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

Fig. 1: Qualitative comparison of the SOTA pipelines predicting both shape and texture with ours on the DeepFashion test set for the images shown on the top. From top to bottom: Input image, PIFu [4], PaMIR [12], PHORHUM [1], and FAMOUS (ours). For each of these methods, we provide front-back slanted visualizations in order. FAMOUS visually outperforms these SOTA approaches in the occluded texture.

1 Distributionally Aligned Hallucinator (DAH)

1.1 Implementation Details

We follow the original training settings of the vanilla hallucinator [3] in this work. However, this method was originally designed for the pose-guided view synthesis

problem, where both the image and target OpenPose COCO skeleton are inputs. Therefore, at inference time, to obtain target SMPLX segmentation and the target silhouette, we had to fit SMPLX using [11] and segment the binary mask from the input image, respectively. The data flow used in each stage of alignment during the training phase is shown in Figure 4. For semantic alignment, we use the original deep fashion images as the image input. For guidance, we sample a subset of GT OpenPose COCO keypoints that represent full-body semantics provided in the original dataset. Using these sampled full body guidance, we generate front/back image pairs $\{(I_f^{\mathcal{SA}}, I_b^{\mathcal{SA}})\}$ as shown in Figure 4(c). These image pairs are then used to retrain the model in a self-supervised manner.

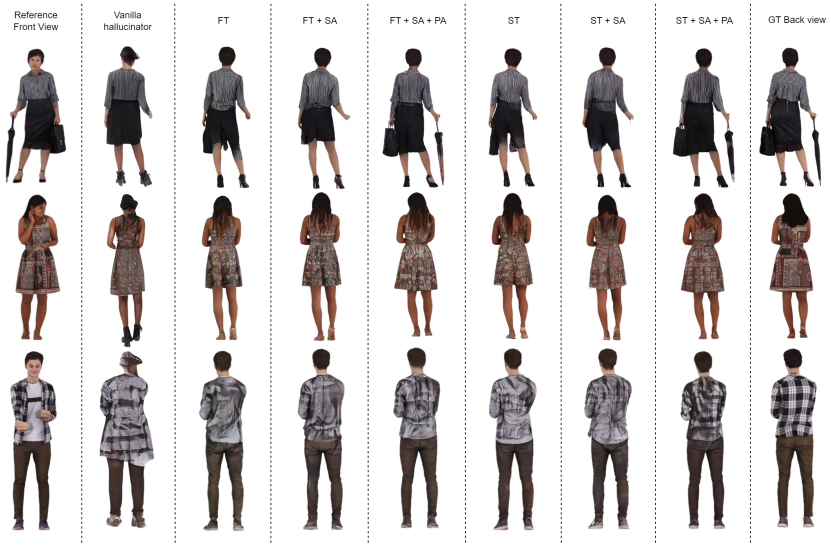

Fig. 2: Qualitative ablation of the disentangled distributional alignment for the 2D hallucinator. From left to right: Input image; Naive finetuning; Semantic Alignment with finetuning; Semantic and Pose Alignment with finetuning; Simultaneous training; semantic alignment with simultaneous training; semantic and pose alignment with simultaneous training, GT back view on target RenderPeople dataset.

Following semantic alignment, we transfer the weights of the network to the pose alignment stage. During this stage, we use the collection of semantically aligned image pairs $\{(I_f^{\mathcal{SA}}, I_b^{\mathcal{SA}})\}$ as image input. For guidance, we sample a subset of GT OpenPose COCO keypoints from the target dataset, in our case, the RenderPeople dataset. The subset is collected based on a thresholding process using discriminator scores. For every guidance sample, we randomly select a few images from $\{(I_f^{\mathcal{SA}}, I_b^{\mathcal{SA}})\}$ and pass them through the network to obtain the corresponding outputs. Then we obtain the discriminator score (from vanilla discriminator checkpoints after vanilla hallucinator training [3]) of these outputs and threshold them to collect the subset. The threshold value is the median of

discriminator scores obtained on the test set of the source deep fashion images. Figure 3 shows the histogram of the discriminator scores and its median, denoted by a dotted red line.

Please note that the goal of this thresholding process is to maintain a high degree of realism and mitigate the artifacts generated in these image pairs $\{(I_f, I_b)\}$. Using this subset of guidance, we generate front/back image pairs $\{(I_f^{\mathcal{PA}}, I_b^{\mathcal{PA}})\}$. Finally, we finetune the semantically aligned hallucinator with these pose-aligned pseudo pairs in a similar way described in the previous stage.

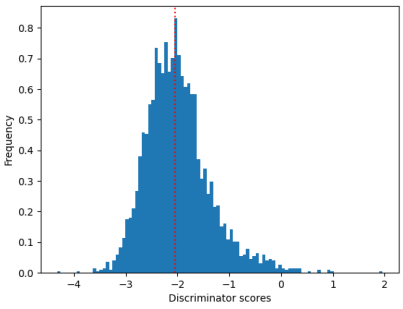

Fig. 3: Histogram of discriminator scores obtained on the source deep fashion test set. The red dotted line denotes the median value of the distribution.

On a related note, we empirically find that off-the-shelf hallucinators often struggle to generate views for highly challenging poses. Therefore, it is advantageous to filter out these poses from our training set during pose alignment. If the target dataset includes poses that are more challenging than those in the 2D source dataset, we hypothesize that the information flow from our aligned 2D prior (in-domain to 3D) to the 3D target space will suffice for reliable textured reconstruction.

For style alignment, we checkpoint the network with weights after finetuning in the previous stage. At the beginning of this stage, we switch guidance to the SMPLX segmentation map and silhoutte. For the finetuning experiments, only the weights are transferred from the previous stages. But for simultaneous training, both weights and the pose-aligned generated image pairs are utilized.

Table 1: Ablation study of the hallucinator on Render People.

Experiments	FID(↓)	LPIPS(↓)	SSIM (↑)	IS (↑)
ST	71.0900	0.0935	0.8678	2.7228
ST + Coral	72.5376	0.0973	0.8652	2.5166
ST + MMD	73.0580	0.0968	0.8651	2.5433

ST ≡ Simultaneous Training

1.2 Experiment Details

Effects of MMD and Coral: Domain adaptation methods like Maximum Mean Discrepancy [8] and Coral [7] aim to reduce the distributional differences between the source and target datasets. But blindly minimizing this difference

in the case of a generative model tends to confuse the network. Table 1 shows the quantitative ablation study for these experiments.

Effects of each alignment: An image can be represented based on semantics, pose, view, and style [2]. For a conditional generative model like NTED [3], we find semantics and pose distributional differences between two datasets play a crucial role in its alignment. Naively fine-tuning the vanilla hallucinator on the target dataset fails to generalize well due to these distributional differences between the datasets. Figure 2 shows the qualitative ablation for each stage of alignment.

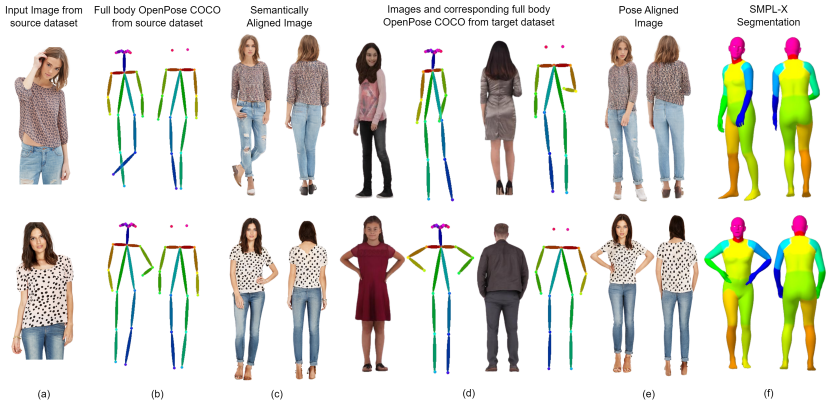

Fig. 4: A few data samples from the source DeepFashion dataset and corresponding data generated during each stage of alignment with the respective guidance used in the previous stage.

2 Occupancy Network

2.1 Experiments

Shape Analysis of ECON and ICON: ICON and ECON take an RBG image and an estimated SMPL-X fit as input. However, recovering SMPL-X fit from a single image is still an unsolved problem, and errors in estimating SMPL-X fit are propagated to ECON and ICON. Furthermore, for fashion images, like in the renderpeople dataset, ICON tends to overfit to the SMPL-X mesh, as shown in 7. ECON suffers from stitching artifacts and depth ambiguity compared to PIFu-HD [9]. So, we empirically find PIFu-HD to be the best occupancy network for fashion images and build our occupancy network on top of it. Figure 5 shows the superior performance of the modified PIFu-HD implemented in our framework.

2.2 Evaluation metrics

We primarily evaluate FAMOUS with SOTA methods that predict surface colors: PIFu, PaMIR, and PHOROUM. We quantitatively evaluate texture quality

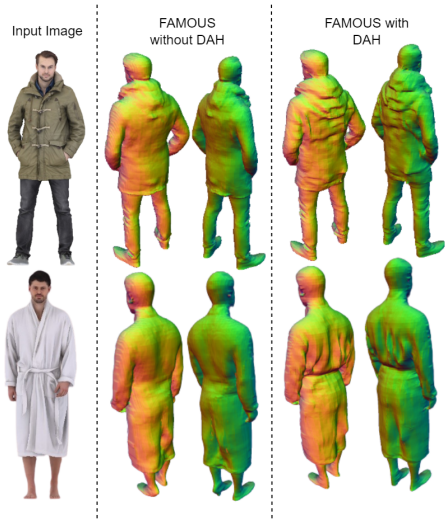

Fig. 5: Qualitative ablation of shape prediction. From left to right: Input image; The front/back surface normals from our framework without (middle) and with (right) the aligned hallucinator prediction DAH. Note the high-precision details recovered in the estimated normals after using aligned hallucinator prediction is compared to the vanilla case in the middle.

by rendering the predicted textured mesh with the respective camera model by rotating the camera by [0, 90, 180, 270] yaw angle and comparing it with corresponding GT renderings using image reconstruction metrics. For image reconstruction metrics, we report Learned Perceptual Image Patch Similarity (LPIPS), Inception Score (IS), structural similarity index (SSIM) and Frechet Inception Distance (FID). For evaluating the reconstruction quality of geometry, we measure the difference between the reconstructed and GT meshes using Chamfer distance and Point to Surface (P2S). Additionally, we render the normal images between the predicted and GT mesh in the same way as colored images and report the MSE error. Note that all the evaluations are performed on the original predicted meshes from the respective models. But for visualizations, the artifacts from predicted meshes are removed. Only the biggest connected, water-tight mesh is kept.

3 Texture Completion

3.1 Implementation Details

The backbone for the texture completion part is based on UNet Architecture 8 encoder-decoder layers with bilinear interpolation in between the filters. The first layer contains 64 channels, and then the channel size is doubled on each layer of the encoder. The decoder reduces the channel size by half on each layer.

Table 2: Ablation study of the Texture Completion on Render People.

Experiments	FID(↓)	LPIPS(↓)	SSIM (↑)	IS (↑)
Tex-PIFu*	103.4353	0.1655	0.8953	3.4092
Tex-PIFu* + SIREN	78.3775	0.1593	0.9022	3.7334
FAMOUS	55.0711	0.1425	0.9140	3.8655

FAMOUS ≡ Tex-PIFu* with SIREN and UNet backbone
Tex-PIFu* ≡ modified Tex-PIFu with predicted back view from DAH

The final layer contains 128 channels. We use Leaky-ReLU activation for both the encoder and decoder. Once we extract features from both the front view and the generated back view, we project them onto the queried points in a pixel-aligned manner. In comparison to Tex-PIFu, where the same pixel features from the encoders are used for color prediction of multiple 3D points along the line of projection (i.e. pixel-aligned). We only use pixel features of the corresponding view in each side. The MLP for predicting per-vertex color is similar to the implementation in Tex-PIFu except for activation at the last layer, which was changed to sine activation inspired by [6]. Table 2 shows the quantitative ablation study for the texture completion part.

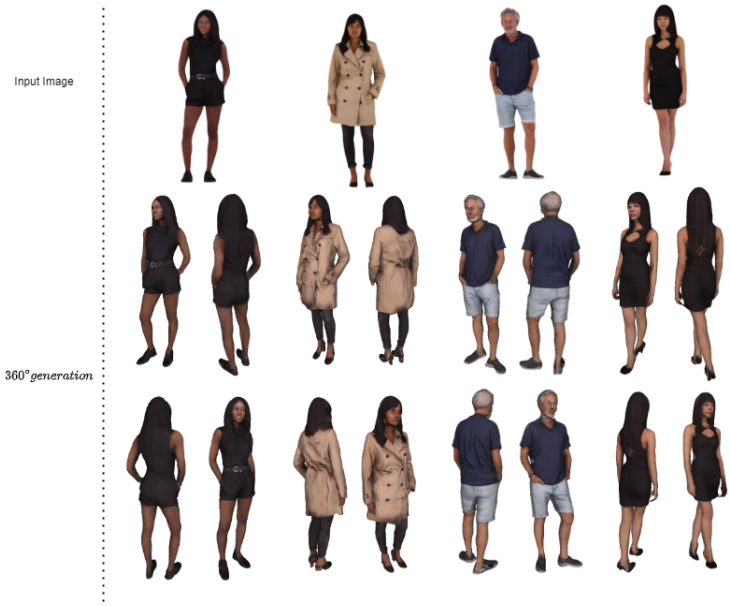

Fig. 6: Additional 360° visualization of our results on the RenderPeople test set.

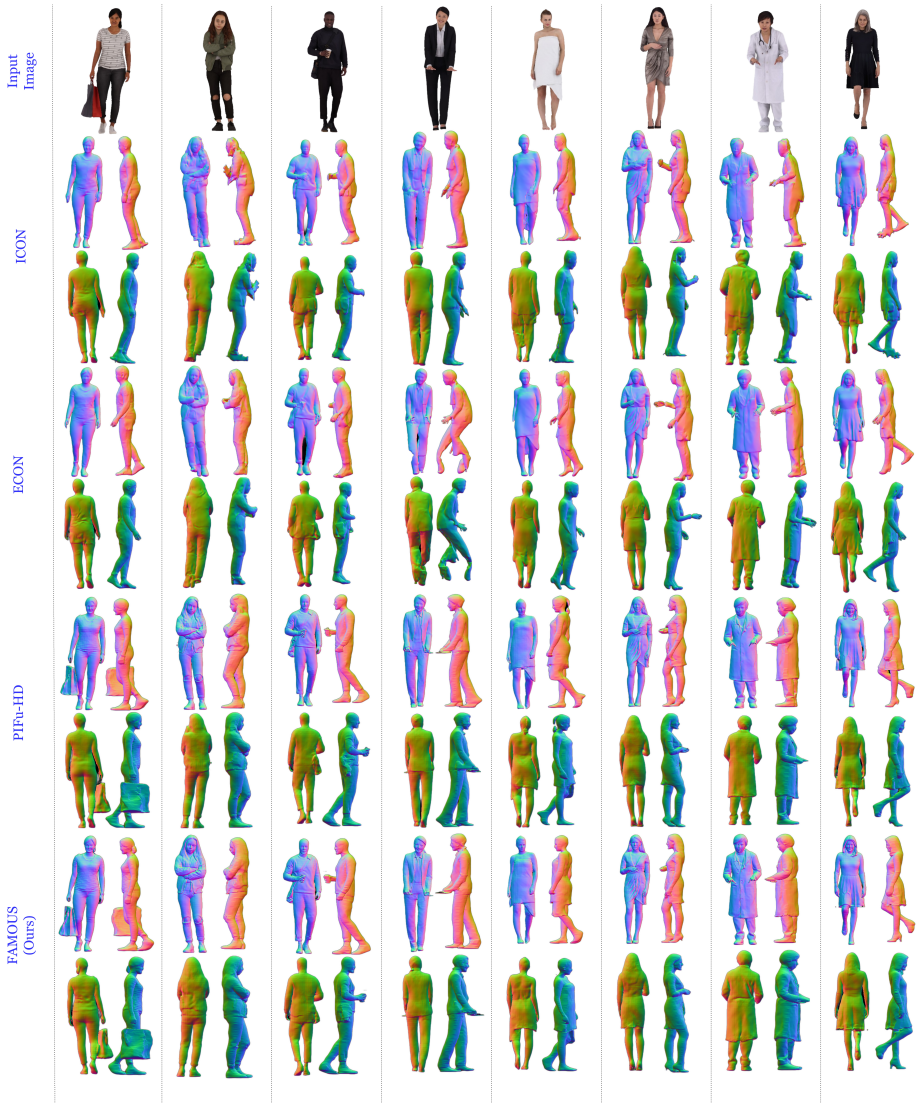

Fig. 7: Qualitative comparison of ICON, ECON and PIFu-HD predicting shape with ours on the RenderPeople test set for the images shown on the top. From top to bottom: Input image, ICON [10], ECON [9], PIFu-HD [5], and FAMOUS (ours). For each of these methods, we provide four visualizations: surface normals at yaw angle 0°, 90°, 180° and 270°. FAMOUS visually outperforms these SOTA approaches in predicting surface details.

References

1. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In: CVPR (2022) 1
2. 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) 4
3. Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. In: CVPR (2022) 1, 2, 4
4. 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) 1
5. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: CVPR (2020) 7
6. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: NeurIPS (2020) 6
7. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV Workshops (2016) 3
8. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. CoRR **abs/1412.3474** (2014) 3
9. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: ECON: Explicit clothed humans optimized via normal integration. In: CVPR (2023) 4, 7
10. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: CVPR (2022) 7
11. 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) 2
12. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE TPAMI **44**(6) (2022) 1