UnitedHuman: Harnessing Multi-Source Data for High-Resolution Human Generation

Supplementary Material

Jianglin Fu[1*], Shikai Li[1*], Yuming Jiang[2], Kwan-Yee Lin[1,3], Wayne Wu[1†], Ziwei Liu[2†]
[1] Shanghai AI Laboratory [2] S-Lab, Nanyang Technological University [3] CUHK
{fujianglin, lishikai}@pjlab.org.cn
{yumingj80, wuwenyan0503}@gmail.com ziwei.liu@ntu.edu.sg junyilin@cuhk.edu.hk

In this supplement, we first show how we construct the synthetic hand dataset (Sec. A). Then we provide the experimental settings for the Multi-source Spatial Transformer (MST), Continuous GAN, as well as the overall training configuration (Sec. B). Additionally, we provide more qualitative results of ablation studies (Sec. C). We also conduct a user study to evaluate the performance of our method and the other three SOTA methods (Sec. D). Finally, an analysis in terms of the poses and garments diversity in the generated images is given in Sec. E.

A. Synthetic Hand Dataset

Due to the lack of high-quality hand datasets for generative tasks, we construct a new dataset based on the DART dataset and SMPL-X model. Specifically, we first randomly sample $7,000$ images from the full-body SHHQ dataset for SMPL-X estimation, where the estimated parameters include hand pose and position. Then, we apply the sampled DART texture to the palm and arm, remove the mesh except for the arm, and introduce a photo studio HDR as the ambient light. The Eevee renderer [2] is used to render the hand images in the full-body image space with the average camera α of SHHQ. As shown in Fig. A1, we finally composite the synthesized hand with an image patch from the DeepFashion-HR dataset.

Figure A1: Synthetic Hand Dataset. We generate the training images by compositing the hand images rendered from the DART dataset with image patches from DeepFashion-HR datasets.

B. Experiment settings

In this section, we provide the experiment settings of our framework. Multi-source Spatial Transformer aims to transform the partial-body image into the full-body image space by estimating the camera parameters of SMPL. We divide the objective function of SMPLify-p into three parts as:

$$E(\alpha_{opt}) = \lambda_v L_{vis} + \lambda_i L_{invis} + \lambda_r L_{reg} \qquad (1)$$

where λ_v, λ_i, λ_r is 1.0, 10.0 and 10.0 respectively. The loss L_{invis} is based on the variational autoencoder (VAE) trained on full-body SHHQ dataset. In particular, we follow the architecture of VPoser [6] that contains a specific decoder for continuous rotation representation. The loss function of VAE consists of reconstruction loss on vertices and KL divergence. We use the pose parameters estimated by PARE [5] as ground-truth for training.

	SG-Human	InsetGAN*	AnyRes	Ours
Batch Size	32	/	16	32
G_Learning rate	0.002	/	0.0025	0.0025
D_Learning rate	0.002	/	0.002	0.002
Total #Params. (M)	30 / 29	/	29 / 31	28 / 55
Train (GPU Days)	7	/	23	43
Test (GPU Sec.)	0.48	200	2.6	2.8

Table A1: Training details of experiments. InsetGAN* is an optimization-based method and generates results without retraining.

As for the Continuous GAN, we follow the training setting of the original StyleGAN3-T except for the R1 gamma, which is set to 2 in our experiments. Both *Stage 1* and *Stage 2* training use a batch size of 32 and are trained on 8 GPUs. Training configurations for all the compared methods can be found in Tab. A1.

with MST w/o MST

Figure A2: Visualization results of ablation on Multi-source Spatial Transformer.

C. Additional qualitative results

Ablation on multi-source spatial transformer. Fig. A2 displays the partial-body images that were transformed into the full-body image space, with and without Multi-source Spatial Transformer (MST). In the right column, the reconstructed parametric mesh does not match the reference image when MST is not used. In contrast, the spatial distribution of images transformed by MST is closer to that of the full-body dataset.

Ablation on multi-source datasets. Fig. A3 demonstrates the improvement of local details as using more multi-source partial datasets. The full-body images are 1024px, while the local patches are cropped from the 2048px images. When only trained on SHHQ, our model is capable of generating coherent full-body images with a resolution of 1024 pixels. However, at higher scales, the generated patches are blurry and have artifacts. With SHHQ^{SR} and DF$_p$ datasets, both the face and hand patches are clearer than before. By adding the face dataset CelebA, more details such as the illumination on the face are captured. Finally, with the addition of a constructed synthetic hand dataset, the details of hand patches are more accurate.

Ablation on alignment strategy. We provide visual comparisons (Fig. A4) of the ablation study conducted on different human alignment strategies. The top row of the figure contains the images generated by the mean latent of each model, and the images in the bottom row are generated by a random latent. It can be observed from the figure that the visual outcomes are consistent with the kFIDs reported in the paper. Aligning humans using only 2D keypoints yields superior face and hands details compared to using an auxiliary "pose-mapping" MLP network. Nonetheless, our proposed approach outperforms these two methods in producing finer details at a higher resolution.

D. User Study

We conduct a user study to evaluate the clarity and realism of the results generated by our method in comparison to three SOTA methods: StyleGAN-Human [4], InsetGAN [3] and AnyRes [1]. Our user study involves a total of 12 volunteers. We randomly selected 10 high-resolution (2048px) images from each model and extracted the face or hand regions from those images as well. During the user study, the participants were presented with four images at a time, along with the corresponding cropped patches. Every participant was asked to evaluate the authenticity and sharpness of both the full images and their cropped parts. One of the example appeared in our questionnaire is displayed in Fig. A5. Based on global realism, global clarity, and local realism, the images generated by our method received over 82% of the votes, while the clarity of the local patches was slightly lacking, receiving a score of 78%. Nonetheless, compared to the other three SOTA methods, our approach is still far ahead in visual quality. More detailed results are shown in Tab. A2.

	SG-Human	InsetGAN	AnyRes	Ours
F-Realism	13.33	0.83	0.83	85
F-Clarity	8.33	3.33	0	88.33
P-Realism	6.66	9.17	1.67	82.5
P-Clarity	2.5	17.5	1.67	78.33

Table A2: Voting scores for images from both SOTAs and our model, which are evaluated based on four criteria: the full-image's realism and clarity, as well as the cropped regions' realism and clarity. The table uses "F-" and "P-" to denote the full images and cropped patches, respectively. All the numbers in the table are presented in percentages (%). Higher values indicate better realism or clarity.

E. More results

The showcased human images demonstrate neutral poses and relatively less diverse garments. In response to this situation, we conducted an additional experiment using $10K$ SHHQ images on StyleGAN-Human to confirm that the limited diversity of pose and clothing is mainly due to the training dataset (see Fig. A6). Specifically, the majority of the poses and garments in our model are sourced from the $10K$ SHHQ and $10K$ DeepFashion images, while StyleGAN-Human utilizes re-balancing techniques and a more extensive dataset of $230K$ images.

Figure A3: Visualization of ablation on datasets. From left to right, the visualizations are the generated results after continuously adding datasets from different sources to the model.

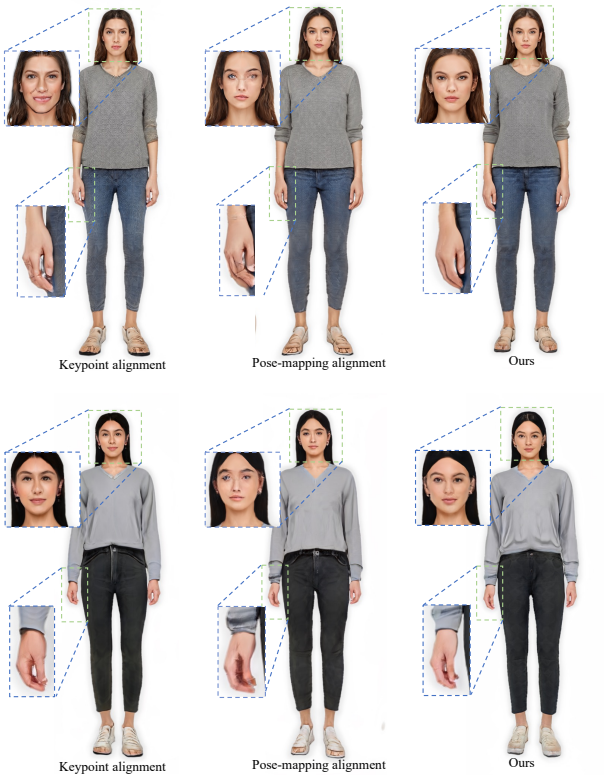

Figure A4: Visualization of ablation on alignment strategy.

Please pick the preferred image among four full-body images and four patches from the following perspectives:

	A	B	C	D
F-Realism:				
F-Clarity:				
P-Realism:				
P-Clarity:				

Figure A5: This is an example of the questionnaire used in our user study.

References

[1] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *ECCV*, 2022. 2

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

Figure A6: Left: StyleGAN-Human (10K). Right: Ours (10K). Experiments show that the constrained diversity of garments and poses is attributed to the training data.

[3] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. InsetGAN for full-body image generation. In *CVPR*, 2022. 2

[4] 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. In *ECCV*, 2022. 2

[5] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 1

[6] 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. In *CVPR*, 2019. 1