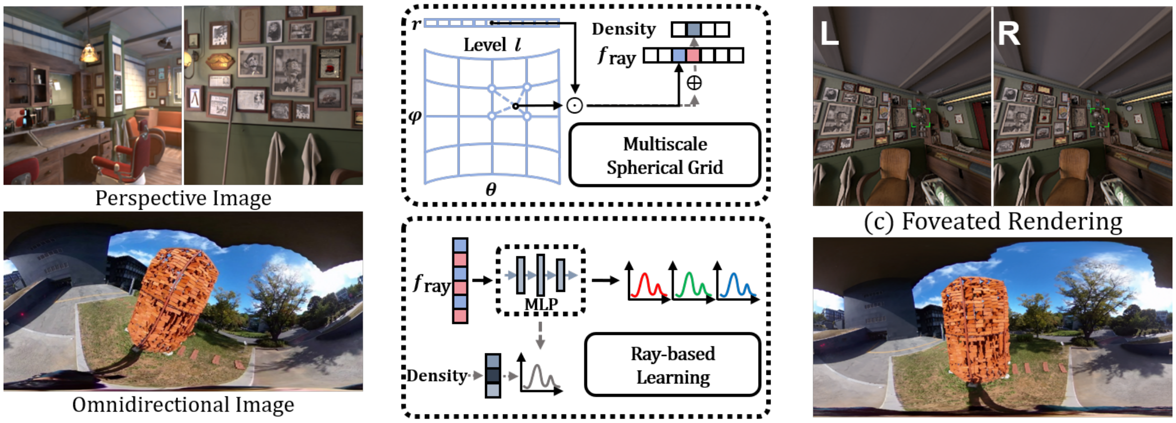
2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3315-1484-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/VRW66409.2025.00265

Ray-based Multiscale Spherical Grid for Egocentric Viewing

Weichao Song* Bingyao Huang†
Southwest University



(a) Supported Input (b) MSTD (c) Foveated Rendering (d) 360° Rendering

Figure 1: Illustration of proposed method and target tasks. (a) Proposed method supports both of perspective and omnidirectional images. (b) Using multiscale spherical tensor decomposition (MSTD) to encode the features of sampled rays and ray-based learning for acceleration. (c) Foveated rendering for real-time VR viewing. (d) Synthesizing omnidirectional images.

Abstract

Virtual reality (VR) and augmented reality (AR) demand real-time and high-quality egocentric viewing. Neural rendering can generate extremely high-quality novel views but requires expensive training time and renders too slow. To address these challenges, this paper proposes a method named multi-scale spherical tensor decomposition (MSTD). Its ability to represent egocentric neural 3D scenes also enables potential further applications in relevant domains. Our method can reconstruct scenes from multi-view images or omnidirectional images, outperforming baseline methods in training and rendering time as well as rendering quality.

Keywords: Neural Representation, Tensor Decomposition, Foveated Rendering, Egocentric Viewing

1 Introduction

Virtual reality (VR) applications demand high field-of-view (FoV) display performance, but existing methods [1,2,8] for image-based rendering often struggle with issues such as high latency, low image quality, and hardly handle occlusions. While neural rendering methods like NeRF [9] have shown promise in representing 3D scenes, traditional MLP-based approaches are slow for both training and rendering. Methods utilizing Cartesian feature grids [4,7] face challenges with egocentric scenes, and although spherical coordinate-based solutions [1,5,6] improve upon this, they still incur significant rendering time due to the large number of sampled points along rays. To address this, we introduce MSTD, which organizes sampled features in spherical coordinates. Unlike triplane methods, we employ a single feature grid, which reduces memory and encoding costs. While previous methods with coarse-density grids [4,5,7] are unstable and sensitive to scene content, our approach ensures faster and more stable training across different scenes. As shown in Fig. 1, our method supports real-time binocular foveated rendering [3,6,10], which accelerates rendering without sacrificing consistency. As table 1 shows, our method shows advantages in egocentric viewing.

Table 1: Comparison of egocentric viewing methods.

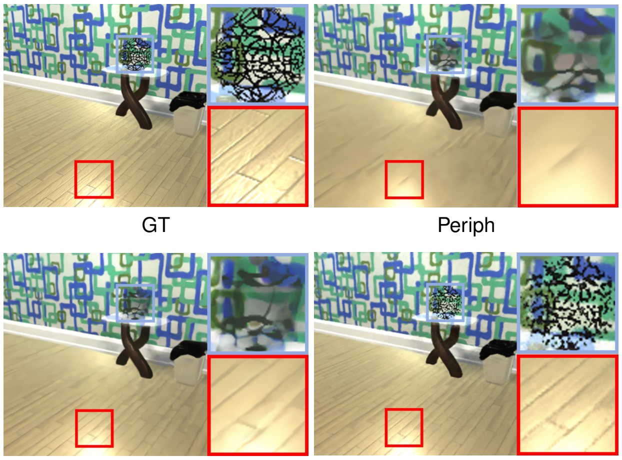
Method	Real-time rendering	Grid utilization rate	Training speed	Single MLP	Stable
NeRF [9]	No	Null	Slow	No	Yes
FoV-NeRF [6]	Yes	Null	Medium	No	Yes
TensoRF [4]	No	Low	Slow	Yes	No
NRFF [7]	No	Low	Slow	Yes	No
EgoNeRF [5]	No	High	Slow	Yes	No
MTSD (ours)	Yes	Medium	Fast	Yes	Yes

2 Method

Our method represents the scene using multi-scale spherical grids and a fully connected deep network. The input consists of the spherical coordinates and corresponding radii of the intersections of the sampled ray with each spherical layer, while the output is the color and density of each point along the sampled ray. To address the requirements for real-time rendering and egocentric viewing, our approach focuses on foveated rendering and omnidirectional rendering, incorporating stratified sampling, MSTD, and ray-based network inference. The proposed MSTD leverages concentric spherical coordinates to encode the features of spatial points from coarse to fine, gradually increasing the resolution of spherical grids to resolve the scale disparity inherent in spherical coordinates. To optimize training and rendering efficiency, we utilize a single network for calculating RGB and density, avoiding the need for multiple networks

*e-mail: swc4869@swu.edu.cn
†e-mail: bhuang@swu.edu.cn

979-8-3315-1484-6/25/$31.00 ©2025 IEEE
DOI 10.1109/VRW66409.2025.00265

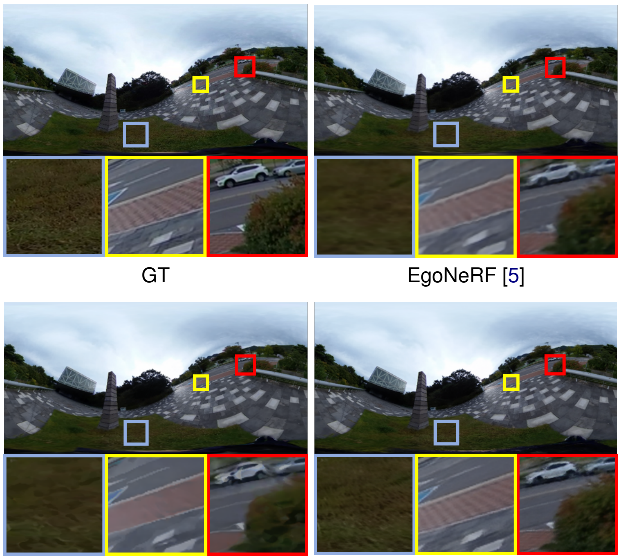

GT

Periph

Fovea

Ours

Figure 2: Comparative results between FoV-NeRF [6] and our method on the synthetic dataset with 40°.


GT

EgoNeRF [5]

FoV-NeRF [6]

Ours (4 + Field 4)

Figure 3: Comparative results of novel view synthesis on the outward-looking real-world omnidirectional dataset.

and ensuring consistent rendering across rays without the requirement of view direction input. Our experimental results demonstrate that MSTD outperforms traditional sine-cosine encoding in preserving high-frequency information. Regarding foveated rendering, our approach simplifies the process by using a single fine network for rendering different regions, eliminating the need for separate networks for coarse and fine details. This reduces training costs, improves rendering quality, and accelerates rendering speed, making it more suitable for our method.

3 Results

Table 2: Average time (ms) for each section. The perceptual model is proposed by FoV-NeRF [6], which synthesizes and renders three images. The data fluctuates within a range of 2ms.

Layers	Method	Encode (ms)	Infer (ms)	Render (ms)	Sum (ms)
3	FoV-NeRF	10.49	23.96	**6.10**	42.72
	Ours (1+1)	2.75	7.87	10.57	**22.76**
	Ours (2+2)	7.85	10.79	10.47	30.93
	Ours (1)	**2.32**	**7.80**	12.26	24.07
	Ours (2)	6.80	10.34	12.21	31.37

We test on two datasets: The synthetic datasets are provided by FoV-NeRF [6]. The real-world 360 video datasets are provided by EgoNeRF [5]. We compare our method with several baselines, including TensoRF [4], NRFF [7], FoV-NeRF [6], and EgoNeRF [5], using PSNR, SSIM [11], and LPIPS [12] as evaluation metrics. Our method demonstrates superior performance in robustness to variations in FoV, outperforming TensoRF [4] and NRFF [7] in all metrics for the synthetic datasets. It achieves faster convergence and requires significantly less training time. As Fig. 2 shows, our method can capture the high-frequency information. On real-world omnidirectional datasets, our method balances speed and quality, outperforming EgoNeRF [5] in SSIM [11] and LPIPS [12] and achieving faster training times. Fig. 3 shows that our method preserves fine details in both near and far regions, with minimal noise near poles compared to FoV-NeRF [6]. In terms of speed, our method reduces training and rendering time by up to 80% compared to FoV-NeRF [6] and achieves real-time rendering with less than half the rendering time of FoV-NeRF [6], showing in table. 2.

4 Conclusion and Limitations

We propose MSTD, an efficient method for real-time foveated rendering and omnidirectional rendering.

References

[1] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *(ECCV)*, 2020. 1

[2] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec. Immersive light field video with a layered mesh representation. *(TOG)*, 2020. 1

[3] P. Chakravarthula, Z. Zhang, O. Tursun, P. Didyk, Q. Sun, and H. Fuchs. Gaze-contingent retinal speckle suppression for perceptually-matched foveated holographic displays. *(TVCG)*, 2021. 1

[4] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. Tensorf: Tensorial radiance fields. In *(ECCV)*, 2022. 1, 2

[5] C. Choi, S. M. Kim, and Y. M. Kim. Balanced spherical grid for egocentric view synthesis. In *(CVPR)*, 2023. 1, 2

[6] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, and Q. Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *(TVCG)*, 2022. 1, 2

[7] K. Han and W. Xiang. Multiscale tensor decomposition and rendering equation encoding for view synthesis. In *(CVPR)*, 2023. 1, 2

[8] J. Li, Z. Feng, Q. She, H. Ding, C. Wang, and G. H. Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *(ICCV)*, 2021. 1

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 1

[10] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *(TOG)*, 2016. 1

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *(TIP)*, 2004. 2

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *(CVPR)*, 2018. 2