A Sequential Learning-based Approach for Monocular Human Performance Capture

Jianchun Chen[1] Jayakorn Vongkulbhisal[2] Fernando De la Torre Frade[1]
[1]Robotics Institute, Carnegie Mellon University
[2]Independent Researcher
jianchuc@alumni.cmu.edu jayakornv@gmail.com ftorre@cs.cmu.edu

Abstract

*Human performance capture from RGB videos in unconstrained environments has become very popular for applications that require generating virtual avatars or digital actors. SOTA methods use neural network (NN) techniques to estimate the shape directly from photos, yielding a simplified model of the human body. While effective, NN techniques frequently fail under challenging poses and do not preserve temporal consistency. On the other hand, optimization-based methods like shape-from-silhouette can produce more precise reconstruction; however, they typically require a good initialization and are computationally more intensive than NN. To address issues of previous methods, this work proposes a learning-based approach for optimizing fine-grained shape representation from a monocular RGB video. Our main idea is to sequentially recover different shape details (*i.e. *average shape, clothing, wrinkles) using separate neural networks. At each level, our network takes the sparse/noisy gradients of body mesh vertices w.r.t. the shape, and predicts dense gradients to update the body shape. Despite being trained on synthetic data, these networks have surprisingly good generalization to real images. Experimental validation shows that our approach outperforms NN approaches in recovering shape details while also being an order of magnitude faster than optimization-based methods and robust across varied poses and novel views.*

1. Introduction

Capturing high-fidelity human shape is essential for many modern applications, including virtual/augmented reality, telepresence, gaming, and digital actors in movies. To address these demands, capturing individuals and their clothes from RGB images [30, 49] or videos [41, 44] with the goal of creating 3D virtual humans has become a prominent field of research. Popular 3D shape estimation techniques use the SMPL body model or its variations [16,23] to successfully estimate 3D body shapes from images/video. However, estimating fine-granular shape such as clothing or wrinkles from images or monocular videos is a challenging task due to the depth ambiguities and the highly deformable nature of clothing.

Several approaches estimate the detailed shape using optimization techniques. Particularly, the 3D clothing shape is optimized based on the contour of the person by progressively reducing the disparity between the rendered and detected silhouette images. Due to the ambiguity caused by the single-view silhouette loss, these methods rely on either highly constrained priors of mesh smoothness [2], cloth shape models [41], highly constrained scenarios of self-rotating video [2, 15, 47], or a pre-scanned avatar [22, 44]. Further, the runtime of the aforementioned optimization is frequently unacceptable for many relevant applications. Alternatively, other researchers, including [9, 13, 14, 30, 31, 42, 43, 48], 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 RGB image as an input to learn pixel-aligned features for predicting an implicit function of a 3D person with intricate textile geometry. These learning-based approaches offer a fast inference speed and a surprising generalization to in-the-wild pictures. However, existing algorithms lack accuracy and robustness in presence of difficult poses, textures, or perspectives. Moreover, when applied to video, there is no consistency on the temporal smoothness of the 3D output.

To improve the efficiency of the optimization methods and the accuracy of the NN methods, this paper proposes a sequential shape recovery method, where a set of networks learn different shape details. Fig.1 illustrates how we are able to estimate the shape in a coarse-to-fine manner. Given an input image, we first estimate the underlying body shape using the SMPL model (Fig.1(a)). Later, two independent networks estimate the average clothing shape and pose-dependent clothing deformations (Fig.1(b-c)). Lastly, the wrinkles are extracted by the final network (Fig.1(d)).

Our approach follows the *learning to optimize* framework that learns a mapping from image features to shape

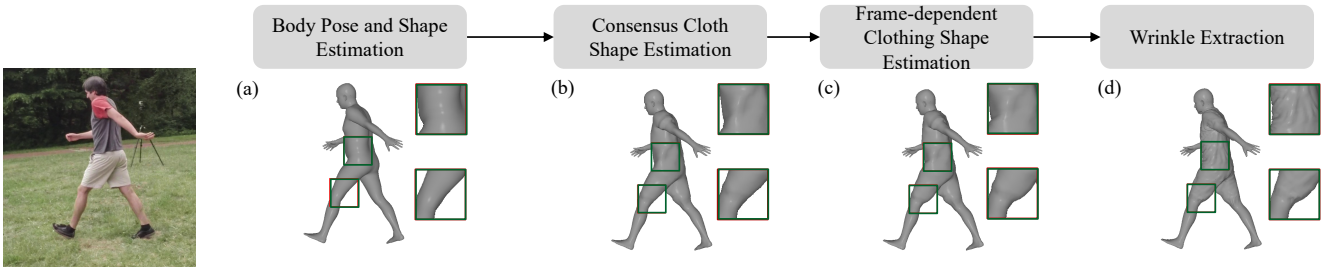

Figure 1. Given a monocular RGB video in the wild, we propose a sequential shape estimation method. The method recovers different shape details sequentially. First, we estimate a naked SMPL body (a), then an average clothing shape (b) is estimated from multiple frames sampled from sequence. (c) represents a frame-dependent shape estimation based on (b). Finally, (d) uses surface normals to estimate the wrinkles. (b-c) uses a *gradient rectified network* (GRN) to robustly estimate each shape deformation.

deformations in a linear iterative fashion [36] or through non-linear regression with neural networks [7,8,32,45]. For each sequential step to improve the shape, the *gradient rectification network* (GRN) takes the sparse/noisy gradient of the cost function w.r.t. the shape and estimates a dense rectified per-vertex gradient to reliably update the vertices. We learn these GRN using synthetic 3D dynamic human motion datasets. More importantly, we just need the 3D geometry to replicate the input gradient, and no high-quality texture data are required. Compared with dense optimization algorithms, our predicted gradient is more accurate and converges within a few iterations, which greatly accelerates the inference time. Extensive results under different settings show the improved accuracy and effectiveness of our approach for 3D human performance capture from monocular video. Our contributions can be summarized as follows:

- We propose the *gradient rectification network* (GRN) which takes as input a sparse and noisy gradient and outputs a robust and dense gradient that is much more informative of shape deformation. Despite the fact that the GRN is trained with 3D synthetic dynamic human data, it shows robustness to generate tracked 3D human models from videos under a variety of scenarios.

- We propose a human performance capture method from monocular video that sequentially recover different shape details (*i.e*. average shape without clothes, clothing, wrinkles) using a group of neural networks. The technique achieves SOTA results while drastically reducing the optimization runtime, potentially enabling real-time execution.

2. Related Works

This section examines the work on predicting body shape from video, with a focus on the effective clothing modeling.

2.1. Human Pose and Shape Estimation

Initially, human pose estimation refers to localize 2D [6, 38] or 3D [26, 29] keypoints from humans in images. Beyond locating 2D or 3D keypoints, recent models [16,23,28] are able to fit high-dense mesh reconstructions into images. While these methods have performed great results, they are typically not robust and efficient. With the rise of deep learning, inference methods that regress pose parameters [17, 19, 21, 34] led to faster inference and robustness to pose initialization. Among all 3D human model fitting approaches, a recently trend called "learned gradient descent" [7, 8, 32, 45] that iteratively refines SMPL parameter prediction, is closer to our idea of sequential clothing shape fitting. However, our emphasis on fine-grained shape recovery by estimating per-vertex clothing deformation is more challenging compared with estimating SMPL controlling parameters, due to the large degree of freedom brought by non-rigid cloth motion. Other researchers [20, 35] align body/hand mesh by regressing a per-vertex "offset map" from image. We argue that gradient features that we take as input from 2D image alignment loss are more robust for detailed mesh refinement than taking image features directly.

2.2. Optimization-based Clothing Capture

Traditional human clothing capture approaches model the task as an optimization problem, where a non-rigid deformation of the body is estimated repeatedly to best suit the input image. Since directly estimating per-vertex offsets from silhouette image is highly unstable, early methods [12,44] start with a pre-scanned template and support a small range of deformation. For in-the-wild videos, such as videos taken from commercial cameras, researchers impose different constraints to obtain plausible cloth shape. Particularly, VideoAvatar [2] hand-crafts multiple regularization terms and decomposes the clothing shape into a global consensus shape and a frame-wise deformation. SelfRecon [15] and Video2Avatar [11] proposed to jointly opti-

mize cloth geometries with textures by leveraging a backend neural renderer. Another popular trend is to parameterize the non-rigid deformation using Principal Component Analsyis [41], a deformation graph [10], or garment parameters [33]. These methods typically converge and provide spatial-temporal coherence. Unfortunately, they only work in constrained cases and are not computationally efficient.

2.3. Learning-based Clothing Capture

Inspired by the success of deep neural networks, pioneer works [5, 49] start to predict 3D human model with cloth directly from input RGB image. Concretely, Bhatnagar et al. [5] regress PCA controlling parameter of multiple types of garments, and Zheng et al. [49] 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 [14, 30, 31] 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 [9, 13, 43, 48] greatly alleviate this problem by adding SMPL shape prior. Overall, these methods allow fast inference speed with plausible reconstruction results; however, the performance of these methods are still limited under challenging poses, and typically it is hard to impose spatial-temporal consistency.

3. SMPL and SMPL+D Models

The SMPL model [23] is a parametric model for naked body shapes. It is parameterized by two groups of parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and $\boldsymbol{\theta} \in \mathbb{R}^{24\times 3}$ to control the naked body shape and pose respectively. Once the parameters, β, θ are estimated, the canonical human shape is deformed from a template shape $\bar{T}$, by adding the body pose and shape dependent deformation $B^P(\boldsymbol{\theta}), B^S(\boldsymbol{\beta})$. On top of the naked body, we follow SMPL+D [23] model that provides extra degrees of freedom $\mathbf{D} \in \mathbb{R}^{6890\times 3}$ to deform the canonical body vertices and generate clothing. We use $\mathbf{X}$ in Eq. 1 to denote the total human body and clothing shape in canonical space:

$$\mathbf{X} = T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = \bar{T} + B^S(\boldsymbol{\beta}) + B^P(\boldsymbol{\theta}) + \mathbf{D}. \quad (1)$$

Then, we transform the T-pose shape X to the posed space M using Linear Blend Skinning (LBS), driven by the pose parameter θ and body joint $J(\beta)$ following [23]. Formally,

$$\mathbf{M} = M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(\mathbf{X}, J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathcal{W}). \quad (2)$$

SMPL Body Pose and Shape Estimation from Video: As shown in Fig. 1(a), given a video with L frames, the first stage of our pipeline estimates the SMPL parameters $\bar{\boldsymbol{\beta}}, \{\boldsymbol{\theta}\}_1^L$, the camera intrinsics K and global translations $\{\mathbf{t}\}_1^L$. We first use the deep neural networks [18] to provide an initial estimate of SMPL parameters $\bar{\boldsymbol{\beta}}, \{\boldsymbol{\theta}\}_1^L$, then jointly refine them with camera intrinsics K and global translations $\{\mathbf{t}\}_1^L$ using gradient methods [41]. Please refer to Appendix A for details.

In the following sections, we introduce our sequential approach to reconstruct the clothing shape $\mathbf{D}$.

4. *Learning-to-Optimize* Clothing Vertices

This section introduces our main contribution, which is a learning-based method to compute accurate gradients to optimize over the vertices in the SMPL+D model [23].

Classic optimization-based methods (*e.g.* shape from silhouette) reconstruct 3D garments by minimizing the disparity between the rendering output of predicted 3D shape and the input image. However, capturing the 3D clothing shape from monocular video is an ill-posed problem due to the scale ambiguity and non-linear projection. Therefore, the previously proposed optimizations [2, 44] are highly non-convex and typically lead to inconsistent 3D solutions. A major issue is that the gradient of the 2D cost function w.r.t. the visible vertices ($\mathbf{X}$) is sparse, noisy, and the gradient w.r.t. vertices that are not observed (in the image) is inexistent. To address this issue, we propose a *gradient rectification network* $\mathcal{F}$, that is learned from synthetic 3D data. At inference time, $\mathcal{F}$ will provide robust directions for *all* vertices (including the unobserved ones) to minimize the 2D disparity between the 3D render model and the image.

The general training and inference scheme of our *learning-to-optimize* framework is shown in Fig. 3. In each iteration τ, the T-posed shape $\mathbf{X}^{(\tau)}$ is deformed by the estimated SMPL parameters θ and translation t to the posed space M, Eq. 2. With the camera parameters $\mathbf{K}$, we used a rasterizer to render a silhouette map in $\mathbb{R}^{h\times w}$ and a normal map in $\mathbb{R}^{h\times w\times 3}$ from the mesh $\mathbf{M}$, see Fig. 1. Off-the-shelf methods provide reliable silhouette and surface normal map prediction given the input image. Our method tries to minimize a cost function, E_c , that generally minimizes the difference between the image features (*e.g.* silhouette, surface normals) and the rendered 3D shape. To minimize E_c through gradient descent, we first compute the derivative of E_c w.r.t. the canonical shape X, thus obtain an initial gradient. Like the classical optimization methods, this gradient from solely the 2D energy term E^c is sparse (only values in observed vertices), noisy, and ambiguous. Given the initial gradient as input, our proposed *gradient rectification network* (GRN) $\mathcal{F}$ predicts a dense and smooth gradient to update the canonical shape $\mathbf{X}$. Eq. 3 describes the "gradient descent" in τ-th step with step size α.

$$\mathbf{X}^{(\tau+1)} = \mathbf{X}^{(\tau)} - \alpha\mathcal{F}(\mathbf{X}^{(\tau)}, \frac{\partial E^c}{\partial \mathbf{X}^{(\tau)}}) \quad (3)$$

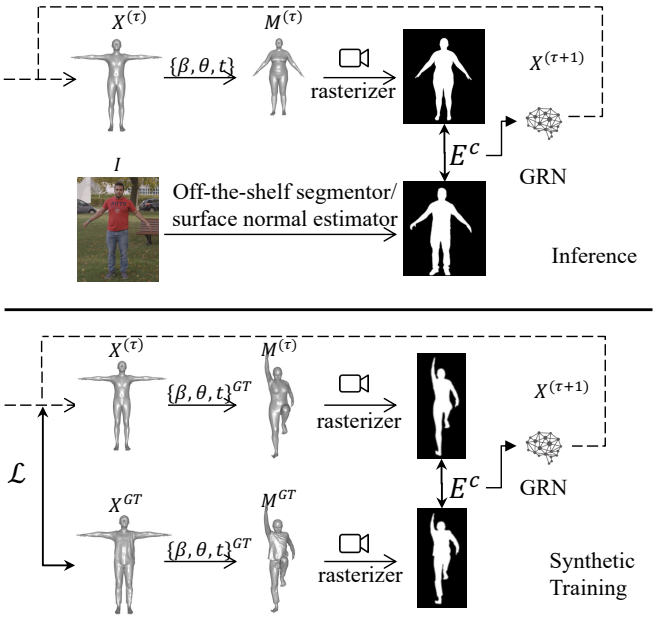

Figure 2. Generation of target silhouette/surface normal map in training and inference phase.

Fig. 2 illustrates the differences in training and testing phases with GRN. Particularly, during the inference time, we obtain the target silhouette/normal map prediction via off-the-shelf segmentation and normal estimation methods. Instead, we leverage a SMPL+D registered 3D human scan dataset in training time. Each 3D human shape $\mathbf{X}$ has a paired ground truth 3D scan $\mathbf{X}_{GT}$. The rendering output of $\mathbf{X}_{GT}$ is considered as target silhouette/surface normal map. GRN learns to update the clothing shape $\mathbf{X}$ such that it could minimize the disparity between the rendered and the target silhouette/surface normal map. To train this network $\mathcal{F}$, we propose a loss function $\mathcal{L}$ that supervises $\mathbf{X}$ to align with ground truth shape $\mathbf{X}_{GT}$.

5. Sequential Human Performance Capture

Our goal is to have an accurate estimation of clothing deformation D from a monocular video. In this paper, we follow a sequential approach to optimize this very high-dimensional deformable shape. Sequential optimization, that progressively provides at each step an additional constraint with a better initialization, has provided better results that jointly optimize all the constraints at once starting from random initializations. We propose to decompose the total clothing shapes into three separate parts and leverage a sequential pipeline as shown in Fig.1.

For the n-th frame, the clothing deformation can be represented as the combination of three deformations:

$$\mathbf{D}_n = \bar{\mathbf{D}} + \hat{\mathbf{D}}_n + \widetilde{\mathbf{D}}_n. \tag{4}$$

where $\bar{D}$ is a personalized clothing template that is estimated from all the frames in the video (Sec. 5.1). $\hat{D}$ is a frame-dependent clothing deformation (Sec. 5.2). Finally, $\widetilde{D}$ represents the high-frequency wrinkle details (Sec. 5.3). In each subsection, we describe the data, the formulation of the 2D energy term E^c and training loss $\mathcal{L}$ to train GRN $\mathcal{F}$.

5.1. Consensus Clothing Shape Estimation

This section describes the consensus shape estimation ($\bar{\mathbf{D}}$) that uses all frames in the sequence, as shown in Fig. 1(b). In this stage, we compute one clothing shape that can minimize the rendering error in the silhouette across all sampled frames in the sequence. Recall that with the estimated pose $\{\boldsymbol{\theta}\}_1^L$ in Sec. 3, we can deform the consensus shape in T-pose to target poses in different frames.

2D Energy Function The energy function of the consensus shape aggregates the single-frame silhouette energy. Specifically, for each vertex of the mesh in the posed space X, we count the per-frame silhouette energy E^c_{sil}. The total energy is the average of a silhouette terms from all L frames that are sampled.

$$E^c = \frac{1}{L}\sum_{n=1}^{L} E^c_{sil\_n} \tag{5}$$

$$E^c_{sil} = \sum_i ||y_j - \Pi(\mathbf{M}_i)||^2 \tag{6}$$

E^c_{sil} describes the disparity between the rendering output and target silhouette image. Particularly, we align two silhouette maps via searching the correspondences $i \rightarrow j$ between pixels of segmented image silhouette y_j and projected mesh silhouette $\Pi(\mathbf{M}_i)$, and minimize the distance between corresponding pixels. Differentiable renderer is able to find correspondence for pixels in the rendered silhouette, but converges after dense iterations. As we target on an instant estimation of cloth shape within a few iterations, we design an ICP algorithm to directly align boundary pixels, but adapt to corner cases when the boundary of a silhouette map degenerates due to self-occlusion. Please refer to Appendix B for our correspondence searching algorithm.

Training Loss for the Gradient Rectification Network: In the training phase, the network learns to predict gradients that correct the naked body shape to the average clothing shape. We make use of an existing dynamic 3D human dataset [25] with ground truth SMPL fitting to simulate the silhouette images of naked people and people with clothes in a temporal sequence. Given 2D silhouette observations from multiple frame, we aim at recovering the averaged clothing shape $\mathbf{X}_{GT}$. Therefore, the output of the network, i.e. the rectified gradient, is supervised by the ground truth deformation from canonical clothing shape $\mathbf{X}$ to target clothing shape $\mathbf{X}_{GT}$. We define the data term E_{data} in Eq. 7 as a L-2 distance between predicted gradient and ground

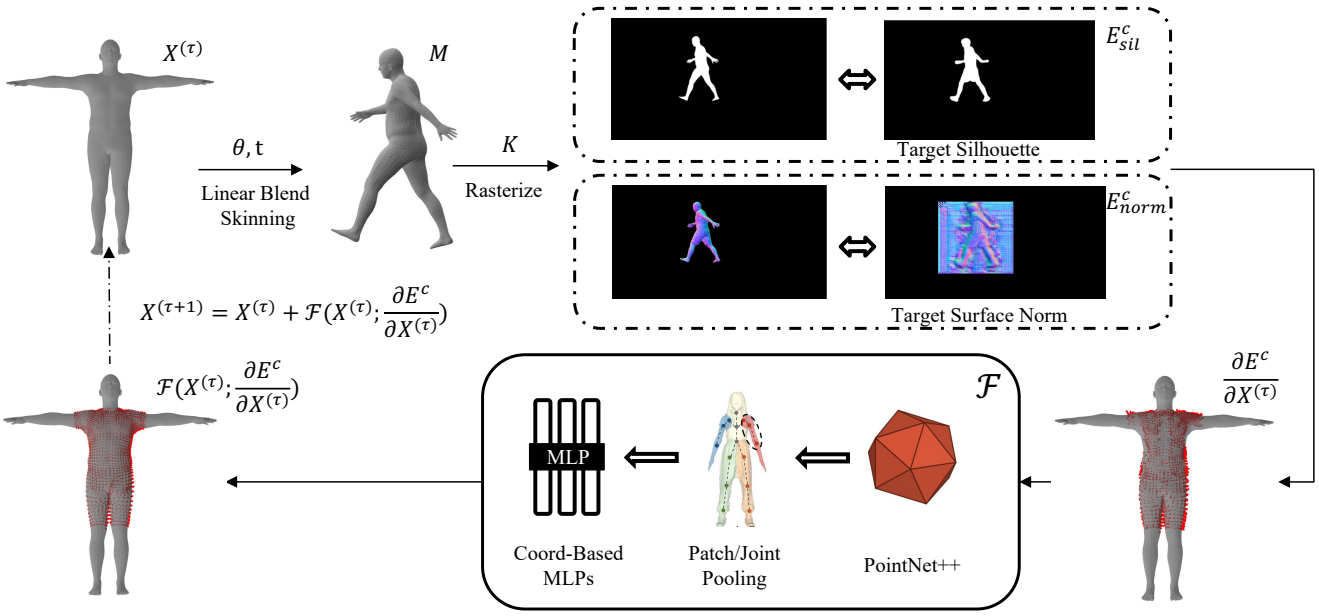

Figure 3. General training and inference procedure of our *gradient rectification network* in iteratively estimating the consensus shape, frame-dependent deformation and wrinkle details of the clothing. The thin red arrows denote the direction of input/output gradient.

truth gradient.

$$E_{data} = ||\mathbf{X}_{GT} - \mathbf{X}^{(\tau)} - \mathcal{F}(\mathbf{X}^{(\tau)}, \frac{\partial E^c}{\partial \mathbf{X}^{(\tau)}})||^2 \quad (7)$$

Additionally, we add a 2D energy term E_{2d} equivalent to E^c in Eq. 5 that supervises the updated mesh to align silhouette maps. E_{2d} reinforces the consistency between the input gradient $\frac{\partial E^c}{\partial \mathbf{X}^{(\tau)}}$ and the output gradient $\mathcal{F}(\mathbf{X}^{(\tau)}, \frac{\partial E^c}{\partial \mathbf{X}^{(\tau)}})$ of the network for visible vertices and thus prevents the overfiting. The loss function is formulated as:

$$\mathcal{L} = E_{data} + \omega_{2d} E_{2d}. \quad (8)$$

where ω_{2d} is the weight parameter to balance two terms.

5.2. Frame-dependent Clothing Shape Estimation

The consensus shape $\bar{\mathbf{D}}$ (Sec. 5.1) provides a good average (over frames) estimation of the shape. However, we miss frame-dependent details, recall that when the person moves slightly over time, the clothing shape also changes. In this section, we introduce a method to predict a frame-dependent cloth motion $\hat{\mathbf{D}}_i$ for each frame. Another GRN is separately trained with a similar scheme as Sec. 5.1. Compared with the consensus shape, in this section, the 2D energy term E^c only counts the single frame silhouette loss E^c_{sil} in Eq. 6.

$$E^c = E^c_{sil} = \sum_i ||y_j - \Pi(\mathbf{M}_i)||^2 \quad (9)$$

With temporal 3D human scans dataset, we simulate the input gradient of the training data as in Sec. 5.1, but use our estimated consensus shape as initial shape. Given a silhouette observation in a single frame, the GRN corrects the clothing shape of this frame from the average shape prediction in Sec. 5.1. Since single frame refinement introduces more ambiguity, extra regularizations including a L-2 term E_{l2} and a Laplacian term E_{lap} on X prevent the gradient predictions of invisible vertices from overfitting the training data. We reweight the data term, 2D silhouette term and regularization terms with $\omega_{sil}, \omega_{reg}, \omega_{smooth}$.

$$\mathcal{L} = E_{data} + \omega_{sil} E_{sil} + \omega_{reg} E_{l2} + \omega_{smooth} E_{lap} \quad (10)$$

5.3. Wrinkle Extraction

In previous section, we provided an estimation of the clothing shape to match the silhouette. However, the silhouette does not convey information to model wrinkles. This section introduces a separate GRN $\mathcal{F}$ to extract wrinkle details for the clothing shape from Sec. 5.2.

Traditional Shape from Shading methods [1, 39] for wrinkle extraction require complicated albedo and illumination interaction, which become problematic to apply in our in-the-wild scenario. Recently, [37, 43] propose deep learning approaches for robust cloth normal estimation from monocular image. With the predicted normal map as a supervisory signal, researchers [41, 43] are able to achieve plausible wrinkle generation by aligning the normal map of the generated clothing shape to the target normal map.

Though directly minimizing the distance of two normal maps results in plausible 3D wrinkle details, one severe problem is the runtime of the optimization process, due

to the heavy use of a differentiable renderer. Motivated by [27, 31, 49], we exploit the potential of neural networks in generating realistic wrinkle details from flat clothing surfaces and benefit from NN inference speed. However, unlike PiFUHD [31] that recovers wrinkles from a target 2D normal map, we follow our framework in Sec. 4 and take the gradient of normal energy w.r.t. 3D vertices as an input. By taking the derivative of E^c w.r.t. clothing shape X in canonical space, 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 testing scenario.

2D Energy function The 2D energy term aligns the rendered normal map of our predicted clothing shape with target normal map. Specifically, we formalize the normal energy term E^c in this stage as Eq.11.

$$E^c = E_{norm} = \sum_i ||n_i - I_i||^2 \tag{11}$$

where n_i is the surface normal of vertex i and I_i is its corresponding surface normal sampled from the ground truth normal map. In practice, n_i is gathered by a bilinear interpolation on the target normal map $N(\cdot)$.

$$I_i = N(\Pi(\mathbf{M}_i)) \tag{12}$$

Training loss for the GRN To train this network, we simulate the wrinkle extraction process from static 3D human scans. Given 2D normal signals, the network learns the high frequency deformations from a Laplacian smoothed mesh to a target clothing mesh with wrinkles. The supervision of the network is composed of L-2 distances E_{l2} (in Eq. 7) and a normal distance E_{norm} (in Eq. 11) between corresponding vertices from the predicted mesh and ground truth mesh. E_{norm} ensures the reconstructed mesh to generate visually plausible wrinkles (*i.e.* accurate surface normals of the clothing shape).

$$\mathcal{L} = E_{data} + \omega_{norm} E_{norm} \tag{13}$$

5.4. Training Data Synthesis

The main challenge to apply our *learning-to-optimize* clothing shape recovery is the domain gap between synthetic training data and in-the-wild videos. This subsection describes our sequential training data synthesis that mimics the input of GRN at each stage of our inference pipeline. For consensus shape estimation in Sec. 5.1, we leverage a temporal 3D human clothing sequence and sample 20 frames per sequence and aggregate individually computed gradients from silhouette energy. The scan under A-pose with minimal pose-dependent deformations is considered as the target consensus shape of the sequence. For frame-dependent deformation estimation, the network learns to deform our learned consensus shape to target scans in a sampled frame. For these two stages, since the 2D silhouette loss is insufficient to guide the wrinkle generation, we smoothen the target shape with a Laplacian filter to eliminate wrinkles. Sec. 5.3 is separately trained on the Tailornet dataset [27] to recover wrinkle details of ground truth clothing shape from a smoothed input shape, since [27] contains higher resolution SMPL registered 3D scans. As training data contains SMPL registrations, in the training phase, we compute the silhouette/normal energy E_c directly from 2D projection of corresponding vertex pairs, instead of the correspondence searching in Sec. 5.1. All target silhouette/normal maps in the training phases are rendered from 3D geometry data without the need of texture information.

The concrete network architecture (in Fig. 3) and our implementation details are described in Appendix C.

6. Experiments

This section provides the accuracy and computational cost evaluation of our proposed method on public benchmarks and self-captured videos against optimization-based and learning-based methods.

6.1. Experimental Setting

Dataset: We evaluate quantitatively the performance of our method on the *Pablo* sequence from the MonoPerfCap dataset [44]. This sequence contains a 156-frame multi-view (8 camera) video and reconstructed 3D scan sequence as ground truth. Following the previous work [41, 44], we selected a single view of the *Pablo* sequence as the input for testing. Besides, we collected a set of online and smartphone shot videos for qualitative results. Our methods do not impose any assumption on the input video including visibility, body pose, and garment type (except outfits *e.g.* dress, hoodie which SMPL+D model could not represent). We also synthesize testing videos by rendering mesh sequence from BUFF dataset [46] into different views for extra quantitative comparison.

Baselines: We make comparisons with both optimization-based video human performance capture methods [41, 44] and learning-based single image 3D human reconstruction methods [3, 30, 31, 48, 49]. In particular, [3] generates clothing shape via a UV displacement map. [30, 31, 43, 48] predict an implicit function for the body and clothing from pixel aligned features. Both optimization based methods [41] and most recent learning-based methods [43, 48] integrate a SMPL fitting procedure to obtain an accurate naked body shape as a prior for garment reconstruction.

Evaluation Metrics: We align the predictions of baseline methods to ground truth scans and report the average point-to-surface distance following the evaluation protocol in MonoClothCap [40]. Concretely, since different methods have different camera settings, we first center and scale the predicted scans according to the height. Then we use

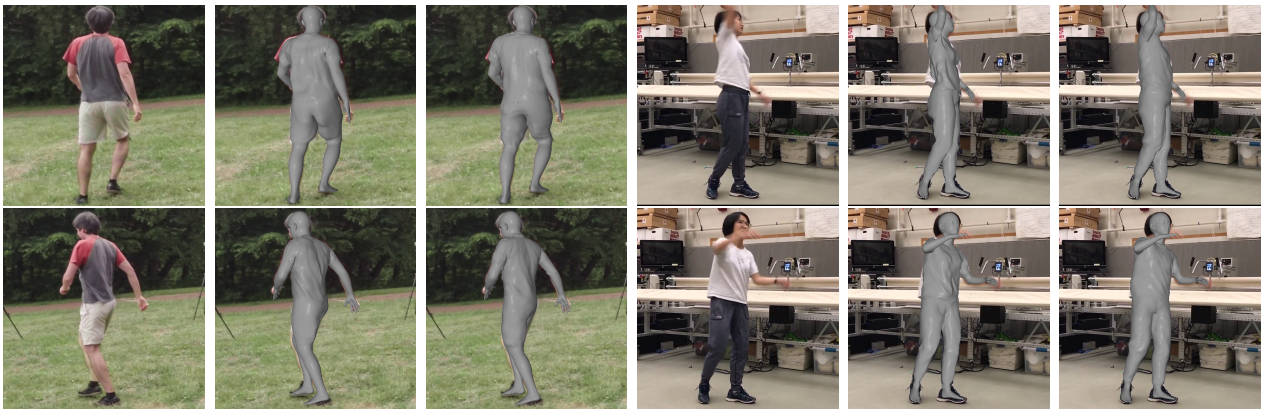

Figure 4. Qualitative results on human performance capture from the *pablo* sequence [44] and a video taken by smartphone. Note that we consider [41] as the best-performing baseline since [44] cannot be applied to wild scenario without a pre-scaned template.

Methods	Point-to-Surface Error (mm)
MonoPerfCap [44]	**14.7**†
MonoClothCap [41]	17.9
Tex2Shape [3]	27.7
DeepHuman [49]	24.2
PIFu [30]	30.5
PIFuHD [31]	26.5
PaMIR [48]	28.3
ICON-filter [43]	24.5
Ours	**17.4**

Table 1. Quantitative comparisons with SOTA methods in the *pablo* sequence [44]. The first two methods are optimization-based methods and the rest are neural network predictions. †MonoPerfCap leverages a pre-scaned template mesh of the video.

ICP [4] to register the predicted scans to clothing regions of ground truth scans. The point-to-surface error is defined as the minimal distance between the estimated clothing vertices and the surface of ground truth scans (after alignment).

6.2. Evaluation on in-the-wild Videos

Table 1 illustrates our performance compared with all learning-based single image human shape reconstruction techniques [3, 30, 31, 43, 48, 49], including methods [43, 48] that take the SMPL pose as a prior. While producing reasonable visual results in the frontal view, PiFU methods [30, 31, 48] generate noisy outliers due to the depth ambiguity, which results in a large quantitative error. We even outperform [41] without dense optimizations, and approach the performance of [44] that leverages a pre-scanned template. Compared with [41], for invisible vertices, we regularize the frame-dependent deformation to predict a more compact and flat shape. We visualize our clothing capture result on both the *Pablo* sequence and smartphone captured video in Fig. 4. With less iterations (see Table 3), we generate a similar amount of details as MonoClothCap, and performs a more robust tracking for the flying cloth under a challenging pose and view (in the right video).

View Point	PaMIR [48]	ICON-filter [43]	Ours
front	31.6	30.7	**29.5**
frontleft	29.68	31.20	**22.10**
left	39.33	33.8	**29.58**

Table 2. Quantitative comparisons in the synthetic sequence *short-long_hip* from BUFF [46] dataset. We report Point-to-Surface Error (mm) as described in Sec.6.1.

Besides, we conducted experiments on challenging online videos with fast body motion and loose long sleeve cloth, which is not supported by MonoClothCap [41]. In Fig. 5, we demonstrate the robustness of our approach against novel poses and even inaccurate 2D segmentation result thanks to our sequential human capture pipeline. In contrast, image-based human reconstruction methods [43, 48] are sensitive to the pose and viewpoint with no guarantee on temporal consistency, which results in the failure reconstructions such as bodies with missing arms and legs.

6.3. Evaluation on synthetic data

Due to the lack of realistic full body capture sequence despite the training data, we conduct experiments on synthetic video sequence rendered from BUFF dataset [46] to further test the performance of our method especially for robustness under different view point. We exclude MonoClothCap [41] from comparison since they use ground truth BUFF scans to train the PCA model. As shown in Table 2, our method outperforms all image-based 3D human reconstruction baselines [43, 48] under all perspectives. Partic-

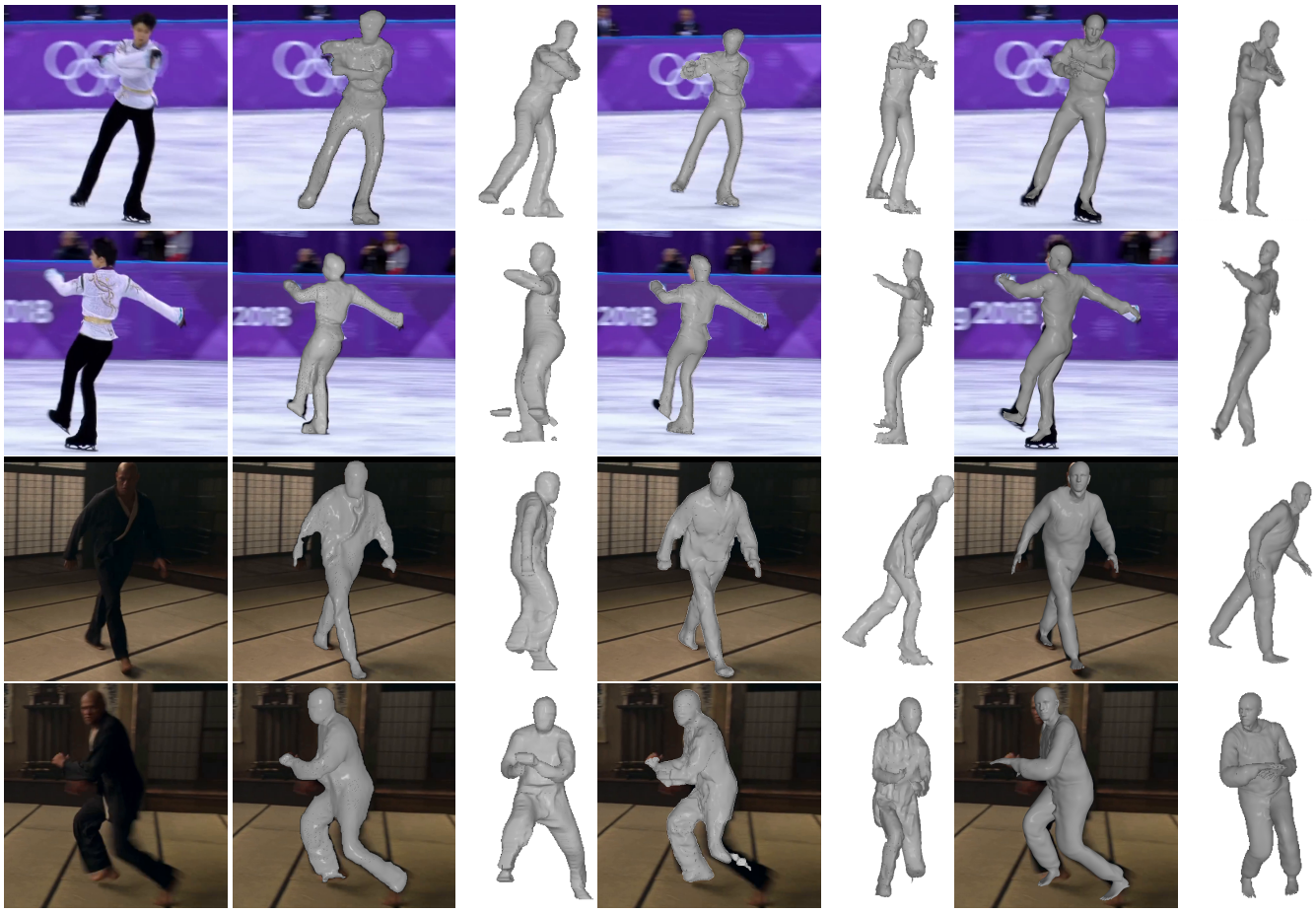

Figure 5. Qualitative results on human performance capture from challenging YouTube video with fast body motion and loose sleeves.

Stage	MonoClothCap [41]	PaMIR [48]	ICON [43]	Ours
Consensus Shape Estimation	89	7.5	1.9	0.06
Frame Refinement				1
Wrinkle Extraction	327		22	0.17
Total	416	7.5	23.9	1.23

Table 3. Per-frame runtime comparison between our method and baselines (in seconds). I/O and pose estimation times are excluded.

ularly, our method does not encounter severe performance downgrading under challenging side views.

6.4. Runtime Analysis

This section performs a runtime comparison with MonoClothCap [41], PaMIR [48] and ICON [43] on the 253-frame *pablo* sequence. These three baselines were selected as the representatives for optimization-based and PiFU-based approaches. In Table 3, our approach can achieve result faster than other baselines. 1) in inference time, our GRN converges within 3 iterations, 2) we avoid using the differentiable renderer, which consumes a huge amount of time and memory to render a high-res image, and 3) we directly predict a 3D mesh instead of an implicit function, without an extra Marching-Cube [24] step to extract mesh.

7. Conclusion

We present a human performance capture approach, which generates a temporal 3D human sequence by sequentially predicting the consensus shape, the frame-dependent clothing deformation and wrinkle details. At each stage, we leverage a *learning-to-optimize* technique to iteratively correct the clothing shape given the gradient of 2D energy w.r.t 3D clothing vertices. Trained on synthetic data, the network shows fast convergence speed and strong generalization ability to in-the-wild videos. Experiments demonstrate the accuracy, robustness and efficiency of our model in reconstructing clothing shape from a monocular RGB video.

References

[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. 5

[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, 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. 6, 7

[4] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987. 7

[5] 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. 3

[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

[7] Vasileios Choutas, Federica Bogo, Jingjing Shen, and Julien Valentin. Learning to fit morphable models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 160–179. Springer, 2022. 2

[8] Enric Corona, Gerard Pons-Moll, Guillem Alenyà, and Francesc Moreno-Noguer. Learned vertex descent: a new direction for 3d human model fitting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 146–165. Springer, 2022. 2

[9] 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, 3

[10] 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. 3

[11] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[12] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video, 2019. 2

[13] 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, 3

[14] 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, 3

[15] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 1, 2

[16] 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, 2

[17] 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

[18] 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. 3

[19] 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

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 2

[21] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu 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. 2

[22] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *European Conference on Computer Vision*, pages 322–341. Springer, 2022. 1

[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, 2, 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. 8

[25] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learn-

ing to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[26] 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

[27] 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. 6

[28] 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

[29] 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

[30] 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, 3, 6, 7

[31] 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, 3, 6, 7

[32] 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

[33] 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. *IEEE Transactions on Visualization and Computer Graphics*, 28(4):1862–1879, 2020. 3

[34] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 2

[35] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11698–11707, 2021. 2

[36] Jayakorn Vongkulbhisal, Fernando De La Torre, and Joao Paulo Costeira. Discriminative optimization: Theory and applications to computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[37] 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. 5

[38] 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

[39] 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. 5

[40] 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. 6

[41] 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, 3, 5, 6, 7, 8

[42] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 1

[43] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 1, 3, 5, 6, 7, 8

[44] 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, 2, 3, 6, 7

[45] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14484–14493, 2021. 2

[46] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 6, 7

[47] 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

[48] 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, 3, 6, 7, 8

[49] 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, 3, 6, 7