RICH: Robust Implicit Clothed Humans Reconstruction from Multi-Scale Spatial Cues

Yukang Lin[2], Ronghui Li[2], Kedi Lyu[1], Yachao Zhang[2(✉)], and Xiu Li[2(✉)]

[1] College of Computer Science and Technology, Jilin University, Changchun, China
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
yachaozhang@stu.xmu.edu.cn, li.xiu@sz.tsinghua.edu.cn

1 Supplementary material

1.1 Datasets

The 3D scans of the THuman2.0 and CAPE datasets, along with their corresponding SMPL-X [2] body models, are represented as meshes and stored in the obj file format. In our experiments, we utilize the SMPL-X model, which consists of 20,908 triangles and 10,475 vertices. Fig. 1 shows some test samples.

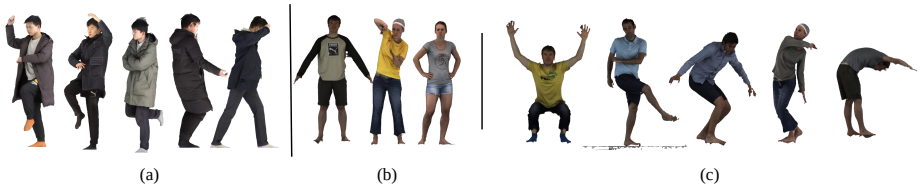

Fig. 1. Testing samples: (a) THuman2.0 test set, (b) samples with fashion poses from CAPE, (c) samples with challenging poses from CAPE.

We employ the OpenGL scripts provided by ICON [6] to generate data samples. By using the differentiable renderer $\mathcal{DR}$ from Pytorch3D [3] and a weak perspective camera, we render 3D human scans and their corresponding SMPL-X fits from 36 viewpoints. The angular spacing between any two adjacent viewpoints is 10 degrees. Each data sample is composed of a 3D clothed-human scan, its SMPL-X fit, an RGB image, camera parameters, 2D normal maps for the clothed human scan and the SMPL-X body (captured from two opposing views), as well as visibility for the SMPL-X mesh w.r.t. the camera. Fig. 2 illustrates all the data items included in a data sample.

1.2 Implementation

For sampling the query points for training, we adopts the stratage of PIFu [4] and sample a total of 8000 query points. Specifically, we perform dense sampling around the surface of the human scan and uniform sampling in space at

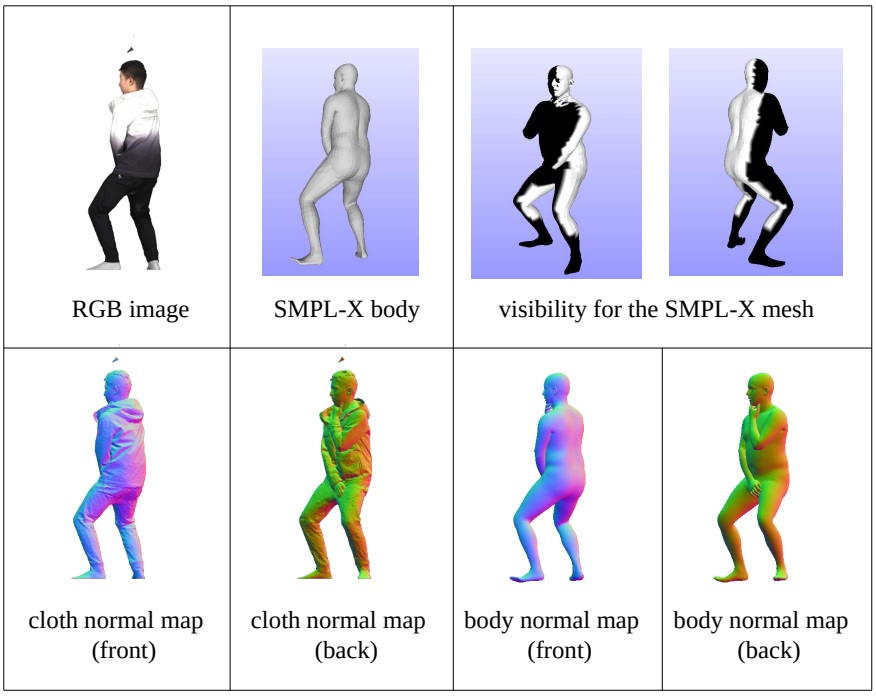

Fig. 2. All the data items included in a data sample.

a sampling density of 16:1. RICH is implemented in PyTorch Lightning. The cloth normal map prediction network $\mathcal{G}_n$ adopts the residual blocks used in PIFuHD [5], and the iamge encoder $\mathcal{G}_g$ adopts the stacked hourglass architectures [1]. The enhanced local feature, well-aligned semantic global feature and attention-based relative spatial feature are represented by 11-dim vector, 12-dim vector and 8-dim vector respectively, and the multi-scale features are represented by 31-dim vector. The implicit function $\mathcal{IF}$ is represented by multi-layer perceptrons (MLPs). The number of neurons in each MLP layer is: 31, 512, 256, 128, and 1, with skip connections at the 2nd, 3rd, and 4th layers. We train the neural network model using the RMSprop optimizer with a learing rate starting from 1e-4. The learning rate is updated using an MultiStepLR scheduler in 3rd and 8th epcoh by multiplying with the factor 0.1. Training the model takes about 35 hours on one NVIDIA 3090 GPU for 10 epoches. During inference, we use PyMAF [7] for pose and shape recovery, and use the optimization approach of ICON [6].

1.3 Training

We need to separately train three networks, namely a cloth normal map prediction network $\mathcal{G}_n$, an image encoder $\mathcal{G}_g$ and a comprehensive network $\mathcal{G}_{MS}$.

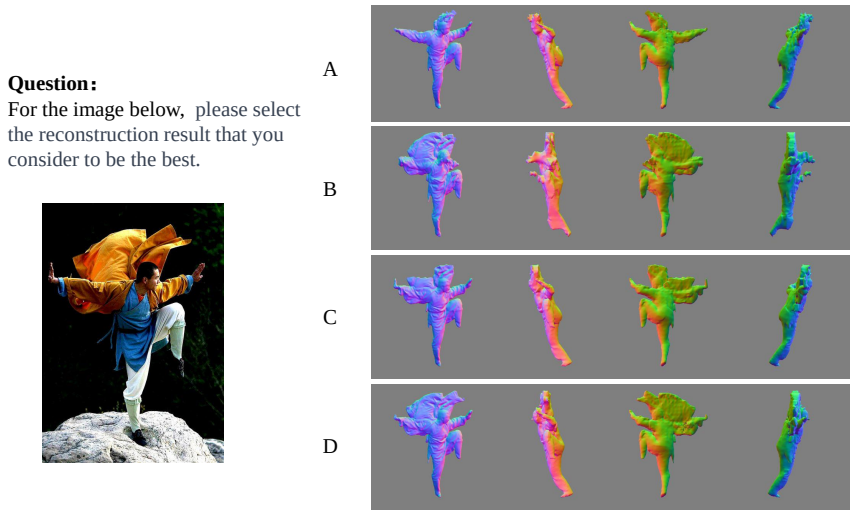

Fig. 3. The design of the questionnaire question.

Training loss. Like ICON [6], we train the cloth normal map prediction network $\mathcal{G}_n^v$ by minimizing the following loss:

$$\mathcal{L}_N^v = \mathcal{L}_{pixel}^v + \lambda_{VGG}\mathcal{L}_{VGG}^v \tag{1}$$

where $v = \{front, back\}$. $\mathcal{L}_{pixel}^v = \left|\mathcal{N}_c^v - \hat{\mathcal{N}}_c^v\right|$, and $\mathcal{L}_{VGG}^v$ is a perceptual loss weighted by λ_{VGG}.

As the iamge encoder $\mathcal{G}_g$ can be embedded into the architecture of PIFu [4], we choose to train $\mathcal{G}_g$ with the loss in PIFu:

$$\mathcal{L}_g = \frac{1}{n}\sum_{i=1}^{n}\left(f_g\left(f_{img}\left(x_i\right), z\left(X_i\right)\right) - f^*\left(X_i\right)\right)^2 \tag{2}$$

where n is the number of query points. $z\left(\cdot\right)$ returns the depth value of a query point and x_i is the projection of X_i on the 2D image. $f_{img}\left(\cdot\right)$ returns the pixel-aligned feature. $f_g\left(\cdot\right)$ is the implicit function used for occupancy prediction, and $f^*\left(\cdot\right)$ returns the ground truth occupancy of a query point, specified by equation .

$$f^*(X_i) = \begin{cases} 1 & \text{if } X_i \text{ inside the human mesh,} \\ 0 & \text{else.} \end{cases} \tag{3}$$

Once the training of $\mathcal{G}_n$ and $\mathcal{G}_g$ is done, We combine them with $\mathcal{G}_l$, $\mathcal{G}_r$ and $\mathcal{IF}$ to form a complete network, referred to as $\mathcal{G}_{MS}$, which is trained with the following loss:

$$\mathcal{L}_{MS} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{IF}\left(MSF\left(X_i\right)\right) - f^*\left(X_i\right)\right)^2 \tag{4}$$

1.4 Perceptual study

Due to the lack of ground-truth mesh. we perform a perceptual study to evaluate the quality of the reconstructed clothed 3D humans from in-the-wild images. We conduct a perceptual study in the form of a questionnaire. Each question in the questionnaire includes four options representing the reconstruction results of PIFu [4], PaMIR [8], ICON [6], and RICH respectively. The reconstruction is shown from 4 different views, that is $\{0°, 90°, 180°, 270°\}$. Fig. 3 illustrates the design of the questionnaire question. It is important to note that the names of the methods do not appear in the questionnaire to ensure fairness. Participants were asked to choose the best result that represents the clothing and body shape of the human in the image. And we compute the chances that participants prefer the SOTA methods over RICH based on the questionnaire results.

References

1. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3d human body reconstruction from a single image via volumetric regression. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
2. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
3. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501 (2020)
4. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019)
5. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 84–93 (2020)
6. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022)
7. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11446–11456 (2021)
8. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE transactions on pattern analysis and machine intelligence **44**(6), 3170–3184 (2021)