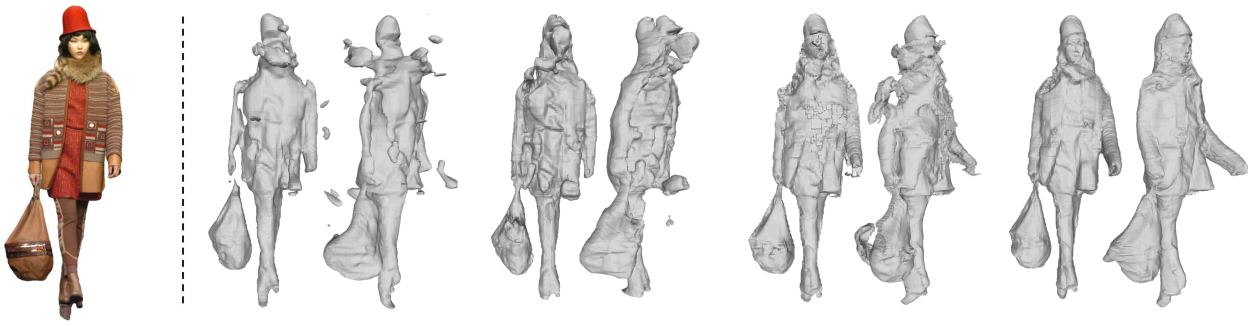
PIDiffu: Pixel-aligned Diffusion Model for High-Fidelity Clothed Human Reconstruction

Jungeun Lee, Sanghun Kim, Hansol Lee, Tserendorj Adiya, Hwasup Lim*
Korea Institute of Science and Technology (KIST)
jeunlee0306@gmail.com, powerkei@naver.com, hansol2651@gmail.com, tdain72@gmail.com,
hslim@kist.re.kr



(a) Input Image (b) PIFu [24] (c) Geo-PIFu [8] (d) IntegratedPIFu [4] (e) Ours

Figure 1. Our proposed PIDiffu could recover high-fidelity reconstruction of clothed human bodies in "in-the-wild" images. Compared with state-of-the-art methods, our reconstruction is more accurate and maintains plausible human topology.

Abstract

This paper presents the Pixel-aligned Diffusion Model (PIDiffu), a new framework for reconstructing high-fidelity clothed 3D human models from a single image. While existing PIFu variants have made significant advances using more complicated 2D and 3D feature extractions, these methods still suffer from floating artifacts and body part duplication due to their reliance on point-wise occupancy field estimations. PIDiffu employs a diffusion-based strategy for line-wise estimation along the ray direction, conditioned by pixel-aligned features with a guided attention. This approach improves the local details and structural accuracy of the reconstructed body shape and is robust to unfamiliar and complex image features. Moreover, PIDiffu can be easily integrated with existing PIFu-based methods to leverage their advantages. The paper demonstrates that PIDiffu outperforms state-of-the-art methods that do not rely on parametric 3D body models. Especially, our method is superior in handling 'in-the-wild' images, such as those with complex patterned clothes unseen in the training data.

1. Introduction

Reconstructing a high-fidelity clothed human body is an important research area to implement virtual reality for online shopping, remote attendance, and entertainment. The research community has shown interest in developing deep learning models for human digitization from a single image. This interest is because specialized devices, such as 3D scanners and multi-view capture environments, are primarily inaccessible to general users.

Pixel-Aligned Implicit Function (PIFu) [24] has remarkably improved in capturing intricate local details by utilizing pixel-aligned image features to determine the occupancy field. PIFu infers points along the same ray from same image features, potentially resulting in redundant and elongated human parts. Recent studies concentrate on conveying rich features to the multilayer perceptron (MLP) while maintaining PIFu's point-wise occupancy estimation strategy built on the MLP to address this issue [5, 13, 15, 16, 34, 41, 42].

Although this MLP-based occupancy network is computationally efficient, it heavily depends on image features without considering the 3D distribution. This limitation can compromise its ability to reconstruct plausible geometry, especially when an input image significantly deviates from the training data distribution.

In this paper, we present PIDiffu, an improved pixel-aligned implicit model designed to refine PIFU's MLP-based occupancy estimation approach. PIDiffu employs a

diffusion model [10], which learns line-wise distribution instead of point-wise occupancy to generate an occupancy field considering 3D distribution. This approach effectively handles depth ambiguity and improves reconstruction capability when faced with new images outside the training domain. In addition, This reconstruction robustness is strengthened by conditioning the diffusion model on image features using the guided attention.

Our contribution can be summarized as follows:

- We introduce the Pixel-aligned Diffusion model (PIDiffu), which is a simple yet advanced pixel-aligned implicit function. PIDiffu improves local details and structural accuracy in 3D human reconstructions without additional data.
- We propose an effective conditioning method, Feature-wise Linear Attention (FiLA), to strengthen the diffusion model's robust generation ability in the reconstruction task. This method allows PIDiffu to reconstruct plausible geometry even with unfamiliar images.
- PIDiffu is designed to incorporate with PIFu-based methods through simple modifications. In our experiments, the reconstruction results with PIDiffu integration surpass the original baseline.

2. Related Work

PIFu-based Human Reconstruction. PIFu [24] architecture comprises two essential components: a feature extraction network that extracts image features and an occupancy network that estimates occupancy fields using pixel-aligned image features. This design leads to high-fidelity 3D reconstruction, but challenges such as depth ambiguity and self-occlusion can arise.

Recent studies based on PIFu mainly focus on the lack of explicit geometric information, an inherent limitation of single images. While retaining the advantages of MLP-based occupancy networks, these studies incorporate extensive additional data into the implicit function to improve performance. For example, depth or LIDAR data are used into the occupancy network to address the issue of depth ambiguity [5, 15, 36]. These additional data are not always available in practical application scenarios. Parameterized 3D human models such as SMPL [17] and SMPL-X [19] enable networks to fuse explicit knowledge about human structure. Geometric information of 3D human such as pose and shape from SMPL is supplied to MLP [41, 42]. ICON [34] selectively utilizes image features based on self-occlusion information obtained from parametric human models. Other studies solve the self-occlusion problem by converting query points into SMPL's canonical space [9, 13, 16].

Another studies propose to extract additional geometric data such as coarse voxels or depth maps. GeoPIFu [8] employs a 3D convolutional network to produce coarse occupancy volumes derived from input images. The intermediate latent voxel features from trained 3D convolutional networks are used as geometric priors. IntegratedPIFu [4] uses a pre-trained network that can infer normal, depth, and region from images. With a slight deviation from previous methods PHORHUM [2] infers normal from SDF gradient predicted from an implicit function and applies this normal to rendering loss to improve geometric accuracy and detail.

Although these methods successfully improve the quality of reconstructed 3D geometries through rich additional data input (or inferred), they still struggle with the challenges of point-wise occupancy estimation, as discussed in the Section 3.

Diffusion Model-based 3D Reconstruction. The diffusion probabilistic model [27] consists of the forward process which incrementally adds noise to the original data and the reverse process which removes this noise. This gradual generation of new samples allows the diffusion model to generate gradients, which enable the model to progressively converge to the data distribution [35]. Due to its ability to capture complex data distributions, the diffusion model currently exhibits outstanding performance in various generative tasks such as image generation [10, 22, 29] and 3D generation [3, 12, 14, 18, 26, 31, 33, 38].

It is a critical concern that network overfitting when applying the diffusion model to 3D clothed human reconstruction from single images. This overfitting often leads to the generation of 3D clothed humans that are not related to the input image. Such overfitting can arise due to the insufficiency of available 3D data, which fails to capture the complete 3D distribution. The severity of this issue amplifies when accounting for varied poses. Rodin [30] can generate plausible 3D models resembling the input image and they require extensive datasets to represent the data distribution. This requirements is extremely difficult to prepare. LAS-DFDIFF [40] is closely related to our work in that it utilizes SDF data to train a diffusion model. This method infers noise using a 3D convolutional U-Net [23] by making the SDF into voxel form. This voxel representation allows direct use of the diffusion strategy in 3D, but resolution is limited due to the computational complexity of the 3D convolution U-Net.

3. Method

To address challenges related to depth ambiguity and sensitivity to out-of-distribution images, we introduce the Pixel-aligned Diffusion model (PIDiffu). In the following Section 3.1, we detail the PIDiffu architecture and discuss the seamless compatibility of PIDiffu with existing PIFu-based methods. Additionally, we introduce how Feature-

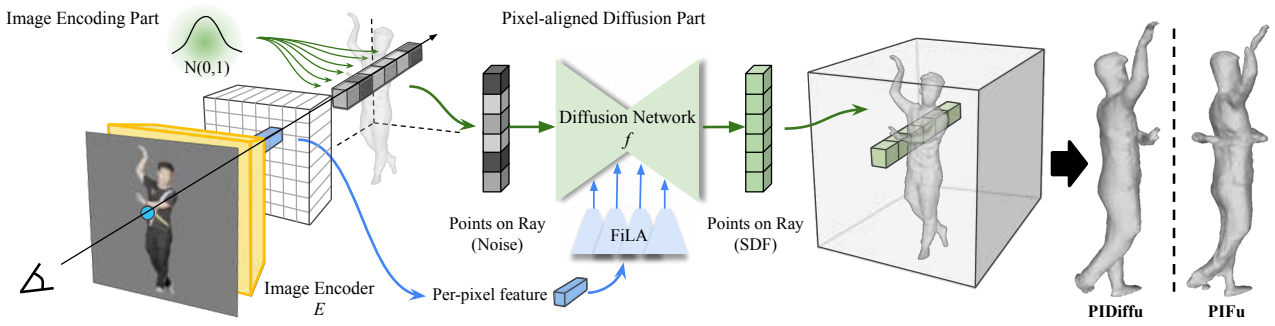


Figure 2. Overview of our PIDiffu framework. A per-pixel feature obtained from the image encoder is propagated to the diffusion model as a conditional feature through the proposed FiLA. The diffusion network generates all points along the z-axis of the outgoing camera ray from a pixel at once.

wise Linear Attention (FiLA) provides localized information to the diffusion model in Section 3.2. Figure 2 shows the overall structure of our pipeline.

3.1. Pixel-aligned Diffusion Model

Tackling the issue of depth ambiguity often involves directly considering the 3D distribution. GeoPIFu [8] employs a similar strategy by learning 3D correlations through a 3D U-Net [23], this approach tends to produce geometries with artifacts or distortions, particularly when handling images that were not part of its training distribution. To overcome this limitation, we propose using diffusion models, known for their robust convergence toward a trained distribution. However, using diffusion models to directly learn the 3D geometry distribution poses specific challenges, particularly in the reconstruction of clothed full-body humans. First, the extensive diversity in poses and appearances necessitates a large dataset to avoid overfitting [30]. Second, architectures designed for learning 3D geometries, such as the 3D U-Net, are computationally expensive, making high-resolution implementations challenging [18].

PIDiffu focuses on learning the 1D geometry distribution along camera rays to bypass these challenges since the image encoder learns spatial correlation in the 2D xy-plane in the training process. In contrast to a 3D structured dataset that struggles to cover the entire 3D distribution, a dataset formulated in 1D rays can encompass a broader distribution. This approach allows for ray-specific sampling that effectively tackles the issue of data overfitting and enhances computational efficiency during the training and inference process.

Image-driven Ray Denoising Diffusion. In this work, PIDiffu learns a 1D geometry distribution based on input image features. We define this 1D geometry distribution as the ray distribution, represented by R. A pixel-aligned ray, denoted by r, is a specific sample of R. To construct this ray, we initially samples a fixed number of points from near to far along the camera ray from a pixel and then calculates the points' signed distance to the 3D mesh surface. A pixel-aligned ray r is defined by these signed distances, represented as $r = [z_{\text{near}}, z_{\text{far}}]$. z is the signed distance at a 3D point to the mesh surface.

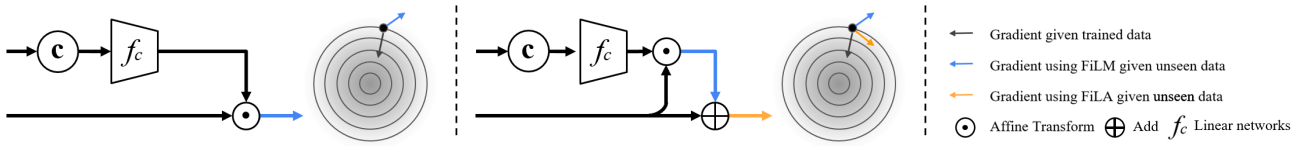
Subsequent to the construction of pixel-aligned rays, PIDiffu learns the ray distribution through the forward and reverse processes. The forward and reverse process of r^i, corresponding to the i^{th} pixel, are described as follows:

$$\begin{aligned} q(r_t^i|r_{t-1}^i) &= \mathcal{N}(r_t^i; \sqrt{1-\alpha_t} r_{t-1}^i, \alpha_t I), \\ p(r_{t-1}^i|r_t^i) &= \mathcal{N}(r_{t-1}^i; \mu(r_t^i, t, E(x)^i), \alpha_t I). \end{aligned} \tag{1}$$

Both q and p represent the forward and the reverse processes, respectively, defined as Markov chains. The predefined variance schedule of diffusion is represented by α. t denotes the timestep, and I stands for the identity matrix. We employ a fully connected convolutional image encoder E, which extracts a per-pixel feature vector from image x. μ is the predicted output from the denoising network. Distinct from the typical diffusion models that predict noise, we directly predict samples as suggested in DALL-E [21] and DiffuStereo [26]. This approach not only results in high-quality generation but also notably reduces the number of iteration steps required during inference. The loss function for our method is defined as follows:

$$L = \mathbb{E}_{x,t}[\frac{1}{N}\sum_{i=1}^{N}(||r_0^i - f(E(x)^i, r_t^i, t)||)] \tag{2}$$

During inference, we sample N rays from Gaussian noise, where N corresponds to the total number of pixels in the input image and then proceed to iteratively denoise these rays. The final mesh is generated through the Marching Cube algorithm. It is essential to use a consistent noise set across all pixels for a smooth surface. This consideration stems from the inherent attributes of the diffusion model. Even with identical input conditions, varying initial noises can lead to drastically different outputs.



(a) FiLM (b) FiLA

Figure 3. Each figure conceptually depicts the architecture and gradient directions of the FiLM (a) and the proposed FiLA (b). Given the unseen images that deviate enormously from the trained images, networks employing Film (a) might struggle to infer the appropriate gradient to converge into trained sample distribution R, leading to potential disorientation. Our proposed FiLA (b) effectively mitigates this issue by retaining the original trajectory.

Integration with Variants of PIFu-based Methods. A major advantage of our method is to integrate seamlessly with existing PIFu-based methods. Integration becomes straightforward with a few adjustments, as PIDiffu is focused on the occupancy estimation network. In the various PIFu-based methods, the occupancy of pixel x can be simply defined as :

$$\text{MLP}(E(x), E'(g), d_j) = s \in \mathbb{R} \tag{3}$$

where d_j is a z-depth of a j^{th} 3D point along the camera ray from pixel x, and E encodes additional information g such as depth maps and normal maps to capture rich features. $E'(g)$ refers to various methodologies used in variants of PIFu-based methods. To incorporate this $E'(g)$, Equation 2 is simply modified as follows:

$$L = \mathbb{E}_{x,t,g}[\frac{1}{N}\sum_{i=1}^{N}(||r_0^i - f(E(x)^i, r_t^i, t, E'(g))||)] \tag{4}$$

The entire network is trained in an end-to-end manner, thereby offering the flexibility to integrate various PIFu-based methods. This integration can be achieved through employing the diffusion model as a surface classifier instead of MLP. Our experiments, detailed in Section 4, demonstrate this compatibility by integrating with one of the state-of-the-art networks, IntegratedPIFu [4].

3.2. Feature-wise Linear Attention

A common approach for conditioning external information into neural networks is through Feature-wise Linear Modulation (FiLM) [20], which can be mathematically described as:

$$\acute{w} = \sigma(\gamma(c) * w + \beta(c)) \tag{5}$$

where w represents the neural network's feature, while γ and β are linear functions that take the condition c as input. The output features are denoted by $\acute{w}$, and σ is a nonlinear function, such as ReLU [1]. FiLM is widely utilized in diffusion methods due to its capability to conditionally adjust the scale and shift of each neural network feature w based on external information c [7, 30].

Despite its advantages, the FiLM approach often fails to generate appropriate geometry for previously unseen in-the-wild datasets, as illustrated in Figure 6. This limitation arises when the distribution X of the training images does not adequately represent the overall distribution $\mathscr{X}$ of in-the-wild images. In such cases, there may be a significant divergence between X and the distribution $\bar{X}$ of unseen in-the-wild images, where $\bar{X}$ is another subset of $\mathscr{X}$. We denote the ray distribution corresponding to X as R and that corresponding to $\bar{X}$ as $\bar{R}$.

The conceptual image of the reason that FiLM fails with given $\bar{X}$ can be found in Figure 3 (a). Given an image from $\bar{X}$, FiLM modulates features to predict the gradient in the ray distribution based on the input image conditions. For clarity, the gradient mentioned here refer the gradient in the ray distribution during the reverse process [35]. When $\bar{X}$ is significantly far from X, FiLM is unlikely to generate favorable condition features. The hadamard product propagates the inherent error in the condition feature to w. Consequently, essential features for gradient inference can be deactivated by the activation function. Diffusion model fails to deduce appropriate gradient to reach R, and drifts toward an unlearned imaginary $\bar{R}$.

We propose to use a conditioning method utilizing an attention mechanism [11, 32] to mitigate this limitation. The equation is as follows :

$$\acute{w} = w + \sigma(\gamma(c) * w + \beta(c)) \tag{6}$$

The pixel-aligned image features to be emphasized are selected and added to the previous path w with this attention mechanism such that output features $\acute{w}$ are not dramatically shifted toward the gradient for $\bar{R}$. This approach smooths the reverse process such that samples are guided to reach R even given an image from $\bar{X}$. Through our observations, this minor modification allows the model to generate geometry within the trained distribution when the in-the-wild input images are given while maintaining the quality when familiar input images are given. More comprehensive results and comparisons with FiLM and our methods are shown in Section 4.3.

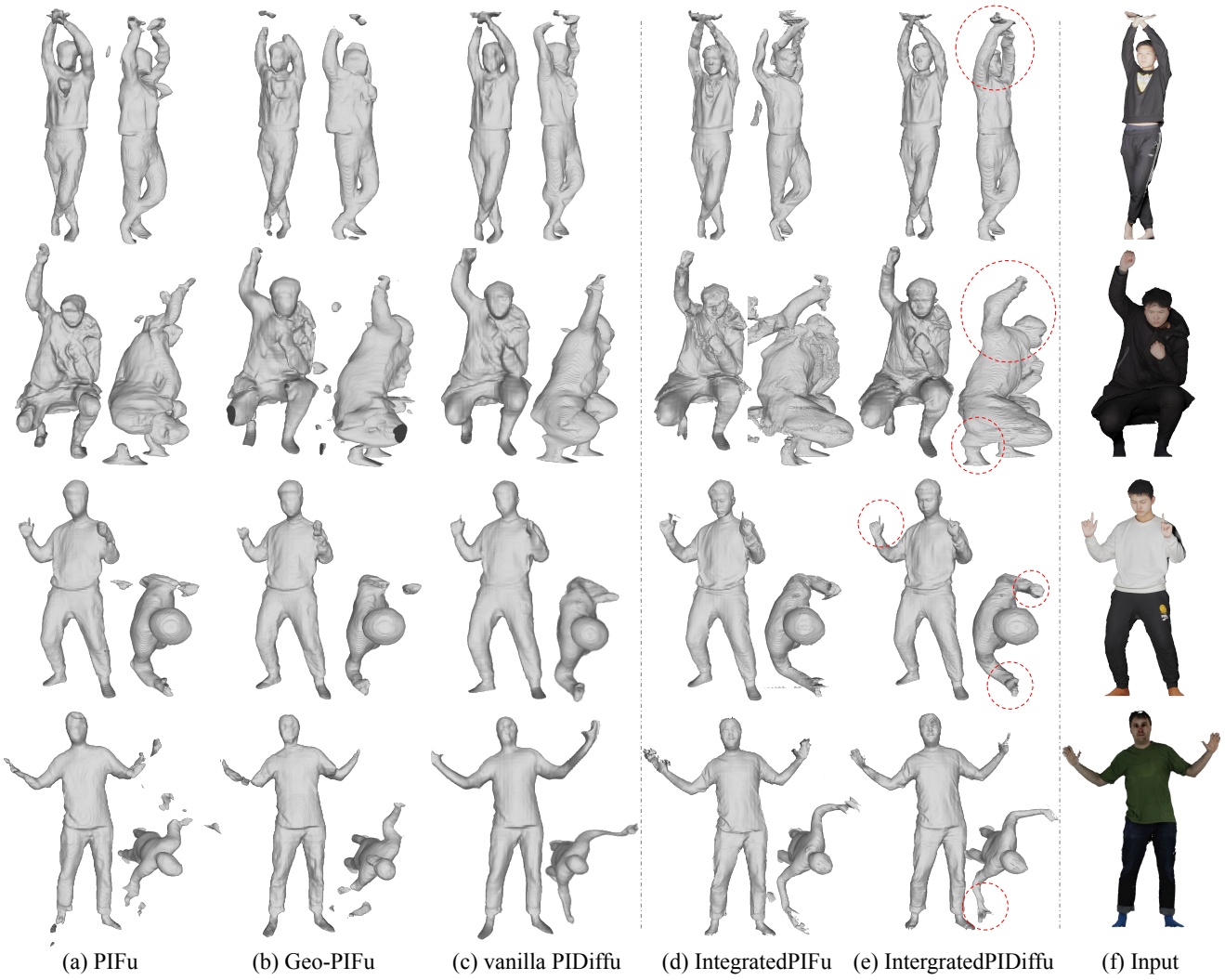


Figure 4. Qualitative evaluation with SOTA methods on Rendering Data (Thuman2.0, BUFF). (a) is a PIFu, (b) is Geo-PIFu, (c) is PIDiffu(ours) implemented on PIFu(a). (d) is IntegratedPIFu using DOS, Normal map prediction and HRI. (e) is PIDiffu(ours) implemented on IntegratedPIFu(d) and (f) is the input RGB image. For each method, we show the frontal view and an alternative view which demonstrates human geometry.

4. Experiment

4.1. Experimental Setup

PIDiffu and Baselines. We compare PIDiffu with three state-of-the-art PIFu-based methods: PIFu [24], Geo-PIFu [8], and IntegratedPIFu [4]. We divide the experiments into two cases: the vanilla PIDiffu case and the Integrated PIDiffu case.

In the vanilla PIDiffu case, we compare our vanilla PIDiffu with PIFu [24] and Geo-PIFu [8]. The resolution of the input image is 512×512, and evaluations are conducted at a resolution of $171 \times 256 \times 171$, consistent with Geo-PIFu. For the sampling method used in the training process, both PIFu and Geo-PIFu employ Discrete Spatial Sampling as described in PIFu [24], where the exterior of the mesh is represented as 0 and its interior as 1. In contrast, our method adopts Depth-Oriented Sampling (DOS), introduced by IntegratedPIFu [4], which expresses points as the signed distance along the camera ray. Because our method generates a fixed number of samples along the ray direction, sampling methods that define points as continuous values are essential. An ablation study on the adaptation of PIFu with DOS is presented in Section 4.3. We note that in the vanilla PIDiffu, the image feature is conditioned using a simplified version of Equation 6: $\acute{w} = w + \sigma(\mathrm{MLP}(c))$. This approach still avoids the direct modulation and activation of w, allowing the model to achieve a comparable effect.

In the IntegratedPIDiffu case, We show PIDiffu can be

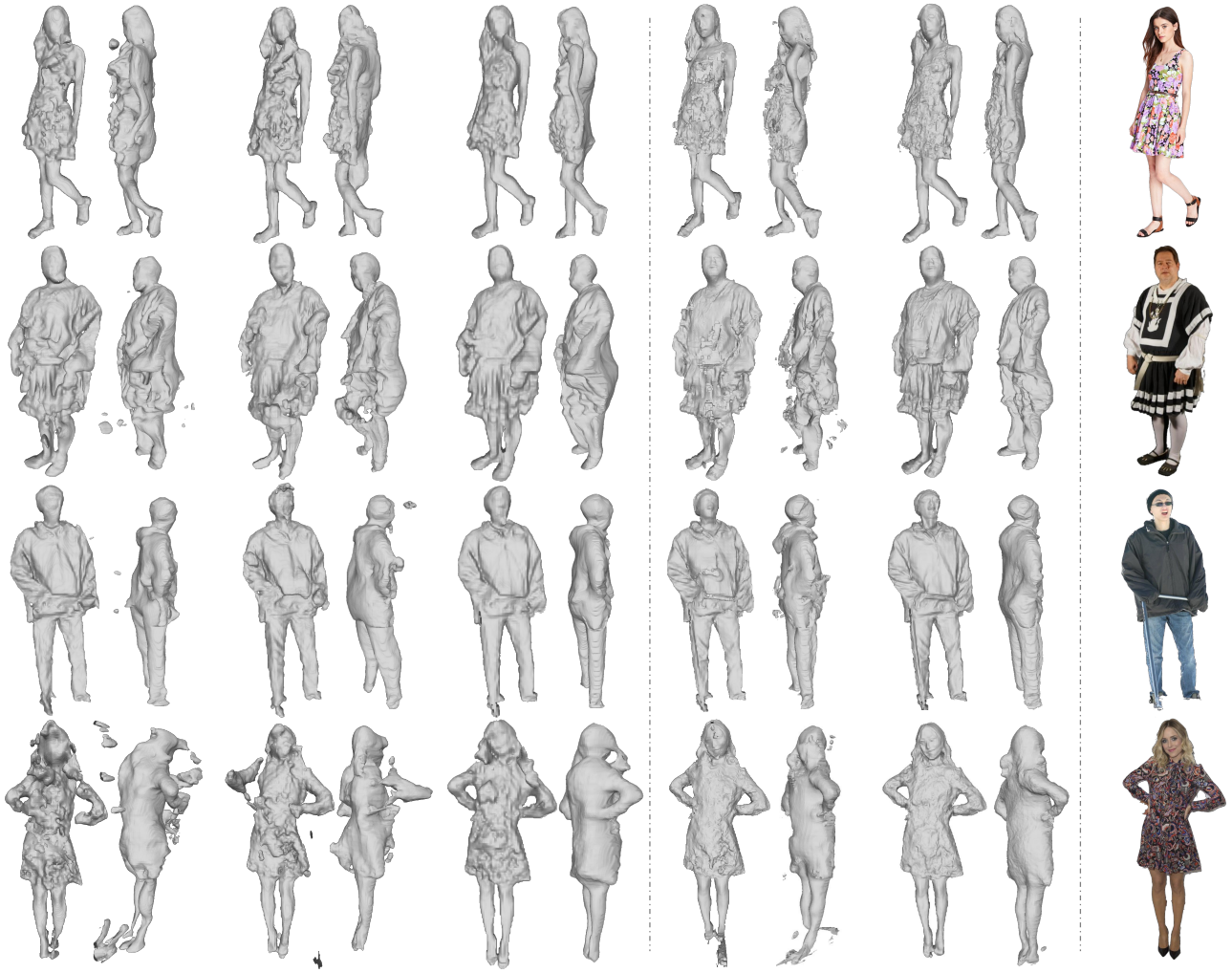


Figure 5. Qualitative evaluation with SOTA methods on in-the-wild data (SSHQ). (a) is a PIFu, (b) is Geo-PIFu, (c) is PIDiffu(ours) implemented on PIFu(a). (d) is IntegratedPIFu using DOS, Normal map prediction and HRI. (e) is IntegratedPIDiffu(ours) implemented on IntegratedPIFu(d) and (f) is the input RGB image. For each method, we show the frontal view and an alternative view which demonstrates human geometry.

incorporated with IntegratedPIFu [4], which is a state-of-the-art method built on PIFuHD [25], and compare our method to IntegratedPIFu. The resolution of an input image is 1024×1024, and evaluations are performed at a resolution of 256^3, following the IntegratedPIFu. IntegratedPIFu has several combinations of its suggested method, and we choose one that achieves the best quantitative score in the paper. This method utilizes DOS, incorporates frontal normal maps predicted from the input image as additional information, and employs the High-Resolution Integrator (HRI). HRI effectively enables high-resolution reconstruction by training the low-resolution path separately from the high-resolution path. For IntegratedPIDiffu, we adhere to the same configurations.

Dataset. We perform experiments using three datasets, only one of which is used for training. We train both PIDiffu and the baseline methods on the THuman2.0 dataset [37], which consists of 525 high-quality human models of Asian individuals. We use all these meshes and adopt random 80-20 train-test splits. For each mesh in both the training and test sets, we render 10 RGB images at various yaw angles.

For evaluation, we use the BUFF dataset [39], following the test split as outlined in IntegratedPIFu [4]. Additionally, we use the SHHQ dataset [6] for only qualitative comparisons, as it absences 3D ground truth meshes. This dataset comprises 2D masked color images of various individuals and accessories. We conduct tests on the first 100 numbered images from this dataset.

	Thuman2.0			BUFF		
	CD (10^4) ↓	P2S (10^4) ↓	Normal ↓	CD (10^{-3}) ↓	P2S (10^{-3})↓	Normal ↓
Geo-PIFu [8]	3.933	3.477	1.190	3.832	4.095	1.603
PIFu [24]	4.607	5.183	1.172	4.575	6.584	1.461
vanilla PIDiffu	**3.680**	**3.181**	**0.992**	**3.453**	**3.986**	**1.162**
IntegratedPIFu [4]	3.337	3.433	0.936	2.968	3.024	1.221
IntegratedPIDiffu	**3.110**	**2.936**	**0.806**	**2.861**	**2.959**	**0.978**

Table 1. Quantitative results of our method with other baselines on THuman2.0 and BUFF dataset. For the CD(↓) and P2S(↓) metrics, we multiplied the Thuman2.0 values by 10^4 and the BUFF values by 10^{-3}. For the Normal(↓) metric, we multiplied the values for all datasets by 10^2.

Metrics. We quantitatively evaluate our method using three metrics: Chamfer Distance (CD), which measures shape similarity between two point sets; Point-to-Surface Distance (P2S), which measures how close the points of the reconstructed shape are to the surface of the ground truth shape; and Normal Reprojection Error, which evaluates the alignment between the estimated and ground truth normals.

4.2. Comparison with Baselines

Qualitative Results. We qualitatively evaluate our method against baselines in both vanilla PIDiffu and IntegratedPIDiffu cases. The geometry reconstruction results for the Thuman2.0 test set and BUFF are shown in Figure 4. Results for the SHHQ dataset are presented in Figure 5.

Figure 4 and Figure 5 show that our method effectively addresses issues such as floating artifacts, duplicated arms, and elongated bodies seen in the baseline while preserving the local detail-capturing ability of traditional PIFu. The comparison between figures (d) and (e) demonstrates that our method can seamlessly incorporate complex structures such as HRI, improving reconstruction results.

Furthermore, Figure 5 demonstrates the robustness of PIDiffu in handling a range of unfamiliar image features, such as complex clothing, hats, and bags. Utilizing the diffusion model that converges to the training data distribution, PIDiffu effectively avoids the floating artifacts commonly encountered in other PIFu-based methods, leading to more plausible geometries. Notably, (d) and (e) in Figure 5 highlight a characteristic of the diffusion model: although PIDiffu produces high-quality facial reconstructions, the faces generated tend to have oriental features, reflecting the feature distribution in the training data.

Quantitative Results. We quantitatively compare our method against baselines in vanilla PIDiffu and IntegratedPIDiffu cases as shown in Table 1. We evaluate the methods using 10 rendered images for each mesh as input, taken at 36-degree yaw intervals for the Thuman2.0 dataset [37]. For the BUFF dataset [39], we use a single frontal image. Normal values are assessed from the same view as the image. As illustrated in Table 1, our method outperforms baselines across all metrics and datasets.

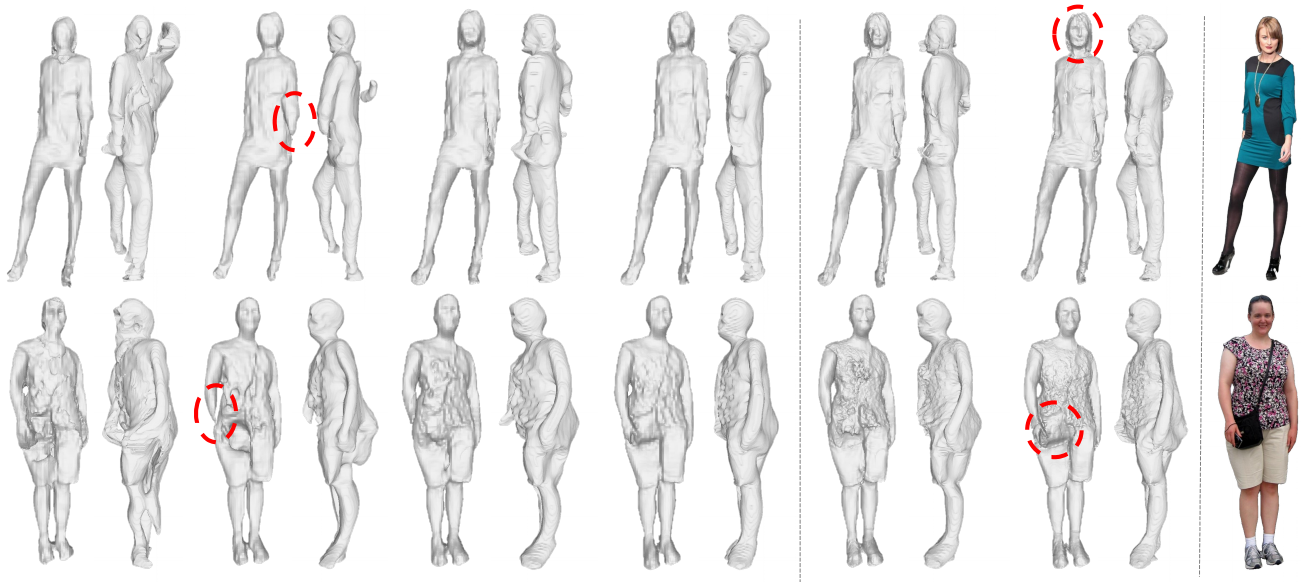
4.3. Ablation Study

We evaluate the effectiveness of the diffusion model and Feature-wise Linear Attention (FiLA) conditioning, both quantitatively and qualitatively. To assess the contribution of the diffusion model, we introduce Ray Prediction PIFu (RayPIFu), a variant that employs a multi-layer perceptron (MLP) to predict sample points along the ray direction, as opposed to using a diffusion model. We also compare the efficacy of FiLA conditioning against Feature-wise Linear Modulation (FiLM) conditioning. For this study, we examine several distinct methods: the baseline PIFu, RayPIFu, vanilla PIDiffu with FiLM, vanilla PIDiffu with FiLA, IntegratedPIDiffu with FiLM, and IntegratedPIDiffu with FiLA. All methods are implemented using PIFu with DOS, taking both an RGB image and a predicted normal map image as inputs.

Qualitative Results. Figure 6 shows the qualitative comparison between the six methods on the SSHQ dataset. RayPIFu(b) improves upon the baseline PIFu(a) by reducing issues, such as duplicate body parts, through the use of 3D correlations. However, it struggles to capture fine details, leading to issues including removing hands and distorted facial features.

In contrast, methods employing diffusion (c, d, e, f) successfully reconstruct more realistic and detailed clothed human geometry. Notably, PIDiffu with FiLA conditioning (d,f) outperforms PIDiffu with FiLM conditioning (c,e). It generates consistently plausible geometries across various scenarios, including in-the-wild images. It is particularly effective in capturing details of the face and bags shown in Figure 6.

Quantitative Results. Quantitative evaluation results are presented in Table 2 and Table 3. As indicated in Table 2, PIDiffu with FiLA conditioning outperforms its FiLM-conditioned except for the CD metric on Thuman2.0 dataset. Importantly, our FiLA-based method excels on the unseen BUFF dataset, indicating its ability to deal with out-of-distribution features. Regarding evaluating the diffu-



(a) PIFu (b) RayPIFu (c) vanilla PIDiffu with FiLM (d) vanilla PIDiffu with FiLA (e) IntergratedPIDiffu with FiLM (f) IntergratedPIDiffu with FiLA (g) Input

Figure 6. The ablation study on in-the-wild data (SSHQ). We evaluate the effect of the diffusion process (a,b,d) and FiLA conditioing (c,d,e,f).

	Thuman2.0 4view (360°)			BUFF 1view		
	CD↓	P2S↓	Normal↓	CD↓	P2S↓	Normal↓
PIDiffu w.FiLM	**3.087**	3.251	0.917	3.188	3.999	0.944
PIDiffu w.FiLA	3.158	**2.964**	**0.880**	**2.949**	**3.399**	**0.904**

Table 2. Quantitative results PIDiffu with Film conditioning and PIDiffu with FilA conditioning in THuman2.0 and BUFF dataset. Both methods trained with low-resolution images (512×512).

	Thuman2.0 4view (360°)			BUFF 1view		
	CD↓	P2S↓	Normal↓	CD↓	P2S↓	Normal↓
PIFu	3.540	3.897	1.015	3.335	4.176	1.170
RayPIFu	3.320	**2.715**	1.040	3.163	**2.830**	1.309
PIDIffu	**3.158**	2.964	**0.880**	**2.949**	3.399	**0.904**

Table 3. Quantitative results of RayPIFu, PIDiffu with Film conditioning, PIDiffu with FilA conditioning in THuman2.0 and BUFF dataset. All methods are trained using DOS and low-resolution images of 512×512 with its predicted normal maps.

sion model in Table 3, PIDiffu outperforms other CD and Normal metrics methods. However, RayPIFu shows better scores in the P2S metric. This discrepancy is partly due to RayPIFu's strategy of eliminating difficult-to-estimate parts. Because the P2S metric measures the distance from generated mesh points to the nearest ground truth surface, the absence of a mesh in complex regions, as observed in the second column of Figure 6, can improve the score.

5. Conclusion

In this paper, we have introduced PIDiffu, a new approach that addresses the existing challenges in PIFu-based human reconstruction. By integrating the robust probabilistic reasoning of diffusion models with spatial image features, PIDiffu offers enhanced human specific geometric structure without compromising local details. In addition, by introducing and utilizing the FiLA mechanism, PIDiffu demonstrates a remarkable ability to generate precise 3D geometries, even when presented with unfamiliar images. Moreover, PIDiffu is designed for easy integration with current PIFu-based methods, which demonstrates its adaptability. While this study did not compare methods using parametric body models such as ARCH++ [9], and DIFu [28], future work could explore such integration.

Limitation. In cases involving extreme self-occlusion, PIDiffu sometimes fails to produce accurate outputs. This limitation is particularly noticeable in high-occlusion scenarios, for example, when one body part entirely obscures another. We believe that this issue could potentially be alleviated by training the model on a more extensive dataset. A broader range of examples might improve the model's ability to learn accurate ray distributions, especially in high-occlusion scenarios.

Acknowledgement

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation)

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 4

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. *CVPR*, pages 1506–1515, 2022. 2

[3] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 2

[4] Kennard Yanting Chan, Guosheng Lin, Haiyu Zhao, and Weisi Lin. Integratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction. *ECCV*, pages 328–344, 2022. 1, 2, 4, 5, 6, 7

[5] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. *CVPR*, 2022. 1, 2

[6] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint*, arXiv:2204.11823, 2022. 6

[7] Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2022. 4

[8] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. 1, 2, 3, 5, 7

[9] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. *CVPR*, 2021. 2, 8

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, pages 6840–6851, 2020. 2

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[12] Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. *arXiv preprint arXiv:2307.06526*, 2023. 2

[13] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. *CVPR*, 2020. 1, 2

[14] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. *arXiv preprint arXiv:2305.11870*, 2023. 2

[15] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. *CVPR*, 2020. 1, 2

[16] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. *CVPR*, 2023. 1, 2

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM TOG*, 2015. 2

[18] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. *CVPR*, 2023. 2, 3

[19] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. *CVPR*, 2019. 2

[20] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. 32(1), 2018. 4

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2022. 3

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 2

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015. 2, 3

[24] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *ICCV*, pages 2304–2314, 2019. 1, 2, 5, 7

[25] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. *CVPR*, pages 84–93, 2020. 6

[26] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. *ECCV*, 2022. 2, 3

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. pages 2256–2265, 2015. 2

[28] Dae-Young Song, HeeKyung Lee, Jeongil Seo, and Donghyeon Cho. Difu: Depth-guided implicit function for clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8738–8747, 2023. 8

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 2

[30] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. *CVPR*, 2023. 2, 3, 4

[31] Zhenzhen Weng, Zeyu Wang, and Serena Yeung. Zeroavatar: Zero-shot 3d avatar generation from a single image. *arXiv preprint arXiv:2305.16411*, 2023. 2

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[33] xiaohui zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *NeurIPS*, 2022. 2

[34] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. Icon: Implicit clothed humans obtained from normals. *CVPR*, 2022. 1, 2

[35] Song Yang and Ermon Stefano. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019. 2, 4

[36] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. *CVPR*, 2021. 2

[37] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. *CVPR*, pages 5746–5756, 2021. 6, 7

[38] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 2

[39] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. *CVPR*, pages 4191–4200, 2017. 6, 7

[40] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *arXiv preprint arXiv:2305.04461*, 2023. 2

[41] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. *CVPR*, 2021. 1, 2

[42] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, 2021. 1, 2