PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization

Shunsuke Saito[1,2*] Zeng Huang[1,2*] Ryota Natsume[3*]
Shigeo Morishima[3] Angjoo Kanazawa[4] Hao Li[1,2,5]

[1]University of Southern California [2]USC Institute for Creative Technologies
[3]Waseda University [4]University of California, Berkeley [5]Pinscreen

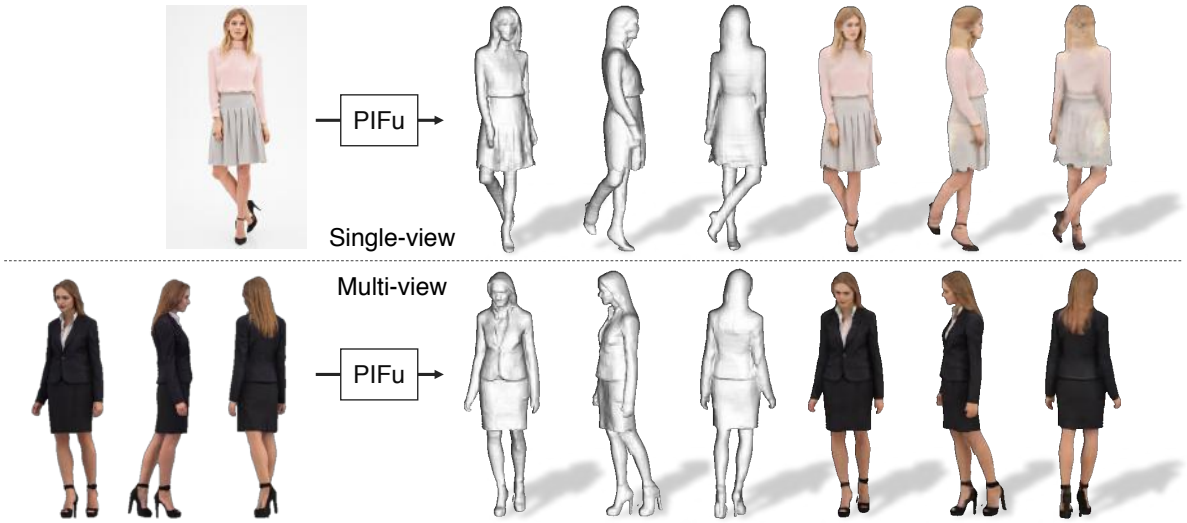

Figure 1: **Pixel-aligned Implicit function (PIFu):** We present pixel-aligned implicit function (PIFu), which allows recovery of high-resolution 3D textured surfaces of clothed humans from a single input image (top row). Our approach can digitize intricate variations in clothing, such as wrinkled skirts and high-heels, including complex hairstyles. The shape and textures can be fully recovered including largely unseen regions such as the back of the subject. PIFu can also be naturally extended to multi-view input images (bottom row).

Abstract

We introduce Pixel-aligned Implicit Function (PIFu), an implicit representation that locally aligns *pixels of 2D images with the global context of their corresponding 3D object. Using PIFu, we propose an end-to-end deep learning method for digitizing highly detailed clothed humans that can infer both 3D surface and texture from a single image, and optionally, multiple input images. Highly intricate shapes, such as hairstyles, clothing, as well as their variations and deformations can be digitized in a unified way. Compared to existing representations used for 3D deep learning, PIFu produces high-resolution surfaces including largely unseen regions such as the back of a person. In particular, it is memory efficient unlike the voxel representation, can handle arbitrary topology, and the resulting surface is spatially aligned with the input image. Furthermore, while previous techniques are designed to process either a single image or multiple views, PIFu extends naturally to arbitrary number of views. We demonstrate high-resolution and robust reconstructions on real world images from the DeepFashion dataset, which contains a variety of challenging clothing types. Our method achieves state-of-the-art performance on a public benchmark and outperforms the prior work for clothed human digitization from a single image. The project website can be found at* https://shunsukesaito.github.io/PIFu/

1. Introduction

In an era where immersive technologies and sensor-packed autonomous systems are becoming increasingly prevalent, our ability to create virtual 3D content at scale

* - indicates equal contribution

goes hand-in-hand with our ability to digitize and understand 3D objects in the wild. If digitizing an entire object in 3D would be as simple as taking a picture, there would be no need for sophisticated 3D scanning devices, multi-view stereo algorithms, or tedious capture procedures, where a sensor needs to be moved around.

For certain domain-specific objects, such as faces, human bodies, or known man made objects, it is already possible to infer relatively accurate 3D surfaces from images with the help of parametric models, data-driven techniques, or deep neural networks. Recent 3D deep learning advances have shown that general shapes can be inferred from very few images and sometimes even a single input. However, the resulting resolutions and accuracy are typically limited, due to ineffective model representations, even for domain specific modeling tasks.

We propose a new Pixel-aligned Implicit Function (PIFu) representation for 3D deep learning for the challenging problem of textured surface inference of clothed 3D humans from a single or multiple input images. While most successful deep learning methods for 2D image processing (e.g., semantic segmentation [51], 2D joint detection [57], etc.) take advantage of "fully-convolutional" network architectures that preserve the spatial alignment between the image and the output, this is particularly challenging in the 3D domain. While voxel representations [59] can be applied in a fully-convolutional manner, the memory intensive nature of the representation inherently restrict its ability to produce fine-scale detailed surfaces. Inference techniques based on global representations [19, 30, 1] are more memory efficient, but cannot guarantee that details of input images are preserved. Similarly, methods based on implicit functions [11, 44, 38] rely on the global context of the image to infer the overall shape, which may not align with the input image accurately. On the other hand, PIFu aligns individual local features at the pixel level to the global context of the entire object in a fully convolutional manner, and does not require high memory usage, as in voxel-based representations. This is particularly relevant for the 3D reconstruction of clothed subjects, whose shape can be of arbitrary topology, highly deformable and highly detailed. While [26] also utilize local features, due to the lack of 3D-aware feature fusion mechanism, their approach is unable to reason 3D shapes from a single-view. In this work we show that combination of local features and 3D-aware implicit surface representation makes a significant difference including highly detailed reconstruction even from a single view.

Specifically, we train an encoder to learn individual feature vectors for each pixel of an image that takes into account the global context relative to its position. Given this per-pixel feature vector and a specified z-depth along the outgoing camera ray from this pixel, we learn an implicit function that can classify whether a 3D point corresponding to this z-depth is inside or outside the surface. In particular, our feature vector spatially aligns the global 3D surface shape to the pixel, which allows us to preserve local details present in the input image while inferring plausible ones in unseen regions.

Our end-to-end and unified digitization approach can directly predict high-resolution 3D shapes of a person with complex hairstyles and wearing arbitrary clothing. Despite the amount of unseen regions, particularly for a single-view input, our method can generate a complete model similar to ones obtained from multi-view stereo photogrammetry or other 3D scanning techniques. As shown in Figure 1, our algorithm can handle a wide range of complex clothing, such as skirts, scarfs, and even high-heels while capturing high frequency details such as wrinkles that match the input image at the pixel level.

By simply adopting the implicit function to regress RGB values at each queried point along the ray, PIFu can be naturally extended to infer per-vertex colors. Hence, our digitization framework also generates a complete texture of the surface, while predicting plausible appearance details in unseen regions. Through additional multi-view stereo constraints, PIFu can also be naturally extended to handle multiple input images, as is often desired for practical human capture settings. Since producing a complete textured mesh is already possible from a single input image, adding more views only improves our results further by providing additional information for unseen regions.

We demonstrate the effectiveness and accuracy of our approach on a wide range of challenging real-world and unconstrained images of clothed subjects. We also show for the first time, high-resolution examples of monocular and textured 3D reconstructions of dynamic clothed human bodies reconstructed from a video sequence. We provide comprehensive evaluations of our method using ground truth 3D scan datasets obtained using high-end photogrammetry. We compare our method with prior work and demonstrate the state-of-the-art performance on a public benchmark for digitizing clothed humans.

2. Related Work

Single-View 3D Human Digitization. Single-view digitization techniques require strong priors due to the ambiguous nature of the problem. Thus, parametric models of human bodies and shapes [4, 35] are widely used for digitizing humans from input images. Silhouettes and other types of manual annotations [20, 72] are often used to initialize the fitting of a statistical body model to images. Bogo et al. [8] proposed a fully automated pipeline for unconstrained input data. Recent methods involve deep neural networks to improve the robustness of pose and shape parameters estimations for highly challenging images [30, 46]. Methods that involve part segmentation as input [33, 42] can produce more accurate fittings. Despite their capability to capture human body measurements and motions, parametric models

only produce a naked human body. The 3D surfaces of clothing, hair, and other accessories are fully ignored. For skin-tight clothing, a displacement vector for each vertex is sometimes used to model some level of clothing as shown in [2, 67, 1]. Nevertheless, these techniques fail for more complex topology such as dresses, skirts, and long hair. To address this issue, template-free methods such as BodyNet [59] learn to directly generate a voxel representation of the person using a deep neural network. Due to the high memory requirements of voxel representations, fine-scale details are often missing in the output. More recently, [39] introduced a multi-view inference approach by synthesizing novel silhouette views from a single image. While multi-view silhouettes are more memory efficient, concave regions are difficult to infer as well as consistently generated views. Consequentially, fine-scale details cannot be produced reliably. In contrast, PIFu is memory efficient and is able to capture fine-scale details present in the image, as well as predict per-vertex colors.

Multi-View 3D Human Digitization. Multi-view acquisition methods are designed to produce a complete model of a person and simplify the reconstruction problem, but are often limited to studio settings and calibrated sensors. Early attempts are based on visual hulls [37, 61, 15, 14] which uses silhouettes from multiple views to carve out the visible areas of a capture volume. Reasonable reconstructions can be obtained when large numbers of cameras are used, but concavities are inherently challenging to handle. More accurate geometries can be obtained using multi-view stereo constraints [55, 75, 65, 16] or using controlled illumination, such as multi-view photometric stereo techniques [62, 68]. Several methods use parametric body models to further guide the digitization process [54, 17, 5, 25, 3, 1]. The use of motion cues has also been introduced as additional priors [47, 70]. While it is clear that multi-view capture techniques outperform single-view ones, they are significantly less flexible and deployable.

A middle ground solution consists of using deep learning frameworks to generate plausible 3D surfaces from very sparse views. [12] train a 3D convolutional LSTM to predict the 3D voxel representation of objects from arbitrary views. [32] combine information from arbitrary views using differentiable unprojection operations. [28] also uses a similar approach, but requires at least two views. All of these techniques rely on the use of voxels, which is memory intensive and prevents the capture of high-frequency details. [26, 18] introduced a deep learning approach based on a volumetric occupancy field that can capture dynamic clothed human performances using sparse viewpoints as input. At least three views are required for these methods to produce reasonable output.

Texture Inference. When reconstructing a 3D model from a single image, the texture can be easily sampled from the input. However, the appearance in occluded regions needs to be inferred in order to obtain a complete texture. Related to the problem of 3D texture inference are view-synthesis approaches that predict novel views from a single image [73, 43] or multiple images [56]. Within the context of texture mesh inference of clothed human bodies, [39] introduced a view synthesis technique that can predict the back view from the front one. Both front and back views are then used to texture the final 3D mesh, however self-occluding regions and side views cannot be handled. Akin to the image inpainting problem [45], [40] inpaints UV images that are sampled from the output of detected surface points, and [58, 22] infers per voxel colors, but the output resolution is very limited. [31] directly predicts RGB values on a UV parameterization, but their technique can only handle shapes with known topology and are therefore not suitable for clothing inference. Our proposed method can predict per vertex colors in an end-to-end fashion and can handle surfaces with arbitrary topology.

3. PIFu: Pixel-Aligned Implicit Function

Given a single or multi-view images, our goal is to reconstruct the underlining 3D geometry and texture of a clothed human while preserving the detail present in the image. To this end, we introduce Pixel-Aligned Implicit Functions (PIFu) which is a memory efficient and spatially-aligned 3D representation for 3D surfaces. An implicit function defines a surface as a level set of a function f, *e.g.* $f(X) = 0$ [50]. This results in a memory efficient representation of a surface where the space in which the surface is embedded does not need to be explicitly stored. The proposed pixel-aligned implicit function consists of a fully convolutional image encoder g and a continuous implicit function f represented by multi-layer perceptrons (MLPs), where the surface is defined as a level set of

$$f(F(x), z(X)) = s : s \in \mathbb{R}, \tag{1}$$

where for a 3D point X, $x = \pi(X)$ is its 2D projection, $z(X)$ is the depth value in the camera coordinate space, $F(x) = g(I(x))$ is the image feature at x. We assume a weak-perspective camera, but extending to perspective cameras is straightforward. Note that we obtain the pixel-aligned feature $F(x)$ using bilinear sampling, because the 2D projection of X is defined in a continuous space rather than a discrete one (i.e., pixel).

The key observation is that we learn an implicit function over the 3D space with pixel-aligned image features rather than global features, which allows the learned functions to preserve the local detail present in the image. The continuous nature of PIFu allows us to generate detailed geometry with arbitrary topology in a memory efficient manner. Moreover, PIFu can be cast as a general framework that can be extended to various co-domains such as RGB colors.

Digitization Pipeline. Figure 2 illustrates the overview of our framework. Given an input image, PIFu for surface

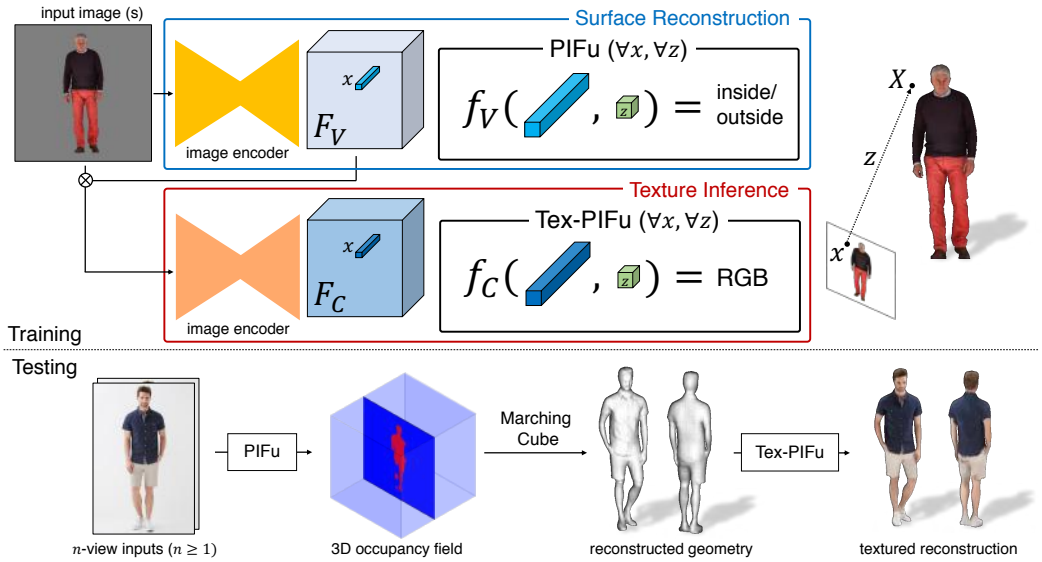

Figure 2: **Overview of our clothed human digitization pipeline:** Given an input image, a pixel-aligned implicit function (PIFu) predicts the continuous inside/outside probability field of a clothed human. Similarly, PIFu for texture inference (Tex-PIFu) infers an RGB value at given 3D positions of the surface geometry with arbitrary topology.

reconstruction predicts the continuous inside/outside probability field of a clothed human, in which iso-surface can be easily extracted (Sec. 3.1). Similarly, PIFu for texture inference (Tex-PIFu) outputs an RGB value at 3D positions of the surface geometry, enabling texture inference in self-occluded surface regions and shapes of arbitrary topology (Sec. 3.2). Furthermore, we show that the proposed approach can handle single-view and multi-view input naturally, which allows us to produce even higher fidelity results when more views are available (Sec. 3.3).

3.1. Single-view Surface Reconstruction

For surface reconstruction, we represent the ground truth surface as a 0.5 level-set of a continuous 3D occupancy field:

$$f_v^*(X) = \begin{cases} 1, & \text{if } X \text{ is inside mesh surface} \\ 0, & \text{otherwise} \end{cases}. \quad (2)$$

We train a pixel-aligned implicit function (PIFu) f_v by minimizing the average of mean squared error:

$$\mathcal{L}_V = \frac{1}{n}\sum_{i=1}^{n} |f_v(F_V(x_i), z(X_i)) - f_v^*(X_i)|^2, \quad (3)$$

where $X_i \in \mathbb{R}^3$, $F_V(x) = g(I(x))$ is the image feature from the image encoder g at $x = \pi(X)$ and n is the number of sampled points. Given a pair of an input image and the corresponding 3D mesh that is spatially aligned with the input image, the parameters of the image encoder g and PIFu f_v are jointly updated by minimizing Eq. 3. As Bansal et al. [6] demonstrate for semantic segmentation, training an image encoder with a subset of pixels does not hurt convergence compared with training with all the pixels. During inference, we densely sample the probability field over the 3D space and extract the iso-surface of the probability field at threshold 0.5 using the Marching Cube algorithm [36]. This implicit surface representation is suitable for detailed objects with arbitrary topology. Aside from PIFu's expressiveness and memory-efficiency, we develop a spatial sampling strategy that is critical for achieving high-fidelity inference.

Spatial Sampling. The resolution of the training data plays a central role in achieving the expressiveness and accuracy of our implicit function. Unlike voxel-based methods, our approach does not require discretization of ground truth 3D meshes. Instead, we can directly sample 3D points on the fly from the ground truth mesh in the original resolution using an efficient ray tracing algorithm [63]. Note that this operation requires water-tight meshes. In the case of non-watertight meshes, one can use off-the-shelf solutions to make the meshes watertight [7]. Additionally, we observe that the sampling strategy can largely influence the final reconstruction quality. If one uniformly samples points in the 3D space, the majority of points are far from the iso-surface, which would unnecessarily weight the network toward outside predictions. On the other hand, sampling only around the iso-surface can cause overfitting. Consequently, we propose to combine uniform sampling and adaptive sampling based on the surface geometry. We first randomly sample points on the surface geometry and add offsets with normal distribution $\mathcal{N}(0, \sigma)$ ($\sigma = 5.0$ cm in our experiments) for x, y, and z axis to perturb their positions around the surface. We combine those samples

with uniformly sampled points within bounding boxes using a ratio of 16 : 1. We provide an ablation study on our sampling strategy in the supplemental materials.

3.2. Texture Inference

While texture inference is often performed on either a 2D parameterization of the surface [31, 21] or in view-space [39], PIFu enables us to directly predict the RGB colors on the surface geometry by defining s in Eq. 1 as an RGB vector field instead of a scalar field. This supports texturing of shapes with arbitrary topology and self-occlusion. However, extending PIFu to color prediction is a non-trivial task as RGB colors are defined only on the surface while the 3D occupancy field is defined over the entire 3D space. Here, we highlight the modification of PIFu in terms of training procedure and network architecture.

Given sampled 3D points on the surface $X \in \Omega$, the objective function for texture inference is the average of L1 error of the sampled colors as follows:

$$\mathcal{L}_C = \frac{1}{n}\sum_{i=1}^{n} |f_c(F_C(x_i), z(X_i)) - C(X_i)|, \qquad (4)$$

where $C(X_i)$ is the ground truth RGB value on the surface point $X_i \in \Omega$ and n is the number of sampled points. We found that naively training f_c with the loss function above severely suffers from overfitting. The problem is that f_c is expected to learn not only RGB color on the surface but also the underlining 3D surfaces of the object so that f_c can infer texture of unseen surface with different pose and shape during inference, which poses a significant challenge. We address this problem with the following modifications. First, we condition the image encoder for texture inference with the image features learned for the surface reconstruction F_V. This way, the image encoder can focus on color inference of a given geometry even if unseen objects have different shape, pose, or topology. Additionally, we introduce an offset $\epsilon \sim \mathcal{N}(0, d)$ to the surface points along the surface normal N so that the color can be defined not only on the exact surface but also on the 3D space around it. With the modifications above, the training objective function can be rewritten as:

$$\mathcal{L}_C = \frac{1}{n}\sum_{i=1}^{n} \left|f_c(F_C(x'_i, F_V), X'_{i,z}) - C(X_i)\right|, \qquad (5)$$

where $X'_i = X_i + \epsilon \cdot N_i$. We use $d = 1.0$ cm for all the experiments. Please refer to the supplemental material for the network architecture for texture inference.

3.3. Multi-View Stereo

Additional views provide more coverage about the person and should improve the digitization accuracy. Our formulation of PIFu provides the option to incorporate information from more views for both surface reconstruction and texture inference. We achieve this by using PIFu to learn a feature embedding for every 3D point in space. Specifically the output domain of Eq. 1 is now a n-dimensional vector space $s \in \mathbb{R}^n$ that represents the latent feature embedding associated with the specified 3D coordinate and the image feature from each view. Since this embedding is defined in the 3D world coordinate space, we can aggregate the embedding from all available views that share the same 3D point. The aggregated feature vector can be used to make a more confident prediction of the surface and the texture.

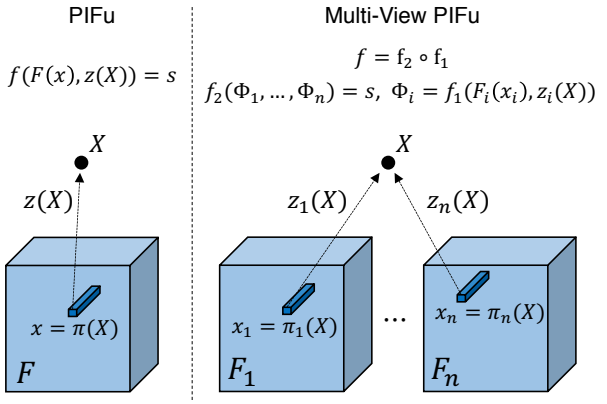

Figure 3: **Multi-view PIFu:** PIFu can be extended to support multi-view inputs by decomposing implicit function f into a feature embedding function f_1 and a multi-view reasoning function f_2. f_1 computes a feature embedding from each view in the 3D world coordinate system, which allows aggregation from arbitrary views. f_2 takes aggregated feature vector to make a more informed 3D surface and texture prediction.

Specifically we decompose the pixel-aligned function f into a feature embedding network f_1 and a multi-view reasoning network f_2 as $f := f_2 \circ f_1$. See Figure 3 for illustrations. The first function f_1 encodes the image feature $F_i(x_i) : x_i = \pi_i(X)$ and depth value $z_i(X)$ from each view point i into latent feature embedding Φ_i. This allows us to aggregate the corresponding pixel features from all the views. Now that the corresponding 3D point X is shared by different views, each image can project X on its own image coordinate system by $\pi_i(X)$ and $z_i(X)$. Then, we aggregate the latent features Φ_i by average pooling operation and obtain the fused embedding $\bar{\Phi} = \text{mean}(\{\Phi_i\})$. The second function f_2 maps from the aggregated embedding $\bar{\Phi}$ to our target implicit field s (i.e., inside/outside probability for surface reconstruction and RGB value for texture inference). The additive nature of the latent embedding allows us to incorporate arbitrary number of inputs. Note that a single-view input can be also handled without modification in the same framework as the average operation simply returns the original latent embedding. For training, we use the same training procedure as the aforementioned single-view cases including loss functions and the point sampling scheme.

While we train with three random views, our experiments show that the model can incorporate information from more than three views (See Sec. 4).

4. Experiments

We evaluate our proposed approach on a variety of datasets, including RenderPeople [48] and BUFF [71], which has ground truth measurements, as well as DeepFashion [34] which contains a diverse variety of complex clothing.

Implementation Detail. Since the framework of PIFu is not limited to a specific network architecture, one can technically use any fully convolutional neural network as the image encoder. For surface reconstruction, we found that stacked hourglass [41] architectures are effective with better generalization on real images. The image encoder for texture inference adopts the architecture of CycleGAN [74] consisting of residual blocks [29]. The implicit function is based on a multi-layer perceptron, whose layers have skip connections from the image feature $F(x)$ and depth z in spirit of [11] to effectively propagate the depth information. Tex-PIFu takes $F_C(x)$ together with the image feature for surface reconstruction $F_V(x)$ as input. For multi-view PIFu, we simply take an intermediate layer output as feature embedding and apply average pooling to aggregate the embedding from different views. Please refer to the supplemental materials for more detail on network architecture and training procedure.

4.1. Quantitative Results

We quantitatively evaluate our reconstruction accuracy with three metrics. In the model space, we measure the average point-to-surface Euclidean distance (P2S) in cm from the vertices on the reconstructed surface to the ground truth. We also measure the Chamfer distance between the reconstructed and the ground truth surfaces. In addition, we introduce the normal reprojection error to measure the fineness of reconstructed local details, as well as the projection consistency from the input image. For both reconstructed and ground truth surfaces, we render their normal maps in the image space from the input viewpoint respectively. We then calculate the L2 error between these two normal maps.

Single-View Reconstruction. In Table 1 and Figure 5, we evaluate the reconstruction errors for each method on both Buff and RenderPeople test set. Note that while Voxel Regression Network (VRN) [27], IM-GAN [11], and ours are retrained with the same High-Fidelity Clothed Human dataset we use for our approach, the reconstruction of [39, 59] are obtained from their trained models as off-the-shelf solutions. Since single-view inputs leaves the scale factor ambiguous, the evaluation is performed with the known scale factor for all the approaches. In contrast to the state-of-the-art single-view reconstruction method using implicit function (IM-GAN) [10] that reconstruct surface from one global feature per image, our method outputs pixel-aligned high-resolution surface reconstruction that captures hair styles and wrinkles of the clothing. We also demonstrate the expressiveness of our PIFu representation compared with voxels. Although VRN and ours share the same network architecture for the image encoder, the higher expressiveness of implicit representation allows us to achieve higher fidelity.

In Figure 6, we also compare our single-view texture inferences with a state-of-the-art texture inference method on clothed human, SiCloPe [39], which infers a 2D image from the back view and stitches it together with the input front-view image to obtain textured meshes. While SiCloPe suffers from projection distortion and artifacts around the silhouette boundary, our approach predicts textures on the surface mesh directly, removing projection artifacts.

Multi-View Reconstruction. In Table 2 and Figure 7, we compare our multi-view reconstruction with other deep learning-based multi-view methods including LSM [32], and a deep visual hull method proposed by Huang et al. [24]. All approaches are trained on the same High-Fidelity Clothed Human Dataset using three-view input images. Note that Huang et al. can be seen as a degeneration of our method where the multi-view feature fusion process solely relies on image features, without explicit conditioning on the 3D coordinate information. To evaluate the importance of conditioning on the depth, we denote our network architecture removing z from input of PIFu as Huang et al. in our experiments. We demonstrate that PIFu achieves the state-of-the-art reconstruction qualitatively and quantitatively in our metrics. We also show that our multi-view PIFu allows us to increasingly refine the geometry and texture by incorporating arbitrary number of views in Figure 8.

4.2. Qualitative Results

In Figure 4, we present our digitization results using real world input images from the DeepFashion dataset [34]. We demonstrate our PIFu can handle wide varieties of clothing, including skirts, jackets, and dresses. Our method can produce high-resolution local details, while inferring plausible 3D surfaces in unseen regions. Complete textures are also inferred successfully from a single input image, which allows us to view our 3D models from 360 degrees. We refer to the supplemental video[2] for additional static and dynamic results. In particular, we show how dynamic clothed human performances and complex deformations can be digitized in 3D from a single 2D input video.

5. Discussion

We introduced a novel pixel-aligned implicit function, which spatially aligns the pixel-level information of the input image with the shape of the 3D object, for deep

[2] https://youtu.be/S1FpjwKqtPs

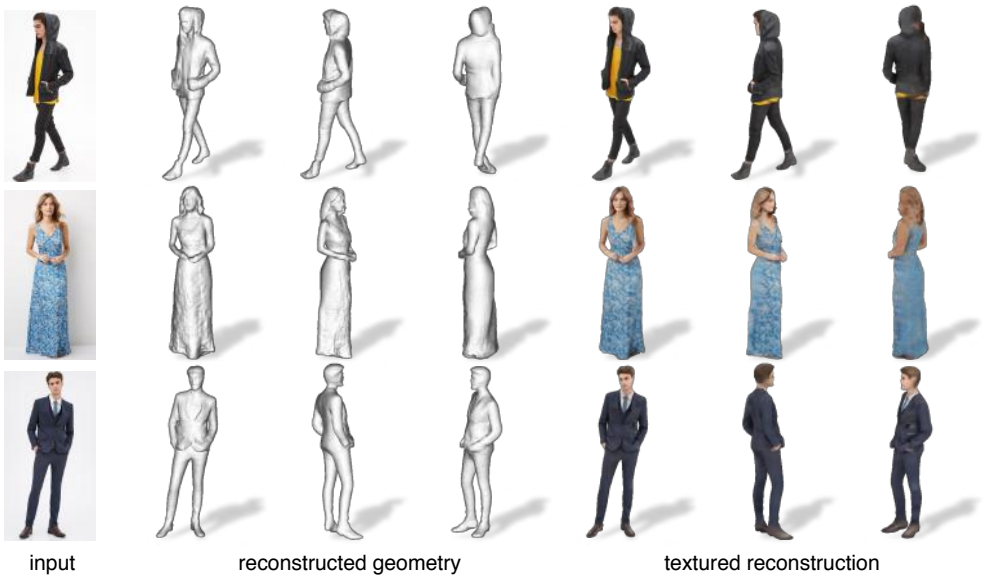

Figure 4: **Qualitative single-view results on real images from DeepFashion dataset [34].** The proposed Pixel-Aligned Implicit Functions, PIFu, achieves a topology-free, memory efficient, spatially-aligned 3D reconstruction of geometry and texture of clothed human.

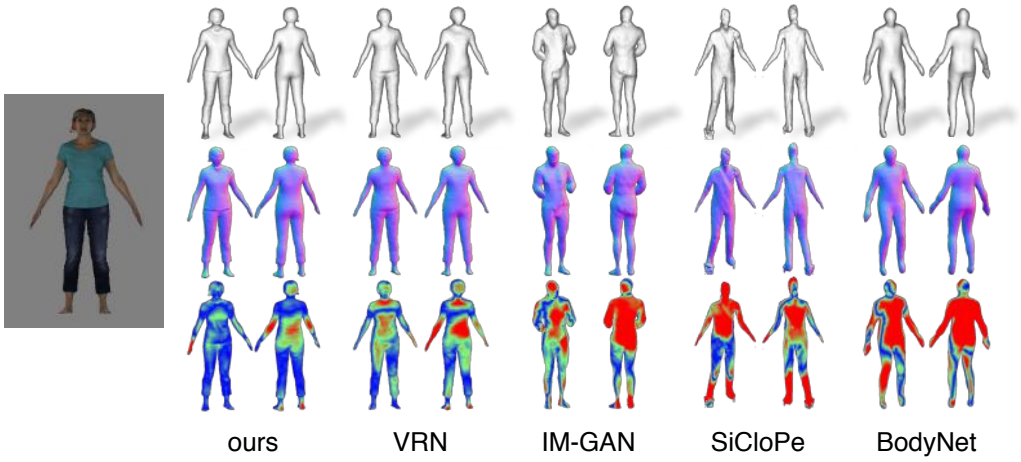

Figure 5: **Comparison with other human digitization methods from a single image.** For each input image on the left, we show the predicted surface (top row), surface normal (middle row), and the point-to-surface errors (bottom row).

	RenderPeople			Buff		
Methods	Normal	P2S	Chamfer	Normal	P2S	Chamfer
BodyNet	0.262	5.72	5.64	0.308	4.94	4.52
SiCloPe	0.216	3.81	4.02	0.222	4.06	3.99
IM-GAN	0.258	2.87	3.14	0.337	5.11	5.32
VRN	0.116	**1.42**	1.56	0.130	2.33	2.48
Ours	**0.084**	1.52	**1.50**	**0.0928**	**1.15**	**1.14**

Table 1: Quantitative evaluation on RenderPeople and BUFF dataset for single-view reconstruction.

	RenderPeople			Buff		
Methods	Normal	P2S	Chamfer	Normal	P2S	Chamfer
LSM	0.251	4.40	3.93	0.272	3.58	3.30
Deep V-Hull	**0.093**	0.639	0.632	0.119	0.698	0.709
Ours	0.094	**0.554**	**0.567**	**0.107**	**0.665**	**0.641**

Table 2: Quantitative comparison between multi-view reconstruction algorithms using 3 views.

learning based 3D shape and texture inference of clothed humans from a single input image. Our experiments indicate that highly plausible geometry can be inferred including largely unseen regions such as the back of a person, while preserving high-frequency details present

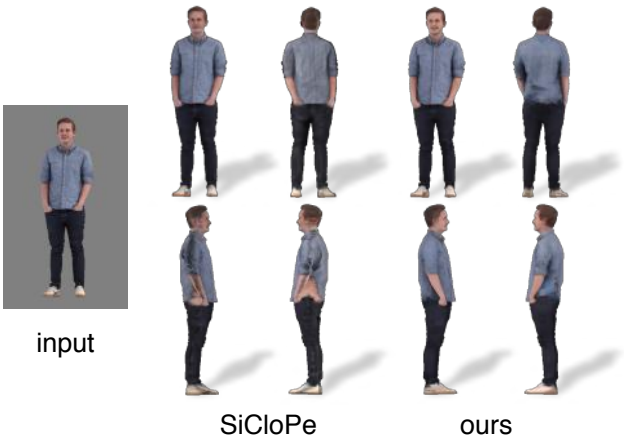

Figure 6: **Comparison with SiCloPe [39] on texture inference.** While texture inference via a view synthesis approach suffers from projection artifacts, proposed approach does not as it directly inpaints textures on the surface geometry.

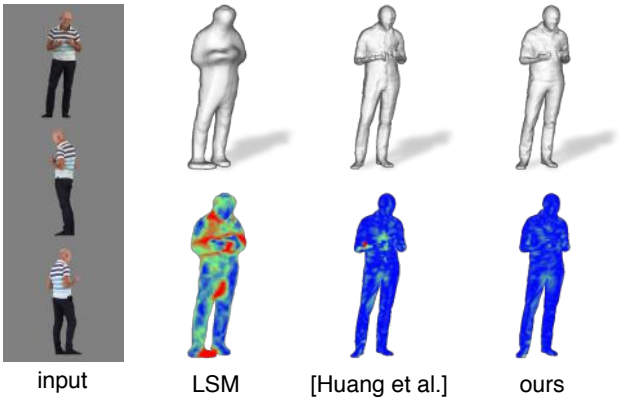

Figure 7: **Comparison with learning-based multi-view methods.** Ours outperforms other learning-based multi-view methods qualitatively and quantitatively. Note that all methods are trained with three view inputs from the same training data.

in the image. Unlike voxel-based representations, our method can produce high-resolution output since we are not limited by the high memory requirements of volumetric representations. Furthermore, we also demonstrate how this method can be naturally extended to infer the entire texture on a person given partial observations. Unlike existing methods, which synthesize the back regions based on frontal views in an image space, our approach can predict colors in unseen, concave and side regions directly on the surface. In particular, our method is the first approach that can inpaint textures for shapes of arbitrary topology. Since we are capable for generating textured 3D surfaces of a clothed person from a single RGB camera, we are moving a step closer toward monocular reconstructions of dynamic scenes from video without the need of a template model. Our ability to handle arbitrary additional views also makes our approach particularly suitable for practical and efficient 3D modeling

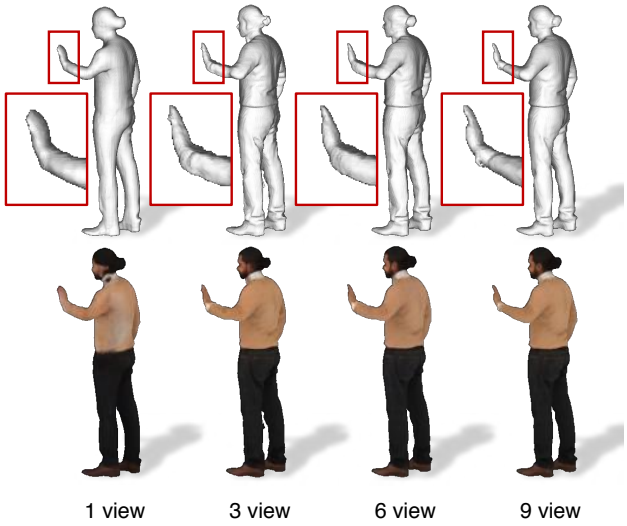

Figure 8: Our surface and texture predictions increasingly improve as more views are added.

settings using sparse views, where traditional multi-view stereo or structure-from-motion would fail.

Future Work. While our texture predictions are reasonable and not limited by the topology or parameterization of the inferred 3D surface, we believe that higher resolution appearances can be inferred, possibly using generative adversarial networks or increasing the input image resolution. In this work, the reconstruction takes place in pixel coordinate space, aligning the scale of subjects as pre-process. As in other single-view methods, inferring scale factor remains an open-question, which future work can address. Lastly, in all our examples, none of the segmented subjects are occluded by any other objects or scene elements. In real-world settings, occlusions often occur and perhaps only a part of the body is framed in the camera. Being able to digitize and predict complete objects in partially visible settings could be highly valuable for analyzing humans in unconstrained settings.

Acknowledgements Hao Li is affiliated with the University of Southern California, the USC Institute for Creative Technologies, and Pinscreen. This research was conducted at USC and was funded by in part by the ONR YIP grant N00014-17-S-FO14, the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation program sponsored by DARPA, the Andrew and Erna Viterbi Early Career Chair, the U.S. Army Research Laboratory under contract number W911NF-14-D-0005, Adobe, and Sony. This project was not funded by Pinscreen, nor has it been conducted at Pinscreen or by anyone else affiliated with Pinscreen. Shigeo Morishima is supported by the JST ACCEL Grant Number JPMJAC1602, JSPS KAKENHI Grant Number JP17H06101, JP19H01129. Angjoo Kanazawa is supported by BAIR sponsors. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019.

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

[3] Thiemo Alldieck, Marcus A Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005.

[5] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[6] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta, and Deva Ramanan. Pixelnet: Representation of the pixels, by the pixels, and for the pixels. *arXiv:1702.06506*, 2017.

[7] Gavin Barill, Neil Dickson, Ryan Schmidt, David I.W. Levin, and Alec Jacobson. Fast winding numbers for soups and clouds. *ACM Transactions on Graphics*, 37(4):43, 2018.

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578, 2016.

[9] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818, 2018.

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.

[12] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644, 2016.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[14] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[15] Yasutaka Furukawa and Jean Ponce. Carved visual hulls for image-based modeling. In *European Conference on Computer Vision*, pages 564–577, 2006.

[16] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[17] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753, 2009.

[18] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *European Conference on Computer Vision*, pages 566–581, 2018.

[19] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[20] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision*, pages 1381–1388, 2009.

[21] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[22] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In *arXiv preprint arXiv:1704.00710*. 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[24] Haibin Huang, Evangelos Kalogerakis, Siddhartha Chaudhuri, Duygu Ceylan, Vladimir G Kim, and Ersin Yumer. Learning local shape descriptors from part correspondences with multiview convolutional networks. *ACM Transactions on Graphics*, 37(1):6, 2018.

[25] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision*, pages 421–430, 2017.

[26] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *European Conference on Computer Vision*, pages 336–354, 2018.

[27] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3D Human Body Reconstruction from a Single Image via Volumetric Regression. In *ECCV Workshop Proceedings*, PeopleCap 2018, pages 0–0, 2018.

[28] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2307–2315, 2017.

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[30] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[31] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *European Conference on Computer Vision*, pages 371–386, 2018.

[32] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems*, pages 364–375, 2017.

[33] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2017.

[34] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248, 2015.

[36] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM siggraph computer graphics*, volume 21, pages 163–169. ACM, 1987.

[37] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. Image-based visual hulls. In *ACM SIGGRAPH*, pages 369–374, 2000.

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *arXiv preprint arXiv:1812.03828*, 2018.

[39] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019.

[40] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *European Conference on Computer Vision*, pages 123–138, 2018.

[41] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

[42] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3D Vision*, pages 484–494, 2018.

[43] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3500–3509, 2017.

[44] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *arXiv preprint arXiv:1901.05103*, 2019.

[45] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[46] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[47] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4):73, 2017.

[48] Renderpeople, 2018. https://renderpeople.com/3d-people.

[49] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

[50] Stan Sclaroff and Alex Pentland. *Generalized implicit functions for computer graphics*, volume 25. ACM, 1991.

[51] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] Peter-Pike Sloan, Jan Kautz, and John Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *ACM Transactions on Graphics*, volume 21, pages 527–536, 2002.

[54] Cristian Sminchisescu and Alexandru Telea. Human pose estimation from silhouettes. a consistent approach using distance level sets. In *International Conference on Computer Graphics, Visualization and Computer Vision*, volume 10, 2002.

[55] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007.

[56] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *European Conference on Computer Vision*, pages 155–171, 2018.

[57] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.

[58] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2626–2634, 2017.

[59] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision*, pages 20–36, 2018.

[60] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[61] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):97, 2008.

[62] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics*, 28(5):174, 2009.

[63] Ingo Wald, Sven Woop, Carsten Benthin, Gregory S Johnson, and Manfred Ernst. Embree: a kernel framework for efficient cpu ray tracing. *ACM Transactions on Graphics*, 33(4):143, 2014.

[64] Weiyue Wang, Xu Qiangeng, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *arXiv preprint arXiv:1905.10711*, 2019.

[65] Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. Scalable 3D video of dynamic scenes. *The Visual Computer*, 21(8):629–638, 2005.

[66] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[67] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. *arXiv preprint arXiv:1812.02246*, 2018.

[68] Chenglei Wu, Kiran Varanasi, and Christian Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. *European Conference on Computer Vision*, pages 757–770, 2012.

[69] Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision*, pages 3–19, 2018.

[70] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, pages 439–454, 2016.

[71] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017.

[72] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. In *ACM Transactions on Graphics*, page 126, 2010.

[73] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, pages 286–301, 2016.

[74] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[75] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, 23(3):600–608, 2004.

Appendix I. Implementation Details

Experimental Setup. Since there is no large scale datasets for high-resolution clothed humans, we collected photogrammetry data of 491 high-quality textured human meshes with a wide range of clothing, shapes, and poses, each consisting of about $100,000$ triangles from RenderPeople[3]. We refer to this database as High-Fidelity Clothed Human Data set. We randomly split the dataset into a training set of 442 subjects and a test set of 49 subjects. To efficiently render the digital humans, Lambertian diffuse shading with surface normal and spherical harmonics are typically used due to its simplicity and efficiency [60, 39]. However, we found that to achieve high-fidelity reconstructions on real images, the synthetic renderings need to correctly simulate light transport effects resulting from both global and local geometric properties such as ambient occlusion. To this end, we use a precomputed radiance transfer technique (PRT) that precomputes visibility on the surface using spherical harmonics and efficiently represents global light transport effects by multiplying spherical harmonics coefficients of illumination and visibility [53]. PRT only needs to be computed once per object and can be reused with arbitrary illuminations and camera angles. Together with PRT, we use 163 second-order spherical harmonics of indoor scene from HDRI Haven[4] using random rotations around y axis. We render the images by aligning subjects to the image center using a weak-perspective camera model and image resolution of 512×512. We also rotate the subjects for 360 degrees in yaw axis, resulting in $360 \times 442 = 159,120$ images for training. For the evaluation, we render 49 subjects from RenderPeople and 5 subjects from the BUFF data set [71] using 4 views spanning every 90 degrees in yaw axis. Note that we render the images without the background. We also test our approach on real images of humans from the DeepFashion data set [34]. In the case of real data, we use a off-the-shelf semantic segmentation network together with Grab-Cut refinement [49].

Network Architecture. Since the framework of PIFu is not limited to a specific network architecture, one can technically use any fully convolutional neural network as the image encoder. For surface reconstruction, we adapt the stacked hourglass network [41] with modifications proposed by [27]. We also replace batch normalization with group normalization [69], which improves the training stability when the batch sizes are small. Similar to [27], the intermediate features of each stack are fed into PIFu, and the losses from all the stacks are aggregated for parameter update. We have conducted ablation study on the network architecture design and compare against other alternatives (VGG16, ResNet34) in Appendix II. The image encoder for texture inference adopts the architecture of CycleGAN [74] consisting of 6 residual blocks [29]. Instead of using transpose convolutions to upsample the latent features, we directly feed the output of the residual blocks to the following Tex-PIFu.

[3] https://renderpeople.com/3d-people/

[4] https://hdrihaven.com/

PIFu for surface reconstruction is based on a multi-layer perceptron, where the number of neurons is $(257, 1024, 512, 256, 128, 1)$ with non-linear activations using leaky ReLU except the last layer that uses sigmoid activation. To effectively propagate the depth information, each layer of MLP has skip connections from the image feature $F(x) \in \mathbb{R}^{256}$ and depth z in spirit of [11]. For multi-view PIFu, we simply take the 4-th layer output as feature embedding and apply average pooling to aggregate the embedding from different views. Tex-PIFu takes $F_C(x) \in \mathbb{R}^{256}$ together with the image feature for surface reconstruction $F_V(x) \in \mathbb{R}^{256}$ by setting the number of the first neurons in the MLP to 513 instead of 257. We also replace the last layer of PIFu with 3 neurons, followed by tanh activation to represent RGB values.

Training procedure. Since the texture inference module requires pretrained image features from the surface reconstruction module, we first train PIFu for the surface reconstruction and then for texture inference, using the learned image features F_V as condition. We use RMSProp for the surface reconstruction following [41] and Adam for the texture inference with learning rate of 1×10^{-3} as in [74], the batch size of 3 and 5, the number of epochs of 12 and 6, and the number of sampled points of 5000 and 10000 per object in every training batch respectively. The learning rate of RMSProp is decayed by the factor of 0.1 at 10-th epoch following [41]. The multi-view PIFu is fine-tuned from the models trained for single-view surface reconstruction and texture inference with a learning rate of 1×10^{-4} and 2 epochs. The training of PIFu for single-view surface reconstruction and texture inference takes 4 and 2 days, respectively, and fine-tuning for multi-view PIFu can be achieved within 1 day on a single 1080ti GPU.

Appendix II. Additional Evaluations

Spatial Sampling. In Table 4 and Figure 9, we provide the effects of sampling methods for surface reconstruction. The most straightforward way is to uniformly sample inside the bounding box of the target object. Although it helps to remove artifacts caused by overfitting, the decision boundary becomes less sharp, losing all the local details (See Figure 9, first column). To obtain a sharper decision boundary, we propose to sample points around the surface with distances following a standard deviation σ from the actual surface mesh. We use $\sigma = 3, 5$, and 15 cm. The smaller σ becomes, the sharper the decision boundary is the result becomes more prone to artifacts outside the decision boundary (second column). We found that combining adaptive sampling with $\sigma = 5$ cm and uniform sampling achieves qualitatively and quantitatively the best results (right-most column). Note that each sampling scheme is trained with the identical setup as our training procedure described in Appendix I.

Network Architecture. In this section, we show comparisons of different architectures for the surface reconstruction and provide insight on design choices of the image encoders. One option is to use bottleneck features of fully convolutional networks [29, 66, 41]. Due to its state-of-the-art performance in volumetric regression for human faces and bodies, we choose Stacked Hourglass network [41] with a modification proposed by [27], denoted as HG. Another option is to aggregate features from multiple layers to obtain multi-scale feature embedding [6, 26]. Here we use two widely used network architectures: VGG16 [52] and ResNet34 [23] for the comparison. We extract the features from the layers of 'relu1_2', 'relu2_2', 'relu3_3', 'relu4_3', and 'relu5_3' for VGG network using bilinear sampling based on x, resulting in 1472 dimensional features. Similarly, we extract the features before every pooling layers in ResNet, resulting in 1024-D features. We modify the first channel size in PIFu to incorporate the feature dimensions and train the surface reconstruction model using the Adam optimizer with a learning rate of 1×10^{-3}, the number of sampling of $10,000$ and batch size of 8 and 4 for VGG and ResNet respectively. Note

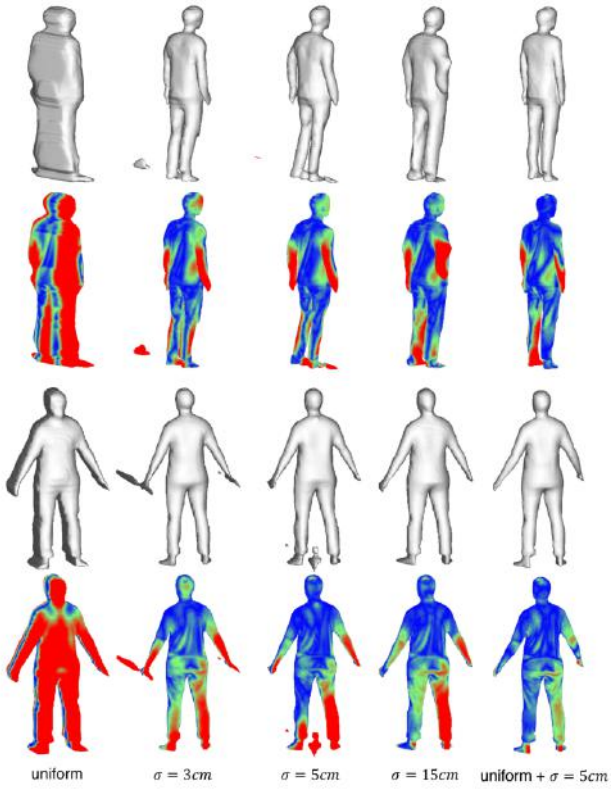

Figure 9: Reconstructed geometry and point to surface error visualization using different sampling methods.

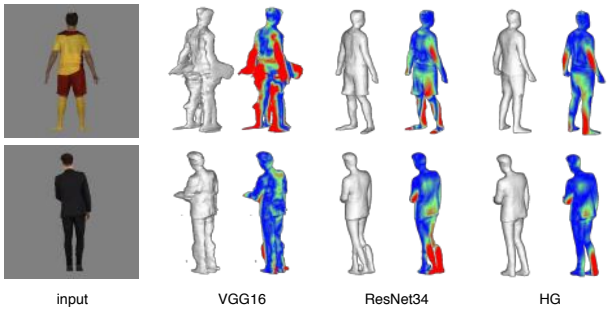

Figure 10: Reconstructed geometry and point to surface error visualization using different architectures for the image encoder.

Methods	RenderPeople Normal	P2S	Chamfer	Buff Normal	P2S	Chamfer
Uniform	0.119	5.07	4.23	0.132	5.98	4.53
$\sigma = 3$cm	0.104	2.03	1.62	0.114	6.15	3.81
$\sigma = 5$cm	0.105	1.73	1.55	0.115	1.54	1.41
$\sigma = 15$cm	0.100	**1.49**	**1.43**	0.105	1.37	1.26
$\sigma = 5$cm + Uniform	**0.084**	1.52	1.50	**0.092**	**1.15**	**1.14**

Table 3: Ablation study on the sampling strategy.

Methods	RenderPeople Normal	P2S	Chamfer	Buff Normal	P2S	Chamfer
VGG16	0.125	3.02	2.25	0.144	4.65	3.08
ResNet34	0.097	**1.49**	**1.43**	0.099	1.68	1.50
HG	**0.084**	1.52	1.50	**0.092**	**1.15**	**1.14**

Table 4: Ablation study on network architectures.

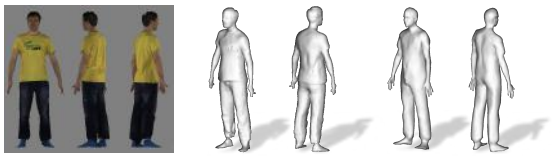

Figure 11: Comparison with a template-based method [3]. Note that while Alldieck et al. uses a dense video sequence without camera calibration, ours uses the calibrated three views as input.

Methods	Buff Normal	P2S	Chamfer
Alldieck et al. 18 (Video)	0.127	0.820	0.795
Ours (3 views)	**0.107**	**0.665**	**0.641**

Table 5: Quantitative comparison between a template-based method [3] using a dense video sequence and ours using 3 views.

that VGG and ResNet are initialized with models pretrained with ImageNet [13]. The other hyper-paremeters are the same as the ones used for our sequential network based on Stacked Hourglass.

In Table 3 and Figure 10, we show comparisons of three architectures using our evaluation data. While ResNet has slightly better performance in the same domain as the training data (i.e., test set in RenderPeople dataset), we observe that the network suffers from overfitting, failing to generalize to other domains (i.e., BUFF and DeepFashion dataset). Thus, we adopt a sequential architecture based on Stacked Hourglass network as our final model.

Appendix III. Additional Results

Please see the supplementary video for more results.

Comparison with Template-based Method. In Figure 11 and Table 5, we compare our approach with a template based method [3] that takes a dense 360 degrees view video as an input on BUFF dataset. From 3 views we outperform the template based method. Note that Alldieck et al. requires an uncalibrated dense video sequence, while ours requires calibrated sparse view inputs.

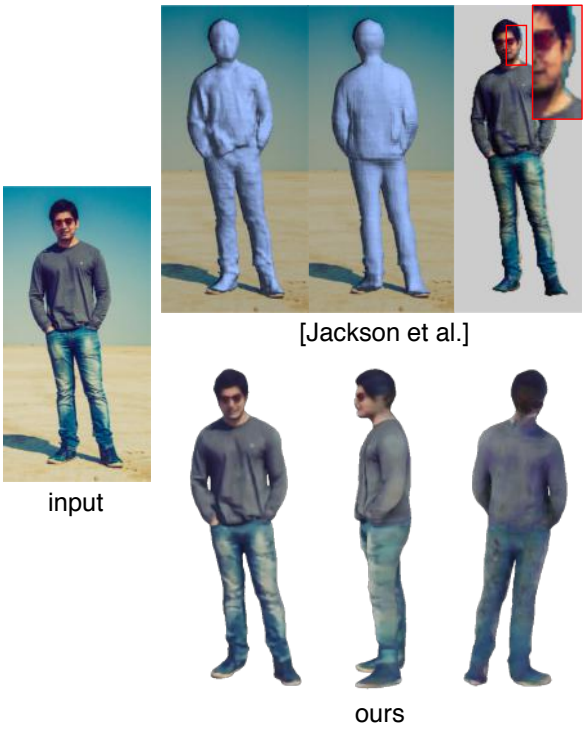

Figure 12: Comparison with Voxel Regression Network [27]. While [27] suffers from texture projection error due to the limited precision of voxel representation, our PIFu representation efficiently not only represents surface geometry in a pixel-aligned manner but also complete texture on the missing region. Note that [27] can only texture the visible portion of the person by projecting the foreground to the recovered surface. In comparison, we recover the texture of the entire surface, including the unseen regions.

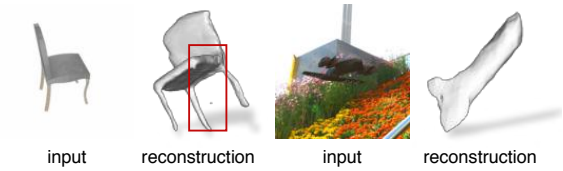

Figure 13: PIFu trained on general objects reveals new challenges to be addressed in future.

Comparison with Voxel Regression Network. We provide an additional comparison with Voxel Regression Network (VRN) [27] to clarify the advantages of PIFu. Figure 12 demonstrates that the proposed PIFu representation can align the 3D reconstruction with pixels at higher resolution, while VRN suffers from misalignment due to the limited precision of its voxel representation. Additionally, the generality of PIFu offers texturing of shapes with arbitrary topology and self-occlusion, which has not been addressed by the work of VRN. Note that VRN only is able to project the image texture onto the recovered surface, and does not provide an approach to do texture inpainting on the unseen side.

Results on General Objects. In this work, we focused largely on clothed human surfaces. A natural question is how it extends to general object shapes. Our preliminary experiments on the ShapeNet dataset [9] in a class agnostic setting reveals new challenges as shown in Figure 13. We speculate that the greater variety of object shapes makes it difficult to learn a globally coherent shape from only pixel-level features. Note that recently [64] extend the idea of PIFu by explicitly combining global features and local features, demonstrating globally coherent and locally detailed reconstruction for general objects is possible.

Results on Video Sequences. We also apply our approach to video sequences obtained from [62]. For the reconstruction, video frames are center cropped and scaled so that the size of the subjects are roughly aligned with our training data. Note that the cropping and scale is fixed for each sequence. Figure 14 demonstrates that our reconstructed results are reasonably temporally coherent even though the frames are processed independently.

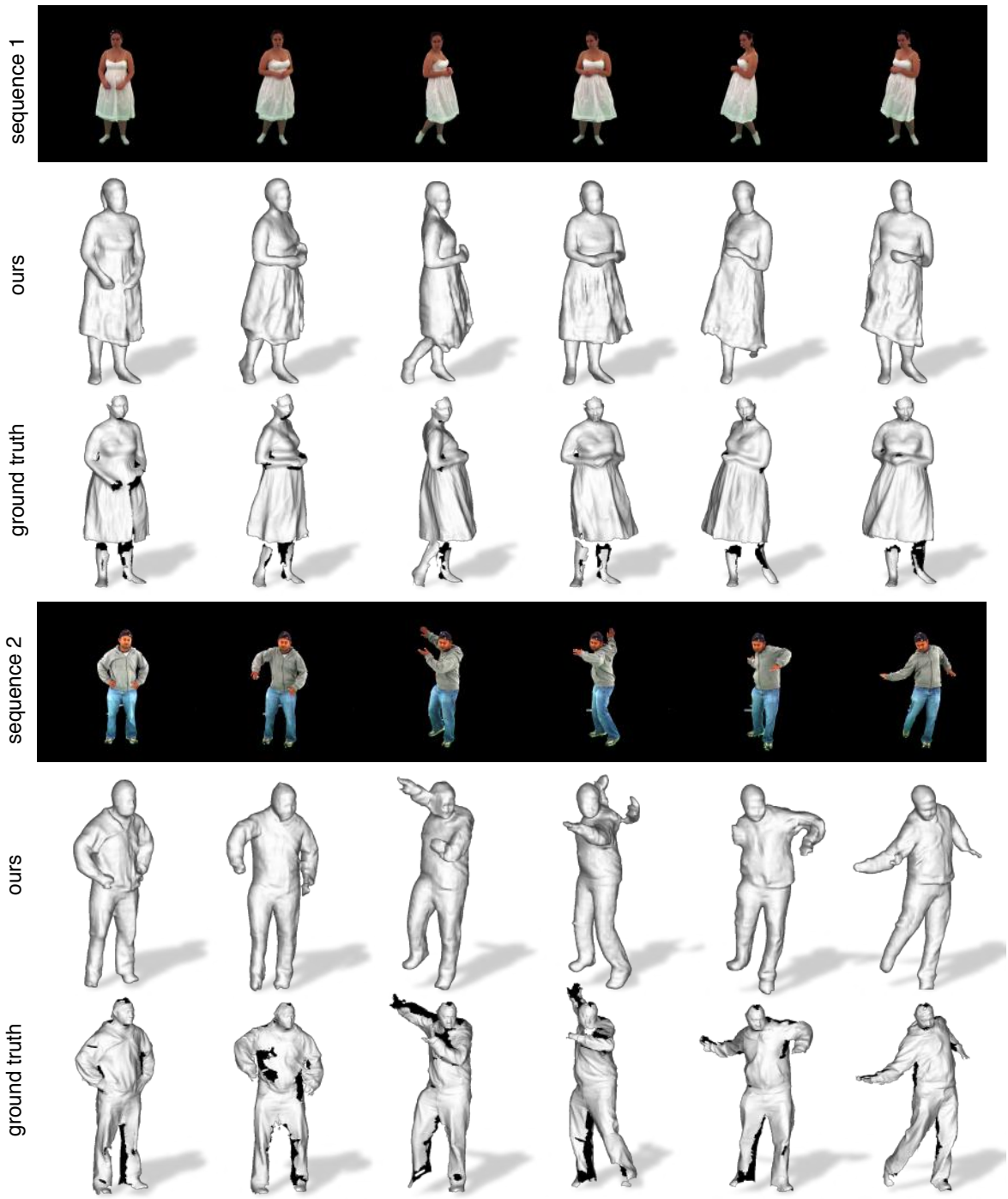

Figure 14: Results on video sequences obtained from [62]. While ours uses a single view input, the ground truth is obtained from 8 views with controlled lighting conditions.