

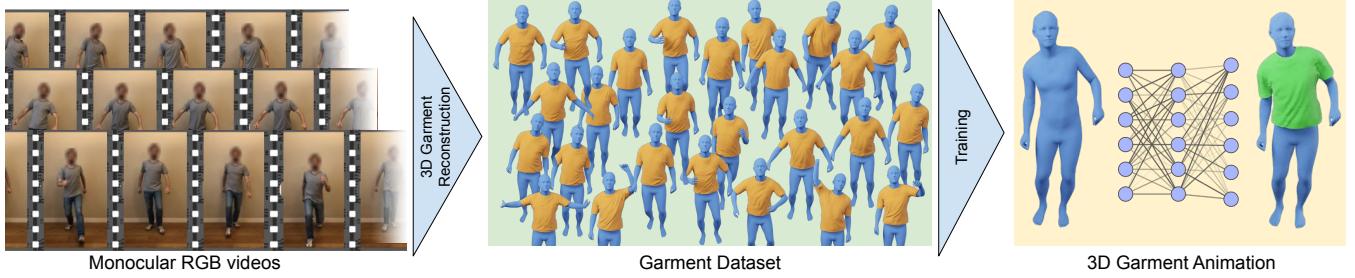
# PERGAMO: Personalized 3D Garments from Monocular Video

Andrés Casado-Elvira 

Marc Comino Trinidad 

Dan Casas 

Universidad Rey Juan Carlos, Madrid, Spain.



**Figure 1:** Using just small a collection of monocular videos captured with a mobile phone (left), we propose a novel approach to reconstruct the 3D garment layer with fine-scale details (center), and use them to learn a deformable model of the captured apparel (right). Our learned model enables to animate the captured garment, exhibiting material-specific details, as a function of the underlying body pose.

## Abstract

Clothing plays a fundamental role in digital humans. Current approaches to animate 3D garments are mostly based on realistic physics simulation, however, they typically suffer from two main issues: high computational run-time cost, which hinders their deployment; and simulation-to-real gap, which impedes the synthesis of specific real-world cloth samples. To circumvent both issues we propose PERGAMO, a data-driven approach to learn a deformable model for 3D garments from monocular images. To this end, we first introduce a novel method to reconstruct the 3D geometry of garments from a single image, and use it to build a dataset of clothing from monocular videos. We use these 3D reconstructions to train a regression model that accurately predicts how the garment deforms as a function of the underlying body pose. We show that our method is capable of producing garment animations that match the real-world behavior, and generalizes to unseen body motions extracted from motion capture dataset.

## CCS Concepts

- Computing methodologies → Computer graphics; Neural networks;

## 1. Introduction

Synthesizing realistic 3D digital humans is a key topic in Computer Animation due to the large number of applications in industries such as video games, films, telecommunications, and more. A fundamental challenge in digital humans is the modeling of dynamic clothing and garments, because it requires solutions capable of representing large diversity of designs, complex materials, and highly deformable surfaces.

The most well-established approach to model clothing is physics-based simulation, which is today a mature field capable of generating highly-detailed results [NSO12, Stu18, CLMMO14]. However, simulation methods suffer from two main difficulties that hinder their deployment in plug-and-play use cases: first, the

significant run-time computational cost, which hinders achieving real-time frame rates. Current solutions for fast simulation need to simplify dynamics [BMO<sup>\*</sup>14] or to adopt custom GPU solutions [TWL<sup>\*</sup>18, LTT<sup>\*</sup>20]; and second, the simulation-to-real gap, which prevents to synthesize simulated cloth that behaves exactly as a desired real-world target. Current methods attempt to fit simulation parameters to real-world observations, but require controlled setups and are limited to small patches of fabrics [MBT<sup>\*</sup>12].

Alternatives to physics-based cloth methods have been proposed to *independently* address these shortcomings. On one hand, impressive advances in data-driven methods based on neural networks [SOC19, PLPM20, CMM<sup>\*</sup>20] can highly-efficiently infer garment deformations for parametric avatars, generating much faster animations compared to physics simulations. However, training data

is usually obtained with simulation, hence the simulation-to-real gap remains. On the other hand, accurate 3D reconstruction methods [PMPHB17, XPB<sup>\*</sup>21] can recover highly-detailed real-world garments, which can be potentially be used to train data-driven approaches. However, the need for multi-camera professional setups prevents to democratize such strategies. All in all, modeling fast *and* real-world realistic garment deformations remains an open challenge with existing methods.

To overcome all these limitations, we introduce PERGAMO, an approach to learn a deformation model for 3D garments from a single monocular video. PERGAMO is based on two key features: it is learned from casual real-world images, hence there is no simulation-to-real gap or need for multi-camera setups; and it is highly-efficient to evaluate, since at inference time it uses a shallow neural network that directly outputs garment deformations. All in all, our main contribution is a 3D clothing reconstruction pipeline that is able to recover the explicit layer of a garment from just monocular RGB input. These reconstructions enable us to train a data-driven model to infer how a *specific* garment deforms.

To formulate PERGAMO, we use a novel two-stage approach where we first build a dataset by reconstructing the 3D geometry of deformed garments (Section 4), and then learn a nonlinear regressor from the reconstructed meshes. More specifically, we initially extract human-related features such as body segmentation, body pose, and body normals, from the input images (Section 4.1), which we leverage to deform a mesh template to reconstruct fine-scale detailed clothing using a differentiable rendering optimization (Sections 4.2 and 4.3). Then, we use the reconstructed garments as ground truth data to train a 3D garment deformation regressor (Section 5). We show that the learned model outputs pose-dependent garment surface details, such as folds and wrinkles, that closely match the real-word behaviour of the garment.

## 2. Related Work

Our work is related to the areas of 3D reconstruction and animation of garments and humans. In this section we discuss the existing literature in these areas.

**Multi-Camera 3D Garment Reconstruction.** Many works on 3D garment reconstruction require multi-camera setups to capture the scene from different viewpoints. Initial approaches use studio setups, controlled environment, and color-coded fabrics [SSK<sup>\*</sup>05, WCF07] to resolve depth and appearance ambiguities. Follow-up methods lifted the need for color patterns by, for example, using photometric stereo and controlled color lighting [HVB<sup>\*</sup>07], or using multi-view stereo [BPS<sup>\*</sup>08]. This enabled the 3D reconstruction of markerless clothing in controlled multi-camera settings, but output meshes lack high-frequency details such as folds and wrinkles. To mitigate this, some methods use postprocessing steps to add wrinkle detail. For example, Popa *et al.* [PZB<sup>\*</sup>09] add spatio-temporal coherent wrinkle details to the 3D reconstructed meshes by detecting edges in the original images. Wu *et al.* [WVL<sup>\*</sup>11] leverage shading information in general unconstrained lighting conditions to add fine-scale dynamic wrinkles. Alternatively, Lähner *et al.* [LCT18] use an image-to-image transla-

tion network based on a Generative Adversarial Network (GAN) to add high-frequency detail to a normal map.

Assuming a detailed point cloud, typically obtained from sophisticated multi-camera setup, it greatly facilitates the 3D garment reconstruction task. For example, ClothCap [PMPHB17] reconstructs fine 3D garment geometry as well as the underlying body. Similarly, Xiang *et al.* [XPB<sup>\*</sup>21] leverage an expensive 140 camera setup to reconstruct dressed humans, including an explicit clothing layer. Bhatnagar *et al.* [BTPM19] demonstrate that, given a large dataset of 3D scans, it is also possible to learn to infer 3D garments geometry from images. Bang *et al.* [BKL21] estimate accurate garment 2D patterns from full body scans, which enables to re-animate the recovered clothing using simulation. We follow a similar goal, but propose a method that only requires a single uncalibrated camera.

**Monocular 3D Garment Reconstruction.** To democratize the 3D reconstruction of garments, some methods have focused on relaxing the capture requirements and enable the reconstruction from single RGB image [ZCF<sup>\*</sup>13, CZL<sup>\*</sup>15, DDÖ<sup>\*</sup>17, JZH<sup>\*</sup>20, YPA<sup>\*</sup>18, SWY<sup>\*</sup>22]. Zhou *et al.* [ZCF<sup>\*</sup>13] combine human pose estimation, garment outline prediction, and shape-from-shading cues to reconstruct garments from monocular images. Daněk *et al.* [DDÖ<sup>\*</sup>17] propose a learning-based solution to directly estimate garment vertex displacements from single images. Similarly, Jiang *et al.* [JZH<sup>\*</sup>20] also learn to infer parametric garments ready to be worn by a SMPL [LMR<sup>\*</sup>15] human model. Other methods require complex physics simulation steps [YPA<sup>\*</sup>18, SWY<sup>\*</sup>22] to drive the surface reconstruction such that it matches the input image. We also reconstruct a garment layer from single image input, but we achieve significantly higher wrinkle detail and do not require physics simulation. Key to our success is the combination of state-of-the-art normal map prediction with a differentiable rendering scheme to optimize vertex positions.

To reduce the inherent depth ambiguity in monocular RGB images, some methods resort to depth or RGB-D input to reconstruct garments. For example, Chen *et al.* [CZL<sup>\*</sup>15] combine garment parsing and contextual priors to reconstruct garments in rest pose from Kinect-based input. Yu *et al.* [YZZ<sup>\*</sup>19] combine RGB-D input with physics-based simulation to reconstruct 3D garments. Despite the promising results, requiring depth input hinders the use of these approaches in domestic environments.

**Multi-Camera 3D Human Reconstruction.** Our work is also related to the methods that, instead of reconstructing an explicit mesh to represent the garment, aim at reconstructing a full dressed body using a single mesh [SH07, VBMP08, DAST<sup>\*</sup>08, SGDA<sup>\*</sup>10, RCR<sup>\*</sup>16]. This line of research, often referred to *performance capture*, usually employs multi-camera setups, and enables the replay of captured performances from any viewpoint. Methods within this category can be split into model-based [DAST<sup>\*</sup>08, VBMP08, SGDA<sup>\*</sup>10], which deform a template to fit it into the images, and model-free [SH07, CCS<sup>\*</sup>15, WWV<sup>\*</sup>16], which have no prior knowledge of the shape and estimate it using the visual hull given by multi-view silhouettes. Despite the high realism of the reconstructions achieved by these methods, they suffer from two main limitations: first, clothing and body are reconstructed in a single

mesh, which precludes using the garment for animation; and second, the requirement for a multi-camera setup hinders their use in uncontrolled setups. Our work addresses both limitations.

**Monocular 3D Human Reconstruction.** Many methods have been proposed to directly reconstruct dressed 3D humans as a single 3D mesh (*i.e.*, no explicit modeling of a garment layer) from monocular RGB [ZYW<sup>\*</sup>19, HXZ<sup>\*</sup>20, XPWH20, MYR<sup>\*</sup>20, AMB<sup>\*</sup>19] or RGB-D input [BNT21, YGX<sup>\*</sup>17]. Most of these works build on top of the impressive advances in deep learning for human pose estimation [MSS<sup>\*</sup>17] and parametric model fitting [KBJM18, PZZD18] in single images.

A common strategy in RGB methods is to use a 3D template of the subject, and deformed it to match visual cues (*e.g.*, silhouette, 2D joints, etc.) extracted from the input image [XCZ<sup>\*</sup>18]. Custom GPU-based optimizers have enabled such methods to run at interactive frame rates [HXZ<sup>\*</sup>19] and, more recently, data-driven methods have been trained to directly infer the geometry of dressed humans [HXZ<sup>\*</sup>20]. Despite the progress made by these methods, many challenges remain: requiring a subject template is not ideal, output meshes tend to exhibit baked-in garment details that remain static across the poses [GCSH21] and, most importantly, they do not explicitly separate garment from body.

Closest to ours is MonoClothCap [XPWH20], an RGB method that does not require a subject-specific template. Instead, a generic parametric human model [LMR<sup>\*</sup>15] is deformed using a differentiable rendering scheme to fit into per-frame estimated normals of an actor. We follow a similar path, but we are able to reconstruct an explicit layer of the garment, which subsequently allows us to train a regressor to predict how such garment deforms given arbitrary body motions.

As an alternative representation for 3D humans, recent works have explored the use of learning-based implicit functions [PFS<sup>\*</sup>19]. This representation is continuous, compact, and differentiable, which has enabled to efficiently and accurately reconstruct humans from single RGB images [SHN<sup>\*</sup>19, SSSJ20, ZYLD21] or partial point clouds [BSTM20, PBT<sup>\*</sup>21]. Despite the impressive advances in this domain, we focus our work on mesh-based representations and explicit garments, which directly enables animation of clothing in well-established character animation pipelines.

**3D Garment Animation.** Many works exist that address the problem of creating 3D animatable garments and dressed avatars. The key underlying challenge is to learn a model that is able to infer how the surface of the garment deforms as a function of the body pose and/or shape. Earlier works assume multi-camera footage [XLS<sup>\*</sup>11, CVCH14], and do not model clothing layer independently. Assuming accurate 3D reconstructed meshes of dressed humans, some methods have tackled the modeling of garment deformations as a function of the underlying body [YFHW18, NH14, PMPHB17]. Our work tackles a similar problem, but learns to deform a garment layer from monocular RGB.

With the raise of data-driven methods to learn highly nonlinear problems, recent works have attempted to model the deformation of garments leveraging large datasets of simulated data [CMM<sup>\*</sup>20, JZGF20]. Such ground-truth dataset are usually gen-

erated offline using computationally expensive physics-based simulations [NSO12]. Closest to ours, Santesteban *et al.* [SOC19] use a recurrent neural network to learn shape-and-pose dynamic deformations of a single garment. Similarly, TailorNet [PLPM20] predicts garment deformations also as a function of clothing style, and decompose deformation into a high frequency and a low frequency component to improve the detail. Vidaurre *et al.* [VSGC20] show better generalization capabilities to garment design by leveraging a fully convolutional architecture capable of inferring deformations for any mesh topology. Despite great results, these methods require loss supervision at the vertex level. Alternatively, very recent methods demonstrate that it is possible to learn deformations directly from physics-based losses [BME21, SOC22] or using distance fields [CPA<sup>\*</sup>21].

Animation of 3D garments has also been tackled using clothed humans represented as a single mesh. CAPE [MYR<sup>\*</sup>20] uses detailed 4D scans of dressed avatars to learn a graph convolutional neural network to deform a full body mesh template. More recent works use implicit representations [SYMB21, WMM<sup>\*</sup>21] and are capable of encoding finer details. Nonetheless, using a single mesh to encode body and garment is not ideal for animation. Instead, we focus on learning a model for garments using an explicit mesh 3D layer. As we demonstrate, this enables to dress animated characters using existing motion capture datasets such as AMASS [MGT<sup>\*</sup>19].

### 3. Overview

Our goal is to learn a deformable model for 3D garments that faithfully reproduces the behavior of a *real* garment-subject pair. In other words, we want to model how a specific garment deforms when worn by a particular subject. This effectively enables a *personalized* animation of garments for many tasks including: virtual try-on, telepresence, VR, videogames, and more.

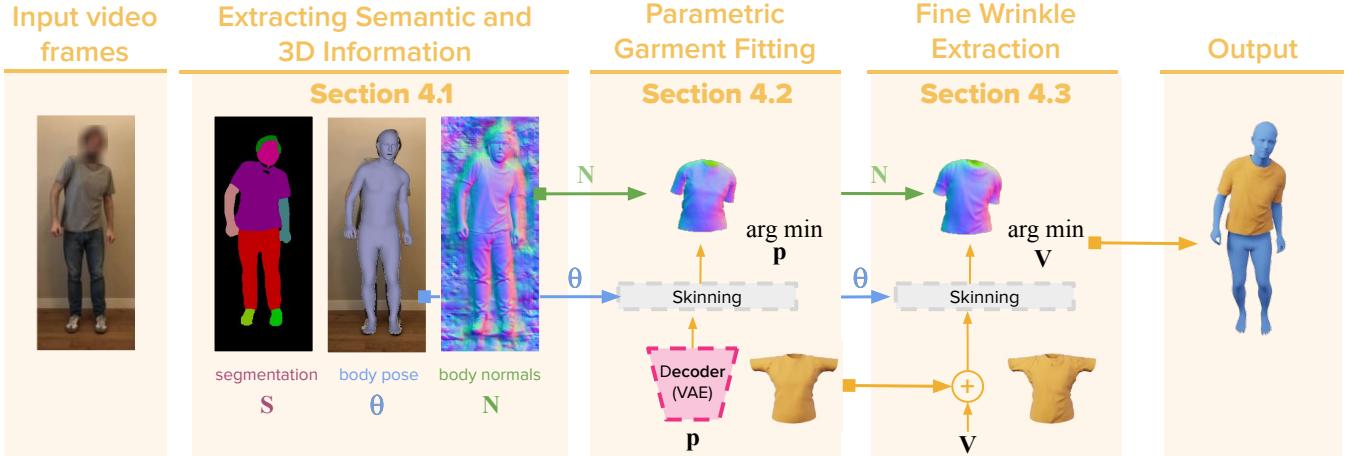
More formally, we want to learn a model  $R$  that outputs a garment

$$\mathbf{X} = R(\theta), \quad (1)$$

where  $\mathbf{X} \in \mathbb{R}^{N_G \times 3}$  encodes the vertices of the deformed garment in rest pose (*i.e.*, it contains pose-dependent wrinkles and folds), and  $\theta$  is the target pose. In practice, the garment  $\mathbf{X}$  is worn by a parametric human body such as a SMPL [LMR<sup>\*</sup>15] that represents the body surface of the target subject. To rig the deformed garment  $\mathbf{X}$ , we borrow the skinning functionalities (*e.g.*, skeleton, rigging weights) from the underlying body model.

Given our goal for personalized garments (*i.e.*, specific real-world behavior), the ground-truth data to learn  $R()$  cannot be obtained with physics-based cloth simulation tools as done in state-of-the-art methods [SOC19, PLPM20, VSGC20, BME21], for two reasons: first, due the unknown mechanical properties of the target garment, which are hard to obtain even in complex laboratory setups [MBT<sup>\*</sup>12]; and, second, due to the unknown soft-tissue model of the underlying body, which also effects how garment behave [ROCP20]. These limitations preclude generating synthetic data that match specific real-world behavior.

Therefore, instead of learning to deform garments using 3D simulations, we propose a image-driven approach that learns to infer



**Figure 2:** Overview of our 3D garment reconstructing approach. Starting from an RGB frame, we first extract human semantic and 3D information. We then fit a coarse garment template parameterized by a latent vector  $\mathbf{p}$  to the estimated body normals, and finally we add fine-scale wrinkles by optimizing per-vertex displacement  $\mathbf{V}$  using differentiable rendering.

the true behaviour of a garment from videos. To this end, in Section 4 we introduce a method to build a database of 3D meshes by directly reconstructing garments from a set of unconstrained monocular videos. Then, in Section 5, we describe a neural regressor that learns to faithfully deform a 3D garment given a target pose of the subject.

#### 4. 3D Garment Reconstruction

To obtain the dataset that we used to learn our personalized 3D deformation model, we directly reconstruct the surface of the target garment captured from monocular videos. Figure 2 illustrates our reconstruction pipeline.

Since recovering 3D geometry from single view is an ill-posed problem, we tackle the task into three steps. First, in Section 4.1, we extract a combination of semantic and 3D features from the input image. This information is key to reduce the complexity inherent to our problem. Second, in Section 4.2, we fit a coarse parametric garment model into the input image by leveraging the extracted features. And finally, in Section 4.3, we add fine wrinkle details into the garment using a differentiable rendering scheme.

##### 4.1. Extracting Semantic and 3D Information

Given a monocular video sequence  $\mathcal{I} = \{\mathbf{I}_t\}_{t=0}^T$ , where  $\mathbf{I}_t$  is an RGB frame, captured in an uncontrolled setting (*i.e.*, unknown camera parameters, unknown illumination, unknown subject pose), we initially extract a set of human-related image features for each frame  $\mathbf{I}_t$ . The extracted information will ease our ill-posed 3D reconstruction task that we describe later in the rest of this section.

First, we leverage a state-of-the-art human parsing approach [LXWY20] to assign a per-pixel label to the input image  $\mathbf{I}$  (to simplify notation, we drop the subindex  $t$  in the rest of this section, but note that all steps are done per each frame). As output, we obtain a set of binary segmentation masks  $\mathbf{S}_i$ , one for each label type,

that encodes what body part or garment type (*e.g.*, t-shirt, trousers, head, right/left arm, etc.) each pixel contains. For the results shown in this work, we use the binary mask  $\mathbf{S}_g$ , which corresponds to the upper body clothing.

Second, we use an image-to-image translation network, based on the work of Saito *et al.* [SSJ20], to estimate the surface normals  $\mathbf{N}$  for each pixel of the input image  $\mathbf{I}$ . Notice that estimated normals for pixels in the background can be unreliable or noisy, but since our goal focuses only on garment regions, we do not suffer from these artifacts.

Finally, we estimate the pose  $\theta \in \mathbb{R}^{69}$  of the subject in the input image  $\mathbf{I}$  using a state-of-the-art human pose estimation method [CPB\*20].

##### 4.2. Parametric Garment Fitting

To reconstruct a coarse approximation of the 3D geometry of a deformed garment that appears on an input frame  $\mathbf{I}$ , we first use a model-based strategy. To this end, we use a parametric garment model

$$G_{\text{coarse}}(\mathbf{p}, \theta) = W(D(\mathbf{p}), \theta, \mathcal{W}), \quad (2)$$

where  $W$  is a skinning function (*e.g.*, linear blend skinning) that articulates a parametric garment template  $D(\mathbf{p}) \in \mathbb{R}^{N_G \times 3}$  based on weights  $\mathcal{W}$ . Our parametric garment template  $D(\mathbf{p})$  is learned from a public dataset of 3D garments [SOC19] using a variational autoencoder network, where  $\mathbf{p} \in \mathbb{R}^{25}$  is the latent space learned with the autoencoder, and  $D()$  the decoder block. Intuitively,  $\mathbf{p}$  encodes a latent representation of T-shirt deformations, and  $D()$  is the mapping from latent variable to vertices position of a template mesh. This parametric template encodes coarse (*i.e.*, not specific to the material of the target garment) deformations of garments in rest pose. Hence, this first fitting step aims at recovering a coarse version of the garment visible in the input frame  $\mathbf{I}$ .

In order to fit the garment model of Equation 2 to a frame  $\mathbf{I}$ , we leverage the human-related image features  $\mathbf{S}_g$ ,  $\mathbf{N}$ , and  $\theta$  described in Section 4.1, and formulate the following optimization problem

$$\arg \min_{\mathbf{p}} \mathcal{E}_{\text{coarse}} + \mathcal{E}_{\text{sil}} + \mathcal{E}_{\text{temp}} + \mathcal{E}_{\text{reg}}. \quad (3)$$

$\mathcal{E}_{\text{coarse}}$  is the main data term, and enforces the normals of the fitted parametric garment to match the predicted normals  $\mathbf{N}$ . This term is formulated as

$$\mathcal{E}_{\text{coarse}} = \lambda_{\text{coarse}_p} \|\phi_N(G(\mathbf{p}, \theta)) - (\mathbf{N} \odot \mathbf{S}_g)\|^2 \quad (4)$$

where  $\phi_N$  is a differentiable rendering function that outputs camera-space per-pixel normals of the garment  $G(\mathbf{p}, \theta)$  and  $\odot$  is the Hadamard product (*i.e.*, element-wise multiplication).

In practice, the function  $\phi_N$  is implemented using the state-of-the-art differentiable rendering library Kaolin [FTSL\*22], which enables the computation of gradients of the image error w.r.t vertices position.

$\mathcal{E}_{\text{sil}}$  is a data-term that enforces the silhouette of the fitted garment to match the predicted mask  $\mathbf{S}_g$ ,

$$\mathcal{E}_{\text{sil}} = \lambda_{\text{sil}_p} \|\phi_S(G(\mathbf{p}, \theta)) - \mathbf{S}_g\|^2, \quad (5)$$

where  $\phi_S$  is a differentiable rendering function that outputs a mask with the silhouette of the garment (with 1s indicating inside, 0s outside and bit of a smooth transition on the edges) in camera space, also implemented using the rendering library Kaolin [FTSL\*22].  $\mathcal{E}_{\text{temp}}$  and  $\mathcal{E}_{\text{reg}}$  are regularizers formulated as

$$\mathcal{E}_{\text{temp}} = \lambda_{\text{temp}_p} \|\mathbf{p}_{t-1} - \mathbf{p}_t\|^2 \quad (6)$$

$$\mathcal{E}_{\text{reg}} = \lambda_{\text{reg}_p} \|\mathbf{V}\|^2 \quad (7)$$

that enforce temporal stability and avoid unnatural deformations, respectively.

#### 4.3. Fine Wrinkle Extraction

The parametric garment model fitted in Section 4.2 is only capable to represent coarse details of the garment. In order to add personalized details (*e.g.*, material-specific wrinkles, pose-and-shape-depending details, etc.), in this final reconstruction step we compute a vector  $\mathbf{V} \in \mathbb{R}^{N_G \times 3}$  of per-vertex 3D displacements to add fine details. To this end, we formulate our final garment model as

$$G_{\text{fine}}(\mathbf{p}, \theta, \mathbf{V}) = W(D(\mathbf{p}) + \mathbf{V}, \theta, \mathcal{W}), \quad (8)$$

and our goal in this last reconstruction step is to find the vector of 3D displacements  $\mathbf{V}$  for each frame we want to reconstruct. Notice that the rest of parameters of the model are already known.

Similar to the parametric garment fitting step from Section 4.2, we find the 3D displacements vector  $\mathbf{V}$  solving an optimization at each frame

$$\begin{aligned} \arg \min_{\mathbf{V}} \mathcal{E}_{\text{fine}} + \mathcal{E}_{\text{edge}} + \mathcal{E}_{\text{temp}} + \mathcal{E}_{\text{reg}} \\ \text{s.t. } \forall \mathbf{v}_i \in \mathbf{V} : \eta_{\max} > v_i > \eta_{\min}, \end{aligned} \quad (9)$$

where  $\eta_{\min}$  and  $\eta_{\max}$  are thresholds for minimum and maximum allowed displacements, respectively.  $\mathcal{E}_{\text{fine}}$  is the main data term, and

it resemble to the term  $\mathcal{E}_{\text{coarse}}$  from Equation 4, but with two important differences. Specifically,

$$\begin{aligned} \mathcal{E}_{\text{fine}} &= \lambda_{\text{fine}_v} \cdot \\ &\|(\nabla \phi_N(G_{\text{fine}}(\mathbf{p}, \theta, \mathbf{V})) - \nabla(\mathbf{N} \odot \mathbf{S}_g)) \odot \phi_S(G(\mathbf{p}, \theta))\|^2 \end{aligned} \quad (10)$$

where  $\nabla$  is the image gradient operator,  $\odot$  is the Hadamard product (*i.e.*, element-wise multiplication), and  $\phi_S(G(\mathbf{p}, \theta))$  a function that renders the silhouette of the parametric garment fitted in the previous section. In practice, the  $\mathcal{E}_{\text{fine}}$  term allows the vertices of the garment to move freely to match the image normal  $\mathbf{N}$ , up to fine details to reproduce wrinkles. The gradient image operator  $\nabla$  ensures attention on the parts of the image normal where normals change (*e.g.*, the wrinkles). Finally, the element-wise multiplication constrains this optimization to focus only on the pixels of the image where the garment  $G(\mathbf{p}, \theta)$  appears.

Equation 10 allows free vertex movement, however, many of the garment vertices are not visible from the input image due to occlusions and self-occlusions. Allowing occluded vertices to move can potentially lead to undesired reconstruction artefacts. To resolve this ill-posed problem, we rely on a set of regularizers to constrain the optimization. Namely, we use the following terms

$$\mathcal{E}_{\text{edge}} = \lambda_{\text{edge}_v} \|\mathbf{E}_{\text{rest}} - \mathbf{E}_t\|^2 \quad (11)$$

$$\mathcal{E}_{\text{temp}} = \lambda_{\text{temp}_v} \|\mathbf{V}_{t-1} - \mathbf{V}_t\|^2 \quad (12)$$

$$\mathcal{E}_{\text{reg}} = \lambda_{\text{reg}_v} \|\mathbf{V}\|^2, \quad (13)$$

where  $E_{\text{rest}}$  and  $E_t$  are the edge lengths of the garment in rest pose and the optimized garment, and  $\mathbf{V}$  the optimized garment vertices. These three terms temporally and spatially regularize the garment deformation, such that the reconstructed mesh does not exhibit unnatural deformations.

#### 5. 3D Garment Regressor

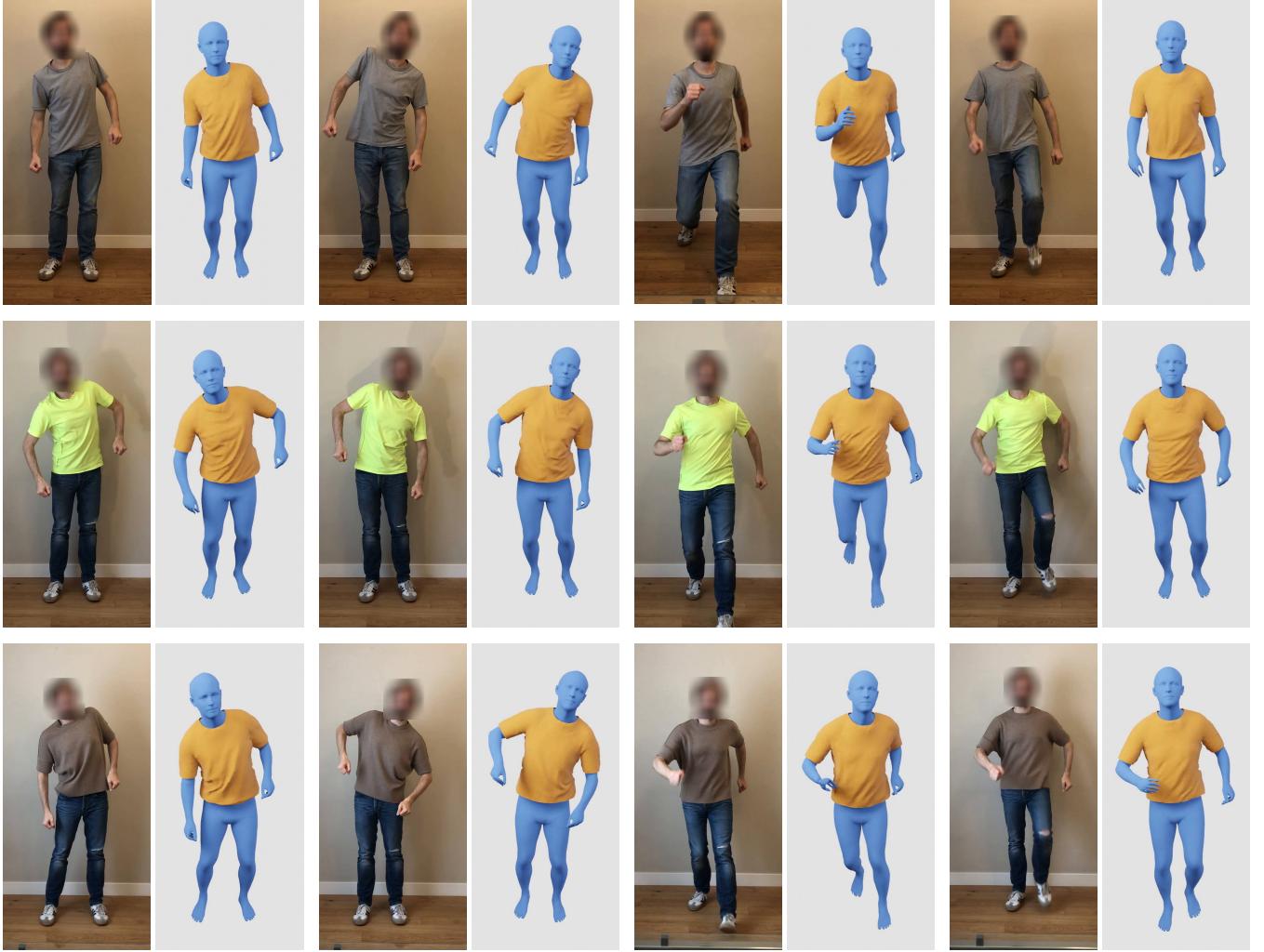
The method introduced in Section 4 allows us to create a dataset of 3D reconstructed meshes  $M$ , each of them with an associated pose parameter  $\theta$ . In this section we describe how to use this data to learn the model defined in Equation 1, capable of inferring pose-dependent deformations. Notice that the model proposed here is independent of the reconstruction method, and it could potentially be used on other similar 3D garment datasets.

We follow the standard approach in data-driven garments [SOC19], and convert our dataset of reconstructed meshes into a dataset of per-vertex displacements with respect to a template mesh in T-pose. Specifically, for each mesh  $M$  of our dataset we compute ground truth displacements

$$\Delta_{\text{GT}} = W^{-1}(M, \theta, \mathcal{W}) - \bar{\mathbf{T}}, \quad (14)$$

where  $W^{-1}$  is the inverse skinning transformation,  $\theta$  the pose parameter, and  $\bar{\mathbf{T}}$  is the average unposed garment of the dataset. The final deformed garment is then defined as  $\mathbf{X} = \bar{\mathbf{T}} + \Delta_{\text{GT}}$ .

To formulate the model defined in Equation 1, we use a regressor that is trained to predict per-vertex garment offsets  $\Delta$  as a function of the target body pose  $\theta$ . In practice, we implement it using a simple MLP network with 3 hidden layers, supervised with an L1 loss



**Figure 3:** Qualitative results of our 3D Garment Reconstruction pipeline described in Section 4. Each row shows representative frames of a different garment of our dataset, and each column a similar pose. Notice how the brown t-shirt in the bottom row, made of a thicker fabric, produces coarser wrinkles than the yellow t-shirt in the middle row. Our reconstructions faithfully capture these subtle differences, hence producing a garment-specific reconstruction details.

on vertex offsets and vertex normals:

$$\mathcal{L} = L_1(\Delta, \Delta_{GT}) + L_1(N(\bar{\mathbf{T}} + \Delta), N(\bar{\mathbf{T}} + \Delta_{GT})) \quad (15)$$

where  $N()$  is a function that computes the per-vertex normals of the garment meshes. Check Section 6.3 for further implementation details.

In contrast to existing data-driven methods that are trained on hundreds of simulated meshes [PLPM20, SOC19, BME20], we only have a reduced set of reconstructed sequences which significantly complicates the learning. To ease the generalization capabilities of our regressor despite relatively small training set, we encode the pose vector  $\theta$  in a compact subspace that better captures key pose features. To this end, we leverage the multi-modal autoencoder introduced by SoftSMPL [SGOC20], which encodes poses in  $\mathbb{R}^{10}$ .

Hence, our final regressor is  $R() : \mathbb{R}^{10} \rightarrow \mathbb{R}^{N_G \times 3}$ , where  $N_G$  is 4,424 for the garment showcased in our results.

Finally, notice that due to residual errors at inference time, a few garment vertices might collide with the underlying body mesh. Similar to existing works in data-driven garments [SOC19, PLPM20], we push problematic vertices the normal direction of the closest body point.

## 6. Results and Evaluation

### 6.1. Garment 3D Reconstruction

#### Qualitative Evaluation

Figure 3 and the supplementary video showcase a large variety of results of our 3D garment reconstruction pipeline introduced in



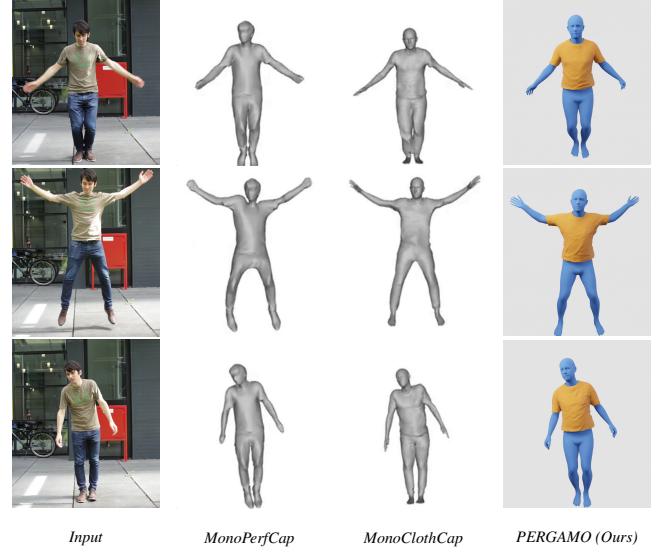
**Figure 4:** Qualitative comparison to MonoClothCap [XPWH20] in CMU Panoptic Dataset [XJS19, JSS18]

Section 4. To visually stress the quality of our results, we show side-by-side the corresponding input frame for each reconstruction.

We have processed a dataset consisting in 3 different garments –grey, yellow, and brown t-shirts–, and 12 motion sequences for each garment. Importantly, each t-shirt is made of a different fabric, hence exhibiting different folds and wrinkles. To highlight this, each row in Figure 3 depicts representative frames of a different garment, and each column similar poses. It can be observed that the yellow t-shirt –a thin 100% polyester sports t-shirt– produces fine wrinkles; the grey t-shirt –100% cotton t-shirt– produces mid-scale wrinkles; and the brown t-shirt –made with thick fabric, similar to a sweater– produces coarser wrinkles.

Additionally, in Figure 4 and in the supplementary video, we show qualitative results of our method in sequences of the Panoptic CMU dataset, and we compare to MonoClothCap [XPWH20]. Notice that MonoClothCap is a monocular reconstruction approach that outputs clothed avatars encoded in a single mesh, which is not ideal to model garments. Results demonstrate that our approach is capable of recovering finer wrinkle detail, while reconstructing an explicit garment layer.

Similarly, in Figure 5 we qualitatively compare our reconstruction results to both MonoClothCap [XPWH20] and MonoPerfCap [XCZ\*18] in an outdoor sequence. The visual quality of our recon-



**Figure 5:** Qualitative comparison to MonoClothCap [XPWH20] and MonoPerfCap in Oleksander\_outdoor sequence from [HXZ\*19]

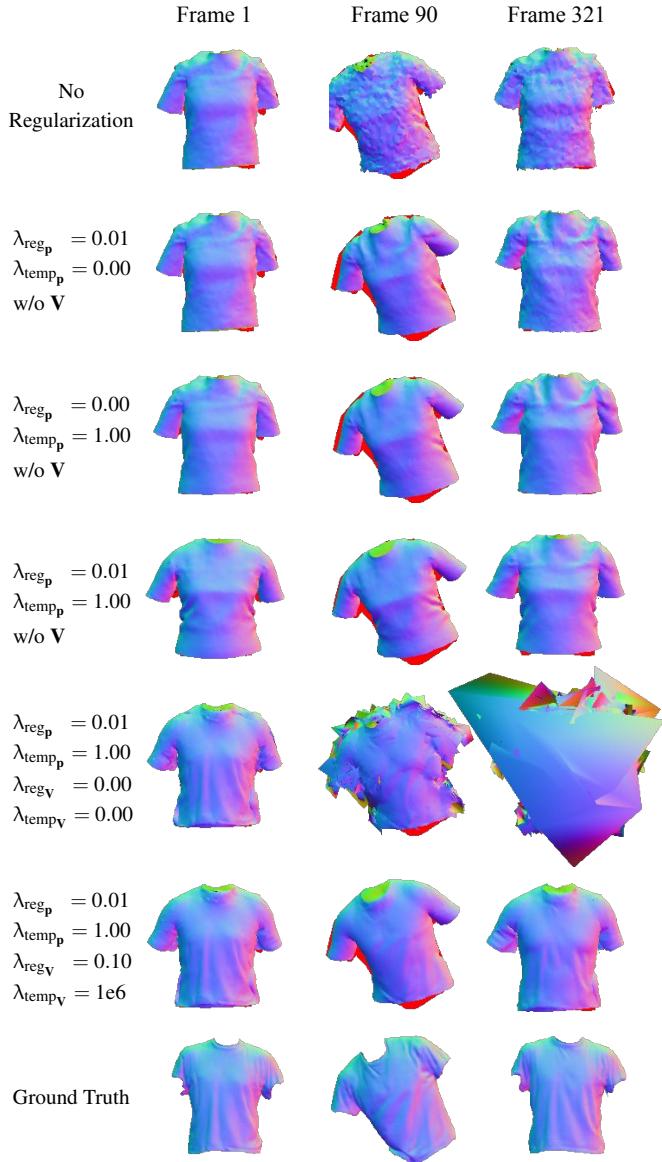
struction outperform these works, while we are able to also recover an explicit layer of the garment.

## Quantitative Evaluation

Recovering 3D geometry from a single view is an ill-posed problem. To tackle it, we use multiple regularization terms in our reconstruction pipeline. In Figure 6 and 8 we depict the effects of these terms, which we discuss below.

We first analyze the effect of the terms  $\mathcal{E}_{\text{temp}}$  and  $\mathcal{E}_{\text{reg}}$  of our parametric garment fitting step (Section 4.2). To this end, we use sequences of the BUFF dataset [ZPBPM17] which provides sequences of 3D human textured meshes. To use such mesh data in our image-based pipeline, we render RGB and ground-truth surface normal in a  $512 \times 512$  pixels image. We then process the resulting RGB frames using our method and a variety of values for  $\lambda_{\text{temp}_p}$  and  $\lambda_{\text{reg}_p}$ . In Figure 6 we show representative frames of the sequence shortlong\_hips\_96 reconstructed using different regularization choices (rows one to four). The first row shows how the mesh may degenerate when no regularization is enforced. The second row shows how a slight regularization of the VAE latent space cannot completely prevent the mesh degeneration, while inconsistencies between consecutive frames may appear causing flickering artifacts (see differences between frame 89 and 90). Temporal regularization (third row) can prevent flickering artifacts at the expense of a worse fit to the ground-truth silhouette. A careful combination of both regularization strategies may prevent degeneration and flickering while not greatly affecting the silhouette fitting.

We also analyze the effect of regularization mechanisms implemented at the fine wrinkle extraction step (Section 4.3). Row 5 at Figure 6 shows how the optimization process can explode when no regularization on the free-vertex movements is enforced. Finally,



**Figure 6:** Effect of the value choice for  $\lambda_{temp_p}$  and  $\lambda_{reg_p}$  (Rows 1 to 4). The lack of regularization causes the resulting meshes to degenerate. Effect of  $\lambda_{temp_v}$  and  $\lambda_{reg_v}$  (Rows 5 to 6). The lack of regularization can causes the vertices to explode.

row 6 show that our final choice for weights closely matches the ground truth normals.

Furthermore, we use the rendered ground-truth normal maps to quantitatively evaluate our reconstructions. In Figure 8 we provide results in terms of RMSE of the normal angle difference and precision and recall of the generated reconstructed silhouettes. Our choice of  $\lambda_{temp_p} = 0.01$  and  $\lambda_{reg_p} = 1$  achieves best results in terms of RMSE and competitive results of recall, meaning our reconstructions overlap well with the ground-truth silhouettes.

## 6.2. Garment 3D Regressor

To validate that the garment deformation regressor introduced in Section 5 generalizes to human poses not present in our dataset, we test the regressor using publicly available motion capture data. Our goal is to demonstrate that our model can produce realistic clothing animations for arbitrary motion input.

To this end, we use motion sequences from the publicly available AMASS [MGT\*19] and BUFF [ZPBPM17] datasets, and feed them into a regressor trained with the grey t-shirt dataset. Figure 7 and the supplementary video show realistic animations generated with our approach.

## 6.3. Implementation Details

In this section we provide additional implementation and performance details of our method. We have implemented PERGAMO in a desktop machine with an Intel Core i7-5820K CPU, a Nvidia RTX 3080 Ti GPU and 16GB of DDR4 RAM.

**Parametric Garment Fitting (Section 4.2).** In our experiments we set  $\lambda_{sil_p} = 1$  and start with  $\lambda_{coarse_p} = 0$  and increase it to 1 after half of the optimization iterations. In particular, we run the optimization loop for a total of 200 iterations for the first frame of a sequence using gradient descent. Afterwards, we always initialize  $\mathbf{p}_t$  using  $\mathbf{p}_{t-1}$  and, thanks to this, we can reduce the number of optimization iterations to 20. We use regularization weights  $\lambda_{temp_p} = 1$  and  $\lambda_{reg_p} = 0.01$ , and explore the effect of these in the ablation study conducted in Section 6.1.

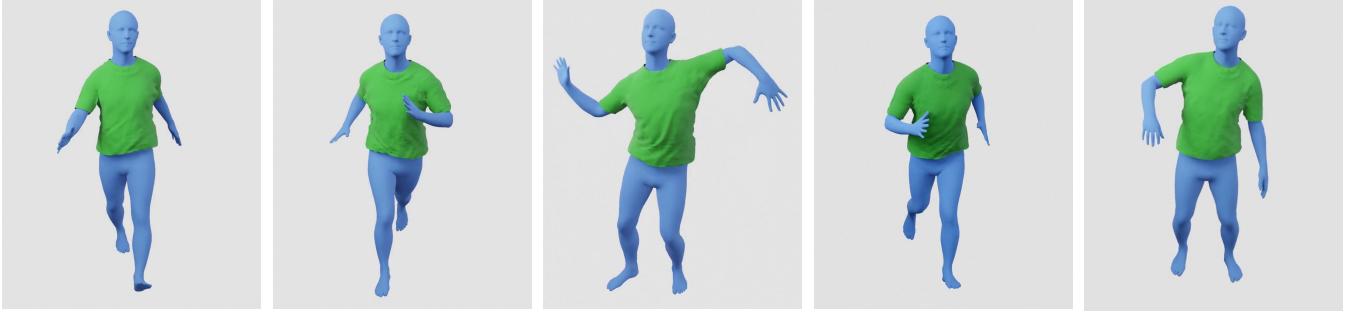
**Fine Wrinkle Extraction (Section 4.3).** We fix  $\lambda_{finev} = 6.5 \cdot 10^2$  and use regularization weights  $\lambda_{temp_v} = 10^6$ ,  $\lambda_{edge_v} = 10^4$  and  $\lambda_{reg_v} = 0.1$ . We explore the effect of these in the ablation study conducted in Section 6.1. In our experiments we run the optimization loop for a total of 200 iterations for the first frame of a sequence using gradient descent. Afterwards, we always initialize  $\mathbf{V}_t$  using  $\mathbf{V}_{t-1}$  which allows us to reduce the number of optimization iterations to 20.

The values for the different loss parameters  $\lambda$  have been set empirically using trade-off between the faithfulness and stability of the reconstructions. Selected values have been used for all reconstructions shown in this paper, including the sequences for the 3 different types of shirts, the Panoptic CMU Dataset [JSS18], and MonoPerfCap [XCZ\*18] sequences used for providing comparisons.

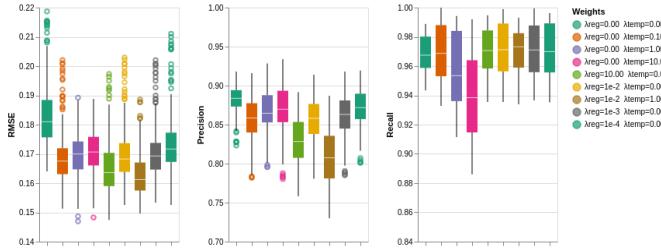
Our full reconstruction pipeline for one input frame takes around  $2.615 \pm 0.039$  seconds.

**Garment Regressor (Section 5).** We use a vanilla MLP ( $R()$ ) :  $\mathbb{R}^{10} \rightarrow \mathbb{R}^{N_G \times 3}$ ) with 3 hidden layers ( $HL_1() : \mathbb{R}^{10} \rightarrow \mathbb{R}^{N_G}$ ,  $HL_2() : \mathbb{R}^{N_G} \rightarrow \mathbb{R}^{N_G}$  and  $HL_3() : \mathbb{R}^{N_G} \rightarrow \mathbb{R}^{N_G \times 3}$ , with  $N_G = 4,424$  for the garment in our results). Each layer is fully-connected with the previous layer, and the first and second layers are followed by a LeakyRELU activation (slope = 0.1) and a Dropout (10%) layer. We use an Adam optimizer for 100 epochs with batch size of 64, and decreasing learning rate from 5e-3 to 1e-5. Finally, we have noticed that normalizing the inputs and outputs of our regressor using the training set statistics can ease the training process.

After training, predicting the deformations for a single pose takes



**Figure 7:** Garment regression results for test motion sequences unseen at training time. PERGAMO is capable of inferring dynamic 3D garment details learning from just the monocular videos. Specifically, here we show test poses from AMASS [MGT\*19] dataset sequences 08\_01, 09\_01, S\_6\_F\_7, 128\_02, and BUFF [ZPBM17]. See supplementary video for animated version of this figure.



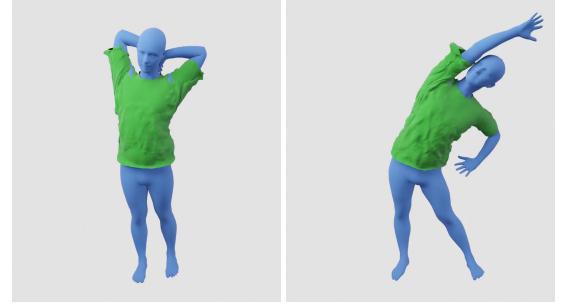
**Figure 8:** Quantitative evaluation for different values of weights  $\lambda_{\text{temp}_p}$  and  $\lambda_{\text{reg}_p}$  used in our garment reconstruction described in Section 4.2. We report RMSE, Precision and Recall scores.

around  $0.5 \pm 10^{-2}$  ms, which is coherent when compared to existing data-driven methods [SOC19, PLPM20]. Posing the resulting vertices took  $1 \pm 0.1$  ms. Lastly, the postprocess to remove potential residual collisions takes  $33 \pm 3$  ms.

## 7. Conclusions

We have introduced PERGAMO, a novel approach to learn a deformable model for 3D garments directly from monocular videos. To the best of our knowledge, PERGAMO is the first method to reconstruct an explicit 3D garment layer from single view, and use the reconstructed geometry to learn a deformable model. Since PERGAMO uses training data that comes directly from real-world images, it circumvents the simulation-to-real gap issue that existing data-driven methods suffer. Comparisons to existing methods demonstrate that our approach is capable of recovering finer wrinkle detail, and it generalizes well to unseen motions at train time.

Despite the step forward that PERGAMO does in the field of 3D garment modeling and animation, it still suffers from a few limitations. First, garment self-collisions are not explicitly modeled, therefore, despite that we regularize 3D deformations to avoid geometry artifacts, residual collisions can occur in highly deformed areas such as armpits. Second, our approach is only capable of reconstructing garments that are close to the coarse deformable mesh used in our reconstructing step. It remains open to future research how to generalize the reconstruction pipeline to a larger variety of



**Figure 9:** Examples of failure cases on extreme body poses.

clothing. Additionally, at the moment we do not model body shape variations, which prevents PERGAMO to generalize to unseen subjects. Third, our method is dependent on the quality of the information used as input. Noisy estimated normals, pose or segmentation can impact the quality of the reconstructed garments. And fourth, the regressor may produce unsatisfactory results when dealing with extreme poses as can be seen in Figure 9.

## Acknowledgments

This work has been partially funded by: the Comunidad de Madrid in the framework of the Multiannual Agreement with the Universidad Rey Juan Carlos in line of Action 1, "Encouragement of research for young PhD", project CaptHuRe (M2736); the Universidad Rey Juan Carlos through the Distinguished Researcher position INVESDIST-04 under the call from 17/12/2020; and by a Leonardo Fellowship from the Fundación BBVA.

## References

- [AMB\*19] ALLDIECK T., MAGNOR M., BHATNAGAR B. L., THEOBALT C., PONS-MOLL G.: Learning to Reconstruct People in Clothing from a Single RGB Camera. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2019). 3
- [BKL21] BANG S., KOROSTELEVA M., LEE S.-H.: Estimating garment patterns from static scan data. In *Computer Graphics Forum* (2021), vol. 40, pp. 273–287. doi:10.1111/cgf.14272. 2
- [BME20] BERTICHE H., MADADI M., ESCALERA S.: CLOTH3D: Clothed 3D Humans. In *Proc. of European Conference on Computer Vision (ECCV)* (2020). 6
- [BME21] BERTICHE H., MADADI M., ESCALERA S.: PBNS: Physically Based Neural Simulation for Unsupervised Garment Pose Space Deformation. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 40, 6 (2021). doi:10.1145/3478513.3480479. 3
- [BMO\*14] BENDER J., MÜLLER M., OTADUY M. A., TESCHNER M., MACKLIN M.: A Survey on Position-Based Simulation Methods in Computer Graphics. *Computer Graphics Forum* 33, 6 (2014), 228–251. 1
- [BNT21] BUROV A., NIESSNER M., THIES J.: Dynamic Surface Function Networks for Clothed Human Bodies. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2021). 3
- [BPS\*08] BRADLEY D., POPA T., SHEFFER A., HEIDRICH W., BOUBEKEUR T.: Markerless Garment Capture. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 27, 3 (2008), 99. doi:10.1145/1360612.1360698. 2
- [BTPM20] BHATNAGAR B. L., SMINCHISESCU C., THEOBALT C., PONS-MOLL G.: Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction. In *Proc. of European Conference on Computer Vision (ECCV)* (2020). 3
- [BTTPM19] BHATNAGAR B. L., TIWARI G., THEOBALT C., PONS-MOLL G.: Multi-garment net: Learning to Dress 3D People from Images. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2019), pp. 5420–5430. doi:10.1109/ICCV.2019.00552. 2
- [CCS\*15] COLLET A., CHUANG M., SWEENEY P., GILLETT D., EVSEEV D., CALABRESE D., HOPPE H., KIRK A., SULLIVAN S.: High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 34, 4 (2015), 1–13. doi:10.1145/2766945. 2
- [CLMM014] CIRIO G., LOPEZ-MORENO J., MIRAUT D., OTADUY M. A.: Yarn-level simulation of woven cloth. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 33, 6 (2014), 1–11. doi:10.1145/2661229.2661279. 1
- [CMM\*20] CHENTANEZ N., MACKLIN M., MÜLLER M., JESCHKE S., KIM T.-Y.: Cloth and Skin Deformation with a Triangle Mesh Based Convolutional Neural Network. *Computer Graphics Forum (Proc. SCA)* 39, 8 (2020), 123–134. doi:doi.org/10.1111/cgf.14107. 1, 3
- [CPA\*21] CORONA E., PUMAROLA A., ALENYÀ G., PONS-MOLL G., MORENO-NOGUER F.: SMPLicit: Topology-aware Generative Model for Clothed People. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2021). 3
- [CPB\*20] CHOUTAS V., PAVLAKOS G., BOLKART T., TZIONAS D., BLACK M. J.: Monocular Expressive Body Regression through Body-Driven Attention. In *Proc. of European Conference on Computer Vision (ECCV)* (2020), pp. 20–40. 4
- [CVCH14] CASAS D., VOLINO M., COLLOMOSSE J., HILTON A.: 4D Video Textures for Interactive Character Appearance. *Computer Graphics Forum (Proc. Eurographics)* 33, 2 (2014), 371–380. doi:10.1111/cgf.12296. 3
- [CZL\*15] CHEN X., ZHOU B., LU F., WANG L., BI L., TAN P.: Garment Modeling with a Depth Camera. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 34, 6 (2015). doi:10.1145/2816795.2818059. 2
- [DAST\*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. In *Proc. ACM SIGGRAPH* (2008). doi:10.1145/1399504.1360697. 2
- [DDÖ\*17] DANĚŘEK R., DIBRA E., ÖZTIRELI C., ZIEGLER R., GROSS M.: DeepGarment: 3D Garment Shape Estimation from a Single Image. *Computer Graphics Forum (Proc. Eurographics)* 36, 2 (2017), 269–280. doi:10.1111/cgf.13125. 2
- [FTSL\*22] FUJI TSANG C., SHUGRINA M., LAFLECHE J. F., TAKIKAWA T., WANG J., LOOP C., CHEN W., JATAVALLABHULA K. M., SMITH E., ROZANTSEV A., PEREL O., SHEN T., GAO J., FIDLER S., STATE G., GORSKI J., XIANG T., LI J., LI M., LEBAREDIAN R.: Kaolin: A Pytorch Library for Accelerating 3D Deep Learning Research. <https://github.com/NVIDIAAGameWorks/kaolin>, 2022. 5
- [GCSH21] GUO C., CHEN X., SONG J., HILLIGES O.: Human performance capture from monocular video in the wild. In *International Conference on 3D Vision (3DV)* (2021), pp. 889–898. doi:10.1109/3DV53792.2021.00097. 3
- [HVB\*07] HERNÁNDEZ C., VOGIATZIS G., BROSTOW G. J., STENGER B., CIPOLLA R.: Non-Rigid Photometric Stereo with Colored Lights. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2007). 2
- [HXZ\*19] HABERMANN M., XU W., ZOLLMÖFER M., PONS-MOLL G., THEOBALT C.: Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 38, 2 (mar 2019). doi:10.1145/3311970. 3, 7
- [HXZ\*20] HABERMANN M., XU W., ZOLLMÖFER M., PONS-MOLL G., THEOBALT C.: DeepCap: Monocular Human Performance Capture Using Weak Supervision. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). 3
- [JSS18] JOO H., SIMON T., SHEIKH Y.: Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2018). 7, 8
- [JZGF20] JIN N., ZHU Y., GENG Z., FEDKIW R.: A Pixel-Based Framework for Data-Driven Clothing. *Computer Graphics Forum (Proc. of SCA)* (2020). doi:10.1111/cgf.14108. 3
- [JZH\*20] JIANG B., ZHANG J., HONG Y., LUO J., LIU L., BAO H.: BCNet: Learning Body and Cloth Shape from A Single Image. In *Proc. of European Conference on Computer Vision (ECCV)* (2020). 2
- [KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end Recovery of Human Shape and Pose. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2018). 3
- [LCT18] LAHNER Z., CREMERS D., TUNG T.: DeepWrinkles: Accurate and Realistic Clothing Modeling. In *Proc. of European Conference on Computer Vision (ECCV)* (2018). doi:10.1007/978-3-030-01225-0\_41. 2
- [LMR\*15] LOPEZ M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 34, 6 (2015), 1–16. doi:10.1145/2816795.2818013. 2, 3
- [LTT\*20] LI C., TANG M., TONG R., CAI M., ZHAO J., MANOCHA D.: P-Cloth: Interactive cloth simulation on multi-GPU systems using dynamic matrix assembly and pipelined implicit integrators. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 39, 6 (2020), 180:1–15. 1
- [LXWY20] LI P., XU Y., WEI Y., YANG Y.: Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020). doi:10.1109/TPAMI.2020.3048039. 4
- [MBT\*12] MIGUEL E., BRADLEY D., THOMASZEWSKI B., BICKEL B., MATSIK W., OTADUY M. A., MARSCHNER S.: Data-driven estimation of cloth simulation models. *Computer Graphics Forum (Proc. Eurographics)* 31 (2012), 519–528. doi:10.1111/j.1467-8659.2012.03031.x. 1, 3

- [MGT\*19] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: AMASS: Archive of Motion Capture as Surface Shapes. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (Oct. 2019), pp. 5442–5451. 3, 8, 9
- [MSS\*17] MEHTA D., SRIDHAR S., SOTNYCHENKO O., RHODIN H., SHAFIEI M., SEIDEL H.-P., XU W., CASAS D., THEOBALT C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. vol. 36. doi:10.1145/3072959.3073596. 3
- [MYR\*20] MA Q., YANG J., RANJAN A., PUJADES S., PONS-MOLL G., TANG S., BLACK M. J.: Learning to Dress 3D People in Generative Clothing. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2020). 3
- [NH14] NEOPHYTOU A., HILTON A.: A layered model of human body and garment deformation. In *Proc. of International Conference on 3D Vision (3DV)* (2014), pp. 171–178. doi:10.1109/3DV.2014.52. 3
- [NSO12] NARAIN R., SAMII A., O'BRIEN J. F.: Adaptive Anisotropic Remeshing for Cloth Simulation. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 31, 6 (2012), 1–10. doi:10.1145/2366145.2366171. 1, 3
- [PBT\*21] PALAFOX P., BOŽIĆ A., THIES J., NIESSNER M., DAI A.: NPMs: Neural Parametric Models for 3D Deformable Shapes. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2021). 3
- [PFS\*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2019). 3
- [PLPM20] PATEL C., LIAO Z., PONS-MOLL G.: The Virtual Tailor: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2020). 1, 3, 6, 9
- [PMPHB17] PONS-MOLL G., PUJADES S., HU S., BLACK M. J.: ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 36, 4 (2017). doi:10.1145/3072959.3073711. 2, 3
- [PZB\*09] POPA T., ZHOU Q., BRADLEY D., KRAEVOY V., FU H., SHEFFER A., HEIDRICH W.: Wrinkling Captured Garments Using Space-Time Data-Driven Deformation. *Computer Graphics Forum (Proc. Eurographics)* 28, 2 (2009), 427–435. doi:10.1111/j.1467-8659.2009.01382.x. 2
- [PZZD18] PAVLAKOS G., ZHU L., ZHOU X., DANIIILIDIS K.: Learning to estimate 3d human pose and shape from a single color image. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 459–468. 3
- [RCR\*16] ROBERTINI N., CASAS D., RHODIN H., SEIDEL H.-P., THEOBALT C.: Model-Based Outdoor Performance Capture. In *Proc. of International Conference on 3D Vision (3DV)* (2016), pp. 166–175. doi:10.1109/3DV.2016.25. 2
- [ROCP20] ROMERO C., OTADUY M. A., CASAS D., PEREZ J.: Modeling and Estimation of Nonlinear Skin Mechanics for Animated Avatars. *Computer Graphics Forum (Proc. Eurographics)* 39, 2 (2020). 3
- [SGDA\*10] STOLL C., GALL J., DE AGUIAR E., THRUN S., THEOBALT C.: Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 29, 6 (2010). 2
- [SGOC20] SANTESTEBAN I., GARCES E., OTADUY M. A., CASAS D.: SoftSMPL: Data-driven Modeling of Nonlinear Soft-tissue Dynamics for Parametric Humans. *Computer Graphics Forum (Proc. Eurographics)* 39, 2 (2020). doi:10.1111/cgf.13912. 6
- [SH07] STARCK J., HILTON A.: Surface Capture for Performance-Based Animation. *IEEE Computer Graphics and Applications* 27, 3 (2007), 21–31. doi:10.1109/MCG.2007.68. 2
- [SHN\*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: PIFU: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2019). 3
- [SOC19] SANTESTEBAN I., OTADUY M. A., CASAS D.: Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum (Proc. Eurographics)* 38, 2 (2019). doi:10.1111/cgf.13643. 1, 3, 4, 5, 6, 9
- [SOC22] SANTESTEBAN I., OTADUY M. A., CASAS D.: SNUG: Self-Supervised Neural Dynamic Garments. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2022). 3
- [SSK\*05] SCHOLZ V., STICH T., KECKEISEN M., WACKER M., MAGNOR M.: Garment Motion Capture Using Color-Coded Patterns. *Computer Graphics Forum (Proc. Eurographics)* 24, 3 (2005), 439–447. doi:10.1111/j.1467-8659.2005.00869.x. 2
- [SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2020). 3, 4
- [Stu18] STUYCK T.: Cloth Simulation for Computer Graphics. *Synthesis Lectures on Visual Computing: Computer Graphics, Animation, Computational Photography, and Imaging* 10, 3 (2018), 1–121. doi:10.2200/S00867ED1V01Y201807VCP032. 1
- [SWY\*22] SU Z., WAN W., YU T., LIU L., FANG L., WANG W., LIU Y.: MulayCap: Multi-Layer Human Performance Capture Using a Monocular Video Camera. *IEEE Transactions on Visualization and Computer Graphics* 28, 4 (2022), 1862–1879. doi:10.1109/TVCG.2020.3027763. 2
- [SYMB21] SAITO S., YANG J., MA Q., BLACK M. J.: SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2021). 3
- [TWL\*18] TANG M., WANG T., LIU Z., TONG R., MANOCHA D.: I-Cloth: Incremental Collision Handling for GPU-Based Interactive Cloth Simulation. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 37, 6 (2018). doi:10.1145/3272127.3275005. 1
- [VBMP08] VLASIC D., BARAN I., MATUSIK W., POPOVIĆ J.: Articulated mesh animation from multi-view silhouettes. In *Proc. of ACM SIGGRAPH* (2008), pp. 1–9. doi:10.1145/1399504.1360696. 2
- [VSGC20] VIDAUERRE R., SANTESTEBAN I., GARCES E., CASAS D.: Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. *Computer Graphics Forum (Proc. SCA)* 39, 8 (2020). doi:10.1111/cgf.14109. 3
- [WCF07] WHITE R., CRANE K., FORSYTH D. A.: Capturing and animating occluded cloth. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 26 (2007). doi:10.1145/1275808.1276420. 2
- [WMM\*21] WANG S., MIHAJLOVIC M., MA Q., GEIGER A., TANG S.: MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images. In *Advances in Neural Information Processing Systems* (2021). 3
- [WVL\*11] WU C., VARANASI K., LIU Y., SEIDEL H.-P., THEOBALT C.: Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2011), pp. 1108–1115. 2
- [WWV\*16] WANG R., WEI L., VOUGA E., HUANG Q., CEYLAN D., MEDIONI G., LI H.: Capturing dynamic textured surfaces of moving targets. In *Proc. of European Conference on Computer Vision (ECCV)* (2016), pp. 271–288. doi:10.1007/978-3-319-46478-7\_17. 2
- [XCZ\*18] XU W., CHATTERJEE A., ZOLLHÖFER M., RHODIN H., MEHTA D., SEIDEL H.-P., THEOBALT C.: MonoPerfCap: Human Performance Capture From Monocular Video. *ACM Transactions on Graphics* 37, 2 (2018). doi:10.1145/3181973. 3, 7, 8
- [XJS19] XIANG D., JOO H., SHEIKH Y.: Monocular total capture: Positing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019). 7
- [XLS\*11] XU F., LIU Y., STOLL C., TOMPKIN J., BHARAJ G., DAI Q., SEIDEL H.-P., KAUTZ J., THEOBALT C.: Video-Based Characters:

Creating New Human Performances from a Multi-View Video Database.  
[doi:10.1145/2010324.1964927](https://doi.org/10.1145/2010324.1964927). 3

[XPB\*21] XIANG D., PRADA F., BAGAUTDINOV T., XU W., DONG Y., WEN H., HODGINS J., WU C.: Modeling Clothing as a Separate Layer for an Animatable Human Avatar. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 40, 6 (2021), 1–15. [doi:10.1145/3478513.3480545](https://doi.org/10.1145/3478513.3480545). 2

[XPWH20] XIANG D., PRADA F., WU C., HODGINS J.: Monocloth-cap: Towards temporally coherent clothing capture from monocular rgb video. In *International Conference on 3D Vision (3DV)* (2020), pp. 322–332. [doi:10.1109/3DV50981.2020.00042](https://doi.org/10.1109/3DV50981.2020.00042). 3, 7

[YFHW18] YANG J., FRANCO J.-S., HÉTROY-WHEELER F., WUHRER S.: Analyzing Clothing Layer Deformation Statistics of 3D Human Motions. In *Proc. of European Conference on Computer Vision (ECCV)* (2018). [doi:10.1007/978-3-030-01234-2\\_15](https://doi.org/10.1007/978-3-030-01234-2_15). 3

[YGX\*17] YU T., GUO K., XU F., DONG Y., SU Z., ZHAO J., LI J., DAI Q., LIU Y.: BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 910–919. 3

[YPA\*18] YANG S., PAN Z., AMERT T., WANG K., YU L., BERG T., LIN M. C.: Physics-Inspired Garment Recovery from a Single-View Image. *ACM Transactions on Graphics* 37, 5 (2018). [doi:10.1145/3026479](https://doi.org/10.1145/3026479). 2

[YZZ\*19] YU T., ZHENG Z., ZHONG Y., ZHAO J., DAI Q., PONS-MOLL G., LIU Y.: SimulCap: Single-view human performance capture with cloth simulation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019). 2

[ZCF\*13] ZHOU B., CHEN X., FU Q., GUO K., TAN P.: Garment modeling from a single image. *Computer Graphics Forum* 32, 7 (2013), 85–91. [doi:10.1111/cgf.12215](https://doi.org/10.1111/cgf.12215). 2

[ZPBPM17] ZHANG C., PUJADES S., BLACK M. J., PONS-MOLL G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proc. of Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 4191–4200. 7, 8, 9

[ZYLD21] ZHENG Z., YU T., LIU Y., DAI Q.: PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). [doi:10.1109/TPAMI.2021.3050505](https://doi.org/10.1109/TPAMI.2021.3050505). 3

[ZYW\*19] ZHENG Z., YU T., WEI Y., DAI Q., LIU Y.: Deephuman: 3d human reconstruction from a single image. In *Proc. of IEEE International Conference on Computer Vision (ICCV)* (2019). 3