Real-time 3D Human Reconstruction and Rendering System from a Single RGB Camera

Yuanwang Yang*
Qiao Feng*
Tianjin University
China
yyw@tju.edu.cn
fengqiao@tju.edu.cn

Yu-Kun Lai
Cardiff University,
United Kingdom
laiy4@cardiff.ac.uk

Kun Li†
Tianjin University
China
lik@tju.edu.cn

Figure 1: Our system achieves real-time high-quality 3D human reconstruction and rendering with a single RGB camera at 28+ FPS and exhibits an end-to-end latency of 75ms. It can be applied to 3D holographic displays and virtual reality environments.

Abstract

Transforming 2D human images into 3D appearance is essential for immersive communication. In this paper, we introduce a low-cost real-time 3D human reconstruction and rendering system with a single RGB camera at 28+ FPS, which guarantees both real-time computing speed and realistic rendering results. We use WebRTC to transmit captured images over the Internet for low-latency remote communication. In addition, we design a lighting-robust rendering approach to enhance the lighting perception and generalization ability of the model. Experimental results show that our system achieves high-quality real-time 3D human reconstruction and rendering for different persons wearing various clothes, even with challenging poses. Our system makes use of only a common USB webcam and a consumer-level GPU with real-time performance and high-fidelity results, which provides a consumer-accessible immersive telepresence solution.

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

*Equal contribution.
†Corresponding author.

SA Technical Communications'24, December 3-6,2024, Tokyo, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Yuanwang Yang, Qiao Feng, Yu-Kun Lai, and Kun Li. 2024. 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, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

3D human reconstruction and rendering effectively enhance the visual quality of camera-captured data by transforming 2D images into immersive 3D representations. With the rise of consumer-level virtual reality (VR) hardware, there is an increasing demand for 3D display technologies to elevate user experiences. It is anticipated that 3D video may eventually supplant 2D video as the mainstream format for public viewing.

However, current 3D systems are often complex and high-cost, exemplified by systems like Project Starline [RUSSELL et al. 2021], FloRen [Shao et al. 2022], and GPS-Gaussian [Zheng et al. 2024], which necessitate multi-view cameras for comprehensive input to achieve high-quality results. This complexity results in high production costs, intricate workflows, and limited accessibility for everyday consumers. In contrast, single-image input methods such as PIFu [Saito et al. 2019] and SHERF [Hu et al. 2023] aim to simplify the process but suffer from slow inference speeds and inadequate support for real-time immersive applications due to extensive computational requirements.

Our goal is to offer a cost-effective real-time solution for high-quality reconstruction and realistic rendering using just a single RGB camera. This demonstration integrates a real-time 3D reconstruction method [Feng et al. 2022] with a 3D appearance rendering framework [Yang et al. 2024], and builds a concise, low-cost, real-time 3D human reconstruction and rendering system, tailored for consumer applications. We also design a lighting-robust data processing method to enhance the model's perception of lighting conditions, ensuring that our results are applicable to real-world scenarios.

2 Approach

The schematic diagram of our system is depicted in Fig. 2. Broadly, our system comprises video capture, reconstruction, novel view rendering, network transmission, and display components. We utilize a USB webcam for video capture, significantly reducing system costs and deployment complexity, thereby enhancing accessibility for ordinary users. Next, we utilize the predicted Fourier Occupancy Field (FOF) [Feng et al. 2022] to extract high-quality human geometry for reconstruction from the foreground of the captured image, as it is pixel-aligned and suitable for efficient processing. We enhance the robustness of FOF by integrating SMPL geometry information corresponding to the person being captured, resulting in a more reasonable human mesh. Subsequently, depth maps, normal maps, and other pertinent data are derived through mesh rendering. We employ the approach described in R^2Human [Yang et al. 2024] to generate high-quality novel view images. Finally, the processed data is transmitted to the display component for presentation in applications. This system enables cost-effective deployment and robust performance suitable for diverse user scenarios.

2.1 Implicit Neural Reconstruction

In order to achieve high-quality geometric reconstruction, we use FOF [Feng et al. 2022] to represent the geometry of human beings in 3D space. Due to the depth ambiguity of a single image input and possible occlusion, relying only on input image prediction often fails to obtain accurate results, which easily leads to limb loss or artifacts. However, human structures are relatively stable, so we enhance the robustness of the network by adopting SMPL parameterized models [Loper et al. 2015]. SMPL template can provide the corresponding pose and shape information of human, effectively reducing the difficulty of network prediction of complex poses.

2.1.1 Image segmentation and SMPL prediction. Given an RGB image, our goal is to quickly and accurately extract human mask and SMPL parameters to meet the high demand for real-time performance of the system. To this end, we built a lightweight and efficient real-time model using TensorRT based on RVM [Lin et al. 2022] and HMR2.0 [Kanazawa et al. 2018]. TensorRT's optimization capabilities allow us to significantly improve inference speed while maintaining precision, thus effectively responding to real-time requirements in dynamic scenarios.

2.1.2 FOF with SMPL prior. For each point in 2D space, FOF [Feng et al. 2022] uses a Fourier series $f(z) = a_0/2 + \sum_{n=1}^{\infty}(a_n cos(nz\pi) + b_n sin(nz\pi))$ to represent the space occupancy of the ray along the z-axis. Here a_n and b_n are coefficients of the basis functions $\cos(nx)$ and $\sin(nx)$. Therefore, by calculating the first $2n$ coefficients of the 1D occupancy signal $f(z)$, we can approximate the occupancy of each position along the z-axis ray. By combining the coefficients of each point in 3D space, we can effectively represent the occupancy field of a 3D model using a 2D image with $2n$ channels.

In order to achieve fast and accurate processing, we set $n = 15$, use a custom CUDA operation to convert the predicted SMPL mesh into a FOF representation, and connect it with the RGB image as input. This pixel-aligned method ensures that the network can more accurately identify the corresponding relationship between the SMPL template and the input image, effectively eliminating depth uncertainty without adding too many additional computational requirements, ensuring the real-time performance of the network.

2.1.3 Mesh extraction. We perform marching cubes algorithm [WE 1987] to extract the mesh surface from the iso-surface of the Fourier occupation field at the threshold of 0.5. To enhance extraction speed, we have also implemented custom atomic operations in CUDA to accelerate the execution of the marching cubes algorithm.

2.2 Novel View Rendering

Our goal is to get a novel view rendering of the whole body from a single RGB image. We employ R^2Human adapted from [Yang et al. 2024] as the basis for our demo. Given that real-world data is often affected by complex lighting conditions, a straightforward shading rendering approach can cause the network to ignore lighting-induced color changes, resulting in distorted rendering results. To address this challenge, we design a lighting-robust data processing method to improve the model's adaptability for complex lighting environments in real-world.

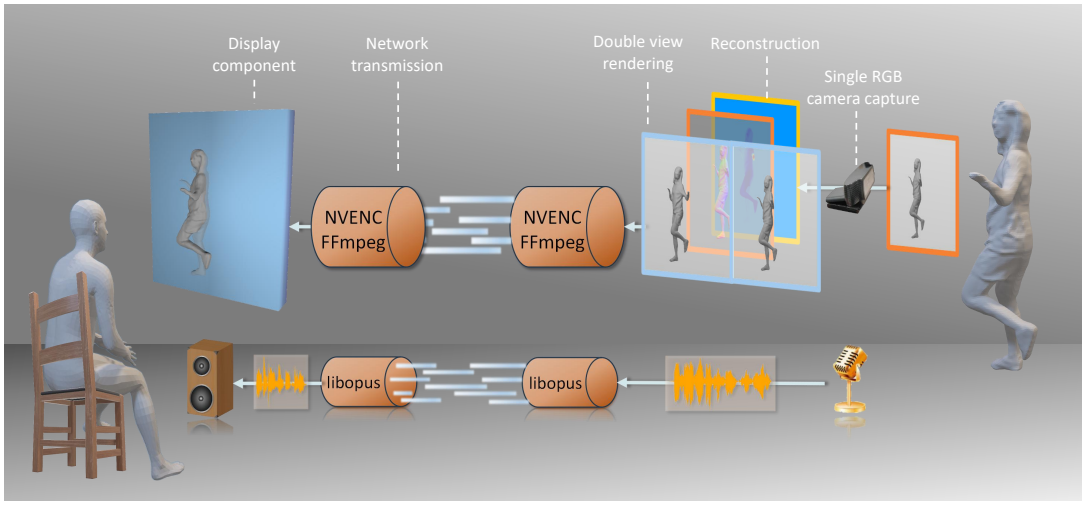

Figure 2: Schematic diagram of the overall system. We only use one USB webcam for video capture, and then perform high-quality human geometry reconstruction and double view rendering. Finally, the results are transmitted over the network to the display component for display to the user.

2.2.1 Description of R^2 Human. In novel view rendering tasks, there are typically two kinds of approaches: One approach [Hu et al. 2023; Saito et al. 2019] encodes image features and depth values from the input view and then decoding the color of each 3D point, but requiring dense sampling, leading to high computational costs and challenges in real-time applications. The other approach warps image features for decoding in 2D, enhancing efficiency but facing ambiguities that necessitate multi-view images to resolve occlusions. R^2Human introduces the Z-map, which captures depth values of visible points from the source view, merging benefits of both methods to improve color inference accuracy in occluded areas while reducing reliance on multi-camera setups. This method allows for efficient simultaneous feature decoding using CNNs, balancing real-time performance and high authenticity with just one RGB input.

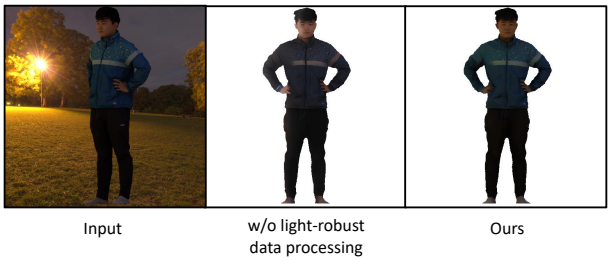

Figure 3: The effectiveness of lighting-robust data processing.

2.2.2 Lighting-robust data processing. Our backbone is trained on synthetic data, collecting 526 high-quality human scans from the THuman 2.0 dataset [Yu et al. 2021], which include a wide range of clothing, poses, and shapes. We randomly selected 368 models as the training dataset and 105 models as the test set, and the remaining models are used as the validation set. We use high dynamic range images (HDRIs) for realistic image-based ambient lighting and as the background, thereby enhancing the model's adaptability to realistic lighting environments. Rendering 32 view high-fidelity images for each model separately in 16 randomly selected HDRi environment textures, yields a total of a dataset containing approximately 269K images. Fig. 3 shows the results without and with apply the lighting-robust data processing. It can be seen that using light-robust data processing can effectively improve the accuracy of rendering colors.

2.3 Data transmission

We employ H.265 encoding to convert the captured images into a high-efficiency format, which is then transmitted over the Internet using WebRTC technology [Johnston and Burnett 2012] for low-latency real-time transmission. We utilize Hardware-based encoder (NVENC) and decoder (NVDEC) to handle the end-to-end encoding and decoding, thus significantly reducing the burden on the system. This allows CUDA cores to prioritize accelerating other critical operations, thereby enhancing overall system performance and fluidity.

Our approach requires only the encoding and transmission of image and audio data. During the encoding process, we transmit images at a resolution of 512×1024 and capture dual-channel stereo audio at a sample rate of 48000Hz. Simultaneously inserting timestamps ensures synchronization between video and audio.

3 Experiments

System setup and results. Our system is implemented in C++. For our prototype capture setup, we utilize a single USB webcam (Logitech C920 PRO) for video capture. Reconstruction and 3D appearance rendering are handled by an Nvidia RTX-4090 graphics card, achieving speeds of up to 28 FPS. Alternatively, a computer equipped with an Nvidia RTX-3090 graphics card achieves speeds of up to 22 FPS. Figure 4 illustrates some of the results.

Figure 4: Examples of the results on the real captured images by our system.

Network details. We mainly train the reconstruction network and the rendering network in our demo system. The reconstruction network is based on HRNet-W32-V2, and the input channel number is changed to 34 to input SMPL template information. The rendering network uses an encoder-decoder structure, where the encoder uses a network based on HRNet-W32-V2, while the decoder uses a UNet-based network implementation. The resolution of the input image is 512^2, and the resolutions of the reconstructed FOF and voxel grid are 512^2 and 512^3, respectively. The final output is a rendered image with a resolution of 512^2.

4 Conclusion and Limitation

Conclusion. In this paper, we present a low-cost monocular real-time immersive telepresence system that can reconstruct and render a 3D human at 28 + FPS in a video stream captured by a single RGB camera, and can serve as a backbone for future low-cost telepresence applications.

Limitation. Although the introduction of the SMPL prior improves the accuracy of the reconstruction, it also introduces the error caused by the SMPL estimation. Due to the depth ambiguity caused by the single image input, sometimes the resulting SMPL will have bent legs. Our pipeline works at 28FPS, and using more GPUs can further improve our frame rate. In the future, techniques such as diffusion models can be used to enhance the rendering results and achieve a more pleasant visual experience.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62122058 and 62171317), and the Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQJC00040).

References

Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. 2022. FOF: learning fourier occupancy field for monocular real-time human reconstruction. *Advances in Neural Information Processing Systems* 35 (2022), 7397–7409.

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.

Alan B Johnston and Daniel C Burnett. 2012. *WebRTC: APIs and RTCWEB protocols of the HTML5 real-time web*. Digital Codex LLC.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7122–7131.

Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2022. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 238–247.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *Acm Transactions on Graphics* 34, Article 248 (2015).

IAN RUSSELL, STEVEN M SEITZ, and KEVIN TONG. 2021. Project Starline: A high-fidelity telepresence system. (2021).

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2304–2314.

Ruizhi Shao, Liliang Chen, Zerong Zheng, Hongwen Zhang, Yuxiang Zhang, Han Huang, Yandong Guo, and Yebin Liu. 2022. Floren: Real-time high-quality human performance rendering via appearance flow using sparse rgb cameras. In *SIGGRAPH Asia 2022 Conference Papers*. 1–10.

LORENSEN WE. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *Computer graphics* 21, 1 (1987), 7–12.

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.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 2024. GPS-Gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19680–19690.