Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 2, 2024

Full-body Texture Reconstruction of Clothed Human Body

Zhipeng Ren[1], Ji Zhang[1, 2, *]

[1] School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China
[2] Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding071003, China
[*]Corresponding author Email: 72zhangji@163.com

Abstract: In order to solve the distortion problem in monocular reconstruction, a new solution is proposed in this paper. In particular, the space occupancy field of the human body can be obtained by combining normal maps and symbol distance field. Meanwhile, transformer encoder and decoder are used to extract feature vectors from input images and feed them into MLP, and finally the whole texture including invisible areas can be inferred. Experiments show that the proposed method makes good progress in the reconstruction of human body based on monocular RGB, even with good robustness for invisible regions.

Keywords: Human reconstruction, Normal mapping, Texture, Implicit function.

1. Introduction

Estimating 3D human pose from monocular RGB images is one of the core techniques for building computational models to understand human behavioral cues. The technology helps to promote a variety of applications, and has a wide range of application prospects in the fields of behavior recognition, AR/VR, attitude tracking, social behavior understanding, and so on. At its core, nuances of human behavior are often conveyed through faces, gestures, and body postures, so accurate estimates of whole body movements are necessary to more accurately capture real behavioral signals.

Thanks to the success of parametric human models such as SMPL[1], STAR[2], SMPL-X[3], etc., parametric human bodies based on deep learning are becoming increasingly popular. With the rise of virtual fitting, simulation games, film and television production and other industries, people are no longer satisfied with simple naked reconstruction, but the material changes of clothing, complex topology and texture changes, coupled with the interaction between clothing and the body in different positions, these factors make the geometric representation space required for the geometric reconstruction of the clothed body become high-dimensional and complex. Therefore, although the traditional parameterization-based representation method can simplify the problem, it can not fully capture the complex geometric structure and detailed characteristics of the human body. SMPLicit[4] Use SMPL's skin weights to transform clothing in T-pose to get a clothing model in the current human pose, but it is not appropriate to use human skin weights for looser clothing. PIFU[5] and PIFUHD[6] used the implicit function based on pixel alignment to reconstruct high-quality human body models with good real-time and robustness, including clothing and details, but only obtained satisfactory results under standing posture. ICON[7] generates a normal graph from RGB in a way similar to pix2pix[8], and based on the predicted human normal graph and SMPL mannequin model, returns the implicit surface of the clothed human body, which also has good robustness in extreme postures. With the birth of Nerf[9], the human body reconstruction method based on Nerf[10-12] has made good breakthroughs in reconstruction speed, quality, illumination consistency and other aspects. However, in general, NERF-related human body reconstruction is more suitable for generating new views rather than directly generating standard human body models.

In order to solve the above problems of missing, artifacts and no texture in human body reconstruction, this paper proposes a new method, which combines normal prediction and symbol distance field, can reconstruct human mesh with texture from a single RGB photo, and still has good expression effect under extreme poses.

2. Clothed Human Texture Model Design

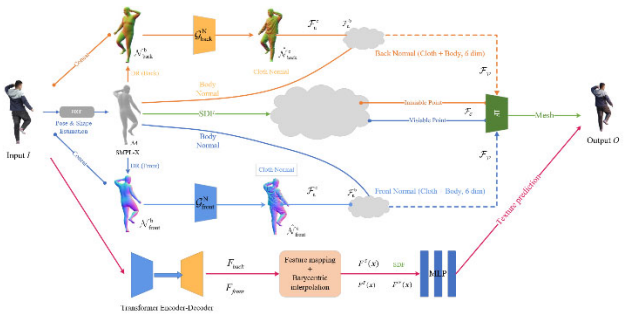

Figure 1. Structure of method

The structure of this paper is shown in Fig 1. A single RGB image I is accepted as input, and after pose and shape Estimation, the SMPL-X model $\mathcal{M}$ is obtained. Then the naked body normal prediction obtained through the Clothed body normal prediction was obtained through the generation network, and the human occupation field was predicted by combining the implicit function and SDF. At the same time ,the input image is encoded and decoded by Transformer to obtain image features, and then the human body texture is inferred by MLP through Feature mapping and Barycentric interpolation

2.1. Clothed body normal prediction.

Trying to infer a full 360-degree 3D graph from a single RGB image of a clothed person is an extremely challenging

task. The main difficulty is that the normal information of the obscured part of the image needs to be obtained by inferring the visible part, which is essentially an ill-posed problem. However, by using a large number of RGB images for training with the corresponding normal images, the image-to-image transformation network structure can be used to accurately estimate the "front" normal map $\mathcal{N}_{\text{front}}$ from the RGB image, and the "back" normal map $\mathcal{N}_{\text{back}}$ is equally important in order to obtain the fine mesh of the clothed human body.

Under the weak perspective camera model with global rotation parameters $R \in \mathbb{R}^{3\times3}$, translation parameters $t \in \mathbb{R}^2$ and scaling parameters $s \in \mathbb{R}$, this paper uses PyTorch3D [13] differentiable renderer, which is expressed as, $\mathcal{DR}$.

$$\mathcal{DR}(\mathcal{M}) \rightarrow \mathcal{N}^{\text{b}} \tag{1}$$

In the formula (1), $\mathcal{N}^{\text{b}} \rightarrow \left\{\mathcal{N}^{\text{b}}_{\text{front}}, \mathcal{N}^{\text{b}}_{\text{back}}\right\}$. In order to guide the normal prediction and make it robust to various body postures, ICON bases the normal graph prediction module on the body normal maps $\mathcal{N}^{\text{b}}$ rendered from the estimated body $\mathcal{M}^{b}$. As shown in formula (2), given $\mathcal{N}^{\text{b}}$ and input image $\mathcal{I}$, the normal-generating network $\mathcal{G}^{\text{N}} = \left\{\mathcal{G}^{\text{N}}_{\text{front}}, \mathcal{G}^{\text{N}}_{\text{back}}\right\}$ predicts the front and back normals of the clothed human body, expressed as $\hat{\mathcal{N}}^{\text{c}} = \left\{\hat{\mathcal{N}}^{\text{c}}_{\text{front}}, \hat{\mathcal{N}}^{\text{c}}_{\text{back}}\right\}$

$$\mathcal{G}^{\text{N}}\left(\mathcal{N}^{\text{b}}, \mathcal{I}\right) \rightarrow \hat{\mathcal{N}}^{\text{c}} \tag{2}$$

In order to predict the normal plot of the invisible region and maintain some consistency in the local region of the normal plot, this paper fine-tuned the back normal generator $\mathcal{G}^{\text{N}}_{\text{back}}$ and increased the MRF loss [14], as shown in formula (3).

$$\mathcal{L}_{\text{N}} = \mathcal{L}_{\text{pixel}} + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}} + \lambda_{\text{MSF}}\mathcal{L}_{\text{MSF}} \tag{3}$$

In formula (3), $\mathcal{L}_{\text{pixel}} = \left|\mathcal{N}^{\text{c}}_{\text{v}} - \hat{\mathcal{N}}^{\text{c}}_{\text{v}}\right|$,where v={front,back}, only the obtained normal line maps $\mathcal{L}_{\text{pixel}}$ is very fuzzy, adding the perception loss containing high-level network feature information $\mathcal{L}_{\text{VGG}}$ helps to restore the normal line graph details.

In addition, in order to accurately estimate the SMPL-X mesh $\mathcal{M}(\beta,\theta) \in \mathbb{R}^{N\times3}$ from the image $\mathcal{I}$, this paper uses PIXIE[15] to obtain better mesh alignment with the image. Specifically, the shape parameters β, pose parameters θ, and translation parameters t of SMPL-X are optimized to minimize:

$$\mathcal{L}_{\text{SMPL-X}} = \min_{\theta,\beta,t}\left(\mathcal{L}_{\text{N\_diff}} + \mathcal{L}_{\text{S\_diff}} + \mathcal{L}_{\text{J\_diff}}\right) \tag{4}$$

$$\mathcal{L}_{\text{N\_diff}} = \left|\mathcal{N}^{\text{b}} - \hat{\mathcal{N}}^{\text{c}}\right| \tag{5}$$

$$\mathcal{L}_{\text{S\_diff}} = \left|\mathcal{S}^{\text{b}} - \hat{\mathcal{S}}^{\text{c}}\right| \tag{6}$$

$$\mathcal{L}_{\text{J\_diff}} = \lambda_{J\_\text{diff}}\left|\mathcal{J}^{b} - \widehat{\mathcal{J}^{\text{c}}}\right| \tag{7}$$

In formula (4), $\mathcal{L}_{\text{N\_diff}}$ is the L1 loss of the normal map; $\mathcal{L}_{\text{S\_diff}}$ is the L1 loss between contours;

$\mathcal{L}_{\text{J\_diff}}$ is the L2 joint loss between 2D joint $\mathcal{J}^{b}$ reprojected by $\mathcal{M}^{b}$ and 2D landmarks D $\widehat{\mathcal{J}^{\text{c}}}$, which can be estimated by 2D pose from RGB images.

In the experiment, set $\lambda_{J\_\text{diff}} = 5.0$, however, if the overlap ratio between the clothing mask and the body mask is less than 0.5, it means that the human is wearing loose clothing. In this case, the 2D joint should be trusted more and λ_{J_diff} will be increased to 50.0. Similarly, occlusion occurs when the overlap between the body mask inside the clothing mask and the full body mask is less than 0.98. In this case, $\lambda_{J\_\text{diff}} = 0.0$ will be set to avoid situations where limbs self-intersect after posture optimization.

The normal map $\mathcal{N}^{\text{b}}$ rendered from the optimized SMPL-X grid is provided to the network $\mathcal{G}^{\text{N}}$. The improved SMPL-X guides $\mathcal{G}^{\text{N}}$ through a grid-image alignment network to infer a more reliable and detailed clothed body normal $\hat{\mathcal{N}}^{\text{c}}$ During the inference process, this paper alternates between (1) refining the SMPL-X grid using inferred normal maps $\hat{\mathcal{N}}^{\text{c}}$ and (2) re-inferring $\hat{\mathcal{N}}^{\text{c}}$ using refined SMPL-X. This feedback loop enables a more reliable positive/negative dressing of the human body normal.

2.2. Implicit spatial reconstruction.

The implicit symbolic distance field is an implicit representation used to represent the 3D shape of the human body in the standard pose space, while encoding the distance information from each point in the space to the human surface. It can better handle topological changes and adapt to changes in human posture, especially when dealing with clothed human bodies. First, the three-dimensional space is discretized, and then the distance between each point and the surface of the object is calculated. Meanwhile, in order to normalize this distance, the algorithm will normalize or truncate the distance that is too large. The distance between the points inside the object is positive, and the distance between the points outside the object and the surface is negative. For surface reconstruction, the algorithm is represented by the isosurface of continuous 3D symbolic distance field, as shown in formula (8). Marching Cubes algorithm can be used to extract the isosurface of the current

implicit symbolic distance field, so as to obtain the continuous surface of the object. Among them, $p \in \mathbb{R}^3$

$$f(p)=\begin{cases} <0, \text{if } p \text{ is inside mesh surface} \\ =0, \text{if } p \text{ on the mesh surface} \\ >0, \text{otherwise} \end{cases} . \quad (8)$$

In this section, a standard template shape $\mathcal{F}_{\mathrm{P}}$ as the zero isosurface of the symbolic distance field. The isosurface of the symbolic distance field is represented by a multilayer perceptron (MLP). Given the predicted clothed body normal maps $\hat{\mathcal{N}}^{\mathrm{c}}$ and SMPL-X-body mesh $\mathcal{M}$, this paper regresses the implicit 3D surface of the clothed person based on local features $\mathcal{F}_{\mathrm{P}}$

$$\mathcal{F}_{\mathrm{P}}=\left[\mathcal{F}_{\mathrm{s}}(\mathrm{P}), \mathcal{F}_{\mathrm{n}}^{\mathrm{b}}(\mathrm{P}), \mathcal{F}_{\mathrm{n}}^{\mathrm{c}}(\mathrm{P})\right] \quad (9)$$

In formula (9), $\mathcal{F}_{\mathrm{s}}$ is the signed distance from the query point P^{b} to the nearest body point $\mathrm{P}^{\mathrm{b}} \in \mathcal{M}$, $\mathcal{F}_{\mathrm{n}}^{\mathrm{b}}$ is the barycentric surface normal of P^{b}, both of which provide strong regularization for self-occlusion，. Finally, $\mathcal{F}_{\mathrm{n}}^{\mathrm{c}}$ is the normal vector extracted from $\hat{\mathcal{N}}_{\mathrm{front}}^{\mathrm{c}}$ or $\hat{\mathcal{N}}_{\mathrm{back}}^{\mathrm{c}}$, depending on the visibility of P^{b}

$$\mathcal{F}_{\mathrm{n}}^{\mathrm{c}}(\mathrm{P})=\begin{cases} \hat{\mathcal{N}}_{\mathrm{front}}^{\mathrm{c}}(\pi(\mathrm{P})) \text{ if } \mathrm{P}^{\mathrm{b}} \text{ is visible} \\ \hat{\mathcal{N}}_{\mathrm{back}}^{\mathrm{c}}(\pi(\mathrm{P})) \quad \text{else,} \end{cases} \quad (10)$$

In formula (10), $\pi(\mathrm{P})$ represents a 2D projection of a 3D point P. It should be noted that $\mathcal{F}_{\mathrm{P}}$ has nothing to do with global body posture. The experimental results show that this is essential to improve the robustness to the transgression pose and to enhance the effectiveness of the training data. This paper inputs $\mathcal{F}_{\mathrm{P}}$ into an implicit function $\mathcal{IF}$, parameterized by a multilayer perceptron (MLP) to estimate the occupancy at point P, denoted as $\hat{o}(\mathrm{P})$. Next, a fast surface positioning algorithm is used to extract the mesh from the three-dimensional occupancy inferred from the $\mathcal{IF}$.

2.3. Front texture mapping

The inference of human surface texture is to realize the inference of the color of points on the model, which is determined by its diffuse reflection coefficient. The goal of the algorithm in this paper is to predict the color value on the human body model, which is also the RGB value. Prediction is a very difficult task, and there are many previous works to solve this task. The algorithm in this paper draws on the idea of IDR and uses MLP to approximate a rendering equation, which is called implicit neural rendering. The input to the network accepts a point to be rendered, the normal vector of that point, the camera Angle direction, and the global feature set vector that the coding network learns to calculate the RGB color value of the corresponding point along the camera ray direction.

As shown in the figure 2, the light is emitted from the center c of the camera along the sampled pixel, its direction v is determined by the inherent parameter t of the camera, and the algorithm calculates its intersection point P on $\mathcal{F}_{\mathrm{P}}$,. At the same time, the gradient calculation $N_P=\nabla f(P;\eta)$ is carried out by the method in Section 3.1 to obtain the point P normal.

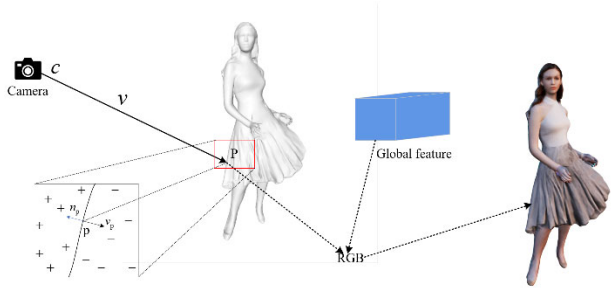

Figure 2. Neural rendering

Then, the Jacobian matrix of $x=\mathcal{D}_i(P)$ of the point after deformation field can be used to transform v to get the camera Angle V_P in standard space. As for global geometric features, the algorithm also uses larger MLP $F(P;\eta)=(f(P;\eta),Z(P;\eta))$ perform additional calculations on it, where $f(P;\eta)$ represents feature information, $Z(P;\eta)$ represents depth-west information, which can be simply understood as feature information with depth information, meaning that geometric information around point P can be used to help predict global shadows. Finally, the algorithm uses MLP M with a learnable weight y to compute the color $C_{\mathbf{p}}\left(\eta,\psi_i,\gamma,\tau\right)$ of P, the implementation here refers to previous work [16]. As shown in formula (11) :

$$C_{\mathbf{p}}\left(\eta,\psi_i,\gamma,\tau\right)=M\left(P,N_P,V_P,Z(P;\eta);\gamma\right) \quad (11)$$

$$N_P=\nabla f(P;\eta) \quad (12)$$

$$x=\mathcal{D}_i(P) \quad (13)$$

$$V_P=J_x(P)^{-1}v \quad (14)$$

In formula (11), it can be seen that the color of the deformation point x along the direction v is determined by the MLP weights η and γ , camera parameters τ and deformation field parameter ψ_i

2.4. Invisible area texture inference

How to learn and recover the texture of the invisible part is a very challenging problem, as with most methods, this article assumes that the texture of the invisible part is similar to that of the visible part. This article's approach takes inspiration

from common features between the back view and the visible front view, such as materials and colors. Despite their different points of view, the similarities in these characteristics are crucial. Therefore, the goal of this paper is to use the SMPL-X model as the front view to effectively separate the features of the back view。

The process starts with a VIT-based global encoder , which encodes the input image I into a latent feature h, capturing the globally relevant features of the image. To decode these features, this article uses two decoders: a front view decoder aligned with h and a back view decoder. The front view decoder utilizes multi-head self-attention within the Vision transformer to process the front view features, representing $F_{\text{front}} \in \mathbb{R}^{H\times W\times C}$, and also obtaining F_{back} using a similar decoder.

Using spatial positioning and prior knowledge of the human body, features are effectively combined at one query point. The query point is projected onto the feature map $F_{\text{j}}, j \in \{front, back\}$ to obtain pixel-aligned feature F_j^S . This paper then combines these features of all planes using a combination of joins as formula (15):

$$F^S(\boldsymbol{x}) = F_f^S(\boldsymbol{x}) \oplus F_b^S(\boldsymbol{x}) \tag{15}$$

In this paper, the vertices of the SMPL-X grid are projected onto two feature maps to obtain feature $F^S(v), v \in \mathcal{M}$. For each query point x, the nearest triangular face $t_x = \left[\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2\right] \in \mathbb{R}^{3\times 3}$ is found, and the features of x are integrated with centroid interpolation, denoted as $F^P(\boldsymbol{x})$:

$$F^P(\boldsymbol{x}) = uF^S\left(\boldsymbol{v}_0\right) + vF^S\left(\boldsymbol{v}_1\right) + wF^S\left(\boldsymbol{v}_2\right) \tag{16}$$

In formula (16), $[u, v, w]$ represents the barycentric coordinates of the query point x projected onto the triangle. This article uses these two query features as the final point features. In addition, we use the signed distance between the query point and the SMPL-X grid $\mathcal{SDF}(x)$ and the pixel-aligned normal feature $F^N(\boldsymbol{x})$ as an input to the multi-layer perceptron (MLP) for predicting colors:

$$\boldsymbol{c} = MLP\left(F^S(\boldsymbol{x}), F^P(\boldsymbol{x}), \mathcal{SDF}(\boldsymbol{x}), F^N(\boldsymbol{x})\right) \tag{17}$$

This paper considers two sets of points as training data, denoted as Go and Gc . Go samples uniformly along the normal of the ground real grid surface with slight perturbations, while Gc samples according to the same strategy as [17]. For points in the middle, this article uses the following loss function:

$$\mathcal{L}_o = \frac{1}{\left|G_o\right|} \sum_{x\in G_o} BCE\left(\hat{o}_{\boldsymbol{x}} - o_{\boldsymbol{x}}\right) \tag{18}$$

In formula (18), $\hat{o}_{\boldsymbol{x}}$ represents the predicted occupancy of the model and $o_{\boldsymbol{x}}$ is the real occupancy of the ground. For sampling points in B, this paper applies the following loss function:

$$\mathcal{L}_c = \frac{1}{\left|G_c\right|} \sum_{x\in G_c} \left|\hat{\boldsymbol{c}}_{\boldsymbol{x}} - \boldsymbol{c}_{\boldsymbol{x}}\right| \tag{19}$$

In formula (19), $\hat{\boldsymbol{c}}_{\boldsymbol{x}}$ represents the predicted color and $\boldsymbol{c}_{\boldsymbol{x}}$ represents the corresponding true color of the ground. The total loss is the sum of these two independent losses and is designed to achieve a comprehensive training goal.

3. Experiments

3.1. dataset

AGORA is a synthetic dataset with highly realistic and highly accurate truth values. Using 4,240 commercially available, high-quality, richly textured body scans in a variety of poses and natural clothing (this included 257 child scans). Create reference 3D poses and body shapes by fitting SMPL-X body models to 3D scans. By rendering 5 to 15 people in each image using image-based lighting or 3D ambient rendering, create training images of about 14K and test images of 3K, and take care to make the images physically reasonable and realistic. In the end, AGORA had 173K rendered images.

THuman2.0 [18] contains 525 high-quality human texture scans, captured by a dense DSLR platform, and corresponding SMPL-X matching.

3.2. Metrics.

In order to quantitatively evaluate the indexes of the algorithm in this chapter and conduct a comparison experiment with other algorithms, the algorithm in this chapter uses three evaluation indexes, namely P2S, Chamfer and Normals. All three indicators are measured in centimeters, and the smaller the better.

P2S (Point-to-surface Euclidean distance) represents the Euclidean distance between the vertex of the reconstructed Surface and the surface of the true value, where the distance from the Point to the surface is calculated.

Chamfer (Chamfer distance) indicates the chamfer distance between the reconstructed surface and the true surface, which is mainly used for point cloud reconstruction or 3D reconstruction. Defined as follows

$$d_{CD}\left(S_1, S_2\right) = \frac{1}{S_1} \sum_{x\in S_1} \min_{y\in S_2} \| x - y \|_2^2 + \frac{1}{S_2} \sum_{y\in S_2} \min_{x\in S_1} \| y - x \|_2^2 \tag{20}$$

In formula (20), S_1 and S_2 of the formula represent two groups of 3D point clouds respectively. The first term represents the sum of the minimum distance between any point x in S1 and S2, and the second term represents the sum of the minimum distance between any point y in S2. If the distance is large, it indicates that the two groups of point clouds are different; if the distance is small, it indicates that the reconstruction effect is better.

Normals (Normals Loss) : The difference between the generated normals and the real normals is assessed to ensure that the generated shape has normals similar to the real shape,

thereby improving the geometric and visual quality of the shape.

3.3. Settings

To perturb the pose and shape parameters of SMPL, random noise is added to θ and β :

$$\begin{aligned} \theta+ &= s_\theta * \mu, \\ \beta+ &= s_\beta * \mu \end{aligned} \tag{21}$$

In formula (21), $s_\theta = 0.15$, $s_\beta = 0.5$ these are the alignment errors generated by simulating the human body pose estimation (Pixie) method based on empirical values.

To verify the necessity of $\mathcal{L}_{\text{N\_diff}}$, $\mathcal{L}_{\text{S\_diff}}$, $\mathcal{L}_{\text{J\_diff}}$ in equation (), an integrity check is performed on AGORA's validation set. Initialized with different attitude noise s_θ (equation (8)), $\mathcal{L}_{\text{N\_diff}}$, $\mathcal{L}_{\text{S\_diff}}$, $\mathcal{L}_{\text{J\_diff}}$ optimize the θ, β, t parameter of the disturbed SMPL-X by minimizing the difference between the rendered SMPL-X body Norma and the clothed body truth Norma for 2K iterations. As shown in Figure 3 , $\mathcal{L}_{\text{N\_diff}} + \mathcal{L}_{\text{S\_diff}} + \mathcal{L}_{\text{J\_diff}}$ always leads to the smallest error at any noise level, which is measured by the Chamfer distance between the disturbed SMPL-X mesh and the real SMPL-X mesh.

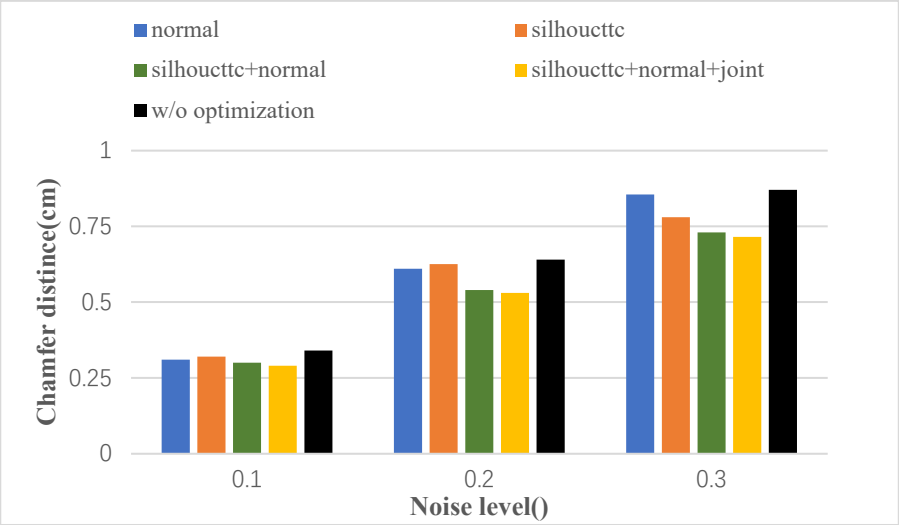

Figure 3. SMPL-X refifinement

Intuitively, a more accurate SMPL-X body model provides a better a priori, helping to infer better clothing - body normal. In practice, however, the human pose and shape (PIXIE) regression variable does not give an SMPL-X fit for pixel alignment. To account for this, during inference, the SMPL-X fits are optimized based on the difference between the rendered SMPL-body normal maps, and the predicted clothed-body normal maps, as shown in Fig. 4

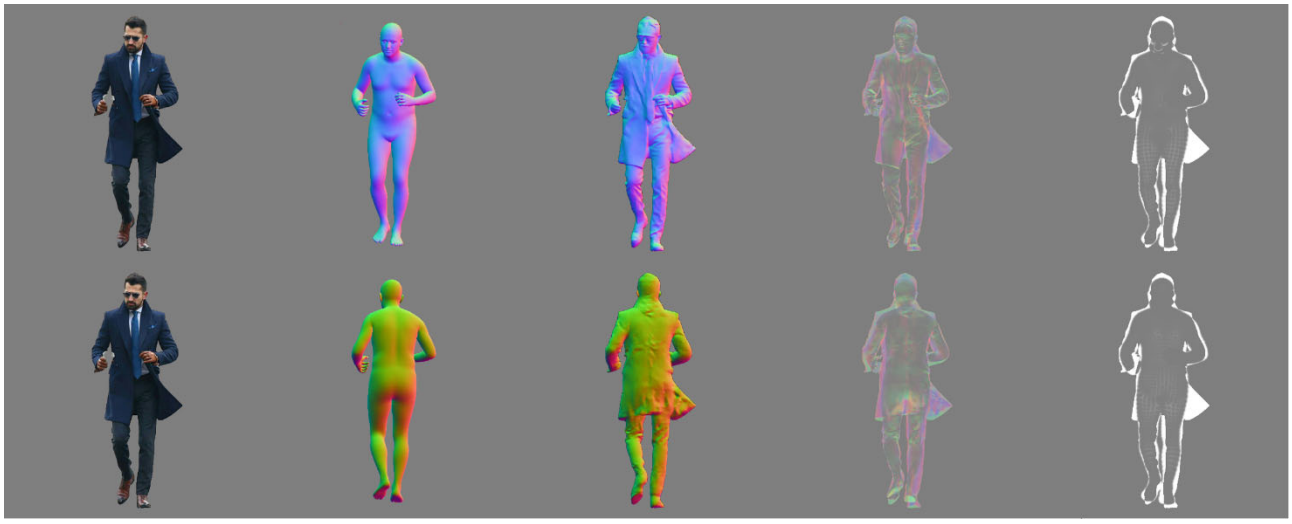

Figure 4. SMPL-X refifinement using a feedback loop.

3.4. Results

As shown in Figure 5, the human body mesh with texture reconstructed from a single image is shown. It can be seen that the human body mesh can be well reconstructed by the method proposed in this paper even in extreme movement posture and only one image is needed, and the texture inference of invisible areas in this paper still has a good effect

without the help of additional input.

Figure 5. The front and back texture of the body

4. Conclusion

This paper presents an end-to-end whole-body reconstruction based on single RGB. Even for invisible areas, it still has a good experimental effect and can complete more detailed textures. Using SMPL-X normals as a guide, the back view features are effectively decoupled during the conversion of 2D features to 3D. Future work is to introduce diffusion models for more refined whole-body texture reconstruction

References

[1] Loper M, Mahmood N, Romero J, et al. SMPL: A Skinned Multi-Person Linear Model[J]. ACM Transactions on Graphics, 2015, 34(6).

[2] Osman A A A, Bolkart T, Black M J. Star: Sparse trained articulated human body regressor[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer International Publishing, 2020: 598-613.

[3] Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3d hands, face, and body from a single image[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10975-10985.

[4] Corona E, Pumarola A, Alenya G, et al. Smplicit: Topology-aware generative model for clothed people[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11875-11885.

[5] Saito S, Huang Z, Natsume R, et al. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2304-2314.

[6] Saito S, Simon T, Saragih J, et al. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 84-93.

[7] Xiu Y, Yang J, Tzionas D, et al. Icon: Implicit clothed humans obtained from normals[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 13286-13296.

[8] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.

[9] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[10] Peng S, Zhang Y, Xu Y, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9054-9063.

[11] Liu L, Habermann M, Rudnev V, et al. Neural actor: Neural free-view synthesis of human actors with pose control[J]. ACM transactions on graphics (TOG), 2021, 40(6): 1-16.

[12] Weng C Y, Curless B, Srinivasan P P, et al. Humannerf: Free-viewpoint rendering of moving people from monocular video[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. 2022: 16210-16220

[13] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. arXiv:2007.08501, 2020. 3

[14] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and JiayaJia. Image inpainting via generative multi-column convolutional neural networks. Conference on Neural Information Processing Systems (NeurIPS), 2018.3

[15] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In International Conference on 3D Vision (3DV), 2021. 3,4

[16] Jiang B, Hong Y, Bao H, et al. Selfrecon: Self reconstruction your digital avatar from monocular video[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 5605-5615.

[17] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In ICCV, pages 2304–2314, 2019. 2, 3, 5, 6, 7, 13, 14, 16

[18] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors. In Computer Vision and Pattern Recognition (CVPR), 2021. 6, 7, 10