MixNeRF: Modeling a Ray with Mixture Density for Novel View Synthesis from Sparse Inputs —Supplementary Material—

Seunghyeon Seo   Donghoon Han*   Yeonjin Chang*   Nojun Kwak

Seoul National Univeristy

{zzzlssh, dhk1349, yjean8315, nojunk}@snu.ac.kr

	Anneal.	LLFF [6]			DTU [3]			Real. Syn. 360° [7]	
		3-view	6-view	9-view	3-view	6-view	9-view	4-view	8-view
λ_C	✓	[4.0, 1e−3]							
λ_D		1e−4	1e−5	1e−6	1e−3	1e−4	1e−5	1e−3	1e−4
$\hat{\lambda}_C$		1e−5	1e−6	1e−7	1e−4	1e−5	1e−6	1e−4	1e−5

Table A. **Overview of our loss balancing weights.** We apply a linear annealing strategy for λ_C to stabilize the training. We divide λ_D and $\hat{\lambda}_C$ by a factor of 10 as more input views are provided for training.

A. Implementation Details

A.1. Hyperparameters

Following RegNeRF [8], we adopt a scene space annealing during the early training stage, an exponential learning rate decay from 2e−3 to 2e−5, and 512 steps of warm up [1] with a delay multiplier of 1e−2. For the Realistic Synthetic 360° [7], we set the initial learning rate as 1e−3 and apply an exponential decay to 1e−5. The Adam [5] optimizer is used and the gradient clippings are applied by value at 0.1 and norm at 0.1 in order. We train our MixNeRF for 500 pixel epochs with 4096 batch size on 2 NVIDIA TITAN RTX, and the training time is measured on the same hardware. For the balancing hyperparameters for our loss terms, we anneal λ_C from 4.0 to 1e−3 over the first 512 iterations, while setting λ_D and $\hat{\lambda}_C$ as different values by the datasets. Tab. A shows the overview of balancing terms by the datasets and the number of input views.

A.2. Architecture

Our MixNeRF is based on the architecture of mip-NeRF [1]. As illustrated in Fig. A, our MixNeRF additionally outputs the scale parameters β using softplus activation and the ray depths μ^d for our mixture model. In practice, we estimate the unnormalized ray directions $\tilde{\mu}^d \in \mathbb{R}^{N\times 3}$, where N indicates the number of samples, and we use its Euclidean norm $\mu^d = \|\tilde{\mu}^d\|_2$ as the estimated ray depths for the training stability.

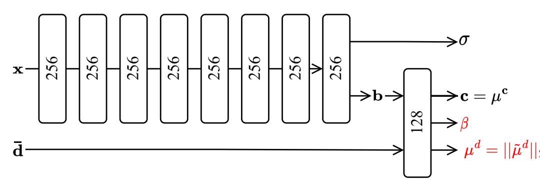


Figure A. **MixNeRF Network Architecture.** The architecture of MixNeRF is implemented based upon the mip-NeRF [1]. It additionally outputs the scale parameter β using softplus activation and the ray depths $\mu^d = \|\tilde{\mu}^d\|_2$, which are denoted in red. $\mathbf{b}$ indicates a bottleneck vector.

B. Experimental Details

B.1. Datasets

We evaluate MixNeRF on the different standard benchmarks: LLFF [6], DTU [3], and Realistic Synthetic 360° [7].

LLFF: It contains realistic forward-facing scenes and is generally used as an out-of-domain test set for pre-training methods. Following the protocol of [7], every 8-th image is used as a held-out test set and input views are chosen evenly from the remaining images. We report the results under the

*D. Han and Y. Chang equally contributed to this work.

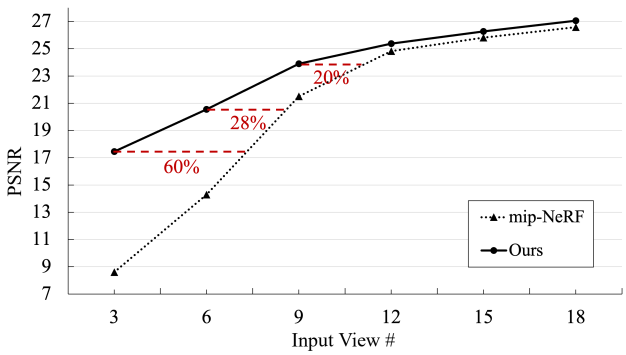


Figure B. **Comparison with baseline by the number of input views.** Our MixNeRF requires up to about 60% fewer input views than mip-NeRF to achieve comparable performance, and outperforms mip-NeRF consistently even when more input views are used for training. Since the reduced test set is used for the experiment following [8], the results can be slightly different from the main table.

scenarios of 3, 6, and 9 input views following [10].

DTU: It consists of images containing objects placed on a white table with a black background. We follow the experimental protocol of [10] and conduct experiments on their designated 15 scenes. As with the LLFF dataset, we conduct the experiments under the scenarios of 3, 6, and 9-view.

Realistic Synthetic 360°: It consists of 8 inward-facing synthetic scenes with different viewpoints, each containing 400 images. Following previous works [2, 4], we conduct the experiments for the scenarios of 4 and 8 views. For a fair comparison with other regularization methods, we sample the first 4 and 8 images from the training set for the scenario of 4 and 8 input views, respectively, for all models and use the 200 test set images for evaluation. Note that the images of the training set are arranged randomly in the first place, and we do not choose the training input views carefully for improving the performance.

B.2. Evaluation Metrics

We adopt a set of evaluation metrics including the mean of PSNR, structural similarity index (SSIM) [9], and LPIPS perceptual metric [11]. Additionally, we report its geometric average [1]: $\text{MSE} = 10^{-\text{PSNR}/10}$, $\sqrt{1 - \text{SSIM}}$, and LPIPS. Following [8], we adopt masked metrics to avoid background bias for DTU.

C. Data Efficiency Experiment

As demonstrated in Fig. B, we observe that our MixNeRF achieves superior data efficiency to the vanilla mip-NeRF. Our MixNeRF requires up to about 60% fewer input views to mip-NeRF to achieve comparable results. Moreover, ours outperforms mip-NeRF consistently even when more than 9 input views are provided. It indicates that our proposed mixture modeling strategy is effective in general scenarios as well as the sparse input setting.

D. Additional Qualitative Results

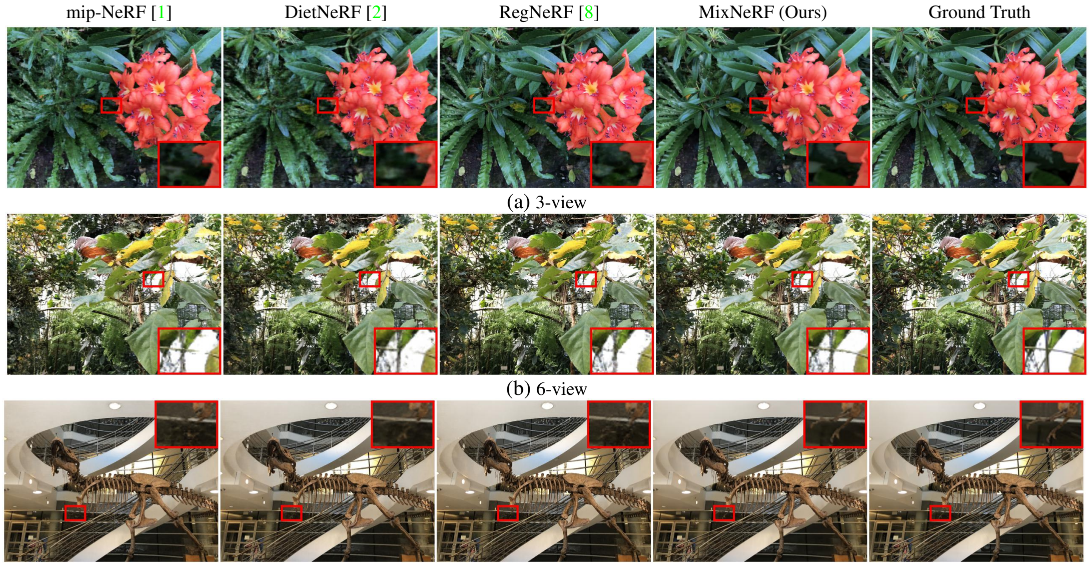
We demonstrate the additional qualitative comparisons in Fig. C, Fig. D, and Fig. E. Moreover, we show the additional qualitative results of our MixNeRF in Fig. F, Fig. G, and Fig. H.

E. Limitations and Future Work

Our MixNeRF achieves the state-of-the-art performance without any extra training resources, *e.g.* additional inference for pre-generated rays from unseen viewpoints, external modules for providing supplemental supervisory signals, or so on. However, it still shows a few degenerate parts in the rendered images under the very sparse scenario as few as 3-view, due to the disturbance from the non-objects, *e.g.* a background or a table, especially on the DTU dataset. To eliminate the artifacts more effectively, developing an algorithm for classifying the pixels into an object or non-object can be a promising future work.

F. Potential Negative Societal Impact

Our method is able to synthesize a photo-realistic image from novel view from the limited training resources. Although it provides much benefits for practical applications where the dense training resources are hard to collect, there exists a possibility of negative consequences with malicious intents, *e.g.* a misleading content made with an intent to either conceal or show some specific views. Therefore, the effort to prevent the malicious usage should be made, *e.g.* strictly checking on the permission to use sensitive data, deep fake detection, and so on.

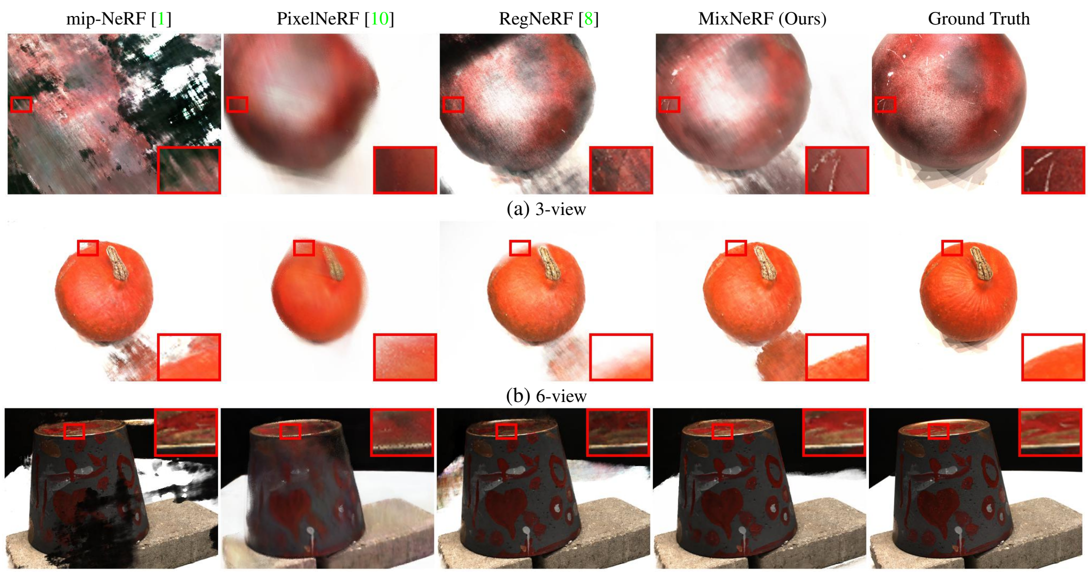


(a) 3-view

(b) 6-view

(c) 9-view

Figure C. **Additional qualitative comparisons on LLFF.**

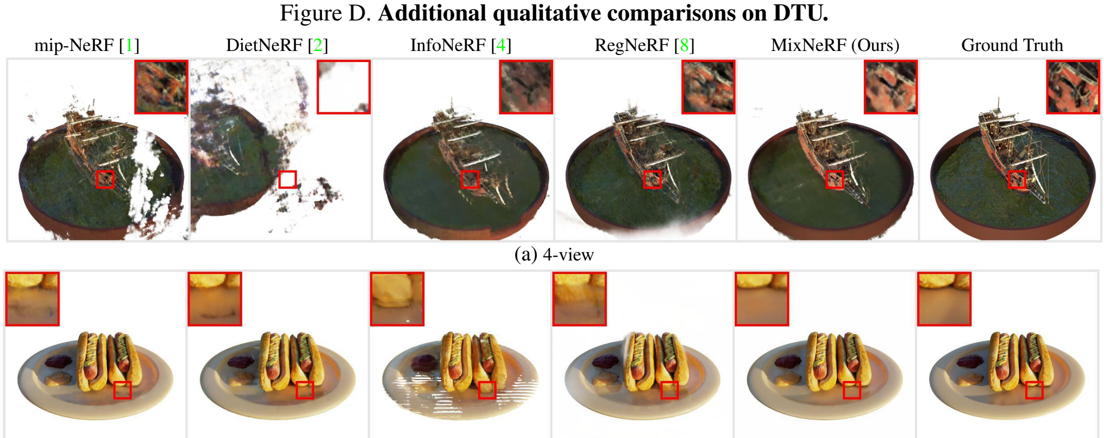


(a) 3-view

(b) 6-view

(c) 9-view

Figure D. **Additional qualitative comparisons on DTU.**

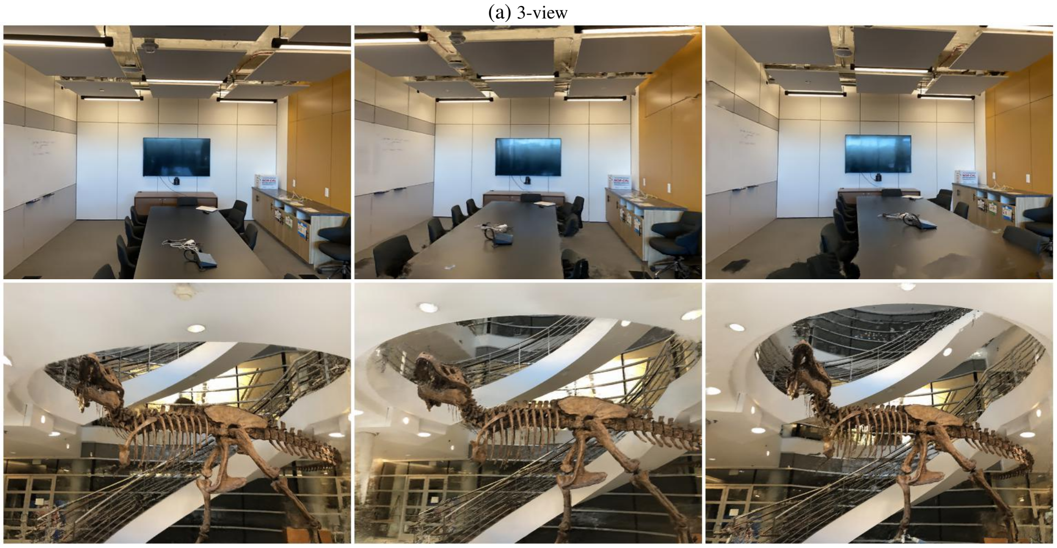


(a) 4-view

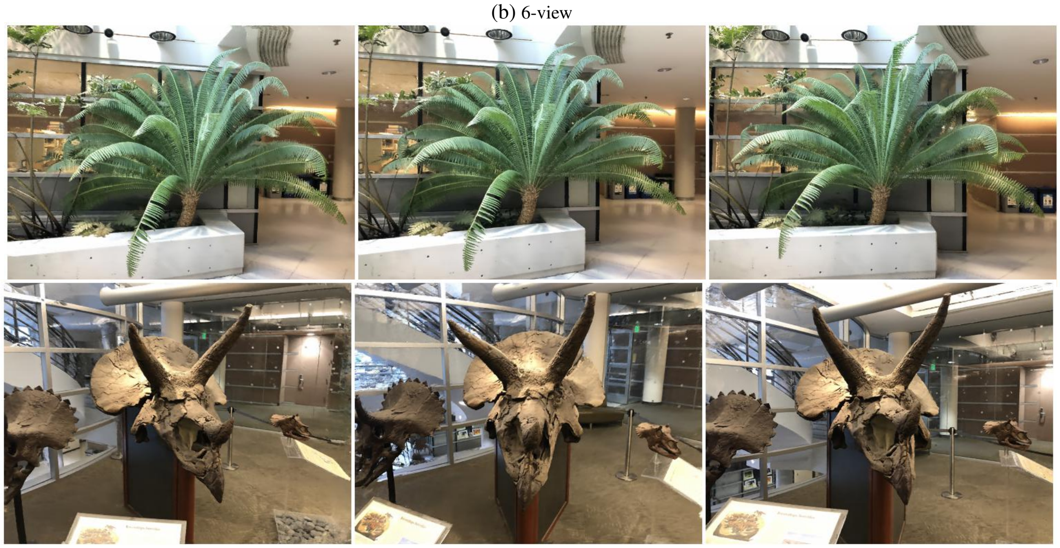
(b) 8-view

Figure E. **Additional qualitative comparisons on Realistic Synthetic 360°.**

(a) 3-view



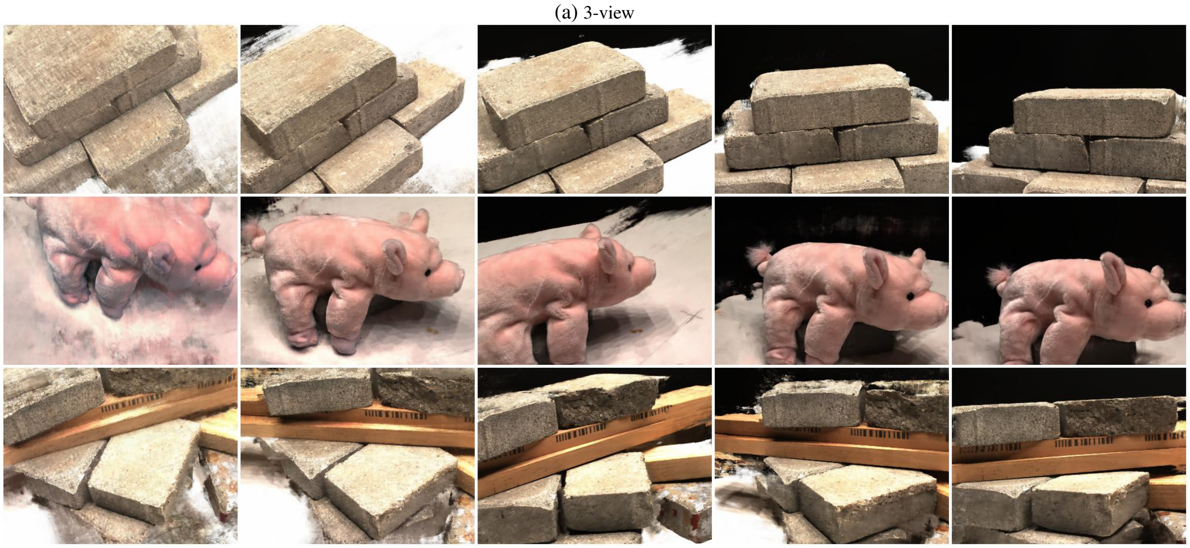
(b) 6-view



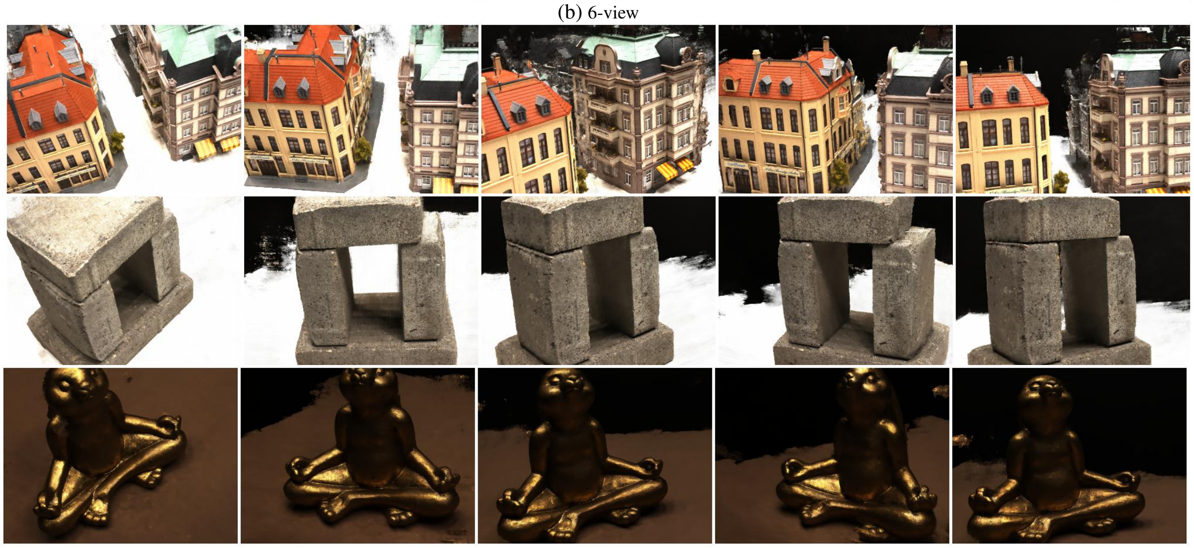
(c) 9-view

Figure F. **Additional qualitative results of our MixNeRF on LLFF.**

(a) 3-view

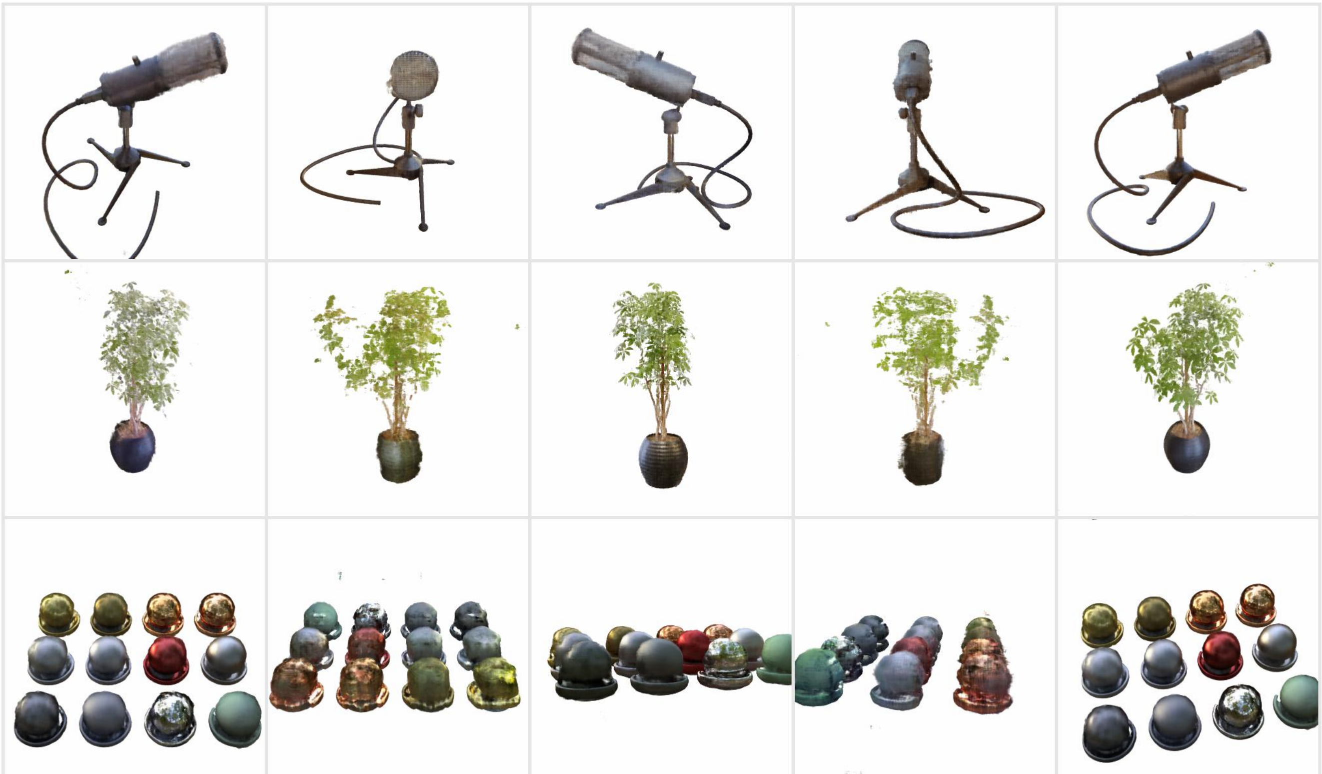


(b) 6-view



(c) 9-view

Figure G. **Additional qualitative results of our MixNeRF on DTU.**



(a) 4-view

(b) 8-view

Figure H. **Additional qualitative results of our MixNeRF on Realistic Synthetic 360°.**

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1, 2, 3

[2] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2, 3

[3] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 1

[4] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, 2022. 2, 3

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1

[7] 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

[8] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 1, 2, 3

[9] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[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*, pages 4578–4587, 2021. 2, 3

[11] 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 conference on computer vision and pattern recognition*, pages 586–595, 2018. 2