SiCloPe: Silhouette-Based Clothed People

Ryota Natsume[1,3*] Shunsuke Saito[1,2*] Zeng Huang[1,2] Weikai Chen[1]
Chongyang Ma[4] Hao Li[1,2,5] Shigeo Morishima[3]

[1]USC Institute for Creative Technologies [2]University of Southern California
[3]Waseda University [4]Snap Inc. [5]Pinscreen

Abstract

We introduce a new silhouette-based representation for modeling clothed human bodies using deep generative models. Our method can reconstruct a complete and textured 3D model of a person wearing clothes from a single input picture. Inspired by the visual hull algorithm, our implicit representation uses 2D silhouettes and 3D joints of a body pose to describe the immense shape complexity and variations of clothed people. Given a segmented 2D silhouette of a person and its inferred 3D joints from the input picture, we first synthesize consistent silhouettes from novel view points around the subject. The synthesized silhouettes, which are the most consistent with the input segmentation are fed into a deep visual hull algorithm for robust 3D shape prediction. We then infer the texture of the subject's back view using the frontal image and segmentation mask as input to a conditional generative adversarial network. Our experiments demonstrate that our silhouette-based model is an effective representation and the appearance of the back view can be predicted reliably using an image-to-image translation network. While classic methods based on parametric models often fail for single-view images of subjects with challenging clothing, our approach can still produce successful results, which are comparable to those obtained from multi-view input.

1. Introduction

The ability to digitize and predict a complete and fully textured 3D model of a clothed subject from a single view can open the door to endless applications, ranging from virtual and augmented reality, gaming, virtual try-on, to 3D printing. A system that could generate a full-body 3D avatar of a person by simply taking a picture as input would significantly impact the scalability of producing virtual humans for immersive content creation, as well as its attainability by the general population. Such single-view inference is extremely difficult due to the vast range of possible shapes and appearances that clothed human bodies can take in natural conditions. Furthermore, only a 2D projection of the real world is available and the entire back view of the subject is missing.

Figure 1: Given a single image of a person from the frontal view, we can automatically reconstruct a complete and textured 3D clothed body shape.

While 3D range sensing [24, 32] and photogrammetry [37] are popular ways of obtaining complete 3D models, they are restricted to a tedious scanning process or require specialized equipment. The modeling of humans from a single view, on the other hand, has been facilitated by the availability of large 3D human model repositories [3, 26], where a *parametric model* of human shapes is used to guide the reconstruction process [6]. However, these parametric models only represent naked bodies and do not describe the clothing geometry nor the texture. Another option is to use a *pre-captured template* of the subject in order to handle new poses [50], but such approach is limited to the recording of one particular person.

In this work, we propose a deep learning based non-parametric approach for generating the geometry and texture of clothed 3D human bodies from a single frontal-view image. Our method can predict fine-level geometric details of clothing and generalizes well to new subjects different from those being used during training (See Figure 1).

While directly estimating 3D volumetric geometry from a single view is notoriously challenging and likely to require a large amount of training data as well as extensive parameter tuning, two cutting-edge deep learning techniques have

*Joint first authors

shown how impressive results can be obtained using 2D silhouettes from very sparse views [20, 42]. Inspired by these approaches based on visual hull, we propose to first algorithm to predict 2D silhouettes of the subject from multiple views given an input segmentation, which implicitly encodes 3D body shapes. We also show that a sparse 3D pose estimated from the 2D input [6, 36] can help reduce the dimensionality of the shape deformation and guide the synthesis of consistent silhouettes from novel views.

We then reconstruct the final 3D geometry from multiple silhouettes using a deep learning-based visual hull technique by incorporating a clothed human shape prior. Since silhouettes from arbitrary views can be generated, we further improve the reconstruction result by greedily choosing view points that will lead to improved silhouette consistency. To fully texture the reconstructed geometry, we propose to train an image-to-image translation framework to infer the color texture of the back view given the input image from the frontal view.

We demonstrate the effectiveness of our method on a variety of input data, including both synthetic and real ones. We also evaluate major design decisions using ablation studies and compare our approach with state of the art single-view as well as multi-view reconstruction techniques.

In summary, our contributions include:

- The first non-parametric solution for reconstructing fully textured and clothed 3D humans from a single-view input image.
- An effective two-stage 3D shape reconstruction pipeline that consists of predicting multi-view 2D silhouettes from a single input segmentation and a novel deep visual hull based mesh reconstruction technique with view sampling optimization.
- An image-to-image translation framework to reconstruct the texture of a full body from a single photo.

2. Related Work

Multi-view reconstruction. Due to the geometric complexity introduced by garment deformation and self occlusions, reconstructing clothed human bodies usually requires images captured from multiple viewpoints. Early attempts in this direction have extensively explored visual hull based approaches [29, 43, 15, 13, 9, 14] due to its efficiency and robustness to approximate the underlying 3D geometry. However, a visual hull based representation cannot handle concave regions nor generate good approximations of fine-scale details especially when the number of input views is limited. To address this issue, detailed geometry are often captured using techniques based on multi-view stereo constraints [39, 56, 46, 37, 44, 16, 49]. A number of techniques [51, 34, 53] exploit motion cues as additional priors for a more accurate digitization of body shapes.

Some more recent research have focused on monocular input capture, with the goal of making human modeling more accessible to end users [50, 2, 1]. With the recent advancement of deep learning, an active research direction is to encode shape prior in a deep neural network in order to model the complexity of human body and garment deformations. To this end, Huang *et al*. [20] and Gilbert *et al*. [17] have presented techniques that can synthesize clothed humans in a volumetric form from highly sparse views. Although the number of input views are reduced, both methods still require a carefully calibrated capture system. In this work, we push the envelop by reducing the input to a single unconstrained input photograph.

Single-view reconstruction. To reduce the immense solution space of human body shapes, several 3D body model repositories, e.g. SCAPE [3] and SMPL [26], have been introduced, which have made the single-view reconstruction of human bodies more tractable. In particular, a 3D parametric model is built from such database, which uses pose and shape parameters of the 3D body to best match an input image [5, 18, 6, 23]. As the mapping between the body geometry and the parameters of the deformable model is highly non-linear, alternative approaches based on deep learning have become increasingly popular. The seminal work of Dibra *et al*. [10, 11] introduces deep neural networks to estimate the shape parameters from a single input silhouette. More recent works predict body parameters of the popular SMPL model [6] by either minimizing the silhouette matching error [40], joint error based on the silhouette and 2D joints [41], or an adversarial loss that can distinguish unrealistic reconstruction output [22]. Concurrent to our work, Weng *et al*. [48] present a method to animate a person in 3D from a single image based on the SMPL model and 2D warping.

Although deformable models offer a low-dimensional embedding of complex non-rigid human body shapes, they are not suitable for modeling of fine-scale clothing details. To address this issue, additional information such as 2D [47, 8] and 3D body pose [30, 52, 19] has been incorporated to help recover clothed body geometry without relying on a template mesh. BodyNet [42] for instance, estimates volumetric body shapes from a single image based on an intermediate inference of 2D pose, 2D part segmentation, as well as 3D pose. The latest advances in novel view synthesis of human pose [27, 4] and 3D shape [55, 54, 35] have demonstrated the ability of obtaining multi-view inference from a single image. In this work, we introduce an approach that combines 3D poses estimation with the inference of silhouettes from novel views for predicting high-fidelity clothed 3D human shapes from a single photograph. We show that our method can achieve reasonably accurate reconstructions automatically without any template model.

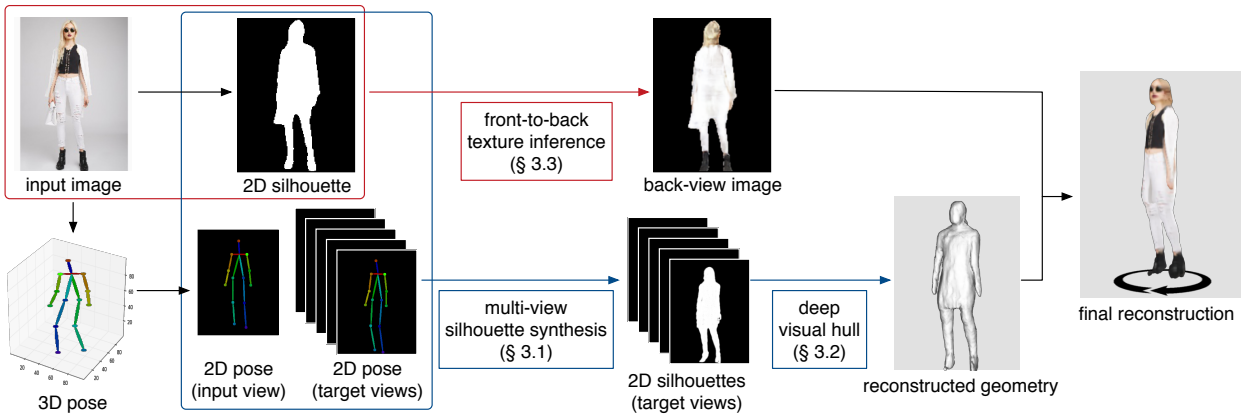

Figure 2: Overview of our framework.

3. Method

Our goal is to reconstruct a wide range of 3D clothed human body shapes with a complete texture from a single image of a person in frontal view. Figure 2 illustrates an overview of our system. Given an input image, we first extract the 2D silhouette and 3D joint locations, which are fed into a silhouette synthesis network to generate plausible 2D silhouettes from novel viewpoints (Sec. 3.1). The network produces multiple silhouettes with known camera projections, which are used as input for 3D reconstruction via visual hull algorithms [43]. However, due to possible inconsistency between the synthesized silhouettes, the subtraction operation of visual hull tends to excessively erode the reconstructed mesh. To further improve the output quality, we adopt a deep visual hull algorithm similar to Huang *et al*. [20] with a greedy view sampling strategy so that the reconstruction results account for domain-specific clothed human body priors (Sec. 3.2). Finally, we inpaint the non-visible body texture on the reconstructed mesh by inferring the back view of the input image using an image-to-image translation network (Sec. 3.3).

3.1. Multi-View Silhouette Synthesis

We seek an effective human shape representation that can handle the shape complexity due to different clothing types and deformations. Inspired by visual hull algorithms [29] and recent advances in conditional image generation [12, 28], we propose to train a generative network for synthesizing 2D silhouettes from viewpoints other than the input image (see Figure 3). We use these silhouettes as an intermediate implicit representation for the 3D shape inference.

Specifically, given the subject's 3D pose, estimated from the input image as a set of 3D joint locations, we project the 3D pose onto the input image and a target image plane to get the 2D pose $\mathbf{P}_s$ in the source view and the pose $\mathbf{P}_t$ in the target view, respectively. Our silhouette synthesis network

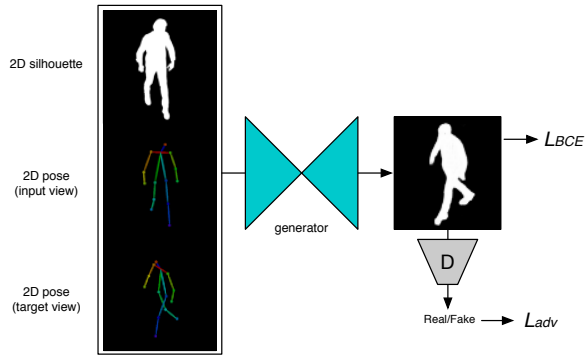

Figure 3: Illustration of our silhouette synthesis network.

$\mathcal{G}_s$ takes the input silhouette $\mathbf{S}_s$ together with $\mathbf{P}_s$ and $\mathbf{P}_t$ as input, and predicts the 2D silhouette in the target view $\mathbf{P}_t$:

$$\mathbf{S}_t = \mathcal{G}_s\big(\mathbf{S}_s, \mathbf{P}_s, \mathbf{P}_t\big). \tag{1}$$

Our loss function for training the network $\mathcal{G}_s$ consists of reconstruction errors of the inferred silhouettes using a binary cross entropy loss $\mathcal{L}_{BCE}$ and a patch-based adversarial loss $\mathcal{L}_{adv}$ [21]. The total objective function is given by

$$\mathcal{L} = \lambda_{BCE} \cdot \mathcal{L}_{BCE} + \mathcal{L}_{adv}, \tag{2}$$

where the relative weight λ_{BCE} is set to 750. In particular, the adversarial loss turns out to be critical for synthesizing sharp and detailed silhouettes. Figure 4 shows that the loss function with the adversarial term generate much sharper silhouettes, while without an adversarial loss would lead to blurry synthesis output.

Discussions. The advantages of using silhouettes to guide the 3D reconstruction are two-fold. First, since silhouettes are binary masks, the synthesis can be formulated as a

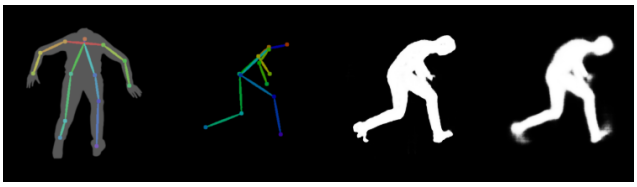

Figure 4: GAN helps generate clean silhouettes in presence of ambiguity in silhouette synthesis from a single view.

pixel-wise classification problem, which can be trained more robustly without the need of complex loss functions or extensive hyper parameter tuning in contrast to novel view image synthesis [27, 4]. Second, the network can predict much a higher spatial resolution since it does not store 3D voxel information explicitly, as with volumetric representations [42], which are bounded by the limited output resolution.

3.2. Deep Visual Hull Prediction

Although our silhouette synthesis algorithm generates sharp prediction of novel-view silhouettes, the estimated results may not be perfectly consistent as the conditioned 3D joints may fail to fully disambiguate the details in the corresponding silhouettes (e.g., fingers, wrinkles of garments). Therefore, naively applying conventional visual hull algorithms is prone to excessive erosion in the reconstruction, since the visual hull is designed to subtract the inconsistent silhouettes in each view. To address this issue, we propose a deep visual hull network that reconstructs a plausible 3D shape of clothed body without requiring perfectly view-consistent silhouettes by leveraging the shape prior of clothed human bodies.

In particular, we use a network structure based on [20]. At a high level, Huang *et al*. [20] propose to map 2D images to a 3D volumetric field through a multi-view convolutional neural network. The 3D field encodes the probabilistic distribution of 3D points on the captured surface. By querying the resulting field, one can instantiate the geometry of clothed human body at an arbitrary resolution. However, unlike their approach which takes carefully calibrated color images from fixed views as input, our network only consumes the probability maps of novel-view silhouettes, which can be inconsistent across different views. Although arbitrary number of novel-view silhouettes can be generated, it remains challenging to properly select optimal input views to maximize the network performance. Therefore, we introduce several improvements to increase the reconstruction accuracy.

Greedy view sampling. We propose a greedy view sampling strategy to choose proper views that can lead to better reconstruction quality. Our key idea is to generate a pool of candidate silhouettes and then select the views that are most consistent in a greedy manner. In particular, the candidate silhouettes are rendered from 12 view bins $\{\mathcal{B}_i\}$: the main orientations of the bins are obtained by uniformly sampling 12 angles in the yaw axis. The first bin only contains the input view and thus has to be aligned with the orientation of the input viewpoint. Each of the other bins consists of 5 candidate viewpoints, which are distributed along the pitch axis with angles sampled from $\{0°, 15°, 30°, 45°, 60°\}$. In the end, we obtain 55 candidate viewpoints $\{\mathcal{V}_i\}$ to cover most parts of the 3D body.

To select the views with maximal consistency, we first compute an initial bounding volume of the target model based on the input 3D joints. We then carve the bounding volume using the silhouette of the input image and obtain a coarse visual hull $\mathcal{H}_1$. The bins with remaining views are iterated in a clockwise order, i.e., only one candidate view will be sampled from each bin at the end of the sampling process. Starting from the second bin $\mathcal{B}_2$, the previously computed visual hull $\mathcal{H}_1$ is projected to its enclosed views. The candidate silhouette that has the maximum 2D intersection over union (IoU) with $\mathcal{H}_1$'s projection will be selected as the next input silhouette for our deep visual hull algorithm. After the best silhouette $\hat{\mathcal{V}}_2$ is sampled from $\mathcal{B}_2$, $\mathcal{H}_1$ is further carved by $\hat{\mathcal{V}}_2$ and the updated visual hull $\mathcal{H}_2$ is passed to the next iteration. We iterated until all the view bins have been sampled.

The selected input silhouettes generated by our greedy view sampling algorithm are then fed into a deep visual hull network. The choice of our network design is similar to that of [20]. The main difference lies in the format of inputs. Specifically, in addition to multi-view silhouettes, our network also takes the 2D projection of the 3D pose as additional channel concatenated with the corresponding silhouette. This change helps to regularize the body part generation by passing the semantic supervision to the network and thus improves robustness. Moreover, we also reduce some layers of the network of [20] to achieve a more compact model and to prevent overfitting. The detailed architecture is provided in our supplementary materials.

3.3. Front-to-Back Texture Synthesis

When capturing the subject from a single viewpoint, only one side of the texture is visible and therefore predicting the other side of the texture appearance is required to reconstruct a fully textured 3D body shape. Our key observation is that the frontal view and the back view of a person are spatially aligned by sharing the same contour and many visual features. This fact has inspired us to solve the problem of back-view texture prediction using an *image-to-image* translation framework based conditional generative adversarial network. Specifically, we train a generator $\mathcal{G}_t$ to

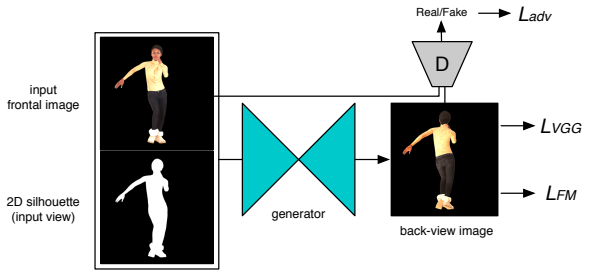

Figure 5: Illustration of our front-to-back synthesis network.

predict the back-view texture $\hat{\mathbf{I}}_b$ from the frontal-view input image $\mathbf{I}_f$ and the corresponding silhouette $\mathbf{S}_f$:

$$\hat{\mathbf{I}}_b = \mathcal{G}_t(\mathbf{I}_f, \mathbf{S}_f). \tag{3}$$

We train the generator $\mathcal{G}_t$ in a supervised manner by leveraging textured 3D human shape repositories to generate a dataset that suffices for our training objective (Sec. 3.4). Adopted from a high-resolution image-to-image translation network [45], our loss function consists of a feature matching loss $\mathcal{L}_{FM}$ that minimizes the discrepancy of intermediate layer activation of the discriminator D, a perceptual loss $\mathcal{L}_{VGG}$ using a VGG19 model pre-trained for image classification task [38], and an adversarial loss $\mathcal{L}_{adv}$ conditioned by the input frontal image (see Figure 5). The total objective is defined as:

$$\mathcal{L} = \lambda_{FM} \cdot \mathcal{L}_{FM} + \lambda_{VGG} \cdot \mathcal{L}_{VGG} + \mathcal{L}_{adv}, \tag{4}$$

where we set the relative weights as $\lambda_{FM} = \lambda_{VGG} = 10.0$ in our experiments.

The resulting back-view image is used to complete the per-vertex color texture of the reconstructed 3D mesh. If the dot product between the surface normal $\mathbf{n}$ in the input camera space and the camera ray $\mathbf{c}$ is negative (i.e., surface is facing towards the camera), the vertex color is sampled from the input view image at the corresponding screen coordinate. Likewise, if the dot product is positive (i.e., surface is facing in the opposite direction), the vertex color is sampled from the synthesized back-view image. When the surface is perpendicular to the camera ray (i.e., $|\mathbf{n} \cdot \mathbf{c}| \leq \epsilon = 1.0 \times 10^{-4}$), we blend the colors from the front and back views so that there are no visible seams across the boundary.

3.4. Implementation Details

Body mesh datasets. We have collected 73 rigged meshes with full textures from aXYZ[1] and 194 meshes from Renderpeople[2]. We randomly split the dataset into a training set and a test set of 247 and 20 meshes, respectively. We apply 48 animation sequences (such as walking, waving, and Samba dancing) from Mixamo[3] to each mesh from Renderpeople to collect body meshes of different poses. Similarly, the meshes from aXYZ have been animated into 11 different sequences. To render synthetic training data, we have also obtained 163 second-order spherical harmonics of indoor environment maps from HDRI Haven[4] and they are randomly rotated around the yaw axis.

Camera settings for synthetic data. We place the projective camera so that the pelvis joint is aligned with the image center and relative body size in the screen space remains unchanged. Since our silhouette synthesis network takes an unconstrained silhouette as input and generate a new silhouette in predefined view points, we separate the data generation for the source silhouettes and the target silhouettes. We render our data images at the resolution of 256×256. For the source silhouettes a yaw angle is randomly sampled from 360° and a pitch angle between -10° and 60°, whereas for the target silhouettes, a yaw angle is sampled from every 7.5° and a pitch angle from $10, 15, 30, 45, 60^\circ$. The camera has a randomly sampled 35mm film equivalent focal length ranged between 40 and 135mm for the source silhouettes and a fixed focal length of 800mm for the target silhouettes. For the front-to-back image synthesis, we set the yaw angle to be frontal and sample the pitch angle from $0, 7.5, 15^\circ$ with a focal length of 800mm. Given the camera projection, we project 13 joint locations that are compatible with MPII [31] onto each view point.

Front-to-back rendering. Figure 6 illustrates how we generate a pair of front and back view images. Given a camera ray, normal rendering of 3D mesh sorts the depth of triangles per pixel and display the rasterization results assigned from the closest triangle. To obtain the corresponding image from the other side, we instead takes that of the furthest triangle. Note that most common graphics libraries (e.g., OpenGL, DirectX) support this function, allowing us to generate training samples within a reasonable amount of time.

Network architectures. Both our silhouette synthesis network and the front-to-back synthesis network follow the U-Net network architecture in [21] with an input channel size of 7 and 4, respectively. All the weights in these networks are initialized based on Gaussian distribution. We use the Adam optimizer with learning rates of 2.0×10^{-4}, 1.0×10^{-4}, and 2.0×10^{-4}, batch size of 30, 1, and 1, the number of iterations of $250,000$, $160,000$, and $50,000$, and no weight decay for the silhouette synthesis, deep visual hull, and front-to-back synthesis, respectively. The deep visual hull is

[1] https://secure.axyz-design.com/
[2] https://renderpeople.com/3d-people/
[3] https://www.mixamo.com/
[4] https://hdrihaven.com/

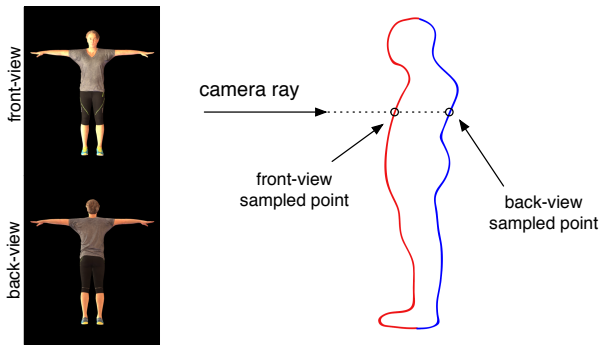

Figure 6: Illustration of our back-view rendering approach.

trained with the output of our silhouette synthesis network so that the distribution gap between the output of silhouette synthesis and the input for the deep visual hull algorithm is minimized.

Additional networks. Although 2D silhouette segmentation and 3D pose estimation are not our major contributions and in practice one can use any existing methods, we train two additional networks to automatically process the input image with consistent segmentation and joint configurations. For the silhouette segmentation, we adopt a stacked hourglass network [33] with three stacks. Given an input image of resolution $256 \times 256 \times 3$, the network predicts a probability map of resolution $64 \times 64 \times 1$ for silhouettes. We further apply a deconvolution layer with a kernel size of 4 to obtain sharper silhouettes, after concatenating $2\times$ upsampled probability and the latent features after the first convolution in the hourglass network. The network is trained with the mean-squared error between the predicted probability map and the ground truth of UP dataset [23]. For 3D pose estimation, we adopt a state-of-the-art 3D face alignment network [7] without modification. We train the pose estimation network using our synthetically rendered body images of resolution 256×256 together with the corresponding 3D joints. We use the RMSProp optimizer with a learning rate of 2.0×10^{-5}, a batch size of 8 and no weight decay for training both the silhouette segmentation and pose estimation networks.

4. Experimental Results

Figure 7 shows our reconstruction results of 3D clothed human bodies with full textures on different single-view input images from the DeepFashion dataset [25]. For each input, we show the back-view texture synthesis result, the reconstructed 3D geometry rendered with plain shading, as well as the final textured geometry. Our method can robustly handle a variety of realistic test photos of different poses, body shapes, and cloth styles, although we train the networks using synthetically rendered images only.

Table 1: Evaluation of our silhouette-based representation compared to direct voxel prediction. The errors are measured using Chamfer distances between the reconstructed meshes and the ground-truth.

Input	Output	IoU (2D)	Error
RGB+2D Pose	Silhouette	0.8260	1.66
Silhouette+2D Pose	Silhouette	**0.8857**	**1.36**
RGB+3D Pose	Voxel	0.4708	2.49
Silhouette+3D Pose	Voxel	0.4615	2.77

Table 2: Evaluation of our greedy sampling method to compute deep visual hull.

Input	Method	Error
Inferred silhouettes	visual hull (random views)	2.12
	visual hull (optimized views)	1.37
	deep v-hull (random views)	1.41
	deep v-hull (optimized views)	**1.34**
GT silhouettes	visual hull (8 views)	0.67
GT images	Huang et al. [20] (4 views)	0.98

4.1. Evaluations

Silhouette Representation. We verify the effectiveness of our silhouette-based representation by comparing it with several alternative approaches based on the Renderpeople dataset, including direct voxel prediction from 3D pose and using RGB image to replace 2D silhouette as input for deep visual hull algorithm. Please refer to our supplementary materials for implementation details of our baseline methods used for comparisons. For all the methods, we report (1) the 2D Intersection over Union (IoU) for the synthetically generated side view and (2) the 3D reconstruction error based on Chamfer distances between the reconstructed meshes and the ground-truth (in centimeter) in Table 1. It is evident that direct voxel prediction will lead to poor accuracy when matching the side view in 2D and aligning with the ground-truth geometry in 3D, as compared to our silhouette-based representation. Figure 8 shows qualitative comparisons demonstrating the advantages of our silhouette-based representation.

Visual hull reconstruction. In Table 2 and Figure 9, we compare our deep visual hull algorithm (Sec. 3.2) with a naive visual hull method. We also evaluate our greedy view sampling strategy by comparing it with random view selection. We use 12 inferred silhouettes as input for different methods and evaluate the reconstruction errors using Chamfer distances. For random view selection, we repeat the process 100 times and compute the average error. As additional references, we also provide the corresponding

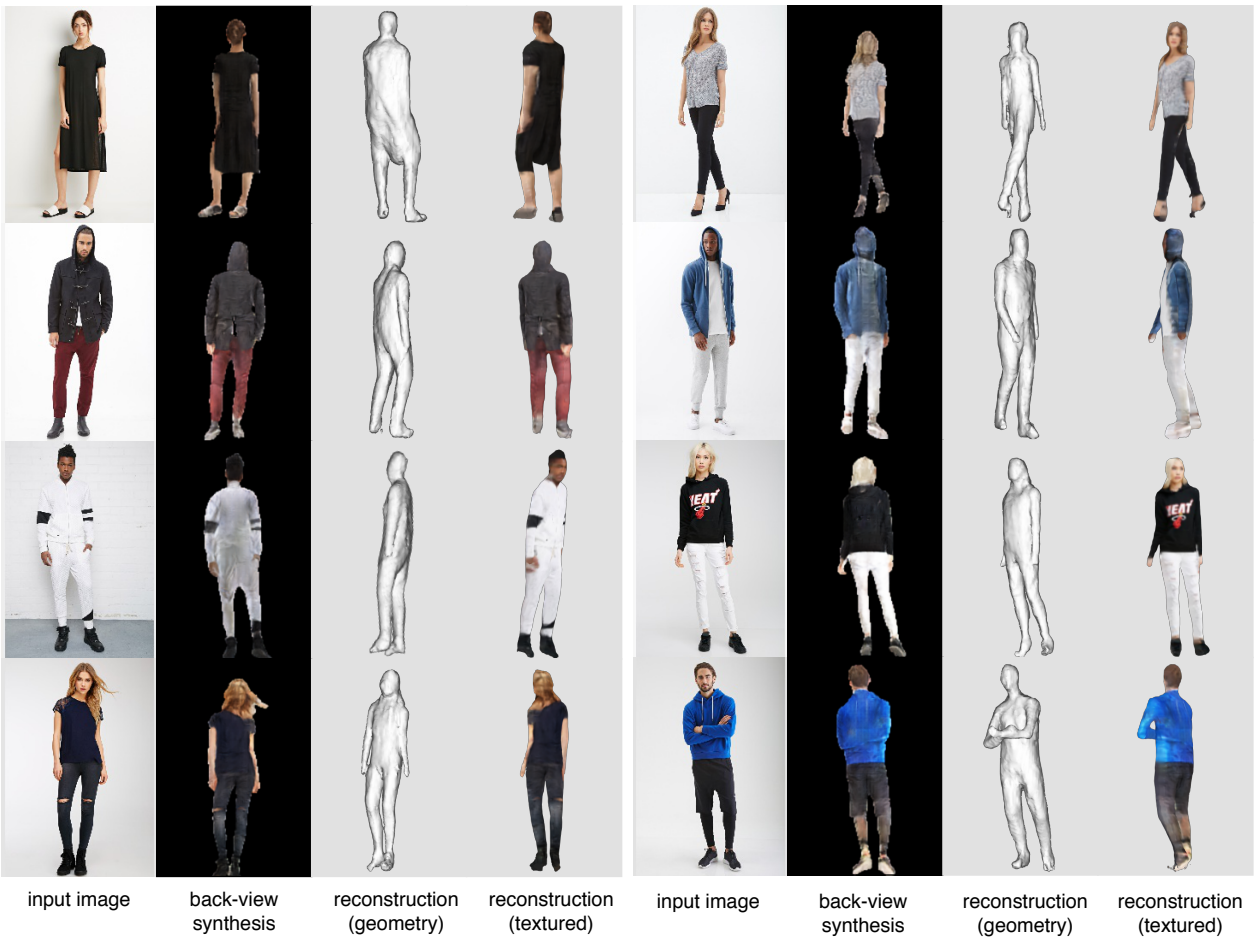

Figure 7: Our 3D reconstruction results of clothed human body using test images from the DeepFashion dataset [25].

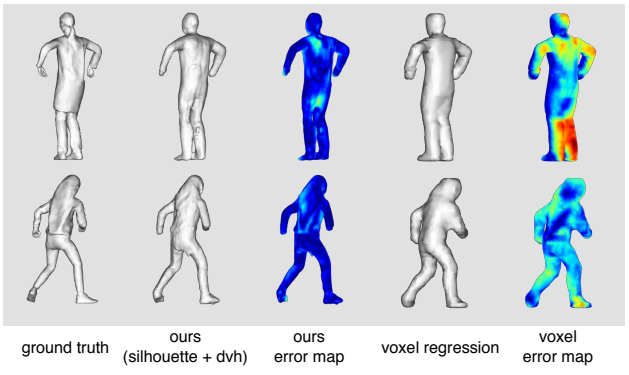

Figure 8: Qualitative evaluation of our silhouette-based shape representation as compared to direct voxel prediction.

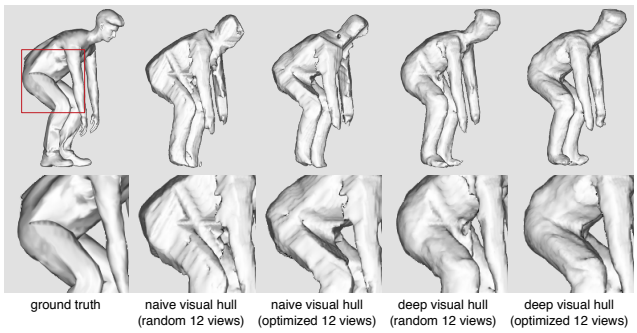

Figure 9: Comparisons between our deep visual hull method with a native visual hull algorithm, using both random view selection and our greedy view sampling strategy.

results using the naive visual hull method with 8 ground-truth silhouettes, as well as the method in [20] using 4 ground-truth images. As shown in Table 2, our deep visual algorithm outperforms a naive approach and our greedy view sampling strategy can significantly improve the results in terms of reconstruction errors. In addition, for the deep visual hull algorithm, our view sampling strategy is better than 69% random selected views, while for the naive visual hull method, our approach always outperforms random view selection. Figure 9 demonstrates that our deep visual hull

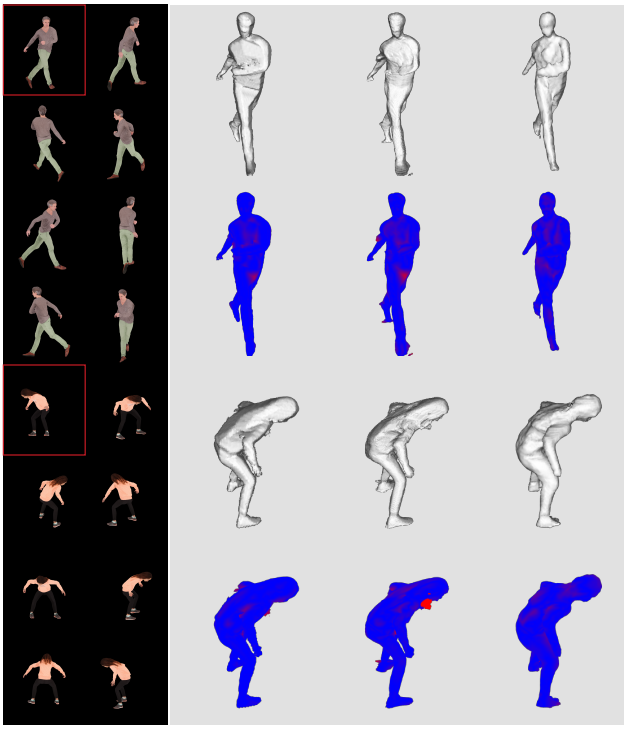

Figure 10: Comparison with multiview visual hull algorithms. Despite the single view input, our method produces comparable reconstruction results. Note that input image in red color is the single-view input for our method and the top four views are used for Huang *et al*. [20].

method helps fix some artifacts and missing parts especially in concave regions, which are caused by inconsistency among multi-view silhouettes synthesis results.

4.2. Comparisons

In Figure 10, we compare our method using single-view input with a native visual hull algorithm using 8 input views as well as Huang et al. [20] using four input views. For each result, we show both the plain shaded 3D geometry and the color-coded 3D reconstruction error. Although using a single image as input each time, we can still generate results that are visually comparable to those from methods based on multi-view input.

In Figure 11, we qualitatively compare our results with state-of-the-art single-view human reconstruction techniques [22, 42]. Since existing methods focus on body shape only using parametric models, our approach can generate more faithful results in cases of complex clothed geometry.

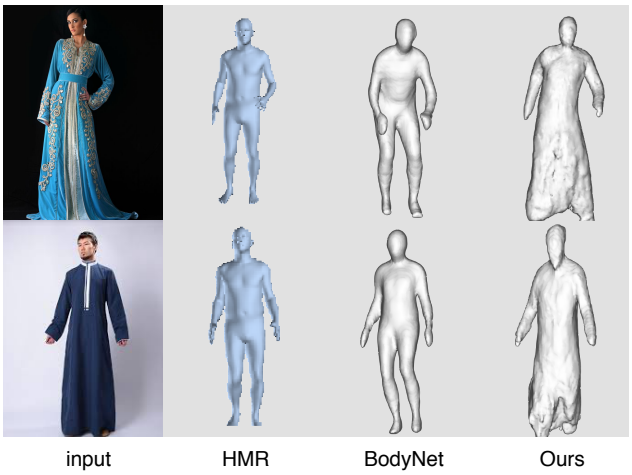

Figure 11: We qualitatively compare our method with two state-of-the-art single view human reconstruction techniques, HMR [22] and BodyNet [42].

Figure 12: Failure cases.

5. Discussions and Future Work

In this paper, we present a framework for monocular 3D human reconstruction using deep neural networks. From a single input image of the subject, we can predict the 3D textured geometry of clothed body shape, without any requirement of a parametric model or a pre-captured template. To this end, we propose a novel-view silhouette synthesis network based on adversarial training, an improved deep visual hull algorithm with a greedy view selection strategy, as well as a front-to-back texture synthesis network.

One major limitation of our current implementation is that our synthetic training data is very limited and may be biased from real images. See Figure 12 for a few typical failure cases, in which the 3D pose estimation may fail or there are some additional accessories not covered by our training data. It would be helpful to add realistic training data which may be tedious and costly to acquire. The output mesh using our method is not rigged and thus cannot be directly used for animation. Also we do not explicitly separate the geometry of cloth and human body. In the future, we plan to extend our method to predict output with high frequency details and semantic labels. Finally, it is interesting to infer relightable textures such as diffuse and specular albedo maps.

Acknowledgements

This work is supported in part by the JST ACCEL Grant Number JPMJAC1602, JSPS KAKENHI Grant Number JP17H06101, the Waseda Research Institute for Science and Engineering, the ONR YIP grant N00014-17-S-FO14, the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, the Andrew and Erna Viterbi Early Career Chair, the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, Adobe, and Sony. 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] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, pages 98–109, 2018.

[2] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005.

[4] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018.

[5] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. 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.

[7] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[9] G. K. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 375–382, 2003.

[10] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *International Conference on 3D Vision*, pages 108–117, 2016.

[11] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4826–4836, 2017.

[12] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[13] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[14] J.-S. Franco, M. Lapierre, and E. Boyer. Visual shapes of silhouette sets. In *International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 397–404, 2006.

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

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

[17] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton. Volumetric performance capture from minimal camera viewpoints. In *European Conference on Computer Vision*, pages 566–581, 2018.

[18] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision*, pages 1381–1388, 2009.

[19] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[20] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li. Deep volumetric video from very sparse multi-view performance capture. In *European Conference on Computer Vision*, pages 336–354, 2018.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[22] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[23] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. 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.

[24] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3D self-portraits. *ACM Transactions on Graphics*, 32(6):187:1–187:9, 2013.

[25] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

[26] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015.

[27] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017.

[28] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[29] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. In *ACM SIGGRAPH*, pages 369–374, 2000.

[30] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics*, 36(4):44:1–44:14, 2017.

[31] A. Mykhaylo, P. Leonid, G. Peter, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[32] R. A. Newcombe, D. Fox, and S. M. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–352, 2015.

[33] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

[34] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4):73:1–73:15, 2017.

[35] K. Rematas, C. H. Nguyen, T. Ritschel, M. Fritz, and T. Tuytelaars. Novel views of objects from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1576–1590, 2017.

[36] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *arXiv preprint arXiv:1803.00455*, 2018.

[37] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2006.

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[40] J. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *British Machine Vision Conference*, pages 6.1–6.11, 2017.

[41] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017.

[42] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision*, pages 20–36, 2018.

[43] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):97:1–97:9, 2008.

[44] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics*, 28(5):174:1–174:11, 2009.

[45] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[46] M. Waschbüsch, S. Würmlin, D. Cotting, F. Sadlo, and M. Gross. Scalable 3D video of dynamic scenes. *The Visual Computer*, 21(8):629–638, 2005.

[47] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[48] C.-Y. Weng, B. Curless, and I. Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. *arXiv preprint arXiv:1812.02246*, 2018.

[49] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *IEEE International Conference on Computer Vision*, pages 1108–1115, 2011.

[50] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics*, 37(2):27:1–27:15, 2018.

[51] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, pages 439–454, 2016.

[52] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3D human pose estimation in the wild by adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018.

[53] C. Zhang, S. Pujades, M. Black, and G. 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.

[54] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301, 2016.

[55] H. Zhu, H. Su, P. Wang, X. Cao, and R. Yang. View extrapolation of human body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2018.

[56] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, 23(3):600–608, 2004.