Supplementary Material:
DiffBody: Diffusion-based Pose and Shape Editing of Human Images

Yuta Okuyama, Yuki Endo, and Yoshihiro Kanamori
University of Tsukuba
okuyama.yuta.sw@alumni.tsukuba.ac.jp, {endo, kanamori}@cs.tsukuba.ac.jp

A. Implementation Details

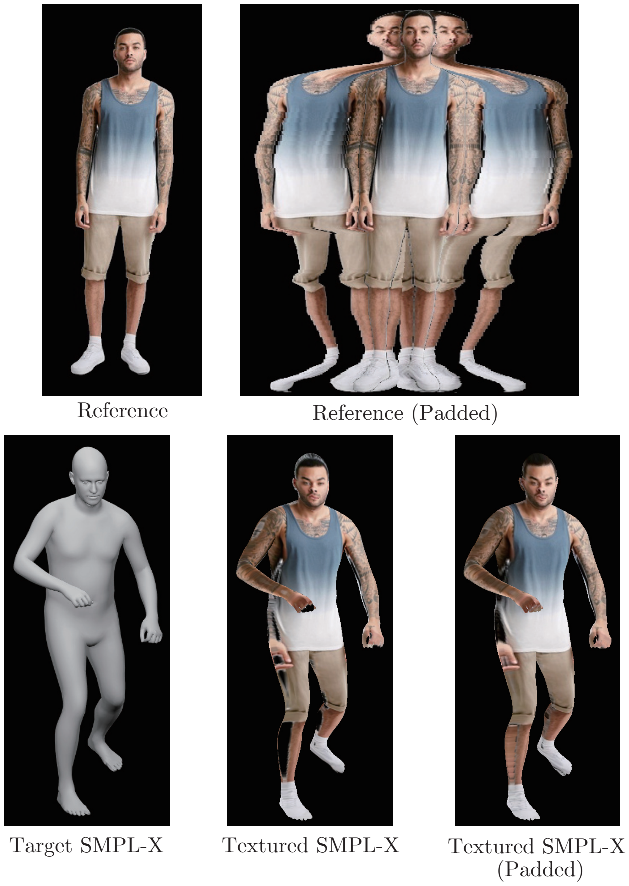
A.1. Initial textured 3D body construction

Projective texture mapping. We explain how to construct an initial textured body model with the desired pose and body shape. First, we fit the SMPL-X model [1] to the reference image using an existing method [2]. For the reference person image and the fitted SMPL-X model, we assign UV coordinates to corresponding vertices via projective texture mapping. We then change the pose and shape of the SMPL-X model to obtain an initial textured 3D body model.

Horizontal reflection padding. Naïvely applying projective texture mapping yields visual artifacts, particularly around the body's silhouette, due to slight misalignment. For example, the black background color appears around the right hand and right leg in the example of Figure 1, lower-middle. As a simple remedy for this, we apply horizontal reflection padding to the original reference image using a binary mask; for each scanline from slightly inside the mask, we copy pixel values at the mirror-symmetric positions about the mask boundary (Figure 1, upper-right). This approach is not a perfect solution but is sufficient to avoid copying the background color (Figure 1, lower-right).

A.2. Loss functions

Here we describe the details of loss functions used in Steps 1 and 2.


Reference

Reference (Padded)

Target SMPL-X

Textured SMPL-X

Textured SMPL-X (Padded)

Figure 1. Horizontal reflection padding is applied to the reference image to prevent the background color from showing up in the texture-projected human model.

Step 1: Fullbody refinement. For the refinement of a fullbody image, we use the Adaptive Wing (AW) loss [3] $\mathcal{L}_{AW}$ and CLIP similarity [4] loss $\mathcal{L}_{CLIP}$. The AW loss $\mathcal{L}_{AW}$ is the adaptive wing loss [3] defined between the joint heatmaps estimated using OpenPose [5] for the output and rendered SMPL-X images. Our heatmap resolution is 128×128. CLIP similarity loss $\mathcal{L}_{CLIP}$ is defined between the output and reference images for each part [6] based on SMPL-X labeling as follows:

$$\mathcal{L}_{\text{CLIP}} = \sum_{p}^{l} \phi(I^{p}_{\text{ref}})^{T} \phi(I^{p}_{\text{out}}), \qquad (1)$$

where l is the number of body part labels of SMPL-X, I_{ref} and I_{out} are the body parts cropped from the reference and

Table 1. Adjectives describing body shape corresponding to BMI.

BMI	adjective
≤ 15.0	"skinny"
≤ 18.5	"under weight"
≤ 25.0	
≤ 30.0	"overweight"
> 30.0	"fat"

output images. ϕ is the normalized embedding function of the CLIP. The total loss function for fullbody refinement is:

$$\mathcal{L}_{Fullbody} = \lambda_{AW}\mathcal{L}_{AW} + \lambda_{CLIP}\mathcal{L}_{CLIP}, \qquad (2)$$

where $\lambda_{Pose} = 0.002$ and $\lambda_{CLIP} = 2$ are the weights.

Step 2: Facial refinement. To optimize the text embedding for refining a face, we use the identity loss using MagFace [7], the keypoint loss using RetinaFace [8], and the CLIP similarity [4]. The keypoint loss $\mathcal{L}_{keypoint}$ is defined as MSE loss between the face keypoints estimated using RefinaFace [8] for the output and rendered SMPL-X images. Unlike fullbody refinement, we simply measure the CLIP similarity between the reference and the output face image. The total loss function for the facial refinement is:

$$\mathcal{L}_{Face} = \lambda_{ID}\mathcal{L}_{ID} + \lambda_{CLIP}\mathcal{L}_{CLIP} + \lambda_{Keypoint}\mathcal{L}_{Keypoint}, \qquad (3)$$

where $\lambda_{Keypoint} = 0.1$, $\lambda_{CLIP} = 10$, and $\lambda_{ID} = 10$ are the weights. When we edit the body shape, we halve λ_{CLIP} and λ_{ID} to tolerate changes in facial features.

A.3. Text prompt

We describe the details of prompts used for conditioning on our refinement module. Our prompts contain "sks," a special token used for text-to-image personalization by DreamBooth [9]. Our method associates this token with a reference person. In addition, our prompts contain information on a target face orientation, such as "facing left." We used the face detection API of Face++ [10] to obtain the face orientation, which is automatically reflected in the prompts. For body shape editing, we use adjectives describing the body shape according to BMI calculated from the input height and weight (see Table 1). For example, when the target model faces to the left with a fat body, we use a prompt "photo of a fat sks man facing left" in Step 1. In Step 2, we use "face" instead of "man".

A.4. Refinement mask

We describe how to create a refinement mask, which indicates areas to be refined. In Step 1, we compute a mask consisting of invisible areas in a reference person image. To do so, we first emit a ray to each triangle's centroid in a SMPL-X [1] mesh from the viewpoint for texture projection. Next, we assign an "invisible" label to the triangles that the rays do not hit. After editing the pose and body shape of the SMPL-X model, we render the edited model to obtain a mask according to the labeled areas. In Step 2, we compute a mask by assigning 0 to pixels within 20% of the mask width from its boundaries, measured using the Manhattan distance, and 1 to the remaining pixels.

B. Additional Results

We show the additional results that are not included in the main paper due to the page limitation. The reference images of the following results were obtained from DeepFashion [11], MonoPerfCap [12], Everybody Dance Now [13], and EHF [1].

B.1. Qualitative evaluation

B.1.1 Evaluation of body shape editing

We conducted a qualitative comparison with the state-of-the-art body shape editing method by Ren et al. [14] in the same way as their paper. Figure 2 shows the results. In the results of their method, increasing the body size often causes significant distortion in the torso. In contrast, our method can create plausible images. Our method can also handle facial appearance changes that occur along with the body weight changes.

B.1.2 Evaluation of pose and body shape editing

Figure 3 shows our unprecedented results in which both poses and body shapes were edited at the same time. Such simultaneous edits have been infeasible with existing methods, to the best of our knowledge. The results demonstrate that our method can edit the target pose and body shape simultaneously while maintaining the subject's identity in terms of clothing and facial features.

B.2. Ablation Study

B.2.1 Facial degradation by VAE

In the LDM used in our method, the face quality is degraded by simply reconstructing the input image with VAE. This is because the VAE used in the LDM cannot reconstruct relatively small faces accurately from low-dimensional latent maps. An example of the degradation is shown in Figure 4, and the quantitative evaluation metrics are shown in Table 2. These results warrant our approach that extracts a face region and refines it separately.

B.2.2 Refinement with weak noise

To find an appropriate noise intensity for our iterative refinement, we experimented with single iterations of refinement

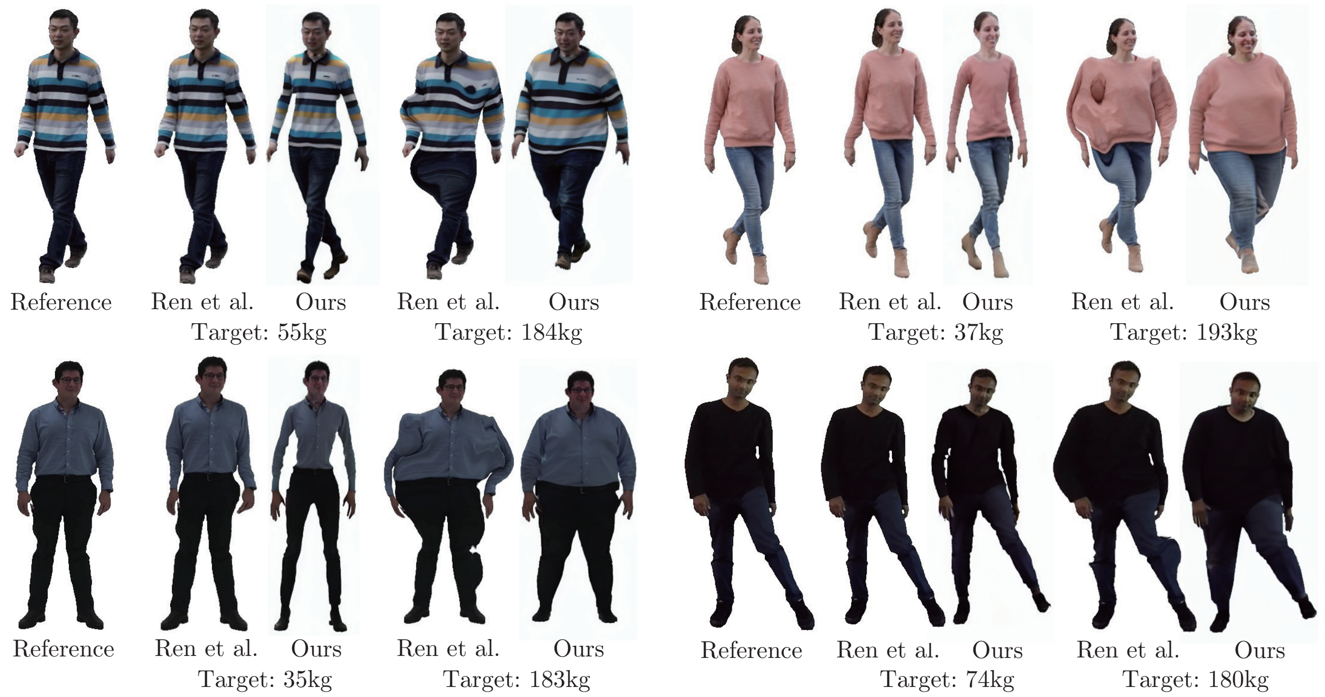


Figure 2. Qualitative comparison of editing the reference images to the target weights with the method by Ren et al. [14] and ours.

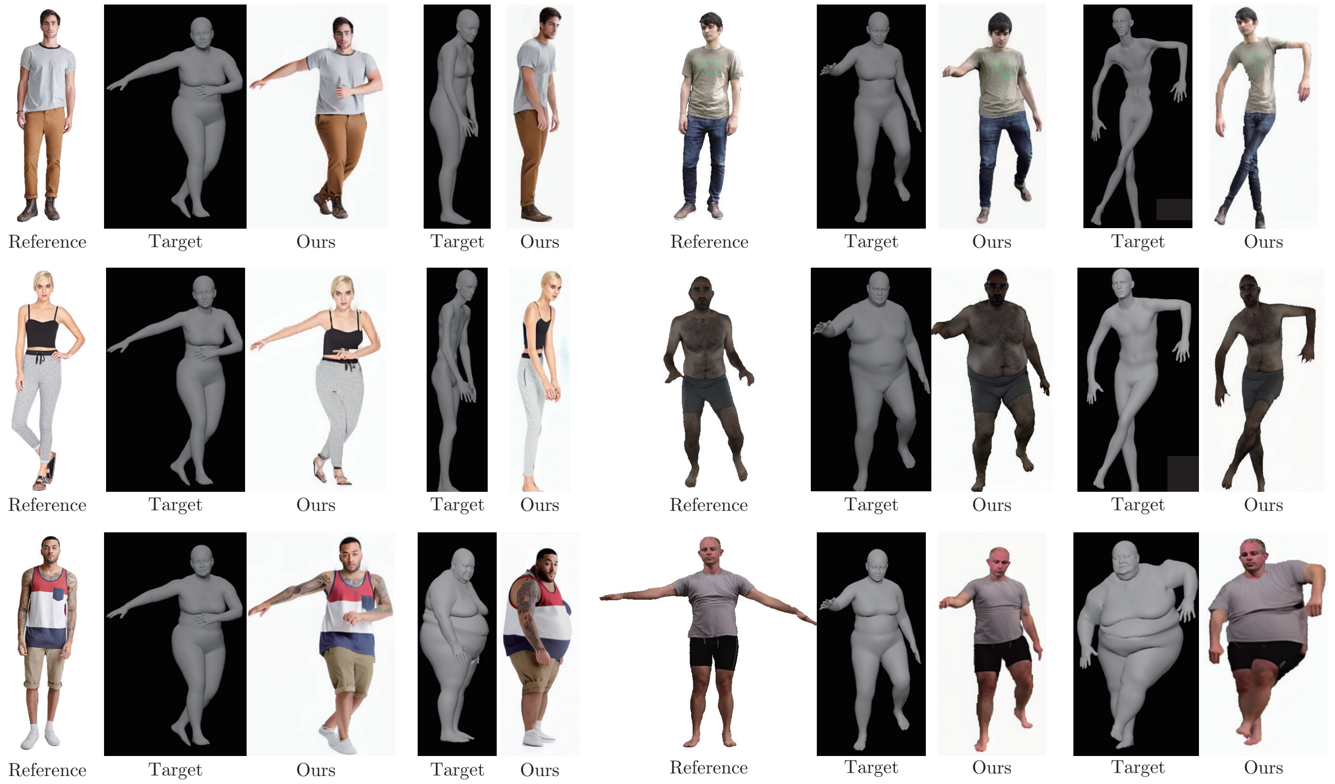
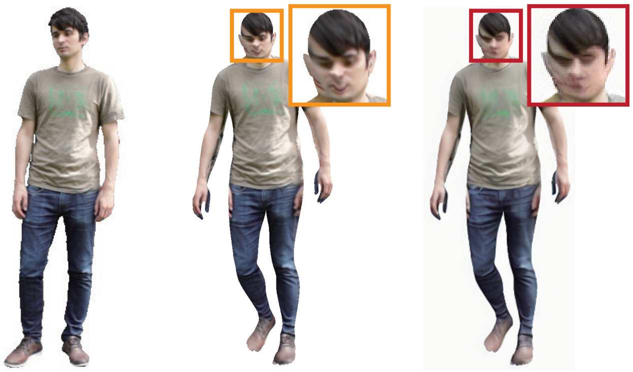


Figure 3. Qualitative comparison of simultaneous edits of both target's poses and body shapes in reference images using our method. The edited results are plausible with identity preservation in terms of clothing and facial features.



Reference Input Reconstruction

Figure 4. Qualitative comparison of decreased facial quality when simply reconstructing images using VAE.

Table 2. Quantitative evaluation when images are simply reconstructed with VAE.

	SSIM ↑	LPIPS ↓	FID ↓	ID ↓
Input	0.714	0.243	**55.862**	**0.232**
Reconstruction	**0.716**	**0.242**	65.484	0.375

Table 3. Quantitative evaluation metrics for a single refinement with varying noise intensities.

	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
10%	19.613	0.717	0.246	59.221
20%	19.650	**0.718**	0.245	57.808
30%	**19.669**	**0.718**	**0.244**	56.093
40%	**19.669**	**0.718**	**0.244**	56.084
50%	19.561	0.716	0.244	**50.748**
60%	19.582	0.714	0.245	52.582
70%	19.519	0.712	0.246	51.816
80%	19.433	0.710	0.248	51.592
90%	19.356	0.708	0.250	52.011

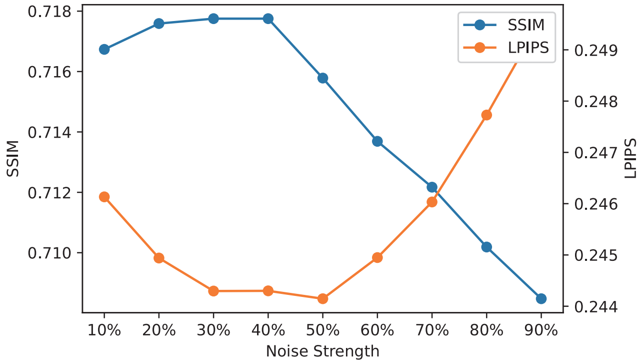


Figure 5. Graphs depicting the variations of SSIM and LPIPS scores with varying noise intensities.

with different noise levels. We increased the noise level from 10% to 90% in increments of 10 percentage points. Table 3 summarizes the qualitative evaluation, and Figure 5 shows graphs of SSIM and LPIPS with varying noise intensities. These results revealed that weaker noise tends to yield better results regarding pixel-level metrics such as PSNR and SSIM because weaker noise preserves the projected textures as they are. On the other hand, for more perceptual metrics such as LPIPS and FID, the values tend to be optimal around 30% to 50% noise intensity, with performance degrading as the noise intensity deviates from this range. This pattern suggests that, around 30% to 40% noise intensity, we can effectively correct unnatural areas while preserving the texture of the input image. Consequently, we conclude that weak noise intensities from 30% to 40% seem effective for refinement. In our approach, we choose 30% noise intensity to balance computational efficiency while maintaining effectiveness.

References

[1] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from

a single image," in *CVPR*, 2019, pp. 10 975–10 985. 1, 2

[2] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop," in *ICCV*, 2021, pp. 11 446–11 456. 1

[3] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *ICCV*, 2019, pp. 6971–6981. 1

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, vol. 139, 2021, pp. 8748–8763. 1, 2

[5] Z. Cao, G. Hidalgo, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *TPAMI*, vol. 43, no. 1, pp. 172–186, 2021. 1

[6] Y. Huang, H. Yi, W. Liu, H. Wang, B. Wu, W. Wang, B. Lin, D. Zhang, and D. Cai, "One-shot implicit animatable avatars with model-based priors," in *ICCV*, 2023. 1

[7] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *CVPR*, 2021. 2

[8] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," in *CVPR*, 2020, pp. 5202–5211. 2

[9] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023, pp. 22 500–22 510. 2

[10] M. Technology, "Face++," https://www.faceplusplus.com/, accessed 1 November 2023. 2

[11] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR 2016*. IEEE Computer Society, 2016, pp. 1096–1104. 2

[12] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt, "MonoPerfCap: Human performance capture from monocular video," *ACM Trans. Graph.*, vol. 37, no. 2, pp. 27:1–27:15, May 2018. 2

[13] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *ICCV*, 2019, pp. 5932–5941. 2

[14] J. Ren, Y. Yao, B. Lei, M. Cui, and X. Xie, "Structure-aware flow generation for human body reshaping," in *CVPR*, 2022, pp. 7744–7753. 2, 3