clothing, etc. Further, it can also be observed that providing part-segmentation prior from PGN leads to some improvement over the RGB-only case. This is because, part segmentation labels provide the network with information about which pixel corresponds to which body part, and acts as a coarse initialization. However, results obtained with PGN prior are still not as good as those obtained with CSE prior, as CSE provides a more meaningful initialization.

4.4.6.2 Anchor Points Selection

As shown in Figure 4.9, the positions and ConVol-E values (RGB colors) of the 19 anchor points on the SMPL mesh are selected manually, to ensure that the extrapolation to remaining vertices does not result in any repetitions, which is not guaranteed by a random selection. We empirically found that manual placement is a good strategy for better efficacy. One can achieve a similar effect by taking a large number of randomly designated anchor points, e.g. designating all 6890 vertices of the SMPL mesh as anchor points. However, neural networks cannot learn such drastically varying embedding values, which will further become dramatic once extrapolated to human scans. On the other hand, placing anchor points symmetrically on the SMPL will ease the learning of the embeddings.

Table 4.6 shows a quantitative study of anchor point selection. We compare the manual placement of 19 anchor points with random placement and also with varying numbers of anchor points using the proposed metric Neighborhood Consistency Score on the meshes from 3DHumans [29] dataset.

No. of Anchor Points	NCS ↓
10 (Random)	0.975
19 (Random)	0.952
50 (Random)	0.913
100 (Random)	0.847
19 (Manual)	**0.838**

Table 4.6 Quantitative study regarding anchor point selection.

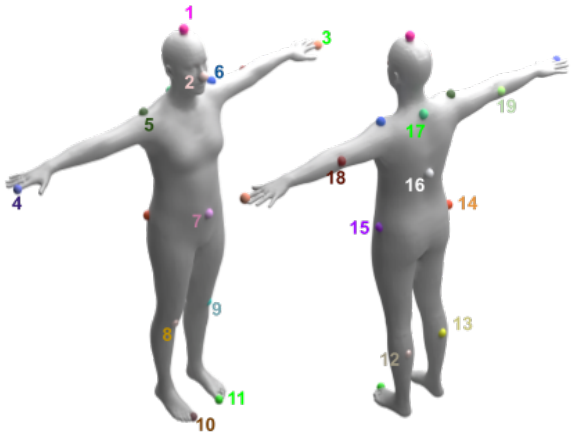


Figure 4.9 Manually selected anchor points on SMPL

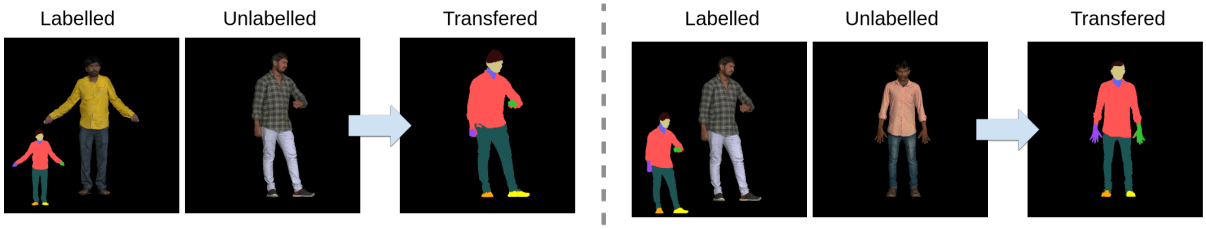


Figure 4.10 Segmentation label transfer performed with dense correspondences obtained with our method on 3DHumans test samples.

4.5 Applications

4.5.1 Segmentation Label Transfer

A potential application of our robust, dense representation is transferring image-based segmentation information across different subjects (given that the style of garments is similar). Given an unlabelled and labelled image of two human subjects I_u and I_l respectively, we can use our method of dense correspondence matching to add labels to the unlabelled image I_u. To do this, we iterate over the pixels of I_u and for each unlabelled pixel $p_u \in I_u$, we identify the *"matching"* pixel $p_l \in I_l$ and label p_u with the same label as p_l. The pixels are *"matched"* using the output of our dense correspondence matching network. Figure 4.10 shows the result of dense pixel-wise semantic label transfer where the two images have different subjects with significantly different body poses and appearances.

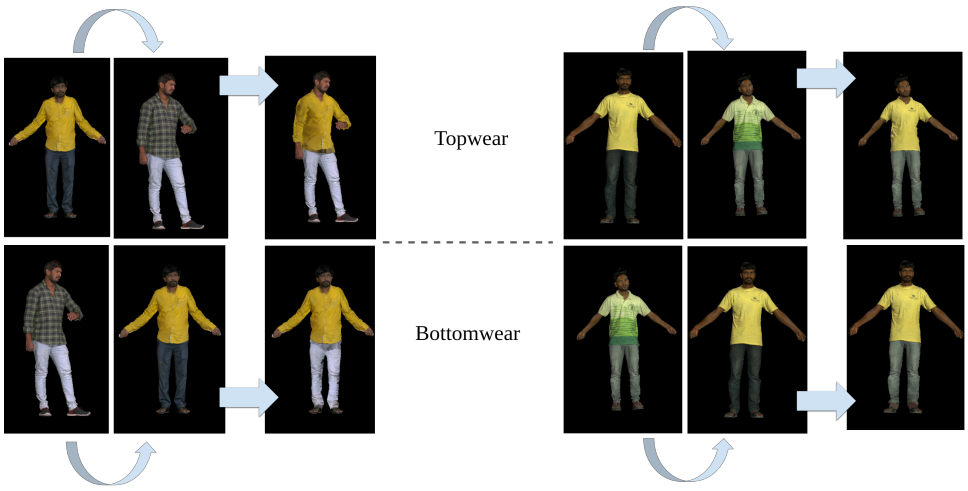


Figure 4.11 Garment appearance transfer performed with dense correspondences obtained from our method on 3D Humans test samples.

4.5.2 Garment Appearance Transfer

Another interesting application of the dense correspondence matching includes *garment appearance transfer* i.e., transferring the appearance of a garment worn by one person to another person. Consider an image I_1 with a human H_1 wearing a garment G_1 and another image I_2 with a human H_2 wearing a garment G_2. We want to transfer the appearance of the garment G_1 worn by human H_1 onto human H_2. The task is closely related to the application of virtual try-on, where we would like to see how a garment draped on a mannequin or worn by any other human would look on us.

This appearance transfer is achieved in the same way as semantic label transfer is performed. We first identify the pixels which belong to the garment in images I_1 and I_2. This segmentation can be obtained either by using a PGN [15]-like method on the input images or alternatively, we can also use the method described above to transfer the segmentation labels from the labelled image, if any. Once we have identified the pixels which belong to the required garment(s) in both images, then, for each pixel of interest $p_2 \in I_2$, we find the corresponding pixel in $p_1 \in I_1$ and transfer the RGB value of p_1 to p_2, as shown in Figure 4.11.

4.6 Limitations of Our Method

Figure 4.12 shows limitation of our method with first case (top row) involving the garment type (south-asian attire *Saree*) that is very loose, wrapped clothing as well as out of training distribution (from the training set). In the second case (bottom row), the failure is due to severe occlusion caused by another body with a similar appearance of the garment.

(a) Input RGB Image (b) Output ConVol-E

Figure 4.12 Failure cases of proposed ConVol-E.

4.7 Conclusion

We present **ConVol-E**, a robust representation for dense correspondence matching across RGB images of different human subjects in different poses/shapes/appearances. Existing methods fail to capture correspondences for points which do not lie in the vicinity of the body model. Our proposed volumetric representation can model arbitrary deviation from the underlying body model by making use of use of carefully chosen anchor nodes and volumetric extrapolation around the parametric body model. The proposed representation is easily learned with a simple U-Net-based architecture demonstrating superior qualitative and quantitative results. Further, we also show qualitative results on internet images, including, loose clothing scenarios. Finally, we discuss two potential applications of this work. Though the proposed embedding is inherently view-invariant, we would like to model the learning process in a way to provide explicit constraint over multi-view consistency. We can also explore explicit solutions for enforcing the embedding to be temporally consistent.

Chapter 5

Conclusion and Future Directions

In this chapter, we provide a summary of our contributions and discuss the potential future research directions that can be explored.

5.1 Summary of Our Contributions

In this thesis, we have focused on the task of 3D human pose, shape, and correspondence estimation from monocular input. This is a crucial problem with extensive applications in various industries like augmented/virtual reality, fashion, entertainment, healthcare, robotics, etc. However, it is also highly challenging due to large variations in the human body's pose, shape, and appearance, (external or self) occlusions, loose clothing details, difficulty in ensuring temporal consistency in results, etc. As part of this thesis, we focused on two key problems in this domain: (1) Temporally Consistent 3D Human Pose and Shape Estimation from Monocular Videos and (2) Dense Human-Centric Correspondence Estimation.

First, we proposed a novel method for temporally consistent 3D human pose and shape estimation from monocular videos. In this work, instead of using the traditionally used, generic ResNet-like features, our method uses a body-aware feature representation and an independent per-frame pose and camera initialization over a temporal window. This is followed by a novel spatio-temporal feature aggregation strategy by using a combination of self-similarity and self-attention over the body-aware features and the per-frame initialization. Together, they yield enhanced spatio-temporal context for every frame by considering the remaining past and future frames. These features are used to predict the pose and shape parameters of the human body model, which are further refined using an LSTM.

Next, we expanded our focus to the task of dense correspondence estimation between humans, which requires understanding the relations (represented using dense correspondences) between different body regions, including the clothing details, of the same or different human(s). In this work, we proposed Continuous Volumetric Embeddings (ConVol-E), a novel robust representation for dense correspondence estimation across RGB images of different human subjects in arbitrary poses and appearances under non-rigid deformation scenarios. Unlike existing representations, ConVol-E captures the devi-

ation from the underlying parametric body model by choosing suitable anchor/key points on the underlying parametric body surface and then representing any point in the volume based on its Euclidean relationship with the anchor points. This allows us to represent any arbitrary point around the parametric body (clothing details, hair, etc.) by an embedding vector. Subsequently, given a monocular RGB image of a person, we learn to predict per-pixel ConVol-E embedding, which carries a similar meaning across different subjects and is invariant to pose and appearance, thereby acting as a descriptor to establish robust, dense correspondences across different images of humans.

Overall, this thesis advances the current state-of-the-art in estimating human pose, shape, and correspondences. We identify and address some of the key limitations of existing methods and introduce novel strategies that overcome these limitations and achieve state-of-the-art performance. We hope our research benefits the broader community and lays the foundation for developing practical computer vision systems with real-world applications in this domain.

5.2 Future Research Directions

Although our work significantly improves the existing state-of-the-art, a few limitations remain. We encourage future research to explore these limitations/challenges as described below:

- **More Efficient Human Pose and Shape Estimation:** Existing human pose and shape estimation methods rely on massive architectures that are memory and compute-intensive. This makes them slow, and limits their applicability for real-time use cases. Future research should focus on methods for obtaining similar (or better performance) using lightweight architectures.
- **Multi-view and Temporally Consistent Correspondence Matching:** As already discussed in section 4.7, our work on dense correspondence matching can be extended to incorporate explicit multi-view and temporal consistency constraints, enhancing the method's applicability in dynamic and multi-view settings.
- **Handling Extremely Loose Clothing:** While our proposed solutions can handle loose clothing to some extent, handling extremely loose clothing like a saree remains a challenge, as such loose clothing make it difficult to localize the underlying human body.
- **Inter-human Evaluation for Dense Correspondence Matching:** For our work (and other existing works) on dense correspondence estimation, training and quantitative evaluation is possible only for intra-human scenarios. This is because different human scans have different topologies, which makes establishing the ground-truth dense correspondences between them challenging.

Publications

Thesis Publications

- **Amogh Tiwari**, Pranav Manu, Nakul Rathore, Astitva Srivastava, Avinash Sharma; *ConVol-E: Continuous Volumetric Embeddings for Human-Centric Dense Correspondence Estimation*; **4th Image Matching Workshop at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPRw), 2023**.

- Sushovan Chanda, **Amogh Tiwari**, Lokender Tiwari, Brojeshwar Bhowmick, Avinash Sharma, Hrishav Barua; *Enhanced Spatio-Temporal Context for Temporally Consistent Robust 3D Human Motion Recovery from Monocular Videos*; **arXiV; under review at the International Journal for Computer Vision (IJCV)**.

Other Publications

- Kanishk Jain, Varun Chhangani, **Amogh Tiwari**, K Madhava Krishna, Vineet Gandhi; *Ground then navigate: Language-guided navigation in dynamic scenes*; **IEEE International Conference on Robotics and Automation (ICRA), 2023**

Bibliography

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.

[2] F. Baradel, R. Brégier, T. Groueix, P. Weinzaepfel, Y. Kalantidis, and G. Rogez. Posebert: A generic transformer module for temporal 3d human modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[3] 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.

[4] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

[5] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016.

[6] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[7] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021.

[8] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016.

[9] K. Crane, M. Livesu, E. Puppo, and Y. Qin. A survey of algorithms for geodesic paths and distances. *arXiv preprint arXiv:2007.10430*, 2020.

[10] K. Crane, C. Weischedel, and M. Wardetzky. The heat method for distance computation. *Communications of the ACM*, 60(11):90–99, 2017.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[12] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *cvpr*, volume 96, page 73, 1996.

[13] G. Georgakis, R. Li, S. Karanam, T. Chen, J. Košecká, and Z. Wu. Hierarchical kinematic human mesh recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 768–784. Springer, 2020.

[14] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023.

[15] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network. In *ECCV*, 07 2018.

[16] R. A. Guler and I. Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019.

[17] R. A. 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.

[18] M. Gyeongsik and K. M. Lee. Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Proceedings of the IEEE/CVF European conference on computer vision*, 2020.

[19] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[20] C. He, J. Saito, J. Zachary, H. Rushmeier, and Y. Zhou. Nemf: Neural motion fields for kinematic animation. In *Advances in Neural Information Processing Systems*, 2022.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[23] C. Hou and K. Behdinan. Dimensionality reduction in surrogate modeling: A review of combined methods. *Data Science and Engineering*, 7, 08 2022.

[24] A. Ianina, N. Sarafianos, Y. Xu, I. Rocco, and T. Tung. Bodymap: Learning full-body dense correspondence map. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13276–13285, 2022.

[25] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European conference on computer vision*, pages 34–50. Springer, 2016.

[26] 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, 2013.

[27] U. Iqbal, K. Xie, Y. Guo, J. Kautz, and P. Molchanov. Kama: 3d keypoint aware body mesh articulation. In *2021 International Conference on 3D Vision (3DV)*, pages 689–699. IEEE, 2021.

[28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[29] S. S. Jinka, A. Srivastava, C. Pokhariya, A. Sharma, and P. J. Narayanan. Sharp: Shape-aware reconstruction of people in loose clothing. *International Journal of Computer Vision*, Dec. 2022.

[30] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[31] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019.

[32] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[34] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021.

[35] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019.

[36] G.-H. Lee and S.-W. Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12375–12384, 2021.

[37] J. Li, S. Bian, C. Xu, G. Liu, G. Yu, and C. Lu. D&d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, 2022.

[38] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021.

[39] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022.

[40] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–248:16, 2015.

[41] Z. Luo, S. A. Golestaneh, and K. M. Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[42] W.-C. Ma, A. J. Yang, S. Wang, R. Urtasun, and A. Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. 2022.

[43] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[44] M. Marsot, S. Wuhrer, J.-S. Franco, and S. Durocher. A structured latent space for human body motion generation. *arXiv preprint arXiv:2106.04387*, 2021.

[45] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[46] J. S. Mitchell, D. M. Mount, and C. H. Papadimitriou. The discrete geodesic problem. *SIAM Journal on Computing*, 16(4):647–668, 1987.

[47] A. Monszpart, P. Guerrero, D. Ceylan, E. Yumer, and N. J. Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Transactions On Graphics (TOG)*, 38(4):1–15, 2019.

[48] N. Neverova, D. Novotny, M. Szafraniec, V. Khalidov, P. Labatut, and A. Vedaldi. Continuous surface embeddings. *Advances in Neural Information Processing Systems*, 33:17258–17270, 2020.

[49] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

[50] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018.

[51] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[52] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021.

[53] D. L. Ripley and T. Politzer. Vision disturbance after tbi. *NeuroRehabilitation*, 27(3):215, 2010.

[54] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017.

[55] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[56] J. A. Sethian. Fast marching methods. *SIAM review*, 41(2):199–235, 1999.

[57] X. Shen, Z. Yang, X. Wang, J. Ma, C. Zhou, and Y. Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8887–8896, 2023.

[58] A. Srivastava, C. Pokhariya, S. S. Jinka, and A. Sharma. Xcloth: Extracting template-free textured 3d clothes from a monocular image. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 2504–2512, New York, NY, USA, 2022. Association for Computing Machinery.

[59] F. Tan, D. Tang, M. Dou, K. Guo, R. Pandey, C. Keskin, R. Du, D. Sun, S. Bouaziz, S. Fanello, P. Tan, and Y. Zhang. HumanGPS: Geodesic PreServing Feature for Dense Human Correspondence. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021.

[60] S. Tripathi, S. Ranade, A. Tyagi, and A. Agrawal. Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In *2020 International Conference on 3D Vision (3DV)*, pages 311–321. IEEE, 2020.

[61] A. VENKAT. *MONOCULAR 3D HUMAN BODY RECONSTRUCTION*. PhD thesis, International Institute of Information Technology Hyderabad, 2020.

[62] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018.

[63] X. Wang, R. B. Girshick, A. K. Gupta, and K. He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2017.

[64] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[65] W.-L. Wei, J.-C. Lin, T.-L. Liu, and H.-Y. M. Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13211–13220, 2022.

[66] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022.

[67] F. Yang, X. Li, H. Cheng, J. Li, and L. Chen. Object-aware dense semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2777–2785, 2017.

[68] Y. You, H. Liu, T. Wang, W. Li, R. Ding, and X. Li. Co-evolution of pose and mesh for 3d human body estimation from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14963–14973, 2023.

[69] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021.

[70] Y. Yuan, U. Iqbal, P. Molchanov, K. Kitani, and J. Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022.

[71] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022.

[72] Z. Zerong, Y. Tao, L. Yebin, and D. Qionghai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2021.

[73] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.

[74] H. Zhu, L. Qiu, Y. Qiu, and X. Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3845–3854, June 2022.

[75] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.