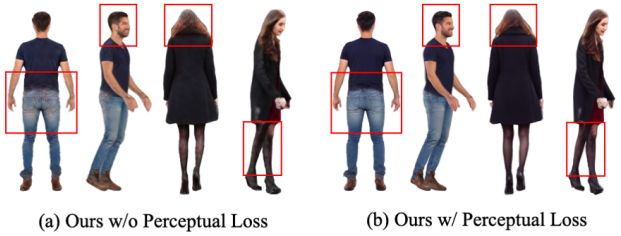
HumanRef: Single Image to 3D Human Generation via Reference-Guided Diffusion

Supplementary Material



(a) Ours w/o Perceptual Loss (b) Ours w/ Perceptual Loss

(c) Ours w/o SDF Loss (d) Ours w/ SDF Loss

Figure 10. Ablation study on analyzing perceptual loss for texture enhancement and SDF loss for body pose preservation.

6. Region Division

The region division process starts by dividing the input image into four regions based on parsing. Next, we divide the 3D space into corresponding regions along the y-axis. As shown in Fig. 12, region masks for each rendered view are determined by intersecting the rendered mask with the projection of the 3D division. Moreover, additional rules are applied to refine the region masks based on parsing semantics. For instance, if a person wears both a coat and a lining, the mask of the lining will be removed in the reference view to eliminate its texture influence on other views of the coat. The benefits of using refined region masks are highlighted in Fig. 9 of the ablation studies.

7. Reference-Guided Region-Aware Attention

In reference branch of Fig. 4, we save the features p_{ref} before each self-attention, which are used to calculate the *query*, *key*, and *value* vectors in attention networks. To inject the reference information into the denoising process of rendering views, we concatenate p_{ref} with corresponding features p_{tar} to calculate *key* and *value* vectors for ref-self-attention, while *query* vector is inferred by p_{tar}. This ensures that the result of ref-self-attention has the same dimension as vanilla self-attention, without affecting subsequent operations. After calculating attention scores, we post-process and normalize the scores using the *query* and *key* masks to achieve region-aware attention.

(a) Janus Problem (b) Fail in Extreme Pose

Figure 11. Our approach may suffer from the Janus problem in the side view because we do not make any constraints on this in the side view. Besides, our method may also fail in some extreme poses where SMPL-X body estimation is wrong.

Methods	**Ours** (Full)	Ours w/o L_{rec}	Ours w/o L_{SDF}	Ours w/o L_{norm}	Ours w/o L_{smooth}
LPIPS ↓	**0.032**	0.104	0.034	0.034	0.033
Contextual ↓	**1.969**	4.181	2.480	2.447	2.304
CLIP Score ↑	**90.0**%	78.9%	88.7%	88.8%	88.5%

Table 3. Quantitative comparison for ablation study on loss functions.

8. Ablation Study on Loss Functions

To assess the effectiveness of our core loss functions, we perform experiments by systematically removing them and evaluating the corresponding results. In the red box of Fig. 10(a), the method without the perceptual loss produces blurred texture details, although it remains realistic and consistent with the input. By incorporating the perceptual loss, we significantly enhance the local details of the generated human, as depicted in Fig. 10(b). Furthermore, Fig. 10(c&d) and Tab. 3 demonstrate the constraining role of the SDF loss in preserving human geometric pose and the necessity of retaining other loss designs, including reconstruction loss, normal loss, and smoothing loss, to improve the robustness and generation quality of our method.

9. Discussion on the Limitations

While our 3D human generation experiments have achieved impressive overall results, it is crucial to acknowledge occasional failures. Fig. 11(a) visually exemplifies the Janus problem, which can impact our results in side views due to the lack of view-specific constraints. Moreover, accurately estimating body poses in extreme cases challenges for our method, potentially resulting in failures, shown in Fig. 11(b). Additionally, in cases where the normal maps estimated by ECON contain artifacts, our method may introduce floating artifacts and produce blurry results, as illustrated in Fig. 13. To mitigate this issue, we propose two

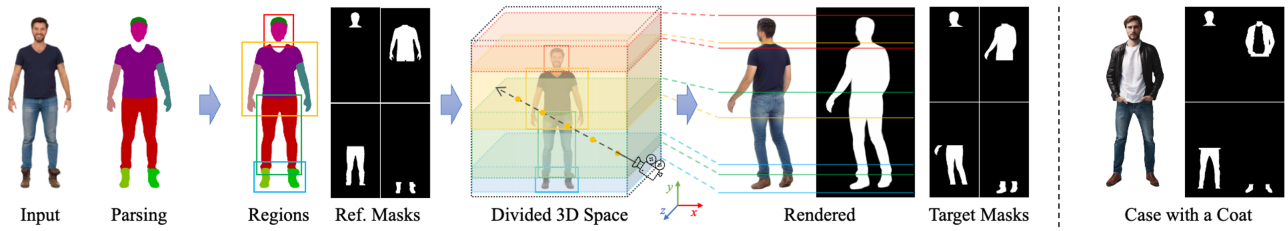


Figure 12. Illustration of 3D region division.

Figure 13. Our approach may introduce floating artifacts when the normal maps estimated by ECON contains artifacts.

potential solutions. Firstly, we can consider removing the normal constraint, albeit at the cost of sacrificing some geometric detail, as depicted in Fig. 13. Alternatively, we can explore the possibility of replacing ECON with a more accurate method for estimating normals.

10. More Results

To showcase the performance of our HumanRef in 3D clothed human generation, we present additional results in Fig. 14.

Figure 14. Additional results produced by our HumanRef.