FAMOUS: High-Fidelity Monocular 3D Human Digitization Using View Synthesis

Vishnu Mani Hema[1] Shubhra Aich[1] Christian Haene[2]
Jean-Charles Bazin[2] Fernando De la Torre[1]

[1] Carnegie Mellon University
[2] Independent Researcher
vmanihem@alumni.cmu.edu

Abstract. The advancement in deep implicit modeling and articulated models has significantly enhanced the process of digitizing human figures in 3D from just a single image. While state-of-the-art methods have greatly improved geometric precision, the challenge of accurately inferring texture remains, particularly in obscured areas such as the back of a person in frontal-view images. This limitation in texture prediction largely stems from the scarcity of large-scale and diverse 3D datasets, whereas their 2D counterparts are abundant and easily accessible. To address this issue, our paper proposes leveraging extensive 2D fashion datasets to enhance both texture and shape prediction in 3D human digitization. We incorporate 2D priors from the fashion dataset to learn the occluded back view, refined with our proposed domain alignment strategy. We then fuse this information with the input image to obtain a fully textured mesh of the given person. Through extensive experimentation on standard 3D human benchmarks, we demonstrate the superior performance of our approach in terms of both texture and geometry. Code and dataset is available at https://github.com/humansensinglab/FAMOUS.

Keywords: Human digitization · 2D prior · fashion dataset

1 Introduction

High-fidelity 3D human digitization (e.g, [34]) is crucial for a wide range of virtual reality (VR) and augmented reality (AR) applications. This technology is extensively used in industries like gaming and entertainment [3], animation [7], visual effects, virtual fashion experiences [6], fitness programs [2], and virtual training simulations. It also has significant applications in healthcare, anthropology [5], and forensics [1, 36] to name a few. This advanced 3D digitization enables the creation of highly realistic avatars, enhancing the immersive experience in virtual environments by providing detailed and lifelike interactions.

Creating such high-precision avatars requires generating both texture and shape in meticulous detail. Traditionally, this involves collecting data using a 3D scanner and/or a multi-camera setup [16]. This data is then processed to create

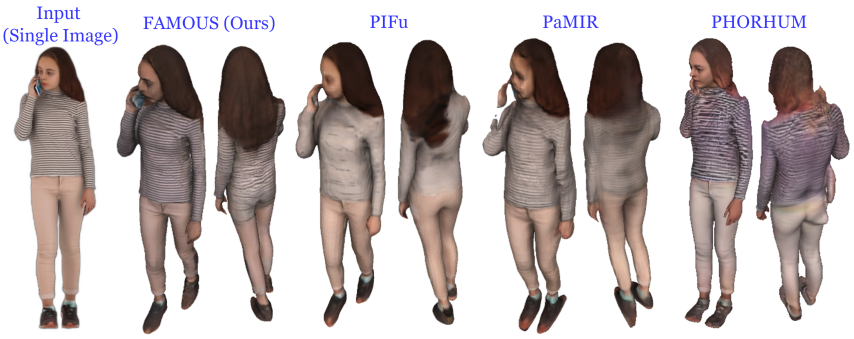

Fig. 1: Visual comparison between monocular 3D human digitization methods. Given a single image input, we demonstrate the renderings of textured mesh generated by our approach (FAMOUS) and SOTA pipelines (*i.e.* predicting both the shape and texture): PIFu [33], PaMIR [52], and PHORHUM [10]. Our approach presents significantly improved results in terms of texture and geometry.

a mesh through algorithms, followed by texture mapping, rigging, and further refinements often involving human artists. However, this semi-manual and multi-step process is not easily scalable for contemporary VR, MR, and AR platforms, which cater to millions of users globally, like the rapidly evolving digital environments of the metaverse.

Creating an automated method that can handle millions of users is crucial for bringing virtual reality into the mainstream. Recent advancements in deep learning have spurred several initiatives towards automation and scalability [33, 34, 43, 44]. However, while the geometric accuracy of these state-of-the-art methods is impressive, the texture quality of unseen areas is still lacking, especially in moderately textured clothing (see Figure 1). Some methods even overlook this aspect entirely [43, 44].

Our approach focuses on achieving both realistic texture quality and geometric accuracy, see Figure 1. Additionally, we address practical concerns like bandwidth limitations in large-scale VR platforms. To tackle this, we propose an approach for 3D human digitization using just a single image per user. Considering the massive user base, ranging from millions to potentially billions, even uploading two images per user would significantly increase bandwidth demands.

PIFuHD [34] is the first attempt regarding full-resolution monocular 3D human reconstruction that builds on top of the pioneering work of PIFu [33] based on implicit function (IF) models. Incorporation of articulated models from the SMPL family [26, 28] to guide the implicit function further improves the geometric reconstruction quality for challenging poses in ICON [44]. However, this guided reconstruction tends to overfit the articulated model and thus fails to produce realistic results for typical fashion poses (details in the supplementary

Section B). Note that among these methods, only PIFu [33] predicts both shape and texture (in a coarse resolution), whereas PIFu-HD, ICON, and ECON focus on the reconstruction task only (i.e. no texture). PHORHUM [10] is one of the most recent entry dealing with both human shape and texture based on the albedo surface color and shading information but still fails to extract semantically accurate texture for the occluded region. DIFU [39] and 2K2K [17] also, effectively reconstructs geometry from high resolution image but DIFU fails to obtain high fidelity textures and 2K2K doesn't focus on texture generation.

We believe that a key obstacle in achieving satisfactory texture quality in existing literature is the limited diversity of textures in the relatively small pool of available 3D samples [4, 46]. Acquiring detailed 3D scans is a time-consuming and costly process. Conversely, there is an abundance of varied cloth textures in 2D images, for example accessible in large-scale 2D fashion datasets [25] and online. This contrast points to an opportunity for leveraging these extensive 2D resources to enhance texture quality in 3D models. Drawing from this insight, our paper leverages extensive 2D fashion datasets to enhance the texture quality of 3D models. As will be shown in the experiments section, this approach not only improves texture fidelity but also boosts the geometric precision of the models.

Our approach FAMOUS integrates the rich textural data from 2D datasets into our 3D modeling process through a technique of view synthesis, utilizing pretrained hallucinator [30]. We then iteratively refine the hallucinator, focusing on disentangled factors as outlined in our methodology section. This self-supervised process, which we term "disentangled domain alignment", effectively aligning to the limited variety found in 3D datasets. Our extensive experiments demonstrate that by merging abundant 2D data with a smaller set of 3D scans in this manner, we can produce 3D models of superior fidelity, both for texture and geometry, compared to existing state-of-the-art methods. To our knowledge, this is the first instance of using 2D datasets in conjunction with limited 3D dataset through the lens of hallucinators, using domain alignment strategy focused on disentanglement factors.

Overall, below is the summary of our contributions:

- We propose FAMOUS, a framework that generates a high-fidelity textured mesh given a single RGB image. We harness the rich context available in the form of 2D fashion datasets through the lens of a hallucinator resulting in improved 3D shape and texture, especially for the occluded views.
- To this end, we also contribute a large-scale 2D fashion dataset, a derivative from Deep Fashion HD [25] but with full body front back image pairs with their COCO-skeleton keypoints annotations. We hope this new dataset will complement future research in the community.
- We introduce a self-supervised disentangled domain alignment approach to iteratively enable the hallucinator generalize to the distribution of the target 3D dataset.
- Finally, we will release our complete codebase along with the dataset splits to facilitate more open-source research in this subdomain.

2 Related Work

2.1 3D Human Digitization

Multi-view reconstruction. Earlier work [42, 54] in visual hull-based surface reconstruction required images to be captured from multiple viewpoints. The reconstructed avatar needed further postprocessing based on multiview silhouettes to compensate for the additional complexity induced by the topology of the cloth and self-occlusion, in particular. However, the visual hull based approaches fall short on fine-grained reconstruction when the number of input views is limited. On this account, deep volumetric stereo [20] attempts to predict the volumetric occupancy that can capture dynamic clothed humans from highly sparse views. A similar approach is used in [15] that uses a trained autoencoder to generate a deep prior to enable high-end volumetric captures. However, the inherent requirement for a multi-camera setup for these systems poses additional constraints regarding scalability to mass users.

Single-view reconstruction. These approaches can be categorized into explicit shape-based methods (i.e. voxel grid techniques, template meshes), implicit function-based approaches, and hybrid methods combining both.

1) Explicit shape based approaches. The introduction of SMPL [26] made the estimation of pose and shape in 3D human reconstruction tractable. SMPL works based on a linear combination of a small set of body shapes and pose parameters, which allows it to generate to a wide range of realistic body shapes and poses. Octopus [9] takes SMPL further by introducing SMPL-D which learns the SMPL body parameter and additional 3D vertex offsets that model clothing, hair, and details. The primary limitation of SMPL and SMPL-D is their fixed topology of the template mesh, which means that these template models are difficult to fit to arbitrary topological changes such as the disappearance of body parts and clothes with substantially different topologies. Other methods use nonparametric depth map [38] or point cloud [47]. However, scaling up these approaches to model the loose or diverse clothing is nontrivial.

2) Implicit function based approaches. This kind of models offer several advantages, such as topology-agnosticism and the ability to represent arbitrary 3D clothed human shapes. PIFu [33] introduces pixel-aligned implicit human shape reconstruction – the first purely implicit function based occupancy and texture predictors for 3D human digitization. PIFuHD [34] improves the geometric details with multi-resolution network architectures and estimated normal maps. However, both these methods tend to overfit to the body poses in training distribution (i.e. fashion poses) [44] GeoPIFu [18], Self-Portraits [23], PINA [14], and S3 [45] overcome this limitation by introducing geometric priors to regularize the deep implicit representation.

3) Hybrid appraoches. PaMIR [52], DeepMultiCap [51], JIFF [12], ARCH [21], ARCH++ [19] and ICON [44] attempt to combine both the explicit, parametric body models with continuous, implicit representations. Despite the promis-

ing performance of these approaches emphasizing primarily on the modeling innovations, the fundamental limitation regarding the scarcity of detailed 3D scans remains somewhat unexplored. This is evident in case the texture pattern of cloth topology is far from the quite limited training distribution of the 3D scans (Figures 1 and 7).

2.2 Pose-Guided Person Image Synthesis

These approaches synthesize novel views of a person from a reference (source) image and a target pose. Controllable GAN based ones [27, 35] extract person attributes from various segmentation maps. Dense deformations and flow-based methods [8, 22] generate aligned features and estimate appearance flow between references and desired targets. GFLA [31] obtains a global flow field and occlusion mask to match the patches from the source image to the target pose. Some other methods [22, 24] use 3D geometric details that fit SMPL mesh onto 2D images and query the source appearance using the 3D context. PISE [49] and CASD [53] parse maps to guide the view synthesis in the target pose. CoCosNet [50] applies cross-attention to extract dense correspondences but are limited in lower resolution due to their quadratic memory footprint. Recently, NTED [30] proposes an efficient attention mechanism based on semantic attributes to achieve promising texture quality but lacks 3D consistency.

3 Our Approach: FAMOUS

FAMOUS aims to infer high-fidelity 3D shape and texture of a clothed human from a single monocular image. Figure 2 illustrates our complete framework comprising of three steps: (1) Distributionally Aligned Hallucinator (DAH) for generation of the back-view image. (2) Geometry prediction, and (3) texture prediction from the reference image and generated back image.

3.1 Distributionally Aligned Hallucinator (DAH)

A pivotal stage in our process involves generating a back image from the frontal view to provide an additional viewpoint for constructing a 3D model. While this task can be achieved easily using methods trained on 3D scans [33, 52], these approaches are typically limited by small training datasets and struggle to generalize well, particularly in term of texture. In this paper, we propose leveraging large-scale 2D datasets, the source (e.g., Fashion dataset) alongside a limited amount of 3D data, the target (e.g., Render-People dataset [4]) to train a robust hallucinator. This hallucinator is designed to generate a back view from a frontal image. Drawing inspiration from recent research [30], we pretrained the hallucinator by augmenting an existing large-scale 2D dataset to account for partial occlusion of the samples (semantic changes) and different body poses (pose changes). Following this augmentation, we conduct simultaneous training and fine-tuning on the target 3D dataset (RenderPeople [4]).

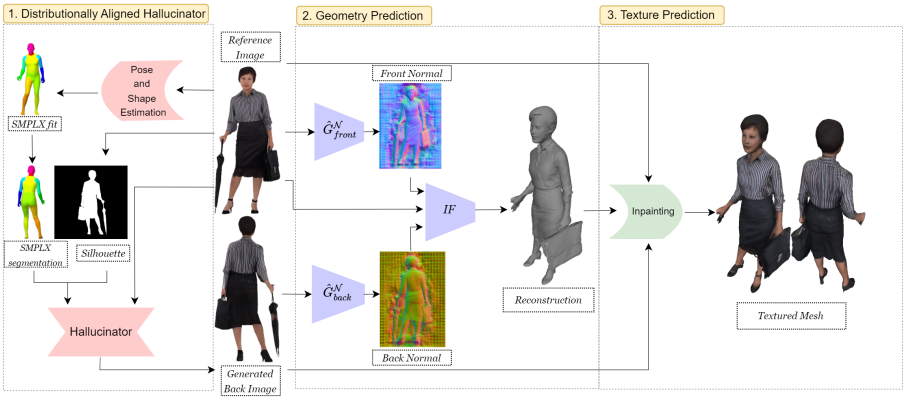

Fig. 2: Method overview. Given a single frontal image of a subject, FAMOUS generates a high-quality textured mesh. Our distributionally aligned hallucinator (Section 3.1) first predicts the back view based on the reference image, SMPL-X segmentation map and silhouette. Next, the normal map generator $\{(\hat{G}_f^{\mathcal{N}}, \hat{G}_b^{\mathcal{N}})\}$ takes the reference image and generated back image as input and output the respective normal maps. The reference image and the normal estimates are leveraged by the implicit function network (IF) for the geometry prediction (Section 3.2). Finally, our learnable texture prediction module refines the 3D texture aggregated from the input reference image and predicted back view (Section 3.3).

It's worth noting that in our current implementation, we utilize the state-of-the-art sparse attention-based StyleGAN hallucinator [30] rather than the most recent diffusion-based approach [11] due to the substantial memory requirements of the latter.

Problems with Existing Approaches. As mentioned above, SOTA methods for monocular 3D human digitization primarily revolve around training models using 3D scans with limited texture variations. Hence, these approaches fail to achieve good generalization in terms of texture. To overcome this bottleneck, a scalable solution is to leverage the abundant and easily accessible fashion 2D datasets. Directly incorporating the hallucinators [11,30] train on source fashion datasets into our pipeline does not yield satisfactory results for the target 3D dataset. That is, if we train our models in 2D Fashion dataset (e.g., DeepFashion [25]), and test it in target 3D dataset (e.g., Render people [4]), the results are not satisfactory. The same occurs when we fine-tune a pretrained model. The domain gap between these two datasets leads to performance degradation of the hallucinator on the target data distribution. Domain adaptation methods such as [40,41] fall short due to the insufficient support [13] of our source and target dataset in terms of pose and texture variations, respectively (see supplementary Section A for examples and more detailed explanation). To address this issues, we propose the Disentangled Domain Alignment (DDA) approach.

Disentangled Domain Alignment (DDA) An image can be represented by its style, semantics, pose, and view [29]. Our hallucinator [30] can learn to dis-

entangle these factors from an image to perform a task like view synthesis. We find that the domain gap between the two datasets in the context of the hallucinator depends on these disentanglement factors, so we introduce a factorized alignment approach rooted in this concept. We refer to this approach as DDA. The key steps of our DDA scheme are outlined as follows.

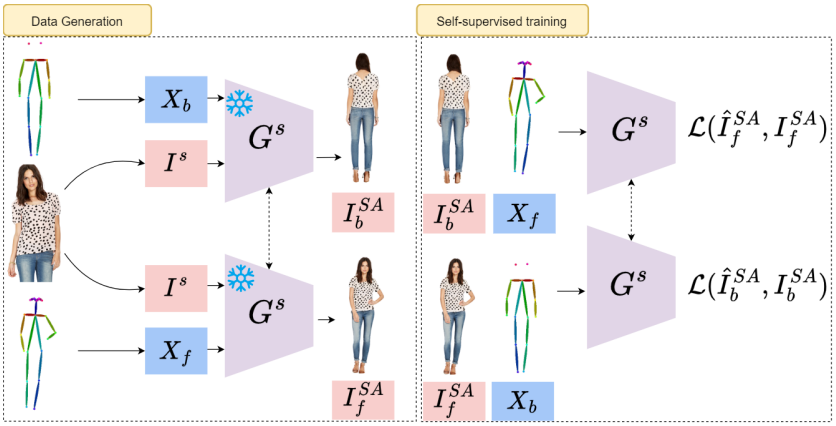

Fig. 3: Semantic Alignment (SA): Given a partial image I^s and full body COCO skeleton (X_f, X_b) from our 2D fashion dataset (i.e. source), we generate the front/back pairs $(I_f^{\mathcal{SA}}, I_b^{\mathcal{SA}})$ similar to our full body target 3D dataset (thus aligning the *semantics*) with a frozen hallucinator (G^s). Next, we finetune the hallucinator on these full body pseudo pairs from one another. The hallucinator is initialized with pretrained weights from [30], indicated by superscript s. The weight sharing in each stage is indicated by the bidirectional dotted arrows. The default loss function $(\mathcal{L})$ proposed in [30] is used.

Semantic Alignment (SA): The goal of SA is to shift the data distribution learned by the hallucinator from the source to the target distribution based on the semantics. The main difference, in terms of semantics, is the obvious lack of a sufficient number of full body image pairs covering the front/back view in the source dataset. More specifically, only about 8.5% of the source fashion dataset [25] contains full body pairs and 10% of which encompasses back views. This is prevalent for the fashion datasets since most of these image pairs only highlight a single cloth (i.e. upper/lower half). To address this limitation, we first generate full body (i.e. target semantics) image pairs for each sample in the source dataset. For this step, we sample image and guidance with target semantics (i.e. full body COCO key points) from the source dataset and generate full body image pairs with a pretrained hallucinator. Then, we finetune the hallucinator with these generated image pairs in a self-supervised manner, as shown in Figure 3. We find this pretraining stage improves the texture prediction, due to the alignment of the hallucinator weights more towards the full body semantic distribution, as will be shown in the experiments section (Table 2).

Fig. 4: Dataset: The sets of images randomly sampled from the source dataset (top) with their corresponding SA (mid) and PA (bottom) pairs.

To this end, we also release these full body image pairs and their COCO-skeleton keypoints annotations to facilitate future research. A few samples of the generated image pairs are shown in Figure 4. Note that the front/back images do not necessarily mean front/back canonical fashion poses.

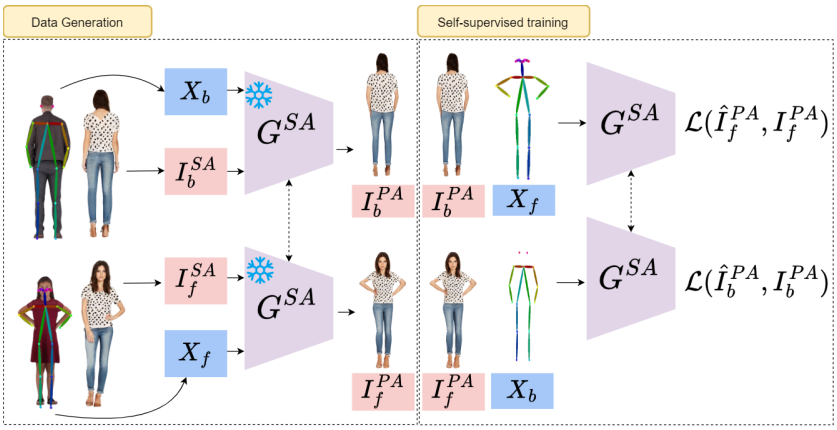

Fig. 5: Pose Alignment (PA): Given the subset of the pseudo pairs $(I_f^{\mathcal{SA}}, I_b^{\mathcal{SA}})$ after SA, we generate the front/back pairs following the pose from the target distribution $(I_f^{\mathcal{PA}}, I_b^{\mathcal{PA}})$ to finetune the semantically aligned hallucinator further. The hallucinator is initialized with weights obtained after SA.

Pose Alignment (PA): Following the semantic alignment, we perform pose alignment to further shift the data distribution learned by our hallucinator based on the target pose distribution. We obtain the pseudo pairs $\{(I_f^{\mathcal{PA}}, I_b^{\mathcal{PA}})\}$

from the SA pairs and guidance (with target pose distribution) from the target dataset, as shown in Figure 5. The sampling for the guidance is based on a discriminator (checkpointed after vanilla hallucinator training [30]) score thresholding to prevent noisy reconstruction of the PA pairs (See supplementary Section A for details). This somewhat conservative threshold allows us to select a subset with a higher degree of realism that eventually prevents degeneracies during further finetuning. Finally, we finetune the semantically aligned hallucinator from the last step with these pose-aligned pseudo pairs. Note that we also switch the guidance structure (used to represent pose) from COCO keypoints (2D) to SMPL-X segmentation map and silhouette [28], which we empirically find in reconstructing the occluded regions more accurately due to its better expressiveness.

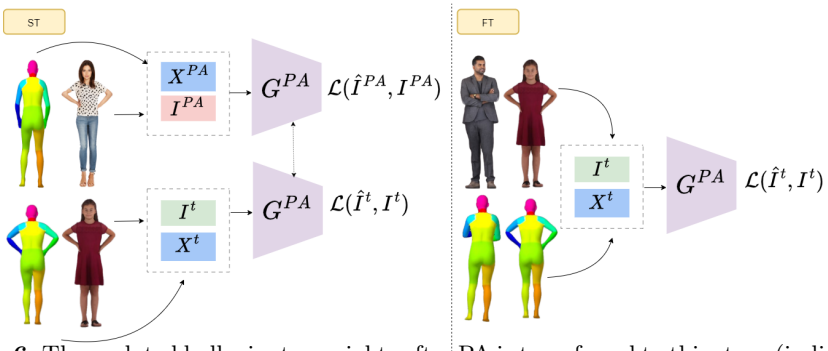

Fig. 6: The updated hallucinator weights after PA is transferred to this stage (indicated by superscript of G). We replace the guidance X from COCO key points to SMPL-X segmentation map for better expressivity. **Simultaneous training:** The pseudo pairs after PA, $I^{\mathcal{PA}}$, and the 2D renderings of our target 3D dataset I^t are trained together to align the style (i.e. cloth texture and topology) from both datasets or domains. **Finetuning:** The hallucinator is finetuned directly on the 2D renderings of our target 3D dataset.

Simultaneous Training (ST) / Finetuning (FT): At this stage, we obtain the pose-aligned pseudo pairs $\{(I_f^{\mathcal{PA}}, I_b^{\mathcal{PA}})\}$. This updated subset represents the source style distribution (i.e. texture/clothing patterns) in the target semantic and pose distribution. For ST, this adapted subset $\{I^{\mathcal{PA}}\}$ along with the 2D renderings $\{I^t\}$ from our target 3D dataset are jointly used to finetune our PA hallucinator for the style alignment (Figure 6). In FT, the 2D renderings $\{I^t\}$ are directly used to finetune our PA hallucinator. In the experiments section, Table 2 shows the ablations of the relative improvements after incorporating these individual alignment stages in the order prescribed above. During training, we use the GT COCO keypoints, and SMPL-X fits, but for testing, we obtain the SMPL-X fits using off the shelf pose and shape estimation model [48].

3.2 Geometry Prediction

This section describes the process for shape generation (using implicit functions) and texture composition.

Background on Implicit Functions For any point $\mathbf{x} \in \mathbb{R}^3$, the implicit function (IF) methods learn the conditional probability for occupancy $p(\mathbf{x}|\Pi_{\mathbf{x}})$, where $\Pi_{\mathbf{x}}$ is the 2D projection of $\mathbf{x}$. PIFu [33], the precursor of PIFuHD [34], estimates the probability as $p(\mathbf{x}|\Pi_{\mathbf{x}}) = f(\mathcal{F}(\Pi_{\mathbf{x}}), \Delta)$, where $\mathcal{F}$ is the pixel aligned feature, and Δ is the depth estimate of x in camera space.

PIFuHD enhances the vanilla PIFu framework with a (low/high) dual-resolution formulation:

$$p_c(\mathbf{x}|\Pi_{\mathbf{x}}) = f^{\theta}(f^{\lambda}(\mathcal{F}_c(\Pi_{\mathbf{x}}), \Delta)) \tag{1}$$

where p_c is the occupancy field at a coarse-resolution, and F_c is the corresponding pixel-aligned feature. Also, the implicit function is split into a composition as $f^{\theta}(f^{\lambda})$, where f^{λ} jointly contributes to both low/high-resolution occupancy predictions, and f^{θ} is employed for low-resolution prediction only. Next, the high-resolution model predicts the final occupancy as:

$$p(\mathbf{x}|\Pi_{\mathbf{x}}) = f(\mathcal{F}(\Pi_{\mathbf{x}}), f^{\lambda}(\mathcal{F}_c(\Pi_{\mathbf{x}}), \Delta)) \tag{2}$$

where $\mathcal{F}$ is a separate fine resolution 2D pixel-aligned feature and $f^{\lambda}(.)$ is the same joint embedding from Eq. 1.

Occupancy Prediction for Shape Estimation Our distributionally aligned hallucinator (DAH), as detailed in Section 3.1, is designed to be versatile and compatible with various shape inference methods. These include purely implicit function (IF) approaches such as PIFu and PIFuHD, as well as hybrid methods that combine implicit functions with explicit articulated model fitting, like SMPL/SMPL-X. Notably, IF approaches generally excel in producing high-quality reconstructions for typical fashion poses, as highlighted in the literature, while hybrid methods tend to shine in handling challenging and acrobatic poses. However, our primary objective is to create lifelike high-quality avatars for AR/VR/MR applications. Once the avatar is initially constructed from a standard pose, further animation of the 3D model becomes feasible, even for complex movements. Consequently, in this paper, we adopt the purely IF-based network, PIFuHD, as the baseline for enhancing DAH.

Examining Equations 1 and 2, it is evident that PIFuHD directly estimates depth ($\Delta(\ldots)$). Additionally, pixel aligned features $\mathcal{F}_c$ and $\mathcal{F}$ are leveraged for the occupancy prediction of query 3D points. Our observations indicate that relying solely on the features derived from the back normal map predicted (from front image) by PIFuHD falls short in estimating fine details in the occluded region. To address this limitation, apart from the improved learning of cloth textures from

the 2D fashion datasets, we reroute our DAH-predicted back view as input to the back normal estimator (as illustrated in Figure 2). This modification ultimately results in superior performance for the occluded back region (Please refer to the supplementary Section B for the visualization).

3.3 Texture prediction

To obtain texture on the reconstructed 3D mesh described earlier, we employ a texture refinement network similar to Tex-PIFu [33], but with some modifications that enhance texture quality. Firstly, we project the front (reference) and back (synthesized) views onto their corresponding pixel-aligned querry point. This differs from the approach used in PIFu [33], where masking was not necessary due to the absence of additional synthesized views in the monocular case. Subsequently, our completion network, based on UNet [32] and MLP, refines the texture at the per-vertex level using an $\mathcal{L}_1$ loss. To achieve more accurate high-frequency texture prediction, the MLP employed for vertex-level feature refinement is equipped with SIREN activation [37]. For further insights and ablations, please consult the supplementary materials Section C.

4 Experiments

Datasets: We utilized the RenderPeople dataset [4] for our evaluation, comprising 950 3D scans. Out of these, 865 scans were allocated for training, while the remaining 85 were reserved for evaluation purposes. To ensure reproducibility, we will make these data splits available alongside our codebase. The 2D renderings derived from these scans were generated at a yaw interval of 10. Additionally, we employed PyMAFX [48] to obtain the SMPLX fits and their corresponding segmentation maps. For our 2D fashion dataset, we turned to DeepFashion [25]. We specifically used the high-resolution pairs from the original training split in DeepFashion for training, following the approach outlined in our vanilla hallucinator [30].

Training and implementation: Our framework is implemented in PyTorch. The disentangled domain alignement scheme is implemented on top of the vanilla 2D hallucinator codebase [30]. Moreover, the occupancy and texture completion subsystems are built on the official releases containing partial implementations only [34]. Unlike these official implementations, we will make our complete source code publicly available to facilitate further research into this direction. For DDA and occupancy prediction, we follow the same hyperparameter settings from the respective literature [30, 33, 34]. For texture completion, however, we use the Adam optimizer with the learning rate of $1e^{-4}$ to train for 20 epochs. For the 3D dataset, we generate 2D renderings at 10 intervals. All the models are trained on a dedicated `8-GPU NVIDIA RTX A4000` server.

4.1 Evaluation

SOTA comparison: Table 1 shows the comparison with the SOTA methods both in terms of texture and shape. The texture scores are not reported for the

Table 1: Quantitative comparison of texture and shape on RenderPeople [4] test set.

Methods	Texture				Shape		
	FID (↓)	IS (↑)	LPIPS (↓)	SSIM (↑)	Chamfer (↓)	P2S (↓)	Nml MSE (↓)
PIFu [33]	67.6852	3.5190	0.1551	0.9101	1.6631	1.4427	0.0554
PAMIR [52]	70.8481	3.6684	0.1542	0.8963	2.6157	2.0091	0.0738
PIFuHD [34]	-	-	-	-	1.5709	1.4025	0.0526
ICON [44]	-	-	-	-	2.4936	2.0389	0.0680
ECON [43]	-	-	-	-	2.7133	2.4414	0.0742
FAMOUS (Ours)	**55.0711**	**3.8655**	**0.1425**	**0.9140**	**1.4102**	**1.2718**	**0.0371**

methods focusing only on shape (i.e. PIFuHD [34], ICON [44], and ECON [43]).

In terms of texture quality, our method, FAMOUS, outperforms its counterparts in 3 out of 4 metrics, achieving approximately 19% and 6% relative improvement in FID and IS, respectively, over the second-best alternative. For SSIM, PIFu [33] exhibits similar performance compared to ours. It is important to note that SSIM places greater emphasis on smoother predictions with reduced high-frequency variations, and this aligns with the characteristics of PIFu's predictions, as evidenced in Figure 7. In the figure, the sharp, localized variations in both normals and textures from PIFu appear as a subdued version of the ground truth, whereas FAMOUS preserves the intricate details that are captured by deep perceptual metrics, such as FID and IS.

Regarding shape reconstruction, our method outperforms the SOTA methods across the board, with PIFuHD [34] as the second-best contender (Table 1). Note that the recent shape-only approaches, i.e. ICON [44] and ECON [43], (with the publicly available codebases open-sourced by the respective authors), perform worse than PIFuHD. Based on further qualitative investigation, we find these SMPLX model based methods excel others for the more difficult acrobatic poses. Such extreme poses are not present in Fashion images, and PIFuHD is preferred for those cases [43]. On a related note, RenderPeople [4], our target 3D dataset, contains mostly casual poses – somewhat similar to the canonical case, and so, more appropriate for our evaluation than the extreme cases. Nonetheless, we include them for completeness. Please refer to the supplementary Section B for additional visualizations in this regard.

4.2 Ablation Studies

Disentangled domain alignment (DDA): Table 2 shows progressive improvements for the back (occluded) view prediction with our DDA alignments on the pretrained hallucinator (Section 3.1). DDA improves the prediction quality for both finetuning (FT) on the 2D renderings of our 3D dataset and simultaneously training (ST) with both **aligned** 2D fashion images (source) and the 2D renderings (target).

Table 2: Ablation study of the hallucinator on Render People.

Experiments	FID(↓)	LPIPS(↓)	SSIM (↑)	IS (↑)
FT	77.1910	0.1007	0.8645	2.7750
FT+SA	73.0653	0.0935	0.8678	2.7222
FT+SA+PA	67.7944	0.0824	**0.8929**	2.6282
ST	71.0900	0.0935	0.8678	2.7228
ST+SA	77.2460	0.0927	0.8683	2.5015
ST+SA+PA	**66.8911**	**0.0822**	0.8899	**3.2810**

FT ≡ Finetuning; ST ≡ Simultaneous Training
SA ≡ Semantic Alignment; PA ≡ Pose Alignment

For finetuning, one thing to note is that IS keeps degrading as we keep doing more alignment for finetuning (FT → FT+SA → FT+SA+PA) whereas the other 3 metrics gradually improve. This is because the distribution of the generated samples in each stage of alignment gets closer to the distribution of 2D renderings on the test set (gradually lower FID). However, the 3D dataset lacks sufficient number of samples with diverse textures, and so, finetuning with just the corresponding 2D renderings lead to an apparent loss of realism reflected by IS degradation.

For simultaneous training (ST), FID and IS degrade while LPIPS and SSIM improve slightly after incorporating the semantic alignment (SA). This is because the SA images from the source (2D fashion dataset) contains predictions with variable amount of realism – some of which are failed cases. Simultaneously tuning the hallucinator with this unfiltered set of generated samples alongside 2D renderings lead to poor convergence which is the reason behind the high perceptual losses for ST+SA in Table 2. However, the additional discriminator scoring in the beginning of pose alignment helps prevent such convergence issues, thus leading to significant improvements in the final stage (ST+SA *vs.* ST+SA+PA). Also, simultaneously training with the aligned 2D fashion images and the 2D renderings equivalent to Style Alignment (TA) as described in Section 3.1. Thus, the improvement of (ST+SA+PA) over (FT+SA+PA) in Table 2 shows the effectiveness of our complete disentangled alignment. Lastly, SSIM for FT is slightly better than ST due to the potential overfitting of the sufficient statistics on the 2D renderings in case of FT where the support of ST [13] will likely generalize better.

5 Conclusion and Future Work

This paper introduces a complete framework for 3D human digitization from one image, emphasizing the generation of high-fidelity textures. In contrast to SOTA methods, we propose a new approach that harnesses the wealth of 2D datasets to predict arbitrary texture patterns, addressing the scarcity of similar

Fig. 7: Qualitative comparison of the SOTA pipelines predicting both shape and texture with ours on the RenderPeople test set for the images shown on the left. From left to right: Input image, PIFu [33], PaMIR [52], PHORHUM [10], and FAMOUS (ours). For each of these methods, we provide 4 visualizations – front orthographic view (2D rendering), front surface normals, back orthographic view (2D rendering), and back surface normals in order. FAMOUS visually outperforms these SOTA approaches in predicting both the occluded texture and surface delicacies.

3D datasets. To achieve this, we leverage recent advancements in 2D hallucinators, incorporating a gradual domain alignment strategy based on disentangled factors. This integration of information from 2D datasets leads to a significant increase in texture prediction accuracy while also improving the accuracy of shape inference. To the best of our knowledge, this represents the first work to primarily focus on enhancing the quality of arbitrary textures for 3D human digitization by utilizing large-scale 2D datasets.

In terms of limitations and future directions, both 3D texture and shape predictions face challenges when dealing with rare clothing types that are not present in either the 2D/3D datasets. Additionally, texture prediction may encounter difficulties with rare or highly acrobatic poses that fall outside the scope of the training distribution.

References

1. Forensic Science Professor Brings Her Innovative VR Tech to Roger Williams University. https://www.rwu.edu/news/news-archive/forensic-science-professor-brings-her-innovative-vr-tech-rwu 1
2. IGOODI Official. https://www.linkedin.com/pulse/virtual-fitness-training-future-realistic-avatars-igoodi-official 1
3. Readyplayer. https://readyplayer.me 1
4. RenderPeople. https://renderpeople.com/3d-people 3, 5, 6, 11, 12
5. Virtual Fieldwork: Using Virtual Reality (VR) in Anthropology. https://www.unf.edu/dhi/projects/current/virtual-fieldwork-using-virtual-reality-in-anthropology.html 1
6. Walmart Virtual Try-On. https://www.walmart.com/cp/virtual-try-on/4879497 1
7. Weta FX. https://www.wetafx.co.nz/films 1
8. AlBahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., Huang, J.B.: Pose with Style: Detail-preserving pose-guided image synthesis with conditional stylegan. ACM ToG (2021) 5
9. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: CVPR (2019) 4
10. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In: CVPR (2022) 2, 3, 14
11. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. In: CVPR (2023) 6
12. Cao, Y., Chen, G., Han, K., Yang, W., Wong, K.Y.K.: JIFF: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In: CVPR (2022) 4
13. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, USA (2006) 6, 13
14. Dong, Z., Guo, C., Song, J., Chen, X., Geiger, A., Hilliges, O.: PINA: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In: CVPR (2022) 4
15. Gilbert, A., Volino, M., Collomosse, J., Hilton, A.: Volumetric performance capture from minimal camera viewpoints. In: ECCV (2018) 4
16. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P., Izadi, S.: The relightables: Volumetric performance capture of humans with realistic relighting. ACM ToG (2019) 1
17. Han, S.H., Park, M.G., Yoon, J.H., Kang, J.M., Park, Y.J., Jeon, H.G.: High-fidelity 3d human digitization from single 2k resolution images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3
18. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In: NeurIPS (2020) 4
19. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: ARCH++: Animation-ready clothed human reconstruction revisited. In: ICCV (2021) 4
20. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Luo, L., Ma, C., Li, H.: Deep volumetric video from very sparse multi-view performance capture. In: ECCV (2018) 4

21. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: ARCH: Animatable reconstruction of clothed humans. In: CVPR (2020) 4
22. Li, Y., Huang, C., Loy, C.C.: Dense intrinsic appearance flow for human pose transfer. In: CVPR (2019) 5
23. Li, Z., Yu, T., Zheng, Z., Liu, Y.: Robust and accurate 3d self-portraits in seconds. IEEE TPAMI (2022) 4
24. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV (2019) 5
25. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016) 3, 6, 7, 11
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graph. **34**(6) (2015) 2, 4
27. Men, Y., Mao, Y., Jiang, Y., Ma, W.Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed gan. In: CVPR (2020) 5
28. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019) 2, 9
29. Prabhudesai, M., Lal, S., Patil, D., Tung, H.Y., Harley, A.W., Fragkiadaki, K.: Disentangling 3d prototypical networks for few-shot concept learning. In: ICLR (2021) 6
30. Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. In: CVPR (2022) 3, 5, 6, 7, 9, 11
31. Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: CVPR (2020) 5
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015) 11
33. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV (2019) 2, 3, 4, 5, 10, 11, 12, 14
34. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: CVPR (2020) 1, 2, 4, 10, 11, 12
35. Sarkar, K., Golyanik, V., Liu, L., Theobalt, C.: Style and pose control for image synthesis of humans from a single monocular view (2021) 5
36. Sieberth, T., Dobay, A., Affolter, R., Ebert, L.C.: Applying virtual reality in forensics – a virtual scene walkthrough. Forensic Science, Medicine and Pathology (2019) 1
37. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: NeurIPS (2020) 11
38. Smith, D., Loper, M., Hu, X., Mavroidis, P., Romero, J.: Facsimile: Fast and accurate scans from an image in less than a second. In: ICCV (2019) 4
39. Song, D.Y., , Lee, H., Seo, J., Cho, D.: Difu: Depth-guided implicit function for clothed human reconstruction (2023) 3
40. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV Workshops (2016) 6
41. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. CoRR **abs/1412.3474** (2014) 6
42. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Trans. Graph. (2008) 4

43. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: ECON: Explicit clothed humans optimized via normal integration. In: CVPR (2023) 2, 12
44. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: CVPR (2022) 2, 4, 12
45. Yang, Z., Wang, S., Manivasagam, S., Huang, Z., Ma, W.C., Yan, X., Yumer, E., Urtasun, R.: S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In: CVPR (2021) 4
46. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: CVPR (2021) 3
47. Zakharkin, I., Mazur, K., Grigorev, A., Lempitsky, V.: Point-based modeling of human clothing. In: ICCV (2021) 4
48. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: PyMAF-X: Towards well-aligned full-body model regression from monocular images. IEEE TPAMI (2023) 9, 11
49. Zhang, J., Li, K., Lai, Y.K., Yang, J.: Pise: Person image synthesis and editing with decoupled gan. In: CVPR (2021) 5
50. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: CVPR (2020) 5
51. Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., Liu, Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In: ICCV (2021) 4
52. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE TPAMI **44**(6) (2022) 2, 4, 5, 12, 14
53. Zhou, X., Yin, M., Chen, X., Sun, L., Gao, C., Li, Q.: Cross attention based style distribution for controllable person image synthesis. In: ECCV (2022) 5
54. Zuo, X., Du, C., Wang, S., Zheng, J., Yang, R.: Interactive visual hull refinement for specular and transparent object surface reconstruction. In: ICCV (2015) 4