Supplementary Material for: Real-time 3D Human Reconstruction and Rendering System from a Single RGB Camera

Yuanwang Yang*, Qiao Feng*, Yu-Kun Lai, and Kun Li[†]

CCS CONCEPTS

• **Computer systems organization** → **Real-time systems**; • **Computing methodologies** → *Computer vision*; Shape modeling.

KEYWORDS

3D appearance, 3D human reconstruction and rendering, a single RGB camera, real-time

ACM Reference Format:
Yuanwang Yang*, Qiao Feng*, Yu-Kun Lai, and Kun Li[†]. 2024. Supplementary Material for: Real-time 3D Human Reconstruction and Rendering System from a Single RGB Camera. In *Proceedings of SA Technical Communications'24.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3681758.3697993

A COMPARISON ON PUBLIC DATASETS

A.1 Implement Details.

Metrics. We use three widely used metrics for images to quantitatively evaluate our method: learned perceptual image patch similarity (LPIPS) [Zhang et al. 2018] that uses AlexNet [Krizhevsky et al. 2012] to extract features, structural similarity (SSIM) [Wang et al. 2004], and peak signal-to-noise ratio (PSNR) [Sara et al. 2019].
Data processing. We selected 105 models from the THuman 2.0 dataset and the 2k2k dataset as test sets, respectively. We used high dynamic range images (HDRIs) for realistic image-based ambient lighting and as backgrounds to simulate complex real-world lighting environments. All the HDRis we used are from Poly Haven[1]. For each model, we rendered images of 16 views as GT images for evaluation.

A.2 Qualitative results.

Figures 1 qualitatively compare our baseline methods on the 2k2k dataset [Han et al. 2023]. PIFu demonstrates the ability to estimate reasonably accurate colors, but its performance is hindered by geometric estimation limitations, resulting in poor results. SHERF can produce results through accurate postures, but it faces challenges in producing desirable results for individuals wearing loose clothing, mainly because it relies heavily on the SMPL prior. R^{2}Human can obtain relatively accurate geometry and apply it to loose clothing, but mistakenly treats the shadows generated by light as the color of the human body itself and rely on precise masks. Our method guarantees accurate geometry and is suitable for a variety of complex lighting environments in real life.

A.3 Quantitative evaluations.

Tab. 1 shows the performance of the method when the novel view is in close proximity to the source view (with a difference of ≤ 45°), and Tab. 2 shows the performance when the novel view is significantly different from the source view (with a difference of ≥ 90°). Some values are inconsistent with those in R^{2}human because the input images are augmented with ambient lighting to simulate the real environment. It is evident that our method outperforms others across all evaluation metrics.

Table 1: Quantitative rendering results in close views.

Method	THuman2.0			2k2k		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
PIFu [Saito et al. 2019]	0.1310	0.8658	23.33	0.1147	0.8672	23.37
SHERF [Hu et al. 2023]	0.0992	0.8890	25.50	0.1067	0.8802	24.52
R^{2}Human [Yang et al. 2024]	0.0673	0.9199	25.85	0.0812	0.8952	24.63
Ours	**0.0393**	**0.9436**	**29.74**	**0.0527**	**0.9214**	**27.85**

Table 2: Quantitative rendering results in divergent views.

Method	THuman2.0			2k2k		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
PIFu [Saito et al. 2019]	0.1571	0.8433	21.90	0.1439	0.8449	21.53
SHERF [Hu et al. 2023]	0.1158	0.8733	23.85	0.1227	0.8663	22.83
R^{2}Human [Yang et al. 2024]	0.0865	0.8890	24.03	0.1052	0.8674	23.57
Ours	**0.0568**	**0.9215**	**28.00**	**0.0796**	**0.8902**	**25.37**

B FAILURE CASES

Due to the absence of an explicit segmentation of skin parts and clothing parts, there are cases where the network may produce inaccurate color estimations on specific body parts. Fig. 2 illustrates some examples of these failure cases.

In addition, there are two main issues encountered in the real-time system.

1) Occasional edge discrepancies, such as those around shoes or heads, stem from errors generated by the open-source segmentation algorithm employed. Addressing this concern can be achieved by integrating more accurate segmentation algorithms in future work.

2) Instances of leg bending when standing may occur due to limitations in the open-source SMPL estimation algorithm when processing single RGB inputs.

*Equal contribution.
[†]Corresponding author.
[1] https://polyhaven.com/

SA Technical Communications'24, December 03-06,2024, Tokyo, Japan

ACM ISBN 979-8-4007-1140-4/24/12
https://doi.org/10.1145/3681758.3697993

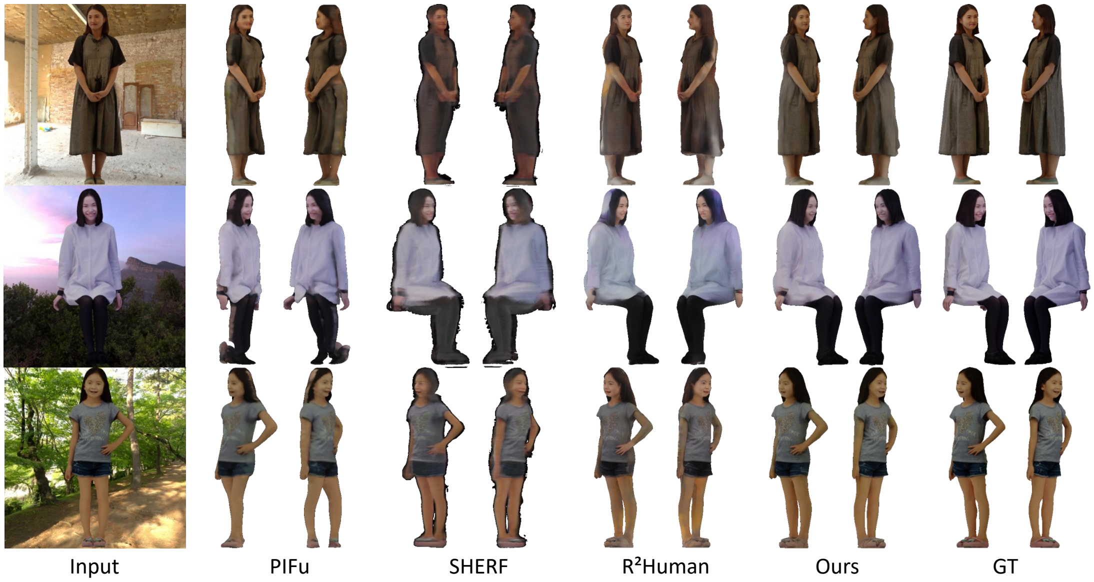

Figure 1: Novel view rendering on 2k2k dataset.

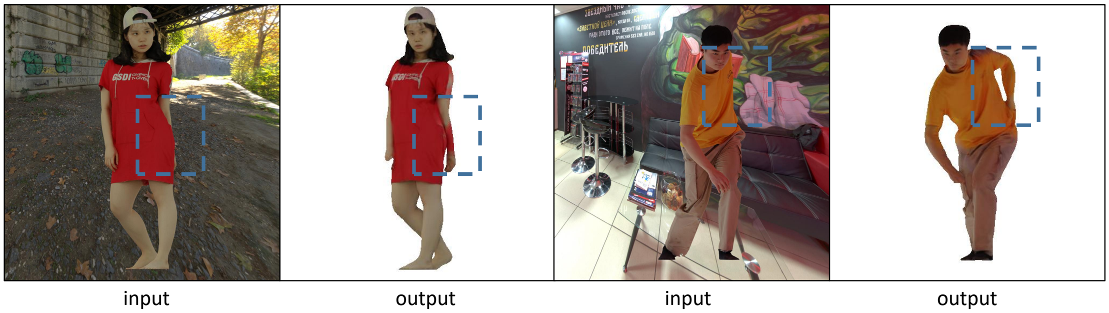

Figure 2: Some examples of failure cases.

In the future, we plan to explore additional supervision methods to enhance the network's ability to recognize human body parts more effectively and thereby improve the quality of the output images. Furthermore, adopt more appropriate algorithms to mitigate pose issues in real-time system, thereby improving the rendering results quality.

C MORE SYSTEM RESULTS

Figures 3 and 4 showcase more real-time reconstruction results. These results demonstrate the effectiveness of our method for a variety of populations in real world.

D MORE VISUAL RESULTS

We provide additional qualitative results produced from a single input image of THuman2.0 [Yu et al. 2021] in Fig. 5. These results demonstrate the effectiveness of our method in producing high-quality images, particularly in handling challenging poses and a variety of body types.

E RESULTS ON 1K RESOLUTION

Our method in the main paper was trained and tested on a resolution of 512 × 512. We also explored the performance of the method at 1K resolution. Specifically, we have remade a 1K dataset and trained our model on it. Fig. 6 shows the results of our method at high resolution. It can be seen that high-resolution images effectively improve the clarity of the results and enhance the visual effect of the results.

REFERENCES

Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. 2023. High-fidelity 3D Human Digitization from Single 2K Resolution Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12869–12879.

Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. 2023. SHERF: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9352–9364.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed

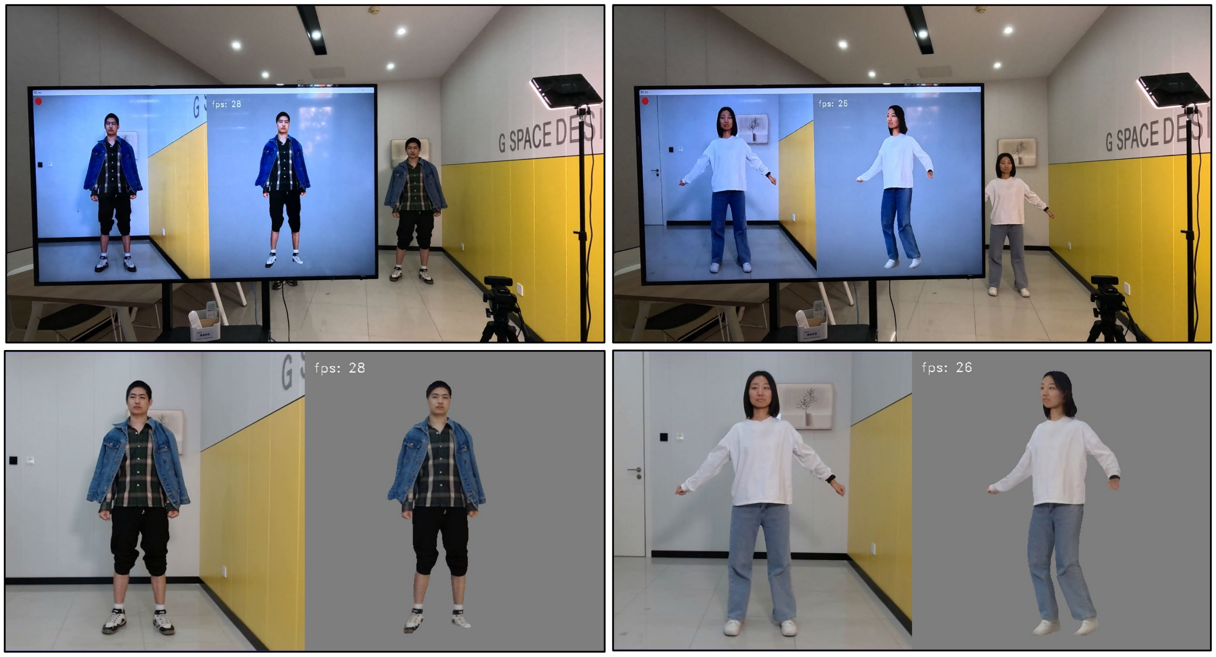

Figure 3: More real-time rendering results by our method.

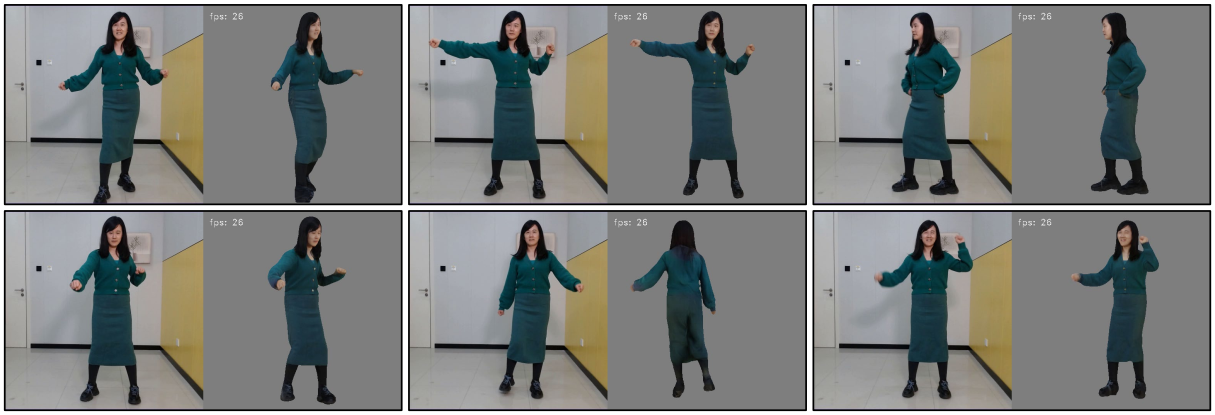

Figure 4: More real-time rendering results by our method.

human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2304–2314.

Umme Sara, Morium Akter, and Mohammad Shorif Uddin. 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications* 7, 3 (2019), 8–18.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Yuanwang Yang, Qiao Feng, Yu-Kun Lai, and Kun Li. 2024. R2Human: Real-Time 3D Human Appearance Rendering from a Single Image. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5746–5756.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

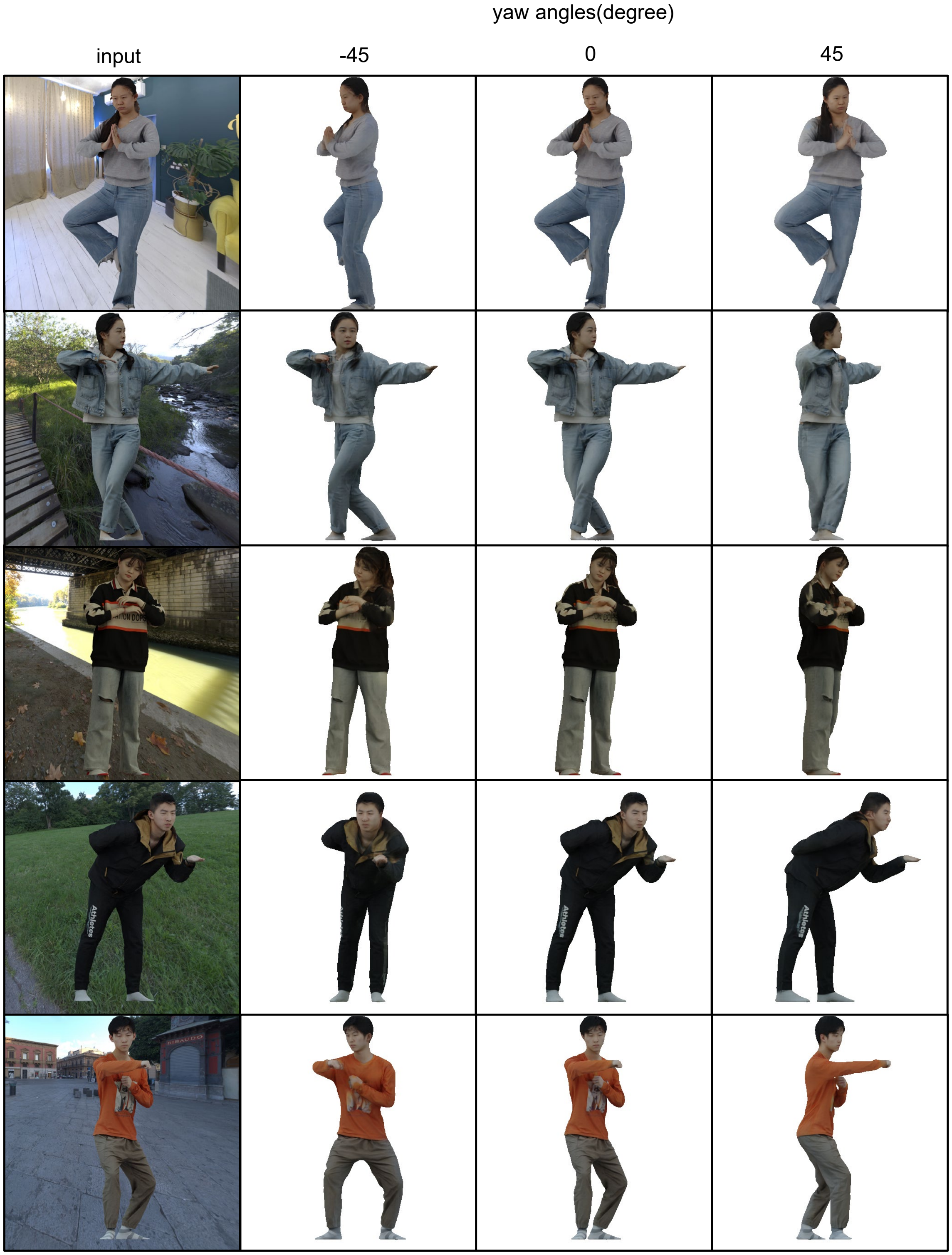

Figure 5: Additional qualitative results generated by our method.

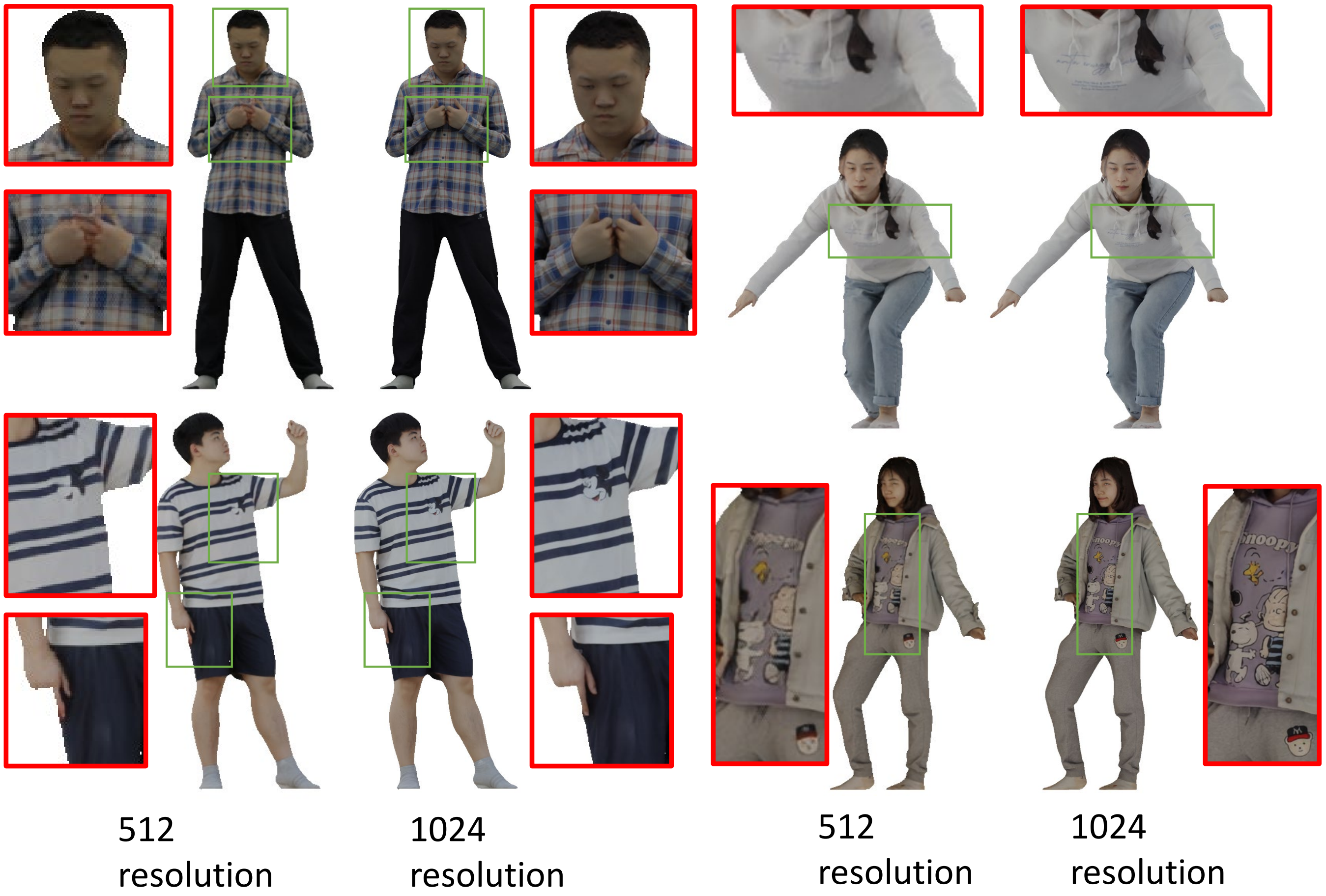

Figure 6: Comparison of results (input view) at low resolution (512×512) and high resolution (1024×1024).