An efficient approach for sequential human performance capture from monocular video

Jianchun Chen

CMU-RI-TR-22-58

August 18, 2022

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Dr. Fernando De la Torre Frade, *chair*
Dr. Dong Huang
Donglai Xiang

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics.

To all advisors, mentors, friends and family who has supported me unconditionally.

Abstract

Human performance capture from RGB videos in unconstrained environments has become very popular for applications to generate virtual avatars or digital actors. Modern approaches rely on neural network algorithms to estimate geometry directly from images, resulting in a coarse representation of the shape of the person. On the other hand, optimization-based approaches that use shape-from-silhouette provide a more accurate reconstruction but they are computationally expensive and require a good initialization. In this work, we propose a learning-based approach for optimizing fine geometry information (e.g., clothes, wrinkles) from monocular RGB cameras. In particular, we sequentially recover different shape details (e.g., average shape without cloths, clothing, wrinkles) using separate neural networks. At each level, our network takes the sparse gradient of body mesh vertices generated from 2D off-the-shelf silhouette/normal supervisions and predicts dense gradients to update the body shape. Our networks are able to converge within a few interactions and achieve pixel-level accuracy. In addition, our method shares the benefit of classical optimization methods under challenging poses and novel views. As demonstrated by the experimental validations, our strategy is both effective and efficient across a wide range of datasets.

Acknowledgments

I would like to express my gratitude to so many people that make up my wonderful memory at CMU.

Firstly, I would like to thank my advisor Prof. Fernando De la Torre Frade. I am fortunate to work with Fernando for two years. I treasure his wisdom in guiding me to virtual human research, which is now proved to be one of the most popular areas in both academia and industry. His research taste and high standard toward research works also impressed me a lot. I firmly believe that the lesson I learned here will benefit my remaining academic career.

I would like to also acknowledge my thesis committee. Prof. Dong Huang gives me valuable comments and encourages me to find solutions for the corner cases of my project in practical use. I am thankful for Donglai's expertise in virtual human and his generous help from high-level idea to practical implementation, and his thoughtful suggestions in enhancing the performance for future work.

Moreover, I am grateful to Dr. Jayakorn Vongkulbhisal for his mentorship ever since I work with Fernando. Jayakorn has given me so much detailed research advice, from whom I learned to be patient and keep my own pace of research. In the meanwhile, I would like to thank all labmates at Human Sensing Lab, particularly Jinqi, Quankai, and many others. I am deeply impressed by the fantastic lab culture, the opportunity to be engaged in diverse research topics, and the accompany at the late night or on weekends.

Finally, I would like to thank my family that gives me consistent support during the whole journey, especially during the first year of pandemic. There's not always an easy moment and it is your love that makes me brave enough to challenge the toughest job to be a better man.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Today's virtual/augmented reality, telepresence, gaming, or digital actors in movies all heavily rely on capturing high-fidelity human geometries and dynamics. A major research area is how to capture persons and clothing from individual RGB images [35, 51] or videos [45, 47] in the hopes of producing 3D virtual humans. Popular 3D pose estimation techniques use the SMPL body model (or variations) [18, 23] to successfully provide accurate 3D body shapes; however, capturing cloth motion is still difficult due to the complex geometry, dynamics, and ambiguities introduced by monocular RGB photos and movies.

To solve this issue, several techniques estimate 3D clothing based on the contour of the naked person by progressively reducing disparities between the rendered and detected silhouette images. Due to the ambiguity caused by single-view silhouette loss, these methods rely on either highly constrained priors of mesh smoothness [2], cloth shape [45], highly constrained scenarios of self-rotating video [2, 17, 50], or a pre-scanned avatar [47]. Furthermore, the aforementioned optimization's runtime is typically unacceptable for many applications of interest.

Other researchers, including [8, 15, 16, 35, 36, 52] extend the capability of deep learning to reconstruct the 3D human body with clothing in a data-driven way. A deep neural network takes a single/sparse-view RGB image as an input to learn pixel-aligned features for predicting an implicit function of a 3D person with intricate

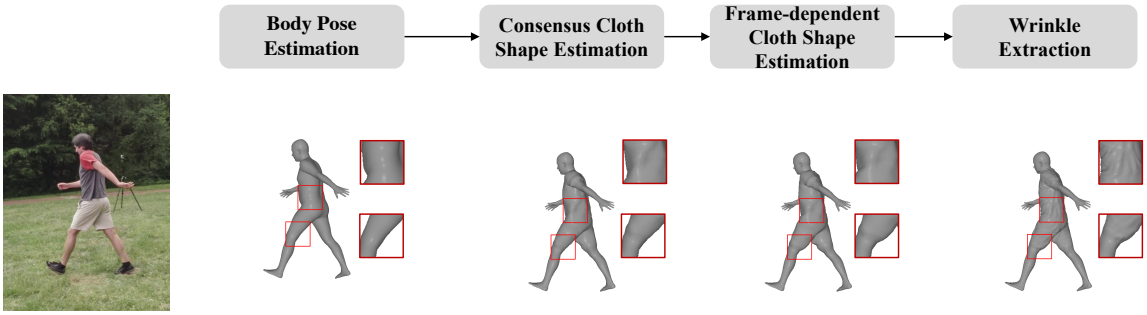

Figure 1.1: Pipeline of our proposed sequential cloth capture method. Given a monocular RGB video in the wild (the upper row), we reconstruct a personalized template shape, a frame-dependent deformation and extract wrinkle details in a coarse-to-fine manner.

textile geometry. These techniques offer a fast inference speed, as well as, a surprising generalization ability to in-the-wild pictures. However, existing efficient algorithms lack accuracy and robustness in presence of difficult poses, textures, or perspectives, even when a good 3D pose is given to the algorithm. Furthermore, the one-step inference cannot ensure the temporal smoothness of the output 3D form as the neural network directly accepts RGB pictures as an input separately.

To reduce the difficulty of directly predicting a holistic cloth shape, this paper proposes a sequential shape recovery method, where a set of networks learn different shape details. Given an input image, we first estimate the underlying body shape using SMPL model. Later, two independent networks estimate the average cloth shape and pose-dependent cloth deformations. The final network is able to extract the wrinkles. Fig.1.1 illustrates how we are able to estimate the shape in a coarse-to-fine manner. Concretely, our approach follows the *learning to optimize* paradigm over different scale resolutions of shapes. Each module independently takes the gradient from 2D human appearance supervision (i.e. silhouette and surface normal) as an input and predicts a rectified gradient per-vertex to update. Such techniques allow us to obtain a personalized human avatar for input video, frame-wise cloth motions and realistic wrinkles. Our approach fully utilizes neural networks' benefit of generalizing clothing knowledge from existing 3D datasets to single/sparse view scenarios in the wild and allows a significantly efficient inference speed. Contrarily, we follow the classical optimization process to separate the input picture from the silhouette and surface

normal loss, enabling a multi-step inference with extra resilience, which increases accuracy and robustness. Compared with dense optimization, our predicted gradient is more accurate and converges within a limited number of iterations, which greatly accelerates the inference time. Further, we are able to simulate the input gradient from purely geometries, without the need of high-quality texture data. Extensive results under different settings demonstrate the superiority of our method in accuracy and efficiency in 3D human performance capture.

1.2 Contribution

We summarize our three-fold contribution as below.

- We propose a sequential human performance capture approach that reconstruct total human cloth shape by progressively predicting an average cloth shape, a frame-dependent deformation and high-frequency wrinkle details.
- Our *gradient rectification network* iteratively predicts cloth deformation given the 2D silhouette/normal alignment loss. The network is trained with only 3D cloth geometry data and achieves robustness in wild inference.
- Our method produces precise human performance capture results with plausible wrinkle details, while significantly reduces the optimization runtime.

Chapter 2

Background

2.1 Optimization-based Human Cloth Capture

The classic human cloth capture methods formalize the problem as an optimization process that iteratively estimates a non-rigid deformation from naked body to fit the input image. Since directly estimating per-vertex offsets from silhouette image is highly unstable, early method [47] can only start with a pre-scanned template and support a small range of deformation. For wild videos, researchers impose different constraints to obtain plausible cloth shape. Particularly, VideoAvatar [2] hand-crafts multiple regularization terms and decompose the cloth shape as a global consensus shape and a frame-wise deformation. For self-rotating videos, SelfRecon [17] proposed to jointly optimize cloth geometries with textures by leveraging a backend neural renderer. Another popular trend is to parameterize the non-rigid deformation by PCA [45], deformation graph [14] or garment parameter [40]. These methods ensure a convergence of 2D image fitting loss and a spatial-temporal coherence, but only proposed in highly constraint cases, and take significant run time.

2.2 Learning-based Human Cloth Capture

Inspired by the success of deep neural networks, pioneer works [4, 51] start to predict 3D human model with cloth directly from input RGB image. Concretely, Bhatnagar

et al. [4] regress PCA controlling parameter of multiple types of garments, and Zheng et al. [51] predict a coarse 3D volume of clothed human then refines the surface normal in the frontal view. The main drawback of these methods is due to the incapability of the global feature to describe highly complex geometry details of the cloth. As implicit function becomes the new fashion of 3D representation, PiFU [16, 35, 36] generates implicit 3D human shape leveraging pixel-aligned features and thereby performs more realistic cloth geometry and wrinkle details. However, due to the lack of large-scale 3D human scans, the generalization ability of these methods are challenged by novel poses and views. Following works [8, 15, 46, 52] greatly alleviate this problem by adding SMPL shape prior. Overall, these methods allow fast inference speed with plausible reconstruction results, but 1) does not guarantee an accurate alignment with input image; 2) hard to control a spatial-temporal consistency.

2.3 Human Pose and Shape Estimation

Initially, human pose estimation refers to localize the body keypoint in 2D [6, 42] and 3D [28, 32]. With 3D body models [18, 23, 33] and regressed joint location, researchers are able to rig the canonical body template to reconstruct the human body shape. Towards model-based human pose estimation, early practice fits the body model into input images, which results in a small re-projection error in the frontal view. With the deep learning fashion, the community is also interested to regress the pose parameters [19, 22], which leads to faster inference speed and robustness against pose initialization. As the large-scale pose annotations are expensive, training such neural networks incorporates a mix of annotated 3D data and unlabeled 2D images as supervision. Researchers nowadays pay more attention on predicting temporally consistent body shapes [20, 21, 49]. The sequential inference mitigates the difficulty in specific frames by considering the body motion. State-of-the-art methods can provide a plausible naked body shape as the underlying model and remain the clothing part as an independent problem.

Among all 3D human model fitting approaches, one recent line of works [9, 39] called “learned gradient descent” iteratively predicts SMPL model parameters, which inspires us to propose our sequential approach approximate the cloth shape in a similar paradigm. However, estimating non-rigid cloth motion is more challenging

due to the large degree of freedom brought by per-vertex deformation. Hence, one iteration loop is insufficient in optimizing total cloth shape.

Chapter 3

Methodology

This section describes the sequential approach for cloth shape recovery as shown in Fig.1.1. We leverage an SMPL+D model [23] described in Sec. 3.1 to capture the total human shape, where the cloth shape is represented as the deformation D over the naked body shape. In Sec. 3.2 we measure body dynamics by optimizing the SMPL parameters. The holistic cloth deformation D is decomposed into a personalized cloth template $\bar{D}$ in Sec. 3.3.2, a frame-dependent cloth deformation $\hat{D}$ in Sec. 3.3.3 and a high-frequency wrinkle details $\widetilde{D}$ in Sec. 3.3.4.

$$D_i = \bar{D} + \hat{D}_i + \widetilde{D}_i \tag{3.1}$$

3.1 SMPL+D Model

With an input video V of length L of a single person, we aim at predicting temporally coherent 3D meshes of clothed human M at each frame. SMPL+D model measures body dynamics and therefore offers vertex correspondence over time for tracking. The SMPL+D model contains two groups of SMPL parameters $\beta \in \mathbb{R}^{10}$ and $\theta \in \mathbb{R}^{24\times3}$ to control the naked shape and pose respectively, as well as per-vertex deformation $D \in \mathbb{R}^{6890\times3}$ from the naked shape to generate cloth. With the estimated parameters, we first reconstruct the canonical human shape consisting of underlying naked body $T(\beta, \theta)$ and cloth deformation D. The naked body is represented as a combination of template shape $\bar{T}$, shape dependent deformation $B^S(\beta)$, pose dependent deformation

$B^P(\theta)$, as shown in Eq. 3.2.

$$T(\beta, \theta, D) = \bar{T} + B^S(\beta) + B^P(\theta) + D \tag{3.2}$$

Then, rigged by the pose parameter θ and body joint $J(\beta)$, we animate the canonical shape to the posed space M using Linear Blend Skinning (LBS). Formally,

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

where W is the LBS function and $\mathcal{W}$ is the skinning weight.

3.2 Body Shape Estimation

As the first stage of our pipeline shown in Fig. 1.1, we first estimate the parameters $\bar{\beta}, \{\theta\}_1^L, \{t\}_1^L$ and camera parameter K from input video through an optimization process. We assume a perspective camera with only focal length f_x, f_y to be measured. Since the naked body shape controlled by SMPL parameter does not depend on the clothes, these methods are fixed in the later section. For the 3D pose estimation problem, deep neural networks [20, 21, 49] provide an initial prediction of parameters. We then refine the pose predictions in the inference time with extra off-the-shelf supervisions following MonoClothCap [45].

$$\min_{K,\beta,\{\theta\}_1^L,t_i^L} E_p^b = E_{2d}^b + E_{dp}^b + E_{sil}^b + E_{pof}^b + E_{reg}^b \tag{3.4}$$

Particularly, E_{2d}^b [5] minimizes the L2 distance between Openpose [6] detected 2D keypoints and projected SMPL joints. E_{dp}^b [12] minimizes the L2 distance between the location of 2D pixels inside the body and projection of corresponding vertices of SMPL body predicted by DensePose [13]. E_{sil}^b [45] maximizes the Intersection-over-Union between differentiably rendered silhouette of SMPL body and 0-1 mask obtained from 2D human parser [11]. E_{pof}^b minimizes the difference between the SMPL joint orientation and predicted Part Orientation Field [44]. E_{reg}^b regularizes $\{\theta\}_1^t$ with the Gaussian prior, β with the L2 loss and the temporal consistency of the SMPL vertices over time.

3.3 Cloth Shape Estimation

Directly capturing the cloth from video under monocular settings is a ill-posed problem due to the scale ambiguity. Therefore, the previously proposed optimization are highly non-convex and unstable. Our sequential decomposition of the total cloth deformation into separate parts under different deformation level greatly alleviate this problem. Concretely, we first aggregate multi-frame observations to predict a consensus shape as a personalized human avatar for the input video. Afterwards, a fine-grained cloth deformation refines the consensus shape to fit each specific frame. We impose strong smoothness assumptions on the cloth shape estimated in these two stages. The wrinkle details are extracted in the last section.

However, even with the three-stage progressive pipeline, optimizing cloth deformation in each step is still challenging, since 1) not all vertices are necessarily observed in at least one frame; 2) the input gradient is noisy due to the inaccurate 2D supervisions. To address these issues, classical method [2] designed other energy terms e.g. Laplacian term, Symmetry term and Regularization term. This yields a trade-off problem between generating smooth shapes and sharp edges in sleeve and cuff.

We address the aforementioned two difficulties in optimization by proposing a *gradient rectification network* $\mathcal{F}$. This network rectifies the gradient obtained from 2D energy term to have a more accurate direction and step size for gradient descent. Compared with the raw gradients, our network inpaints the gradients of non-observed vertices, and adaptively controls the smoothness of the mesh surface after one step of gradient descent.

Since the cloth is decoupled into three component and progressively predicted, we use X to denote the T-posed clothed body shape in current stage, where D^{temp} is the assembly of deformations estimated in current and all previous stage. n_X denotes the normal of T-posed shape X. In the posed space, we use M to denote the body shape and n_M to denote the normal of mesh M.

$$X = T(\beta, \theta, D^{temp}) = \bar{T} + B^S(\beta) + B^P(\theta) + D^{temp} \tag{3.5}$$

$$M = W(X, J(\beta), \theta, \mathcal{W}) \tag{3.6}$$

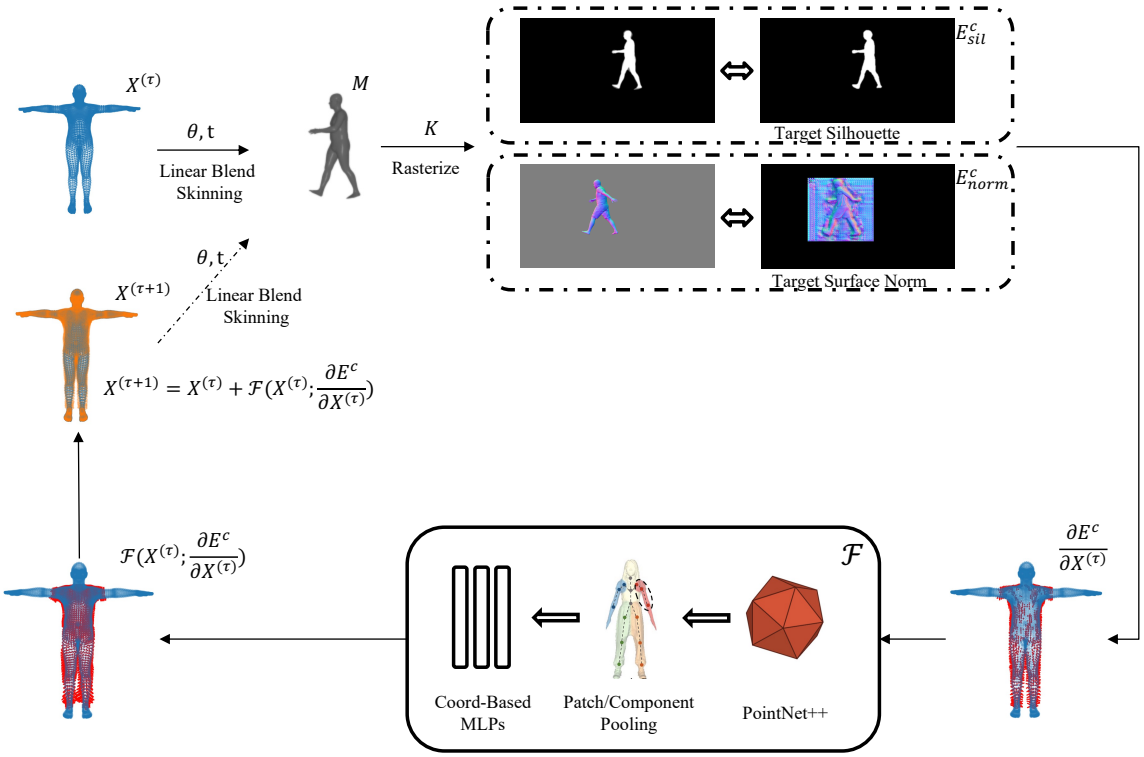

Figure 3.1: General training and inference procedure of our *gradient rectification network* in iteratively estimating the consensus shape, frame-dependent deformation and wrinkle details of the cloth.

Moreover, E^c is used to denote the energy term in each step.

3.3.1 Iterative Training and Inference Scheme

We first introduce the general training and inference scheme of our *learning-to-optimize* framework shared across three independent stages in Sec. 3.3. As shown in Fig. 3.1, in each iteration, the T-posed shape X is animated by estimated SMPL parameters θ and translation t to the posed space M. With the camera parameters K, we use a rasterizer to render a silhouette map and a normal map from mesh M. Off-the-shelf methods provide reliable silhouette and surface normal map prediction given the input image. At each stage, the 2D energy term E_c aligns either our rendered silhouette map or surface map to the target. By taking the derivative of E_c w.r.t. the canonical shape X, we obtain an initial gradient. Like the classical optimization methods, this

gradient from solely 2D energy term is sparse and ambiguous, as only partial vertices are observed. Given the initial gradient as input, our proposed *gradient rectification network* $\mathcal{F}$ predicts a dense and smooth gradient to update the canonical shape X. Eq.3.7 describes the "gradient descent" in τ-th step.

$$D^{(\tau+1)} = D^{(\tau)} - \alpha\mathcal{F}(X^{(\tau)}, \frac{\partial E^c}{\partial X^{(\tau)}}) \tag{3.7}$$

With the access of temporal 3D human dataset, we simulate our inference procedure of obtaining input gradient in the training time and generate ground truth output gradient supervision. We define the data term E_{data} in Eq. 3.8 as a L-2 distance between predicted gradient and ground truth gradient.

$$E_{data} = ||X_{GT} - X^{(\tau)} - \mathcal{F}(X^{(\tau)}, \frac{\partial E^c}{\partial X^{(\tau)}})||^2 \tag{3.8}$$

Moreover, a consistency term E_{2d} is used in supervising the network. E_{2d} is equivalent to the 2D energy term that generates input gradient, but with the updated cloth $D^{(\tau+1)}$. This enforces the consistency between the predicted gradient and the input gradient in observed vertices.

3.3.2 Consensus Shape Estimation

In this stage, we jointly measure one consensus shape in canonical space that matches silhouette shapes in multiple frames with the estimated SMPL pose parameters. The energy function of consensus shape merges losses from different frames. Specifically, for each vertices of the mesh in the posed space, we count the per-frame silhouette energy. The total energy is the sum of silhouette terms from all sampled frames.

$$E^c = \frac{1}{L}\sum_{t=1}^{L} E^c_{sil_n} \tag{3.9}$$

To compute E^c_{sil}, one solution is to extend the silhouette term E^b_{sil} in Sec.3.2 using differentiable renderer. However, the differentiable renderer finds correspondences of silhouette points within the blur radius and update the 3D vertices with a small step size. Therefore, optimization-based methods converges after multiple iterations.

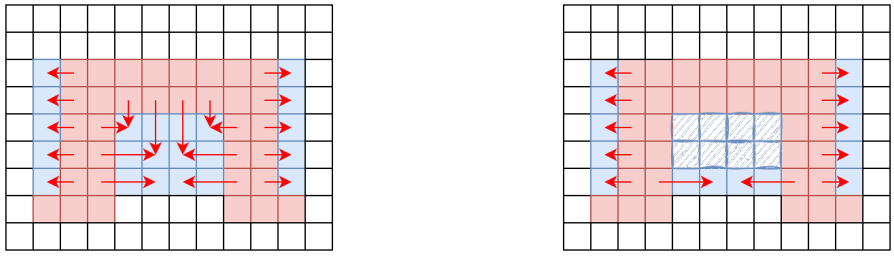

Figure 3.2: An illustration of the advantage of our correspondence searching algorithm (left) over Closest Point (right). Blue pixels denotes the target silhouette and the red pixels denote rendered silhouette.

As we aim at an instant estimation of cloth shape, it's critical to have a one-step approximation of the ground truth correspondence of 2D silhouette points and leverage 2D silhouette term from Eq.3.10, where $\Pi(M_i)$ is the perspective projection of boundary vertex M_i and y_j is its 2D corresponding silhouette point.

$$E_{sil}^{c} = \sum_{i} ||y_j - \Pi(M_i)||^2 \tag{3.10}$$

Early literature [2, 47] assigns correspondences via closest boundary points of silhouette. However, main drawback of ICP occurs when the 2D boundary points of silhouette image degenerates due to self-occlusion. This happens especially with fat pant legs. Inspired by traditional curve matching algorithm [10], we exploit level-set algorithm [7, 37] to mimic the differentiable renderer and find correspondences during the inference time, which better approximates the 2D location y_j that does not necessarily lies in the boundary of silhouette. The advantage of our algorithms in finding correspondences over ICP is illustrated in Fig.3.2. When two parts of the cloth overlaps, the boundary of the target silhouette disappears and ICP therefore failed to find correspondences to cover the unoccupied region (blue sketchy pixels in Fig.3.2). To be consistent in training and inference, all 2D locations are normalized under a calibrated camera. However, we directly use the ground truth correspondence to compute 2D energy term for training efficiency.

We train this network in temporal 3D human scan sequences. For each 3D sequence, we fix the camera location and simulate the merged silhouette energy of

mesh E_c in τ-th step. The loss function is formulated as

$$\mathcal{L} = E_{data} + \omega_c E_{2d} \tag{3.11}$$

where ω_c is the weight parameter to balance two terms. Eq.3.11 learns the dense gradient output given sparse gradient on boundary vertices from 2D supervision, and enforces the predicted gradient in the boundary vertices to be consistent with the input gradient. It deserves notion that the network input $\frac{\partial E_c}{\partial X^{(\tau)}}$ is pre-computed and the gradient is detached to avoid introducing the second-order gradient issue.

3.3.3 Frame-dependent Shape Estimation

With the consensus shape $\bar{D}$ estimated in Sec.3.3.2, we further go through an iteration to predict a frame-dependent deformation $\hat{D}_i$ for each frame. In this section, the 2D loss E^c is equivalent to the single frame silhouette loss E^c_{sil} in Eq. 3.10. With temporal 3D human scans dataset, we simulate the input gradient of the training data in the same manner as Sec. 3.3.2, but use our predicted consensus shape as initialized shape. Since single frame refinement introduces more ambiguity, extra constraints including a L-2 and a Laplacian term on X prevents the non-observed vertices from overfitting the training data.

$$\mathcal{L} = E_{data} + \omega_c E_{2d} + \omega_{reg} E_{l2} + \omega_{smooth} E_{lap} \tag{3.12}$$

3.3.4 Wrinkle Extraction

Traditional methods leverage Shape from Shading (SfS) [1, 43] to add wrinkle details. However, in our scenario of monocular RGB video in the wild, the complex albedo and lighting condition makes SfS becomes impractical to use. The recent deep learning method [41, 46] provides high quality surface normal estimation that is used as a supervision to reconstruct wrinkle details. With the predicted normal map as a prior, MonoClothCap [45] hereby exploits a differentiable renderer to align the rendered normal map of human mesh and the estimated normal map from input RGB image.

Though directly minimizing distance of two normal maps result in plausible 3D wrinkle details, one severe problem is the runtime due to the differentiable renderer.

Motivated by [31, 36, 51], we exploit the potential of neural network in generating realistic wrinkle details from flat cloth surfaces and benefit from speed of neural inference. However, different from PiFUHD [36] which learns features directly from target 2D normal map, we follow our paradigm in Sec. 3.3.2 and take gradient of normal loss as an input. By taking derivative of E^c over T-posed shape X, the input gradient is invariant to poses and rotations, which allows us to train the network with a small set of data and generalize to the wild scenario.

Specifically, we first formalize the normal loss E^c in this stage as Eq.3.13.

$$E^c = E_{norm} = \sum_i ||n_i - I_i||^2 \tag{3.13}$$

n_i is the surface normal of vertex i and I_i is its corresponding surface normal sampled from ground truth normal map. During the inference time, n_i is gathered by a bilinear interpolation on normal map $N(\cdot)$ from inverse rendering networks [41, 46].

$$I_i = N(\Pi(M_i)) \tag{3.14}$$

Then, a neural network $\mathcal{F}$ is derived to deform the visible vertices of SMPL+D mesh to generate wrinkles. In order to pursue realistic wrinkle details, in this section we use high-resolution SMPL model with 27754 vertices.

To train this network, we simulate the wrinkle extraction process from single static 3D human scans with rich wrinkle details. The network learns to generate high frequency part of the human scan from a Laplacian smoothed mesh. The training loss is composed of a L-2 loss and a normal loss between corresponding vertices from predicted mesh and ground truth mesh as shown in Eq.3.15. E_{2d} ensures the network to generate plausible visual results without necessarily predict the exact ground truth gradient.

$$\mathcal{L} = E_{data} + E_{2d} \tag{3.15}$$

3.3.5 Implementation Details

Training data simulation The challenging part of our *learning to optimize* scheme is the domain gap between training and wild inference. For consensus shape estimation,

Section	Dataset	Initialized Shape X^0	Target Shape X_{GT}
3.3.2	CAPE [25], ReSynth [26, 27], MGN [4], Thuman [48, 51]	Naked SMPL body	Smoothed Clothed body in ref. frame
3.3.3	CAPE [25], ReSynth [26, 27], MGN [4], Thuman [48, 51]	Clothed body in ref. frame estimated by Sec. 3.3.2	Smoothed Clothed body in sampled frame
3.3.4	TailorNet [31]	Smoothed Garment Mesh	Original Garment Mesh

Table 3.1: Data preparation in three stages respectively.

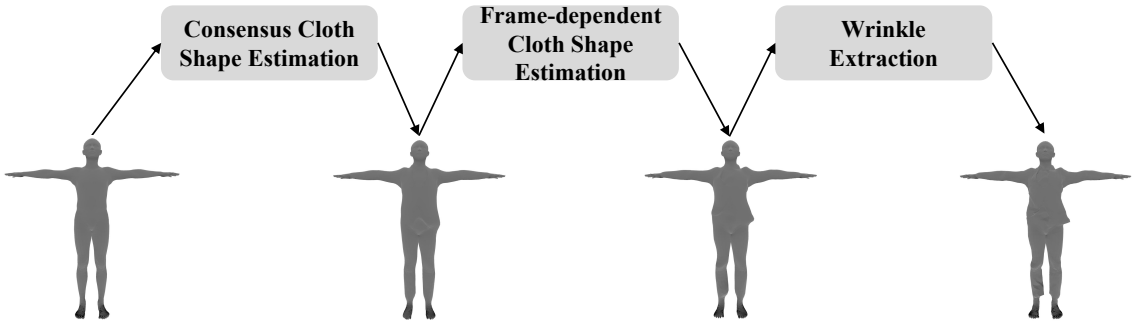

Figure 3.3: Visualization of decomposed training data.

we sample 20 frames per sequence and aggregate individually computed gradient from 2D loss. The scan under A-pose with minimal pose-dependent deformations is recognized as the consensus shape of the sequence. For frame-dependent deformation estimation, the network learns to deform our learned canonical shape to scans in sampled frame. As the 2D silhouette loss is insufficient to guide the wrinkle generation, we smoothen the ground truth shape with Laplacian filter as Tailornet [31] to eliminate wrinkles. As all training data contains SMPL registration, in the training phase, we detect the boundary vertices of the rasterized 2D silhouette map and compute 2D alignment loss given the correspondence. For wrinkle extraction, the network learns to recover high-frequency details from Laplacian smoothed mesh in each frame. Same as the previous section, the input gradient comes from 2D alignment loss between normals of corresponding vertices. All target silhouette/normal map in the training phases are rendered from 3D geometry data without the need of texture information.

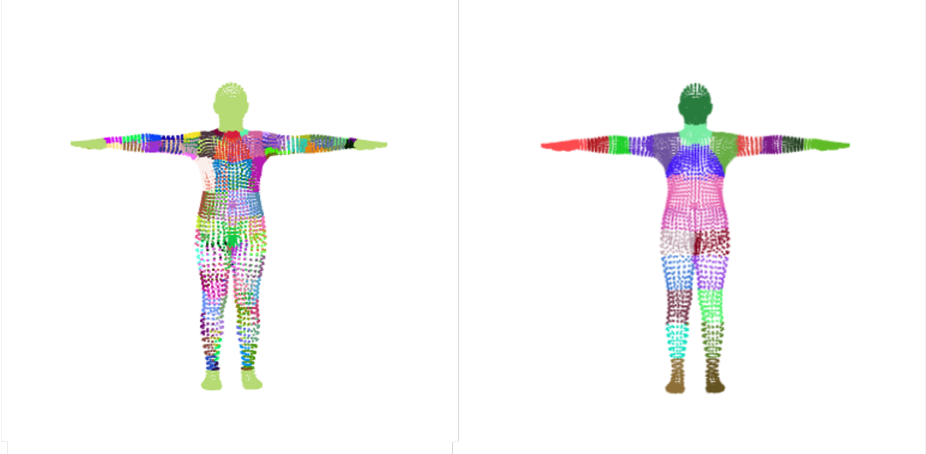

Figure 3.4: Visualization of patch clusters (left) and component clusters (right).

Detailed visualization of training data in each stage is shown in Fig. 3.3.

Network Architecture We leverage a PointNet++ [34] structure to encode the input gradient and a coordinate-based MLPs to decode the multi-scale feature to output gradient. Given $X, \frac{\partial E^c}{\partial X}$, the encoder is consist of two separate MLPs for geometry feature and gradient feature. Geometry feature simply contains 3D coordinate of X and normal n_X. For Sec. 3.3.2 and Sec. 3.3.3, we also consider the symmetric assumption of garments motivated by [2]. To achieve this, each vertex concatenates the gradient of the x-symmetric and z-symmetric vertices to its gradient as auxiliary feature. Since the input point clouds are from canonical T-posed SMPL+D body, the clusters for downsampling in *Set Abstraction Module* and the interpolation weights for upsampling in *Feature Propagation Module* in PointNet++ is pre-defined. Particularly, from UV-map, the vertices are devided into 104 patches and 24 components for two levels of downsampling, as shown in Fig. 3.4. The multiscale feature learned from PointNet++ is concatenated with the positional-encoding [29] of 3D vertex coordinates to predict the output gradient. Following the fashion from [29, 30], the decoder has a skip connection. In order to constraint the output range in each step, we add a *tanh* activation function in the last layer. The overall architecture of our network is visualized in Fig. 3.5.

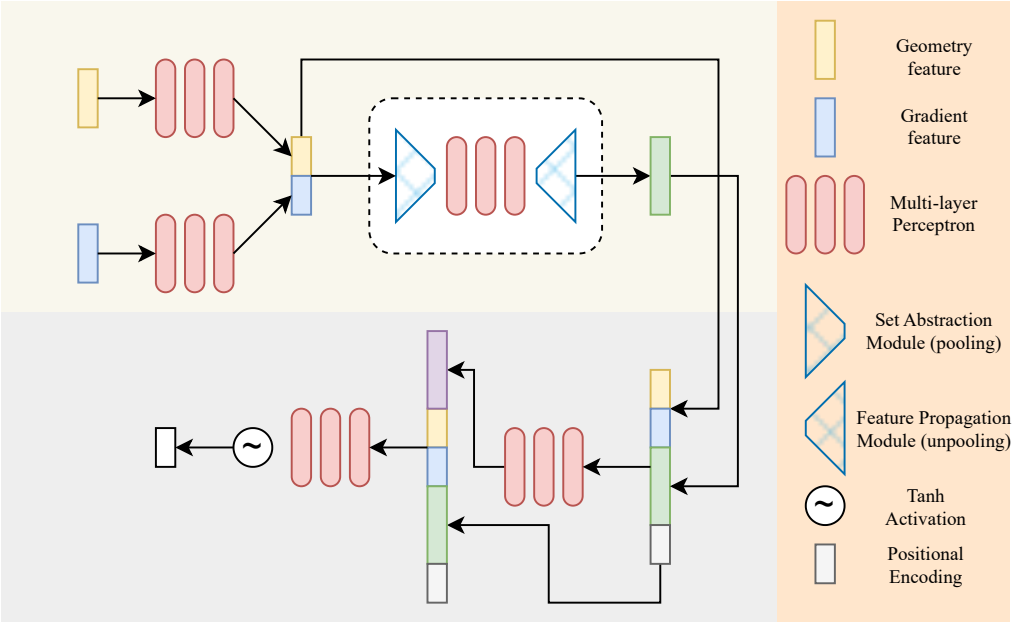

Figure 3.5: General architecture of our proposed *gradient rectification network* $\mathcal{F}$.

Chapter 4

Experiments

In this section, we validate the effectiveness of our proposed methods qualitatively and quantitatively in a set of experiments. Specifically, in Sec. 4.1, we study the overall cloth capture performance of our method on monocular RGB videos in-the-wild. Despite capturing cloth for each frame, we can leverage the consensus shape produced in Sec. 3.3.2 as an animatable avatar. Therefore, we demonstrate the capability of our method in generating accurate human avatars from monocular RGB video in Sec. 4.2. The superiority of our method in runtime is analyzed in Sec. 4.3.

4.1 Results on Human Performance Capture in the Wild

4.1.1 Dataset

We examine the quantitative performance of our method on *Pablo* sequence from MonoPerfCap dataset [47]. This sequence contains a 156-frame multi-view (8 camera) video and a reconstructed 3D scan sequence as ground truth. Following the previous works [45, 47], we select a single view of the *Pablo* sequence as the input for testing. Additionally, we carried out a set of qualitative experiments on online and smartphone shot videos.

Figure 4.1: Qualitative result on human performance capture from *pablo* sequence and video taken by smartphone.

4.1.2 Baselines

We make comparisons with both optimization-based video human performance capture methods [45, 47] and learning-based single image 3D human reconstruction methods [3, 35, 36, 51, 52]. Notably, similar to the most recent learning-based methods [52], we require a few steps of optimization (e.g. in Sec. 3.2) for pose tracking to achieve the reported accuracy. In the aspect of efficiency, we can replace this procedure with an pose tracking network [20].

4.1.3 Evaluation Metrics

We report the average point-to-surface distance following the evaluation protocol in MonoClothCap [44]. Concretely, since different methods have diverse camera settings, we first centralize and scale the predicted scans according to the height. Then we manually segment the ground truth scans and use ICP to register the predicted scans to cloth region of ground truth scans under translation. The point-to-surface is defined as the minimal distance between the cloth vertices of aligned prediction scans to the surface of ground truth scans.

Methods	Point-to-Surface Error (mm)
MonoPerfCap [47]	**14.7**
MonoClothCap [45]	17.9
Tex2Shape [3]	27.7
DeepHuman [51]	24.2
PIFu [35]	30.5
PIFuHD [36]	26.5
PaMIR [52]	28.3
Ours	**17.4**

Table 4.1: Quantitative comparisons with state-of-the-arts on *pablo* sequence. The first two methods are optimization-based methods and the rest are neural network predictions. Note that MonoPerfCap leverages a pre-scaned template mesh of the video as prior.

4.1.4 Result Analysis

Table 4.1 illustrates the distinct advantage of our method over all single image human shape reconstruction techniques [3, 35, 36, 51, 52], including PaMIR [52] that take the SMPL pose as a prior. While producing reasonable visualizations in the frontal view, PIFu series generate noisy vertices that results in large quantitative error. We even outperform template-free optimization methods [45] with significantly less runtime and approaches the performance of [47] using pre-scanned template shape. One of our quantitative advantage is that for the non-observed vertices we predicts a more compact and flat shape. We visualize our cloth capture result on both *Pablo* sequence and smartphone shot video in Fig. 4.1. As we can see in the figure, with significantly less running time, we generate a similar amount of details as MonoClothCap, and performs a more robust tracking for the flying cloth for the right video.

Besides, we conduct extra experiments on challenging online videos with fast body motion and loose long sleeve cloth, which is not supported by MonoClothCap [45]. Fig. 4.2 shows qualitative results of our approach against generic single image reconstruction method PaMIR [52]. As we can see from the figure, our method demonstrates remarkable robustness against novel poses and viewpoints and even inaccurate 2D segmentation result thanks to our sequential human performance capture pipeline. In contrast, PaMIR is sensitive to the pose and viewpoint which results in the failure reconstructions such as bodies with missing arms in specific hard

frames.

4.2 Results on Video-based Human Avatar Generation

4.2.1 Dataset

We validate the ability of our network in generating accurate human avatars by leveraging PeopleSnapshot dataset [2]. This dataset contains real-world 360 self-rotating videos of 6 males and 6 females in different types of cloth. Ground-truth segmentations, camera parameters and accurate SMPL pose parameters are also provided. However, no ground truth 3D scan is captured with these videos. Therefore, we only perform qualitative results for comparison.

4.2.2 Baselines

In the experiment, we mainly compete against the classical optimization-based method VideoAvatar [2]. VideoAvatar also exploits SMPL+D model and deforms the template body mesh with the aggregated gradient from 2D silhouette energy. This comparison clearly demonstrates the benefit of introducing 3D training data and the *learning-to-optimize* scheme.

4.2.3 Result Analysis

The qualitative comparison is shown in Fig. 4.3. For each scenario, we outperform classical optimization-based methods [2] in generating cloth wrinkles. Our wrinkles are extracted from a frontal and a back image. Moreover, our method better captures the cloth shape in the side view instead of generating an over-smooth mesh, e.g. the chest region of the women in the first row.

Stage	MonoClothCap [45]	PaMIR [52]	Ours
Pose Estimation	6	12.5	6
Consensus Shape Estimation	89	7.5	0.06
Frame Refinement			1
Wrinkle Extraction	327		0.17

Table 4.2: Runtime comparison between our method and optimization-based method and PIFu method with pose prior. We compare the average running time (second) per-frame. I/O time is excluded.

4.3 Runtime Analysis

One of our major contributions is the runtime improvement. Therefore, we carry out a runtime comparison tested from the 253-frame *pablo* sequence. Since the consensus shape is one personalized template shape for the given video, we divide the total runtime into the number of aggregated frames to obtain per-frame runtime. The detailed runtime in each stage is shown in Table 4.2. By using learning-based approaches [20, 21, 49] to replace the optimization loop, we can achieve totally 1FPS inference speed.

Our runtime is boosted for three reasons: 1) in inference time, our *gradient rectification network* converges up to 3 iterations. 2) we avoid using the differentiable renderer, which consumes a huge amount of time and memory to render a high-resolution image. 3) we directly predict a 3D human mesh instead of implicit function, which does not require extra marching cube algorithm [24] to extract mesh.

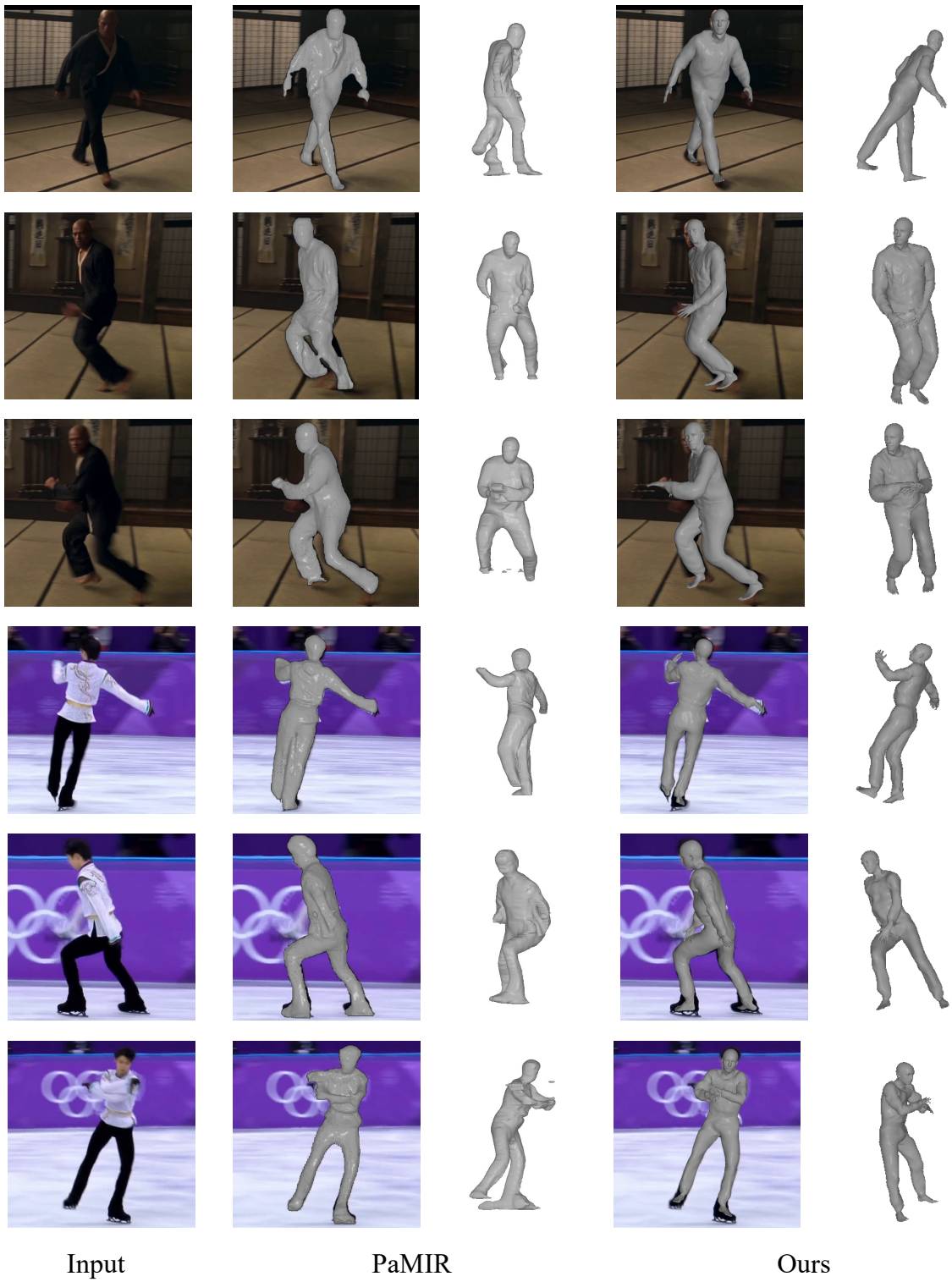

Figure 4.2: Qualitative result on human performance capture from challenging YouTube video with fast body motion and loose sleeves. Note that side-views from different methods are not aligned due to different camera settings.

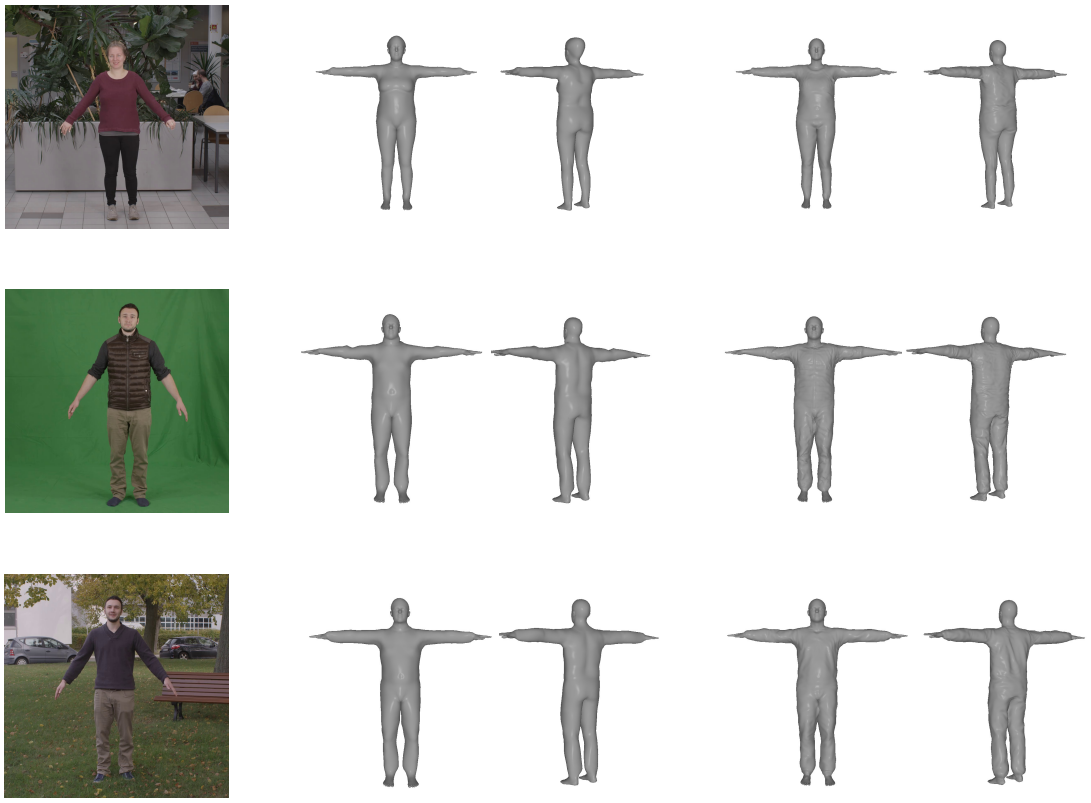

Figure 4.3: Qualitative result on video-based human avatar generation.

Chapter 5

Conclusions

In this thesis paper, we present a sequential human performance capture method to reconstruct 3D human body and cloth from monocular RGB videos in the wild. Our three-stage pipeline independently estimates an average cloth shape, a frame-wise deformation and the plausible wrinkles in a coarse-to-fine manner. At each stage, we leverage a deep neural network to predict cloth deformations from sparse vertices gradient generated from 2D image fitting energy. This *learning-to-optimize* idea ensures our method to benefit from both the robustness and accuracy in classical optimization from classical optimization methods and the efficiency and generalization ability of deep learning methods. The experiments in cloth capture and human avatar generation demonstrate the advantage of our approach in both accuracy, robustness and efficiency.

5.1 Limitations and Future Works

The major drawback of our method is from the underlying SMPL+D model. Since SMPL+D model approximates the skinning weight of cloth vertex by the skinning weight of the nearest skin vertex. Therefore, large error exist when we use such skinning weight to animate loose cloth. Further, the smooth cloth deformation is unable to represent cloth types with different topology such as hoodies or pockets of the cloth. Lastly, the strong assumption about the cloth smoothness we hold in the frame refinement stage limits the ability of our methods in minimizing 2D energy

terms and generate sharp edges and multi-layer cloth. As a result, we are likely to lose track of the sleeves and cuffs in fast motion.

These issues could be address by leveraging the implicit function to represent the shape and some pioneers start to study human body animation under the implicit representation. However, the cost of volume tracking is large and the robustness of these new animation algorithms under monocular setting could be easily challenged. In the future, with more advanced human avatar animation techniques, we expect to extend our idea of *learning-to-optimize* to generate body shape for boarder cloth types.

Appendix A

2D Correspondence Searching Algorithm

Algorithm 1 2D Correspondence Searching Algorithm for Silhouette Map Alignment.

Input Source point set $\mathcal{S}$, Target point set $\mathcal{T}$
Output Correspondence Assignment Matrix $\mathcal{M}$

$\mathcal{D} = \infty, \mathcal{M} = 0$
for $t_j \in \mathcal{T}$ **do**
    k = $\texttt{ClosestPoint}(\mathcal{S}, t_j)$
    $\mathcal{D}_{k,j} = ||s_k - t_j||^2$
end for
for $s_i \in \mathcal{S}$ **do**
    $\mathcal{M}_{i,\texttt{argmax}(\mathcal{D}_i)} = 1$
end for
return $\mathcal{M}$

In the appendix we describe a correspondence searching algorithm to align our rendered silhouette map to target silhouette map from segmentation [11]. Instead of using Closest Point to align boundaries of two silhouette maps, our main motivation is to cope with the corner cases where the boundary of silhouette degenerates under self-occlusion, as illustrate in Fig. 3.2. Since the 3D cloth shape generated from naked body has less boundary degeneration issue, we register the boundary of our rendered silhouette map to the disparity regions between two silhouette maps. Specifically, we extract the edge of our rendered silhouette by a laplacian filter [38] as source

point set $\mathcal{S}$ and record the 2D coordinate of misaligned pixels as target point set $\mathcal{T}$. For $i \in \mathcal{S}$, Alg. 1 searches the correspondence point $j \in \mathcal{T}$. Since the human parsing approach [11] also provides semantic part labels, we apply our correspondence search algorithm independently for upper and lower cloth, which better generates the boundary between two separate clothes.

Bibliography

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 3.3.4

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 1.1, 2.1, 3.3, 3.3.2, 3.3.5, 4.2.1, 4.2.2, 4.2.3

[3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. 4.1.2, **??**, 4.1.4

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 2.2, **??**, **??**

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, October 2016. 3.2

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2.3, 3.2

[7] Thomas H Cormen. Section 24.3: Dijkstra's algorithm. *Introduction to algorithms*, pages 595–601, 2001. 3.3.2

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11875–11885, 2021. 1.1, 2.2

[9] Enric Corona, Gerard Pons-Moll, Guillem Alenyà, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. *arXiv preprint arXiv:2205.06254*, 2022. 2.3

[10] Max Frenkel and Ronen Basri. Curve matching using the fast marching method. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 35–51. Springer, 2003. 3.3.2

[11] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. 3.2, A

[12] Riza Alp Guler and Iasonas 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. 3.2

[13] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3.2

[14] Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pages 889–898. IEEE, 2021. 2.1

[15] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. 1.1, 2.2

[16] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 1.1, 2.2

[17] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. *arXiv preprint arXiv:2201.12792*, 2022. 1.1, 2.1

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 1.1, 2.3

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2.3

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael 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 (CVPR)*, June 2020. 2.3, 3.2, 4.1.2, 4.3

[21] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, October 2021. 2.3, 3.2, 4.3

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2.3

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1.1, 2.3, 3

[24] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 4.3

[25] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. ??, ??

[26] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. ??, ??

[27] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. ??, ??

[28] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2.3

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 3.3.5

[30] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3.3.5

[31] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 3.3.4, **??**, 3.3.5

[32] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2.3

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2.3

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3.3.5

[35] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1.1, 2.2, 4.1.2, **??**, 4.1.4

[36] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1.1, 2.2, 3.3.4, 4.1.2, **??**, 4.1.4

[37] James Albert Sethian. *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*, volume 3. Cambridge university press, 1999. 3.3.2

[38] Mohsen Sharifi, Mahmood Fathy, and Maryam Tayefeh Mahmoudi. A classified and comparative study of edge detection algorithms. In *Proceedings. International conference on information technology: Coding and computing*, pages 117–120. IEEE, 2002. A

[39] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, pages 744–760. Springer, 2020. 2.3

[40] Zhaoqi Su, Weilin Wan, Tao Yu, Lingjie Liu, Lu Fang, Wenping Wang, and Yebin Liu. Mulaycap: Multi-layer human performance capture using a monocular video camera. *arXiv preprint arXiv:2004.05815*, 2020. 2.1

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and

Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3.3.4, 3.3.4

[42] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 2.3

[43] Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics (TOG)*, 32(6):1–11, 2013. 3.3.4

[44] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 3.2, 4.1.3

[45] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 1.1, 2.1, 3.2, 3.2, 3.3.4, 4.1.1, 4.1.2, **??**, 4.1.4, **??**

[46] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. *arXiv preprint arXiv:2112.09127*, 2021. 2.2, 3.3.4, 3.3.4

[47] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2):1–15, 2018. 1.1, 2.1, 3.3.2, 4.1.1, 4.1.2, **??**, 4.1.4

[48] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin 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. **??**, **??**

[49] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2.3, 3.2, 4.3

[50] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li. High-fidelity human avatars from a single rgb camera. In *CVPR*, 2022. 1.1

[51] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 1.1, 2.2, 3.3.4, **??**, **??**, 4.1.2, **??**, 4.1.4

[52] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1.1, 2.2, 4.1.2, **??**, 4.1.4, **??**