BSRF: Better SceneRF

Muhammad Umer Abbasi Flavio Arrigoni Ahmad Elghobashy Tallha Ijaz
Technical University of Munich
umer.abbasi@tum.de, flavio.arrigoni@tum.de, ,ahmed.elghobashy@tum.de , tallha.ijaz@tum.de

Abstract

Single-view 3D reconstruction typically relies on depth supervision, limiting real-world applicability. SceneRF relaxes this requirement by learning from posed image sequences only. We build upon SceneRF with BetterSceneRF (BSRF), which enhances feature representation via Random Fourier Feature (RFF) cosine encodings and refines probabilistic ray sampling using a hierarchical strategy. Evaluations on BundleFusion and TUM RGB-D demonstrate significant improvements in novel depth synthesis and competitive novel view synthesis and scene reconstruction, highlighting the effectiveness of enhanced encodings and hierarchical sampling for single-view 3D reconstruction.

1. Introduction

Recent advancements in neural implicit representations, particularly Neural Radiance Fields (NeRF) [6], have revolutionized novel view synthesis and scene reconstruction. NeRF optimizes a radiance field self-supervisedly from one or more views, achieving remarkable results in 3D scene modeling. However, when constrained to single-view inputs, existing NeRF-based methods often struggle with complex scenes. While some approaches [2, 11] tackle single-image reconstruction, they rely on depth supervision, making large-scale, image-only datasets challenging to acquire. Others either train on synthetic data [8] or require additional geometric priors [7], increasing supervision costs.

SceneRF [1] introduced a fully self-supervised method for large-scale single-view scene reconstruction by leveraging an image-conditioned NeRF and probabilistic ray sampling. However, SceneRF still faces limitations in capturing fine details, efficiently allocating computational resources, and reconstructing large-scale environments. In this work, we introduce BetterSceneRF, an improved framework for self-supervised single-view 3D reconstruction. Our approach builds upon SceneRF but incorporates key advancements:

- **Enhanced Feature Encoding with RFF Cosine Encoding (Sec. 3.1):** We replace standard Fourier feature encoding with Random Fourier Feature (RFF) cosine encoding, allowing better representation of high-frequency scene details while maintaining computational efficiency.
- **Improved Sampling Strategy with Probabilistic and Hierarchical Sampling (Sec. 3.2):** We refine the probabilistic ray sampling method using a Mixture of Gaussians and introduce hierarchical sampling, which balances scene exploration and exploitation, improving both efficiency and accuracy.
- **Stronger Feature Representation with Spherical U-Net and Bottleneck Attention (Sec. 3.3):** We extend the Spherical U-Net with a Multi-Head Self-Attention (MHSA) bottleneck, enabling better feature extraction beyond the camera's field of view for more coherent scene completion.

2. Related Work

2.1. Neural Radiance Fields (NeRF)

NeRF [6] models a scene as a function from 3D coordinates and viewing directions to radiance and density using an MLP, reconstructed from 2D images via differentiable volume rendering. Rays are cast through pixels, sampling points to predict densities σ and colors $\mathbf{c}$, which are then integrated along each ray.

2.2. PixelNeRF

PixelNeRF [10] extends NeRF by introducing image-conditioned scene representations, enabling novel view synthesis from as few as one input image without per-scene optimization. It uses a ResNet encoder to extract multi-scale image features, which are combined with 3D coordinates and processed by an MLP. This allows feed-forward novel view synthesis by leveraging both 2D and 3D information, learning a scene prior across multiple scenes.

2.3. SceneRF

SceneRF, as illustrated in Fig. 1, is a self-supervised monocular scene reconstruction method that trains on posed image sequences without requiring ground-truth depth. During inference, it synthesizes multiple depth views from a single

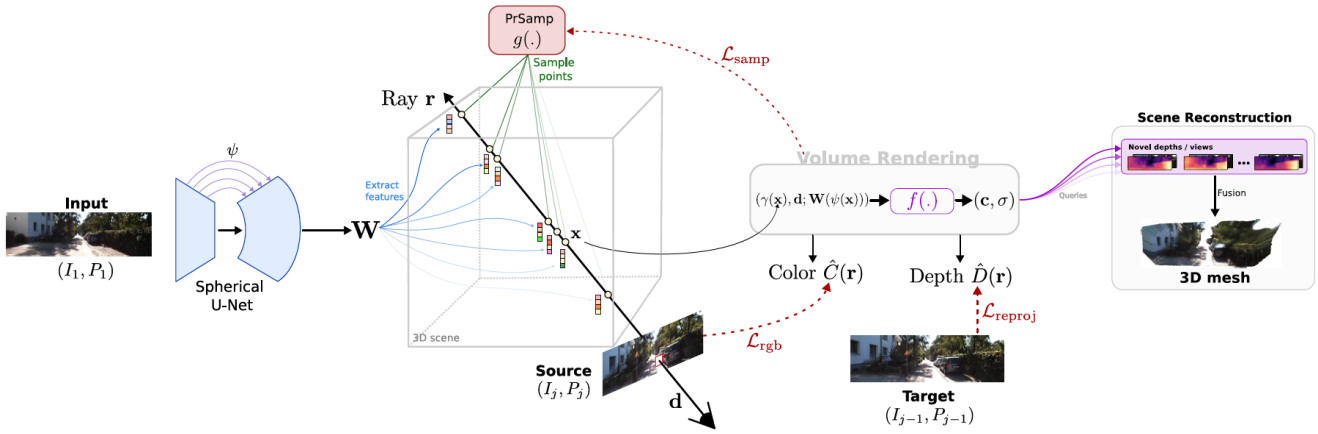

Figure 1. The pipeline of SceneRF.

input image, which are fused into a complete 3D reconstruction.

Building upon PixelNeRF, SceneRF explicitly optimizes for depth using a photometric reprojection loss, ensuring depth consistency without supervision. It introduces a probabilistic ray sampling (PrSamp) strategy, where density along each ray is modeled as a mixture of 1D Gaussians. An auxiliary MLP predicts Gaussian parameters, from which samples are drawn and refined using a Probabilistic Self-Organizing Map (PrSOM) to improve depth estimation.

For training, SceneRF extracts features from the first frame using a Spherical U-Net (SU-Net), then samples points along rays in subsequent frames. SU-Net extends beyond the camera's field of view using spherical convolutions, providing a wider, distortion-free perspective. This structured pipeline enables SceneRF to reconstruct large-scale scenes efficiently while maintaining depth accuracy.

SceneRF reconstructs 3D scenes by fusing novel depth views from a single input frame. Depth is sampled along a straight path and converted into a Truncated Signed Distance Function (TSDF), where the minimum TSDF value across views improves accuracy over weighted averaging. By leveraging SU-Net features, probabilistic density modeling, and self-supervised photometric loss, SceneRF achieves robust reconstruction without explicit depth supervision, even in large, complex environments.

3. Methodology

3.1. RFF Cosine Encoding

We replace the standard Fourier feature positional encoding with a Random Fourier Feature (RFF) cosine encoding [3] to improve the model's capacity and represent high-frequency scene details. This encoding is integrated into the Spherical U-Net's preprocessing stage, which maps spherical coordinates to feature vectors. In this modification, we double the number of sampled frequencies while retaining the same output dimension, enabling the network to capture finer spatial variations without increasing computational overhead. The RFF cosine encoding projects input coordinates into a higher-dimensional space using randomly sampled frequencies $w \sim N(0, \sigma^2)$ and phase shifts $b \sim U(0, 2\pi)$. For an input $x \in R^d$, the encoding is computed as $\gamma(x) = \sqrt{2} \cdot \cos(2\pi x w + b)$, where w and b are fixed after initialization. Crucially, we double the frequency count but use only cosine terms, unlike traditional Fourier encodings that employ both sine and cosine.

3.2. Probabilistic Ray Sampling + Hierarchical Sampling

The SceneRF paper introduces Probabilistic Ray Sampling (PrSamp) and incorporates additionally 32 uniformly distributed points to prevent the model from collapsing into local minima and to explore the whole scene volume. To address the issue of Gaussians converging to suboptimal solutions, we enhance the ray sampling strategy by integrating hierarchical sampling, as illustrated in Fig. 3. Specifically, we modify the original approach by replacing half of the 32 uniformly sampled points with hierarchically sampled points, dividing the process into two distinct stages. In the first *coarse sampling* stage, we employ PrSamp, which operates independently of hierarchical sampling and performs a coarse sampling step by drawing 16 uniformly distributed points along each ray. Using the volume density predictions from this initial stage, we estimate a probability density function (PDF) that informs the fine sampling process. In the second *fine sampling* stage, the remaining 16 points are sampled from this PDF using weighted sampling, ensuring that a higher proportion of samples are allocated to regions of higher density, such as object surfaces. In contrast, fewer samples are placed in empty space.

This combined sampling technique enhances perfor-

	Method	Novel depths synthesis							Novel views synthesis			Scene reconstruction		
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ1 ↑	δ2 ↑	δ3 ↑	LPIPS ↓	SSIM ↑	PSNR ↑	IoU ↑	Prec. ↑	Rec. ↑
BundleFusion	Original	0.1766	0.0940	0.3680	0.2100	72.71	94.89	99.23	**0.323**	**0.853**	**25.07**	**20.16**	**25.82**	**47.92**
	Scaled Down	0.1961	0.1087	0.3911	0.2326	67.86	92.76	98.57	0.337	0.842	24.35	17.72	24.64	38.70
	U_Net Bottleneck Attention	0.1937	0.1009	0.3895	0.2397	65.40	91.97	98.18	0.351	0.840	24.33	15.33	21.50	34.80
	Hierarchical Sampling v1 (n_uni_pts = 8, n_hier_pts =8)	0.1677	0.0761	0.3377	0.2041	72.72	95.11	99.41	0.331	0.845	24.65	18.38	25.00	40.94
	Hierarchical Sampling v2 (n_uni_pts = 6, n_hier_pts =10)	0.1744	0.0780	0.3365	0.2048	72.17	95.47	99.58	0.338	0.840	24.60	18.75	25.48	41.54
	RFF Positional Encoding	0.1733	0.0779	0.3368	0.2035	72.75	95.47	99.46	0.341	0.845	24.70	18.37	25.49	41.91
	RFF + Hier. Samp. v1	**0.1582**	**0.0675**	**0.3209**	**0.1921**	**75.29**	**96.35**	**99.70**	0.327	0.850	24.73	19.06	25.63	42.63
TUM RBG-D	Scaled Down	0.3694	0.4214	1.1309	0.4462	37.05	69.79	86.89	0.515	0.607	18.79	**7.87**	10.59	**23.40**
	Hierarchical Sampling v1 (n_uni_pts = 8, n_hier_pts =8)	**0.3337**	0.3986	**1.0770**	**0.4301**	**41.74**	**72.78**	**87.95**	0.523	0.604	**18.92**	7.80	**10.60**	22.68
	RFF Positional Encoding	0.3622	0.4030	1.0812	0.4357	37.78	70.30	87.41	0.513	0.618	18.84	7.46	10.35	21.10
	RFF + Hier. Samp. v1	0.3503	**0.3955**	1.1229	0.4438	37.92	70.36	87.45	**0.507**	**0.625**	18.76	7.01	10.00	19.00

Figure 2. Quantitative results for BSRF On Bundle Fusion and TUM RGB-D

mance by combining an *exploring* uniform sampling technique, which broadly explores scene geometry, and two efficient *exploiting* techniques, which focus on high-density regions. This balanced approach mitigates suboptimal Gaussian convergence while maintaining computational resource optimization through strategic sample allocation.

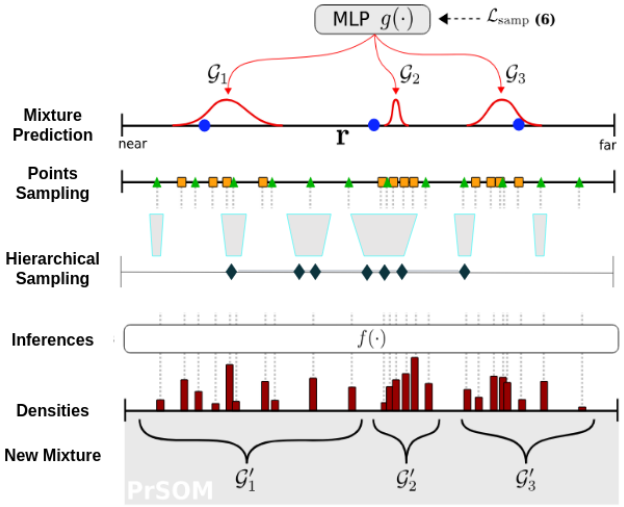

Figure 3. Comparison of three sampling methods: Probabilistic Sampling, Uniform Sampling (Coarse Sampling), and Hierarchical Sampling (Fine Sampling).

3.3. U-Net Bottleneck Attention

The Spherical U-Net extends the standard U-Net by using spherical convolutions to extract features beyond the camera's field of view, enabling color and depth hallucination in unseen regions. To enhance feature representation, we introduce Multi-Head Self-Attention (MHSA) in the bottleneck. While spherical convolutions handle local patterns, MHSA captures long-range dependencies, improving scene coherence beyond the input view. Encoder features are mapped to a spherical latitude-longitude grid, where MHSA processes spatial relationships, enhancing feature aggregation and generalization for novel views.

4. Experiments

Following SceneRF, we evaluated BSRF on two primary tasks, namely novel depth synthesis and scene reconstruction, and novel view synthesis which we refer as a 'subsidiary task' because it is not used for scene reconstruction.

4.1. Datasets

BundleFusion has indoor scenes captured with a handheld device. It has RGB-D images of 640×480 each with an estimated 6-DOF pose. We drop every alternate frame to increase diversity, i.e. getting 9733 images split in sequences of 17 frames. The middle frame serves as input and remaining ones for supervision. We select 7 of the 8 scenes for training and 1 for validation. We evaluate at 1:2 resolution.

TUM RGB-D contains indoor scenes captured with a Microsoft Kinect, providing RGB-D images (640×480) at 30 Hz with 6-DOF ground-truth poses. Unlike BundleFusion, its color and depth images are unsynchronized, requiring alignment with a 0.02s margin due to the 30 fps frame rate. Additionally, depth pixel values differ: 1000 = 1m in BundleFusion, while 5000 = 1m in TUM RGB-D. We train on 7 scenes and validate on 1 scene, evaluating at 1:2 resolution.

4.2. Metrics

To measure our reconstruction quality, we use the intersection over union (IoU), precision, and recall of occupied voxels. For novel depth estimation, we choose usual metrics [4]: relative error absolute (Abs Rel) or squared (Sq Rel), root mean squared error (RMSE), mean log10 error (RMSE log), threshold accuracies ($\delta 1$, $\delta 2$, $\delta 3$). As a common practice, depth is capped to 10m in BundleFusion. Following [5], we measure the quality of synthesized RGB images

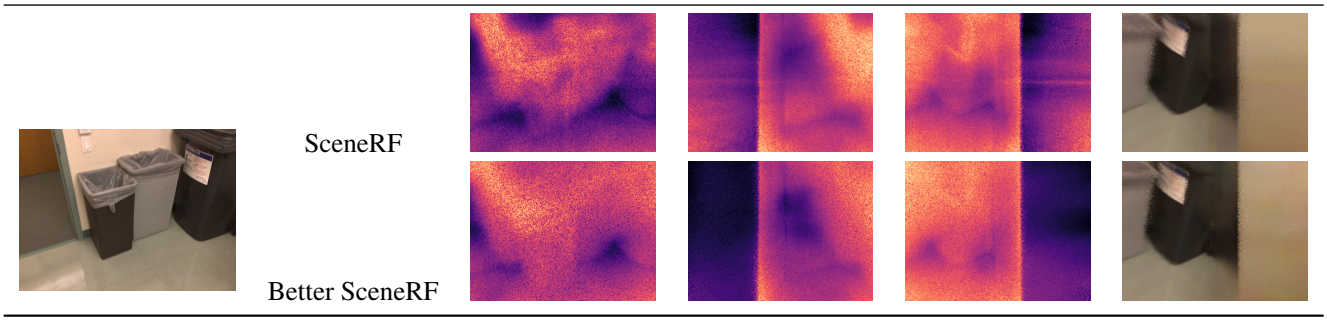

Figure 4. Qualitative results on BundleFusion. For each row, we report novel depths/views at varying positions and viewing angles w.r.t. the input frame.

with: Structural Similarity Index (SSIM) [9], Peak Signal-to-Noise Ratio (PSNR), and LPIPS perceptual similarity [12].

4.3. Training Setup

BSRF trains end-to-end minimizing $L_{\text{total}} = L_{\text{rgb}} + L_{\text{reproj}} + L_{\text{samp}}$. where L_{rgb} is the standard L2 photometric reconstruction loss of NeRFs [6]. We report results for 30 epochs training with batch size of 4 and initial learning rate of 1e-5 with exponential decay at each epoch with gamma 0.95. Training was conducted on a Cluster with 4 Nvidia A100 GPUs.

4.4. Baselines

Since BSRF aims to improve upon SceneRF, we begin by reproducing results for BundleFusion on Novel Depth Synthesis, Novel View Synthesis, and Scene Reconstruction. We evaluate the scaled-down model, a U-Net bottleneck attention model, and models incorporating Hierarchical Sampling and RFF Positional Encoding, both separately and together. The scaled-down model reduces the number of rays from 2048 to 1024 and the number of sample points along the rays from 32 to 16, with 8 points uniformly sampled and 8 using hierarchical sampling. This serves as the baseline for all ablations.

5. Results

The quantitative results for experiments on both BundleFuion and TUM RGB-D Dataset are given in Fig. 2. Also, qualitative results for experiments on BundleFusion are given in Fig. 4. For BundleFusion, We observe that in the Novel Depths Synthesis task a combination of Hierarchical Sampling v1 and RFF give significantly better performance over the baselines and all other ablations. All the other ablations perform better than the standard SceneRF for the same training length. For the Novel Views Synthesis and Scene Reconstruction tasks, we get worse performance than the baseline with the combination of RFF and Hierarchical Sampling giving us the second-best results. For TUM RGB-D, we observe that in the Novel Depths Synthesis task, a combination of Hierarchical Sampling v1 and RFF gives significantly better performance over the baselines and all other ablations. All the other ablations perform better than the standard SceneRF for the same training length. For the Novel Views Synthesis and Scene Reconstruction tasks, we get worse performance than the baseline for all but one combination.

6. Conclusion

In this paper, we presented BSRF, a self-supervised framework that builds upon SceneRF for more effective single-view 3D scene reconstruction. We introduced RFF cosine encodings for capturing high-frequency details and a hierarchical sampling strategy to allocate computational resources more effectively in high-density regions. Although we also experimented with a multi-head self-attention bottleneck in the Spherical U-Net, we found that it did not yield significant gains and thus did not include it in the final model. Through extensive evaluation on the BundleFusion and TUM RGB-D datasets, our enhancements have demonstrated strong performance in novel depth synthesis and competitive results in novel view synthesis and scene reconstruction, addressing several limitations of SceneRF.

Future efforts could extend BSRF to dynamic or deformable objects, potentially integrating temporal coherence or frame-by-frame pose optimization for improved performance. Incorporating additional modalities, such as IMU, LiDAR, or hyperspectral data, could further enrich 3D reconstruction in challenging conditions. Finally, investigating real-time, on-the-fly reconstruction in robotics or AR/VR scenarios—by continuously refining the learned model—would open up new applications for BSRF in interactive systems.

References

[1] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12345–12355, 2023. 1

[2] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34: 8282–8293, 2021. 1

[3] Bharath Bhushan Damodaran, Francois Schnitzler, Anne Lambert, and Pierre Hellier. Improved positional encoding for implicit neural representation based compact data representation. *arXiv preprint arXiv:2311.06059*, 2023. 2

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3828–3838, 2019. 3

[5] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12578–12588, 2021. 3

[6] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 4

[7] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12932–12942, 2022. 1

[8] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Ambrus, Adrien Gaidon, William T Freeman, Frédo Durand, Joshua B Tenenbaum, and Vincent Sitzmann. Neural groundplans: Persistent neural scene representations from a single image. *arXiv preprint arXiv:2207.11232*, 2022. 1

[9] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 4

[10] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 1

[11] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 1

[12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 4