PHOTONS: Pose-Free Human-Centric Photo-Realistic Real-Time Novel View Synthesis from Sparse Views

Yongyang Cheng, Boqin Qin, Zhao Hui, Xu Chen, Tao Zhang, Shang Sun, Haiquan Kang, Xiaojie Xu, Junwei Lv, Lei Yang, Xinyu Liu, Feng Jiang*

International Business Division, China Telecom Cloud Technology Co., Ltd.
chengyy2@chinatelecom.cn (Y. Cheng), qinbq@chinatelecom.cn (B. Qin), jiangfeng@chinatelecom.cn (F. Jiang)

Abstract

We present PHOTONS (**P**ose-Free **H**uman-Centric Ph**o**to-Realistic Real-**T**ime N**o**vel View Sy**n**thesis from **S**parse Views), a real-time framework for novel view synthesis without requiring camera calibration. Our method reconstructs consistent 3D Gaussian point clouds and synthesizes 2K photo-realistic novel views from arbitrary numbers (≥ 2) of freely placed cameras. PHOTONS faithfully renders dynamic human bodies amid complex backgrounds, including interactive object manipulation and fine-grained details (e.g., hair strands), while maintaining 25 FPS throughput on commodity GPU like NVIDIA RTX 4090. By combining pose-free spatial point cloud reconstruction with Gaussian parameter estimation, our method demonstrates strong resilience to occlusions and camera perturbations. Additionally, we develop a 3D stereo system that drastically reduces setup complexity compared to existing solutions. Experiments on public and custom datasets show that PHOTONS outperforms state-of-the-art methods in both efficiency and visual quality.

Introduction

Novel view synthesis (Mildenhall et al. 2021) aims to render photo-realistic images from new viewpoints using multi-view RGB inputs. Recent 3D Gaussian Splatting (3DGS) methods (Kerbl et al. 2023) set new performance standards, yet human-centric telepresence remains challenging due to dynamic viewpoints, occlusions, fine details, and real-time constraints on commodity GPUs.

Existing approaches have key limitations: Optimization-based 3DGS (Zhang et al. 2025) achieves high fidelity but is too slow for real-time use. Feed-forward methods using stereo/multi-view depth (Zheng et al. 2024; Liu et al. 2024; Chen et al. 2024) or human priors (Xiao et al. 2025) run in real time but depend heavily on accurate calibration. Pose-free reconstruction (Ye et al. 2024; Jiang et al. 2025) reduces calibration needs but still struggles with occlusion and scalability. VGGT (Wang et al. 2025) predicts dense 3D from unposed views, but performs poorly on human scenes due to limited human-specific data and requires ground-truth geometry, making it incompatible with RGB-only settings.

*Corresponding author.

We introduce **PHOTONS**, a real-time, pose-free novel view synthesis framework. By integrating point cloud reconstruction and Gaussian parameter estimation into a feed-forward pipeline, PHOTONS supports arbitrary camera configurations and robust multi-view rendering.

Our contributions are as follows:

- A pose-free multi-view synthesis framework producing 2K-resolution novel views with state-of-the-art quality.
- A global feature-driven GPE module enabling RGB-only training while reducing model complexity.
- A high-resolution RGB feature shortcut for fine-grained point cloud upsampling and rendering.
- A simplified 3D stereo setup requiring no calibration, improving deployability compared to Google Beam (Lawrence et al. 2024; Google 2025) and Tele-Aloha (Tu et al. 2024).

PHOTONS eliminates reliance on camera extrinsics and ground-truth geometry, substantially lowering deployment barriers and enabling scalable immersive 3D stereo experiences.

System Architecture

Figure 1 provides an overview of the PHOTONS framework, illustrating the pipeline from unposed source images to stereo novel views.

Dataset and System Setup

To enable novel view synthesis in human-centric scenes, we construct a pose-free multi-view RGB-only dataset. Each training sample contains four source views (see Step 1 in Figure 1): three freely placed input cameras and one designated main camera that defines the global coordinate origin. In addition, five fixed reference cameras capture ground-truth supervision. Only the extrinsics of reference views relative to the main camera are required. All the camera intrinsics are factory-specified and fixed, so no additional calibration is needed. Our training set consists of 100,000 image groups generated from 3D human models from the public THuman dataset (Yu et al. 2021) and 100,000 image groups captured with our own camera rig. During inference, only source views are used. Experiments are conducted on two NVIDIA RTX 4090 GPUs with balanced parallel workload.

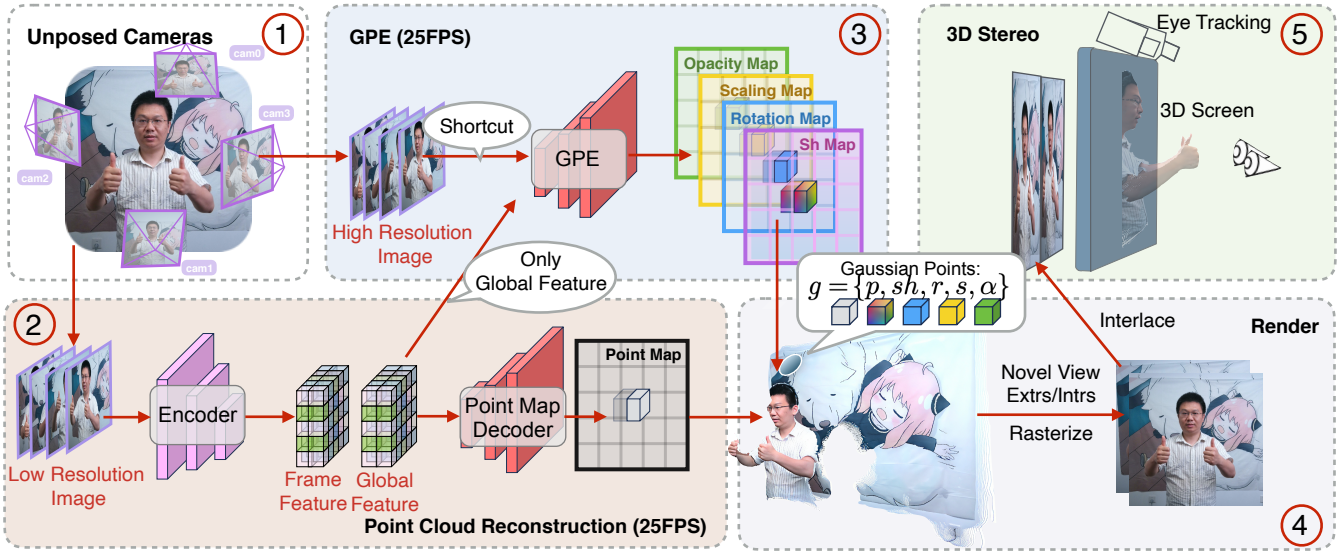

Figure 1: System architecture, illustrating how to achieve the 3D stereo from unposed cameras through steps 1 to 5.

Point Cloud Reconstruction

As shown in Step 2 of Figure 1, we employ a VGGT backbone to encode unposed low-resolution images into frame and global features, which are then decoded into dense point maps (p). Trained solely with RGB supervision from the downstream novel view synthesis task, the backbone enhances human geometry without requiring geometric ground truth. For real-time performance, point maps are predicted at low resolution and bilinearly upsampled, achieving high-quality rendering with an average latency of 34.59 ms.

Gaussian Parameter Estimation

We introduce a GPE module to predict 3DGS parameters — including spherical harmonics (sh), rotation (r), scale (s), and opacity (α) — for 3DGS rendering (Step 3 in Figure 1). The module decodes parameters directly from global features, omitting frame features for efficiency. A shortcut connection encodes high-resolution images into fine-grained features, which are fused with global features to preserve detail while sustaining real-time performance. Running in parallel with point cloud reconstruction on the other GPU, the inference latency is 36.84 ms.

Stereo Rendering

Using high-resolution point maps (Step 2) and 3DGS parameters (Step 3), Step 4 in Figure 1 renders novel-view images for the left and right eyes, specifically determined via eye-tracking (Step 5). The images are interlaced for 3D display, producing a binocular stereo effect, and the system supports motion parallax by dynamically updating views as the viewer's gaze shifts. Running on the same GPU as the GPE module, the rendering stage adds only 3.3 ms, resulting in a total latency of 40.14 ms.

Figure 2: Comparison of rendered and ground truth image.

	Pose-free	PSNR↑	SSIM↑	LPIPS↓	FPS↑
MVSplat	×	27.24	0.869	0.151	3
MVSGaussian	×	28.29	0.907	0.121	23
NoPoSplat	✓	24.63	0.815	0.213	6
AnySplat	✓	25.35	0.864	0.166	2
Ours	✓	**30.65**	**0.924**	**0.101**	**25**

Table 1: Comparisons with state-of-the-art methods.

Results and Conclusions

We evaluate PHOTONS and competing methods by rendering novel views from reference poses and comparing them against ground truth. As shown in Table 1, it outperforms all baselines across every metric while maintaining higher FPS on our dataset captured using our own camera rig. We observe similar trends on the dataset generated from the THuman dataset. Figure 2 highlights its superior rendering quality. By combining pose-free point cloud reconstruction with the GPE module, our approach remains robust to occlusions and camera perturbations, delivering a high-resolution, real-time, and practical 3D stereo system.

Acknowledgments

We thank China Telecom Cloud Technology Co., Ltd. for their continuous support during the development of this work. We also extend our sincere appreciation to Yingting Wang for assisting in the preparation and filming of the demonstration video.

References

Chen, Y.; Xu, H.; Zheng, C.; Zhuang, B.; Pollefeys, M.; Geiger, A.; Cham, T.-J.; and Cai, J. 2024. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, 370–386. Springer.

Google. 2025. Google Beam. https://blog.google/technology/research/project-starline-google-beam-update. Accessed: 2025-08-06.

Jiang, L.; Mao, Y.; Xu, L.; Lu, T.; Ren, K.; Jin, Y.; Xu, X.; Yu, M.; Pang, J.; Zhao, F.; et al. 2025. AnySplat: Feed-forward 3D Gaussian Splatting from Unconstrained Views. *arXiv preprint arXiv:2505.23716*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4): 139–1.

Lawrence, J.; Overbeck, R.; Prives, T.; Fortes, T.; Roth, N.; and Newman, B. 2024. Project starline: A high-fidelity telepresence system. In *ACM SIGGRAPH 2024 emerging technologies*, 1–2.

Liu, T.; Wang, G.; Hu, S.; Shen, L.; Ye, X.; Zang, Y.; Cao, Z.; Li, W.; and Liu, Z. 2024. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference on Computer Vision*, 37–53. Springer.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Tu, H.; Shao, R.; Dong, X.; Zheng, S.; Zhang, H.; Chen, L.; Wang, M.; Li, W.; Ma, S.; Zhang, S.; et al. 2024. TeleAloha: A telepresence system with low-budget and high-authenticity using sparse rgb cameras. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Xiao, J.; Zhang, Q.; Nie, Y.; Zhu, L.; and Zheng, W.-S. 2025. RoGSplat: Learning Robust Generalizable Human Gaussian Splatting from Sparse Multi-View Images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5980–5990.

Ye, B.; Liu, S.; Xu, H.; Li, X.; Pollefeys, M.; Yang, M.-H.; and Peng, S. 2024. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 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.

Zhang, Z.; Kaufmann, M.; Xue, L.; Song, J.; and Oswald, M. R. 2025. ODHSR: Online Dense 3D Reconstruction of Humans and Scenes from Monocular Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21824–21835.

Zheng, S.; Zhou, B.; Shao, R.; Liu, B.; Zhang, S.; Nie, L.; and Liu, Y. 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.