ConTex-Human: Free-View Rendering of Human from a Single Image with Texture-Consistent Synthesis

Supplementary Material

Overview. The supplementary material has the following contents:

- Coarse stage implementation details
- Fine stage implementation details
- Compare with TeCH
- More visual comparison
- Geometry Evaluation
- User Study

A. Coarse Stage Details

Pre-Process Given a single image of a specific person, we first adopt the off-the-shelf background removal tool from https://github.com/danielgatis/rembg to attain the human foreground mask $\mathbf{M}$. Based on the foreground mask $\mathbf{M}$, we create an RGBA image with 648×648 resolution and make sure that the valid human region occupies approximately 70 % of the image's height, ensuring it remains centered.

Besides the mask, we also need the reference image normal map. In practice, we employ the designed normal estimator $\mathbf{N}$ from ECON[53]. Note that, $\mathbf{N}$ is conditioned with an optimized SMPL normal map. Therefore, our optimized geometry also incorporates the human pose information.

Camera Setting. For the coarse stage, we optimize the neural radiance field(NeRF)[32] with 128×128 resolution. The goal of the coarse stage is to supply a coarse geometry with a roughly accurate human pose and boundary for *back view synthesis stage* and *fine stage*. The elevation and azimuth degree of the reference image is set to **0.** as default.

For the camera setting during *Score Distillation Sampling*, the elevation range is set to [-30°, 60°], and the azimuth range is set to [-180°, 180°]. The camera distance is set to 3.8 as default, camera field of view (FOV) is set to 20° which is aligned with Zero-1-to-3[25].

3D Representation. We employ a multi-resolution hash grid from Instant-NGP[33] as the 3D NeRF representation. We use 16 levels of hash dictionaries of size 2^{19}, each entry is with a dimension 2 feature vector. The 3D grid resolution range from 2^4 to 2^{12} with an exponential growth rate of 1.447. A two-layer tiny MLP with 64 hidden units is adopted to decode the concatenated features interpolated from Instant-NGP to RGB color and volume density. The background is a "white" solid color background. We sample 512 points along each ray.

Text Prompt. The text is set by ourselves with a predefined text prompt template. We just need to change some keywords in the template according to the input image information.

Score Distillation Sampling. We sample images with a batch_size of 4 each iteration for *Score Distillation Sampling*. We sample the timestep $t \sim \mathcal{U}(0.2, 0.6)$, the classifier-free guidance weight is set to 5.

The overall $\mathcal{L}_{coarse}$ loss for the coarse stage can be formulated as a combination of $\mathcal{L}_{sds}^{z123}$, $\mathcal{L}_{mask}$, $\mathcal{L}_{rgb}$ and $\mathcal{L}_{normal}$:

$$\mathcal{L}_{coarse} = \lambda_1 \mathcal{L}_{sds}^{z123} + \lambda_2 \mathcal{L}_{rgb} + \lambda_3 \mathcal{L}_{normal} + \lambda_4 \mathcal{L}_{mask} \tag{9}$$

where in practice $\lambda_1 = 1.0$, $\lambda_2 = 1000$, $\lambda_3 = 1000$, $\lambda_4 = 1000$, some additional constrain like density sparsity and normal smoothness are also employed during optimization. We optimize the coarse stage using Adam optimizer for 3000 steps with a learning rate $5\text{×}10^{-3}$.

B. Fine Stage Details

Geometry Optimization. We adopt DMTet[45] in the fine stage, a hybrid SDF-Mesh representation, the DMtet resolution is set to 256×256×256.

The overall $\mathcal{L}_{fine}^{geo}$ loss for the coarse stage can be formulated as a combination of $\mathcal{L}_{sds}^{z123}$, $\mathcal{L}_{mask}$, $\mathcal{L}_{rgb}$ and $\mathcal{L}_{normal}$:

$$\mathcal{L}_{fine}^{geo} = \lambda_1 \mathcal{L}_{normal} + \lambda_2 \mathcal{L}_{mask} + \lambda_3 \mathcal{L}_{lap} + \lambda_4 \mathcal{L}_{smooth} \tag{10}$$

where $\mathcal{L}_{smooth}$ is the mesh normal constraint, $\mathcal{L}_{lap}$ is the mesh laplacian constraint. In practice $\lambda_1 = 10000$, $\lambda_2 = 50000$, $\lambda_3 = 1000$, and $\lambda_4 = 1000$. We optimize the geometry stage using Adam optimizer for 3000 steps with a learning rate $1\text{×}10^{-2}$. In steps 2000∼3000 step, λ_3 and λ_4 are set to 100 for more human geometry details.

Texture Field. We employ another multi-resolution hash grid to represent the texture field. We use 14 levels of hash dictionaries of size 2^{19}, each entry is with a dimension 2 feature vector. Same as the coarse stage, the 3D grid resolution ranges from 2^4 to 2^{12}. A two-layer tiny MLP with 64 hidden units is adopted to decode the concatenated features to RGB color. The background is a "white" solid color background.

Camera Setting. The camera setup of the fine stage is similar to the coarse stage except that the elevation degree range is [-45°, 45°] and the image resolution is 648×648.

Score Distillation Sampling. We sample images with a batch_size of 1 for each iteration for SDS. For Zero-1-to-

Figure 8. **Qualitative comparison with TeCH[16]results on THuman2.0 and SSHQ dataset.** Compared with TeCH, our methods have a consistent texture with input images. Row 1&3 are TeCH results, Row 2&4 are our results. Please ***Zoom in*** for the details.

3 SDS, we sample the timestep $t \sim \mathcal{U}(0.2, 0.6)$, and the classifier-free guidance weight is set to 5. For Stable Diffusion SDS, we sample the timestep $t \sim \mathcal{U}(0.02, 0.5)$, and the classifier-free guidance weight is set to 50.

The overall $\mathcal{L}_{fine}^{tex}$ loss for the coarse stage can be formulated as a combination of $\mathcal{L}_{sds}^{z123}$, $\mathcal{L}_{sds}^{sd}$, $\mathcal{L}_{rgb}$ and $\mathcal{L}_{vpc}$:

$$\mathcal{L}_{fine}^{tex} = \lambda_1 \mathcal{L}_{sds}^{z123} + \lambda_2 \mathcal{L}_{sds}^{sd} + \lambda_3 \mathcal{L}_{rgb} + \lambda_4 \mathcal{L}_{vpc} \quad (11)$$

where in practice $\lambda_1 = 0.002$, $\lambda_2 = 0.5$, $\lambda_3 = 10000$, $\lambda_4 = 10$. We optimize the texture stage using Adam optimizer for 4000 steps with a learning rate $1{\times}10^{-3}$. To maintain the front/back view details and generate consistent side view texture, we optimize another 2000 steps with $\lambda_1 = 0$, $\lambda_2 = 0$, $\lambda_3 = 10000$, $\lambda_4 = 100$.

C. Compare with TeCH

TeCH[16] is our concurrent work, which is also an optimization-based method that employs *Score Distillation Sampling* during the optimization process. As can be observed in Figure 8, TeCH tends to predict a floating human pose and always exhibits a misaligned texture in the hand region. Most importantly, as shown in the back view, TeCH shows an unreasonable texture compared with the input image in terms of texture pattern, texture style, and wrong prediction of the hat in the back head region.

D. More visual comparison

We provide more visual results in Figure 9 on THuman2.0 dataset and Figure 10 on SSHQ dataset. Please Zoom in For more details.

E. Geometry Evaluation

We employ the commonly used *Chamfer Distances* and *Volume IoU* to evaluate the geometry quality on *THuman* dataset. We utilize all 30 cases for evaluation. For fair comparison, we align the shapes of all methods to a unified scale and origin. *Chamfer* is calculated between the vertices of the predicted mesh and ground truth 3D scan. The size of

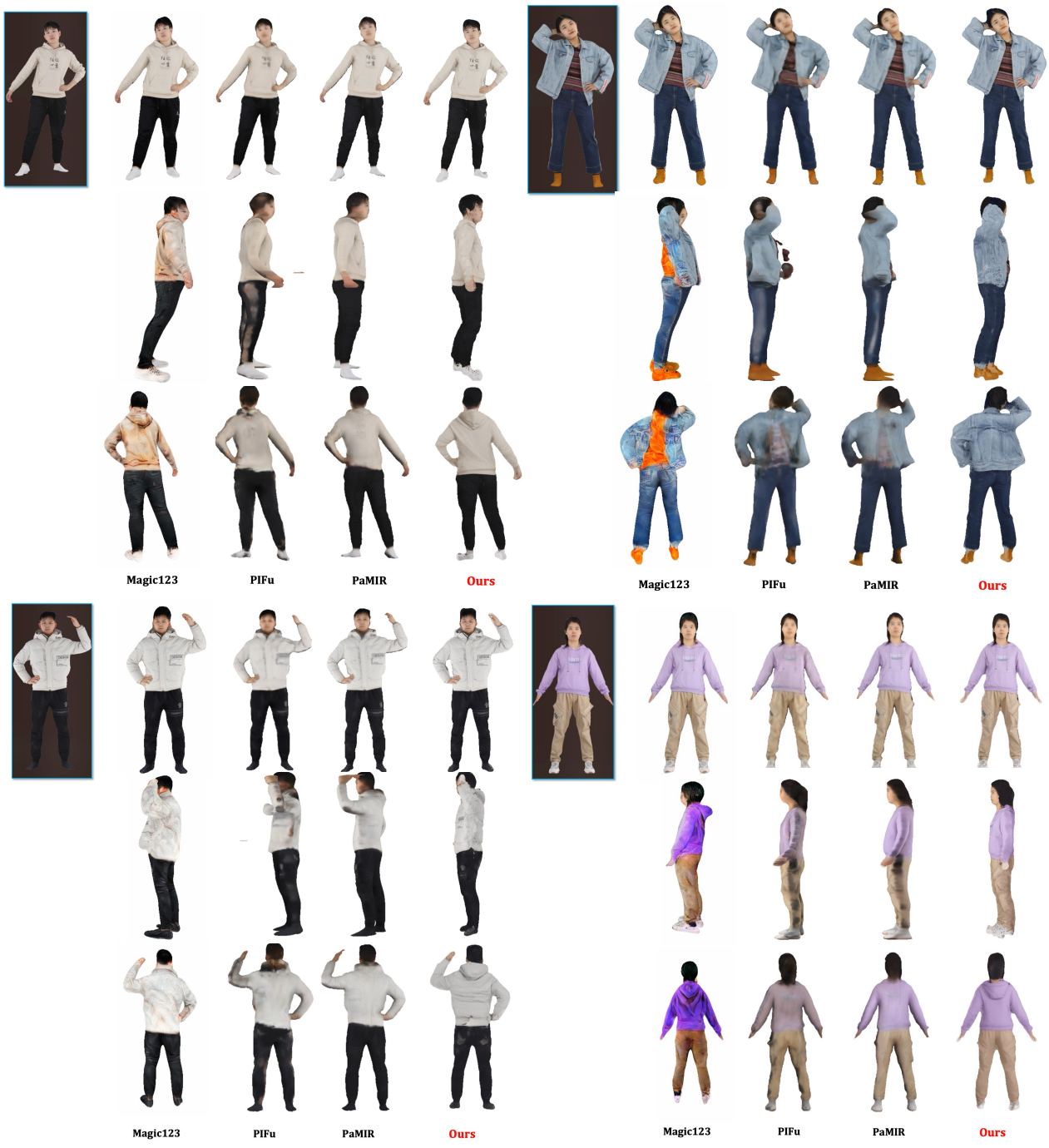

Figure 9. **Qualitative comparison results on THuman2.0 dataset.** Compared with them, our methods can render texture-consistent and high-fidelity novel views.

the voxel grid during *IoU* calculation is set to 64×64×64. Although our primary goal is not to achieve precise geometric reconstruction but high-quality texture synthesis and renderings, quantitative results in Table 3 indicate that our geometric outcomes outperform all the competing methods. We don't show the concurrent work TeCH results here due to its more than 4 hours of training time for each case with 2 A100 GPUs.

Method	PIFu	PaMIR	Magic123	Ours
Chamfer ↓	0.0206	0.0182	0.0251	**0.0177**
IoU ↑	0.5072	0.5391	0.4328	**0.5626**

Table 3. Geometry evaluation

Figure 10. **Qualitative results on SSHQ dataset.** Compared with them, our methods can render texture-consistent and high-fidelity novel views.

F. User Study.

Using CLIP as the evaluation metric is inspired by recent papers in image-to-3D generation, such as Magic123. They use CLIP to evaluate visual *quality* and image *consistency* between the input and novel view. We also provide user study results in Table 4, showing the percentages of user preference. The participants are given multi views and asked to choose the best method in image quality and texture consistency respectively.

Method	Magic123	PIFu	PaMIR	TeCH	Ours
Quality (%) ↑	8.15	4.35	7.61	10.87	**69.02**
Consistency (%) ↑	8.70	5.98	12.50	3.26	**69.57**

Table 4. User study of preference