2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10687533

SID-NERF: FEW-SHOT NERF BASED ON SCENE INFORMATION DISTRIBUTION

*Yuchen Li**, *Fan Wan**, *Yang Long* ✉

Department of Computer Science, Durham University, Durham, UK

ABSTRACT

The novel view synthesis from a limited set of images is a significant research focus. Traditional NeRF methods, relying mainly on color supervision, struggle with accurate scene geometry reconstruction when faced with sparse input images, leading to suboptimal rendering. We propose a Few-shot NeRF Based on Scene Information Distribution(Sid-NeRF) to address this by integrating geometric and color supervision, enhancing the model's understanding of scene geometry. We also implement a data selector during training to identify and utilize the most accurate geometric data, thus improving training efficiency. Additionally, a residual module is introduced to counteract any optimization biases from the selector. Our method was tested on three datasets and showed excellent performance in various environments with limited images. Notably, compared to other novel view synthesis methods based on fewer views, our method does not require any prior knowledge and thus does not incur additional computational and storage costs.

***Index Terms*—** Novel view synthesis, Limited input images, Without prior knowledge

1. INTRODUCTION

Neural rendering, exemplified by Neural Radiance Fields (NeRF [1, 2]), utilizes neural networks to learn scene representations, encoding scene volume density and color, enabling high-quality novel view synthesis from input images. Its versatility has led to applications in various fields, including complex environments, dynamic scenes, and editable NeRF. However, NeRF struggles when input images are scarce, limiting perspective coverage and detail capture and diminishing its novel view synthesis performance.

In order to improve the quality of novel view synthesis for NeRF with limited input images, several studies [3, 4] have introduced prior knowledge. By introducing prior knowledge, the model's learning and understanding ability of the scene can be enhanced by providing the model with additional viewpoint, geometry and depth information. However, this approach introduces additional computational resource consumption and storage resource consumption. Specifically, large pre-trained models with prior knowledge increase storage requirements, while model tuning increases computational requirements, thus affecting the practical implementation and scalability of the model.

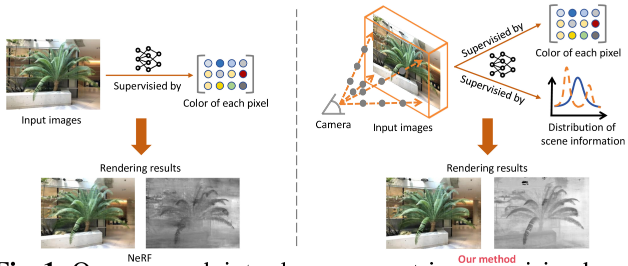

Fig. 1: Our approach introduces geometric supervision based on the scene information distribution, which complements the traditional NeRF color-based supervision and shows superior rendering quality in input-limited environments, as demonstrated in the comparison.

In this study, we propose Few-shot NeRF Based on Scene Information Distribution(Sid-NeRF) to enhance the novel view synthesis quality of NeRF in scenarios with insufficient input images without incurring additional storage or computational resource costs. The original NeRF model primarily relies on color information for supervision. Our approach extends this by incorporating geometric information into the supervision process. Specifically, this geometric supervision is based on the distribution of scene information. We adopt ray entropy as a scene information distribution constraint. Ray entropy is a metric to assess the uncertainty of opacity distribution across sampling points along a ray's path. In simple environments, scene information exists mainly near the surfaces of non-transparent objects. In this case, the opacity distribution of the sampling points should be concentrated near the surface of the non-transparent objects, when the ray entropy is low. Thus, by minimizing ray entropy during the training process, we can more effectively direct the network's focus to areas that most significantly contribute to the final image. This optimization is particularly effective in simple environmental conditions, where the scene's geometry is more pronounced, making the opacity distribution of the sampling points more concentrated can be an excellent way to learn the geometric details of the scene.

However, in more complex environments, characterized by multiple materials, varying lighting, or intricate geome-

*Contributed Equally.

Corresponding Author: Yang Long (yang.long@durham.ac.uk).

This work is supported by the UK Medical Research Council (MRC) Innovation Fellowship under Grant MR/S003916/2, International Exchanges 2022 IEC\NSFC\223523.

979-8-3503-9015-5/24/$31.00 ©2024 IEEE

tries, the distribution of scene information becomes more dispersed. In such scenarios, solely relying on ray entropy minimization is insufficient. To enhance the model's ability to learn scene information in complex environments, we introduce KL divergence of depth difference and weight. Because there is a correlation between the depth difference distribution and the weight distribution even in complex environments, to this end, we introduce additional supervisory elements: the depth difference distribution and the weight distribution. Since part of the data introduced for geometric supervision is obtained during the training process, we implement a selector mechanism to evaluate and utilize data with higher precision during the training process. Additionally, recognizing that this selector might introduce an optimization bias, we incorporate a residual module to mitigate this effect. This integrated approach aims to refine the training and rendering capabilities of NeRF to ensure high-quality output even under conditions of insufficient input images, without increasing the demand on storage and computational resources.

In summary, our contributions are three-fold:

- **Geometric Supervision:** We integrate geometric information into NeRF's training, focusing on the opacity distribution at sampling points. This enhances the model's attention to key areas, improving novel view synthesis quality, especially in simpler scenes.
- **Complex Environment adaptation:** To address the limitations of ray entropy in complex environments, we introduce depth difference and weight distribution as additional supervisory factors to help models accurately capture and render scene details in environments with varying materials, lighting, and geometries.
- **Selective Optimization with Residual Module:** The depth difference and weight distribution we introduce are obtained during training, therefore, we use a selector to filter the data in which the accuracy is higher. And we introduce the residual module to address the problem of optimization bias caused by the selector.

2. RELATED WORK

Novel View Synthesis. Novel view synthesis, a pivotal technique in computer vision[5, 6, 7, 8, 9], leverages generative models and Neural Radiance Fields (NeRF) to generate unseen perspectives of a scene by intricately modeling its 3D structure. It has progressed from geometry-based methods [10, 11] to depth map techniques [12] and light field approaches [13]. The advent of deep learning, especially with Neural Radiance Fields (NeRF), marked a significant shift. NeRF has pioneered innovations in novel view synthesis, enhancing realism and detail in diverse scenarios with its advanced deep learning capabilities.

NeRF For Few Images. In scenarios where obtaining diverse images is challenging, the lack of extensive viewpoint data hampers the model's ability to learn surface details comprehensively, especially in zero-shot and few-shot learning contexts [14, 15, 16, 17]. To mitigate these limitations, several works have turned to utilizing spatial structure information, such as depth or 3D point clouds, to enhance the training and inference processes. Notably, [18] employs depth information prior to supervise NeRF training, thereby improving both training efficiency and rendering quality in environments with fewer viewpoints. In a similar vein, [19] leverages point-based 3D data as a prior, specifically to bolster NeRF's performance in complex scenes. Collectively, these papers collectively demonstrate the efficacy of using varied types of prior information (depth data and 3D point clouds) to optimize NeRF models in environments constrained by limited data. Additionally, [20] introduces multi-view geometric constraints and a depth consistency loss to enhance rendering results without relying on prior information. [21, 22] employ neural networks for depth estimation, using this as prior information to guide the sampling process and supervise training. [23] takes a different approach, optimizing single-view NeRF synthesis by incorporating a general image prior from 2D diffusion models and linguistic guidance from visual-language models. This methodology not only bolsters multi-view content consistency but also further refines NeRF's 3D geometry through estimated depth maps, thus enhancing novel view synthesis quality and demonstrating wide applicability across various scenarios. These methods highlight a spectrum of reliance on the accuracy of prior information and the associated computational resources.

Our Method. In contrast to existing techniques, our method introduces geometric supervision based on the distribution of scene information, augmenting the traditional color-supervised NeRF. Notably, our approach does not require any prior knowledge to achieve high-quality novel view synthesis with insufficient input images, thus effectively avoiding the consumption of additional computational and storage resources that may be caused by introducing prior knowledge.

3. METHODOLOGY

Our method is mainly focused on the study of improving the novel view synthesis effect in the case of an insufficient number of input images. By combining color supervision with geometric supervision in the case of an insufficient number of input images, our method is able to achieve high-quality novel view synthesis without introducing any prior knowledge. This section focuses on the overall architecture of our method.

3.1. Original NeRF

NeRF uses a neural network to learn scene information, facilitating 3D scene rendering from various viewpoints. It inputs 3D coordinates (x,y,z) and view direction (θ, ϕ) to output color (r,g,b) and density σ. The color computation process is as follows:

$$\begin{gathered}\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i \delta_i))\mathbf{c}_i, \\ T_i = exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j),\end{gathered} \tag{1}$$

where σ_i indicates density, δ_i represents the distance between adjacent sampling points, $\mathbf{c}_i$ is the emitted color, $\mathbf{r}$ indicates a ray, $\hat{C}(\mathbf{r})$ shows the expected color.

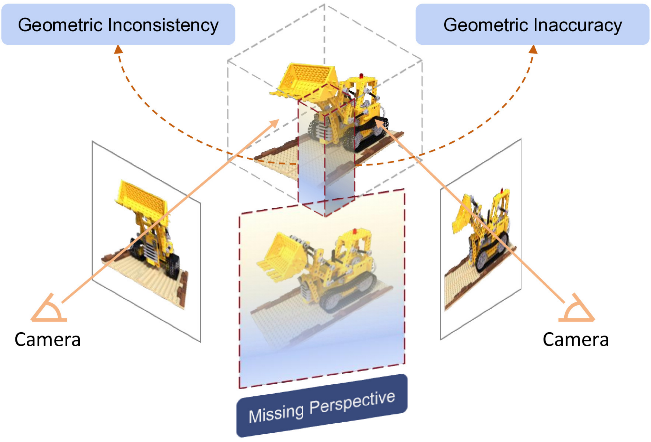

Fig. 2: NeRF's limitations with insufficient input images result in incomplete viewpoints, geometric discontinuities, and reduced novel view synthesis quality.

3.2. Motivation

NeRF struggles with few input images, impacting its 3D scene understanding. Limited data leads to imprecise geometry and lighting details, affecting novel view synthesis, as shown in Fig. 2. Adding geometric supervision improves accuracy, especially with limited views, enhancing NeRF's ability to generate reliable novel views.

3.3. The Pipeline of Our Method

Ray entropy in Scene Modeling. In simple environments, important scene information resides primarily on opaque object surfaces. To optimize data acquisition, we employ 'ray entropy' [24] to guide sampling towards these critical regions. Ray entropy quantifies opacity distribution at sampling points, directing sampling to enhance model learning on opaque surfaces. The formula for ray entropy is presented below:

$$p(\mathbf{r}_i) = \frac{\alpha_i}{\sum_j \alpha_j} = \frac{1 - exp(-\sigma_i \delta_i)}{\sum_j 1 - exp(-\sigma_j \delta_j)}, \tag{2}$$

$$H(\mathbf{r}) = -\sum_{i=1}^{N} p(\mathbf{r}_i) log p(\mathbf{r}_i), \tag{3}$$

where H($\mathbf{r}$) denotes ray entropy, $p(\mathbf{r}_i)$ represents ray density, σ_i denotes density, and δ_i indicates the distance between adjacent samples. Opacity distribution α and ray entropy H($\mathbf{r}$) are inversely related: a dispersed α yields higher H($\mathbf{r}$), and a focused α results in lower H($\mathbf{r}$). Minimizing ray entropy promotes concentrated opacity distribution, leading to denser sampling points. The pertinent loss term for this process is outlined below:

$$\mathcal{L}_{entropy} = min(H(\mathbf{r})), \tag{4}$$

Depth difference and weight. In simple environments, models should focus on opaque object surfaces, while in complex scenes, comprehensive scene information is crucial. We guide models by aligning depth difference with weight distributions to ensure even sampling across the scene and equal value given to each sample. This balance prevents bias towards specific scene areas. The computation is detailed in the formula below:

$$\begin{gathered} W = T(1 - exp(-\sigma\delta)), \\ D(\mathbf{r}) = \sum_{i=1}^{N} W_i Z_i, \end{gathered} \tag{5}$$

where W denotes the weight, D($\mathbf{r}$) denotes the pixel depth, and Z denotes the sampling point depth. $Z - D(\mathbf{r})$ denotes the depth difference. We employ KL divergence as a metric for comparing depth difference and weight similarity. Consequently, we use KL divergence between these distributions as a loss term:

$$\mathcal{L}_{KL} = D_{KL}(w||d), \tag{6}$$

where w represents weight, and d represents depth difference. A higher KL divergence indicates greater distribution disparity, while a lower value indicates less disparity.

Using a selector to filter the data. Depth information and weight information are computed during the training process, and we need to evaluate the accuracy of their data, selecting the more accurate of them for use and discarding the less accurate of them. Therefore, we introduce a selector for data filtering. The formula is shown below:

$$Err(\mathbf{r}) = C_{gt}(\mathbf{r}) - C_{pd}(\mathbf{r}), \tag{7}$$

$$\text{selector} = \begin{cases} 0 & Err(\mathbf{r}) > threshold \\ 1 & Err(\mathbf{r}) < threshold \end{cases}, \tag{8}$$

where $C_{pd}(\mathbf{r})$ and $C_{gt}(\mathbf{r})$ represent predicted and ground truth colors, while Err($\mathbf{r}$) quantifies their difference. Err($\mathbf{r}$) evaluates pixel depth prediction accuracy and consistency with depth difference and weight distributions. We use a threshold to improve model optimization: Err($\mathbf{r}$) above it implies low accuracy and data exclusion, while Err($\mathbf{r}$) below it signifies high accuracy and data retention.

Residual module. Since the selector we use acts directly on the loss term, this may cause an optimization bias that focuses the model on optimizing the loss term filtered by the selector for the poor distribution of depth difference and weight. To this end, we introduce a residual module to mitigate this problem. The introduction of a residual module improves the overall model optimization, mitigating the impact of optimization bias on performance.

Loss function. The loss function consists of supervision of color, ray entropy, and supervision of KL divergence between

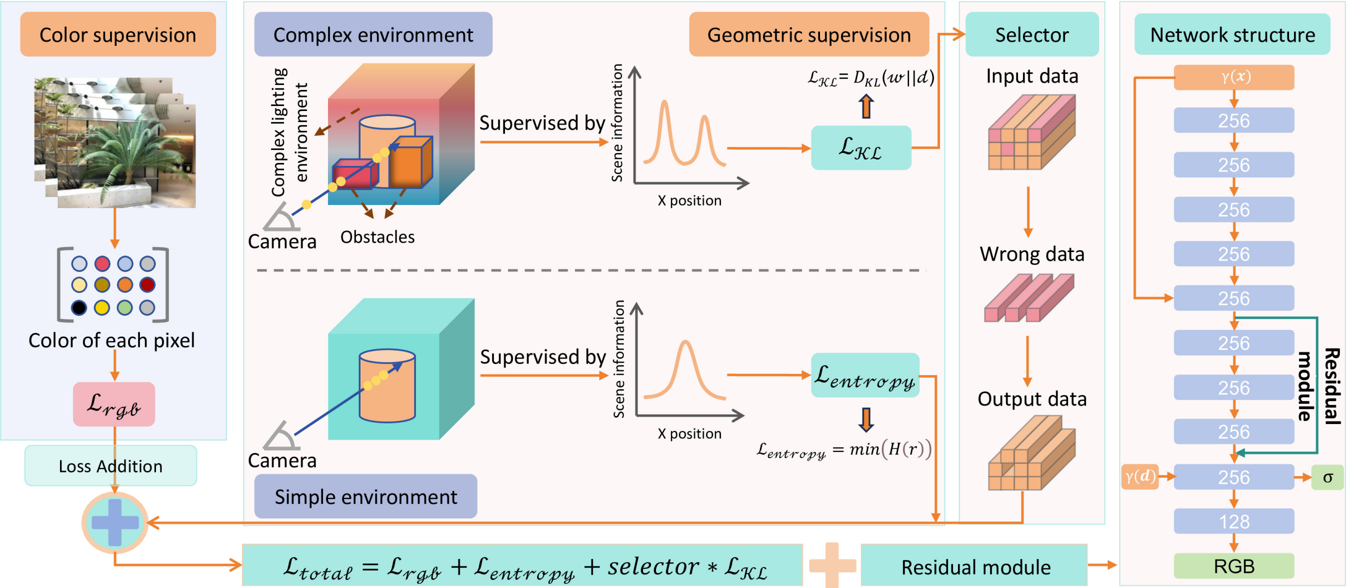

Fig. 3: Our method involves adding the residual module and using color, ray entropy, and KL divergence of depth difference and weight as supervision. Color supervision is for color accuracy, while ray entropy and KL divergence of depth difference and weight provide geometric supervision. We also utilize a selector to filter the KL divergence of depth difference and weight.

the depth difference and weight. The loss function of our method is shown below:

$$\mathcal{L}_{total} = \mathcal{L}_{rgb} + \mathcal{L}_{entropy} + S * \mathcal{L}_{KL}, \tag{9}$$

Here $\mathcal{S}$ indicates selector, $\mathcal{L}_{KL}$ is filtered by the selector. **Overall framework.** Our method, illustrated in Fig. 3, comprises color supervision $\mathcal{L}_{rgb}$ and geometric supervision ($\mathcal{L}_{entropy}, S * \mathcal{L}_{KL}$), along with a residual module to mitigate optimization bias. It effectively achieves high-quality novel view synthesis in NeRF, requiring no prior knowledge and incurring no additional computational or storage overhead, even with limited inputs.

4. EXPERIMENTS

In this section, we evaluate our method's performance across various datasets. And then focus on analyzing the impact of each module of our method and assessing our method's resistance to overfitting.

Baseline. We use InfoNeRF [24] as our backbone, and compare our method with baseline InfoNeRF and multiple state-of-the-art(SOTA) methods: PixelNeRF [25] and DietNeRF [4].

Dataset and metric. We assess our method on NeRF llff (real world foreground scenes), NeRF synthetic (synthetic images), and NeRF real 360 (varied perspectives scenes in real world), using four images per dataset for training. Evaluation metrics include PSNR (for pixel differences), SSIM (structural disparities), and LPIPS (perceptual discrepancies). Higher PSNR and SSIM values, and lower LPIPS values, indicate better performance.

4.1. Main Results

NeRF synthetic. We compared our method with NeRF, PixelNeRF, DietNeRF, and InfoNeRF, all trained with four images on the NeRF synthetic dataset. Table 2 presents quantitative results, highlighting our method's superior performance in PSNR, SSIM, and LPIPS metrics. Table 1 provides detailed results for eight different scenes, demonstrating our method's superiority across all scenes. In terms of qualitative results (see Fig. 4), our method outperforms baseline InfoNeRF in both RGB and depth images.

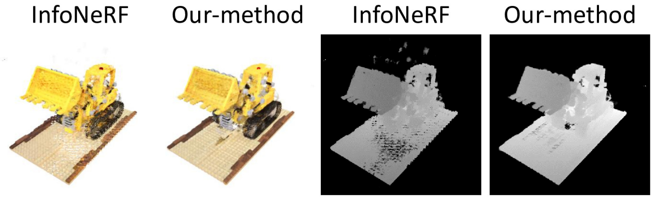

Fig. 4: These figures compare rendering results for the NeRF synthetic dataset across baseline InfoNeRF, with our approach outperforming baseline InfoNeRF.

NeRF llff. We compared our method with NeRF and baseline InfoNeRF on the NeRF llff dataset, which contains real-world images with complex environmental factors. All methods were trained with four images. Fig. 5 displays qualitative results, showcasing our method's superior rendering compared to others. Table 3 presents quantitative results, indicating our method's better performance in PSNR, SSIM, and LPIPS metrics in the NeRF llff dataset, demonstrating its effectiveness in complex environmental scenes.

Methods	Lego	Chair	Drums	Ficus	Hotdog	Materials	Mic	Ship	avg
NeRF, 100views	32.54	33.00	25.01	30.13	36.18	29.62	32.91	28.65	31.01
PixelNeRF* [25]	15.14±0.75	18.87±1.38	15.10±0.63	16.60±0.70	19.37±1.78	12.31±1.02	16.35±0.97	14.96±0.75	16.09±0.78
NeRF [1]	15.61±4.53	18.57±1.64	12.50±0.98	16.37±2.24	19.64±2.26	15.65±4.16	14.78±2.37	14.30±4.04	15.93±1.06
DietNeRF [4]	17.13±4.77	19.37±3.12	13.74±1.55	15.76±3.56	18.24±5.28	15.00±5.18	17.71±1.55	11.51±4.27	16.06±1.13
InfoNeRF [24]	18.92±0.51	20.06±1.11	14.33±0.62	19.41±0.07	21.30±2.31	18.34±0.88	**18.55±1.71**	**18.27±0.71**	18.65±0.18
Our-method	**19.62±0.22**	**21.52±0.54**	**14.47±0.54**	**20.14±0.16**	**22.60±0.29**	**18.96±0.14**	18.28±1.96	17.06±0.85	**19.08±0.59**

Table 1: The table shows the PSNR value of different methods in 8 different scenes of the NeRF synthetic dataset. And it can be observed that our method has better performance than other methods. The asterisk (*) indicates the model is pretrained on an external training dataset with dense input views and fine-tuned on this dataset with 4 input views.(comparative data of other methods referenced from [24]).

Methods	PSNR↑	SSIM↑	LPIPS↓
NeRF, 100views	31.01	0.947	0.081
PixelNeRF* [25]	16.09±0.78	0.738±0.012	0.390±0.030
NeRF [1]	15.93±1.06	0.780±0.014	0.320±0.049
DietNeRF [4]	16.06±1.13	0.793±0.019	0.306±0.050
InfoNeRF [24]	18.65±0.18	0.811±0.008	0.230±0.008
Our-method	**19.08±0.59**	**0.818±0.006**	**0.214±0.016**

Table 2: The table lists average PSNR, SSIM, and LPIPS values for different methods tested on NeRF synthetic dataset (comparative data of other methods referenced from [24]).

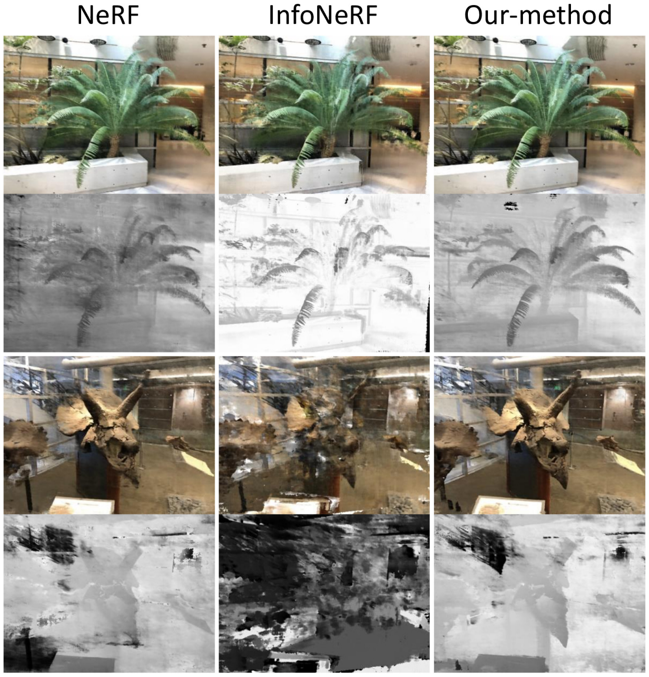

Fig. 5: These figures compare the rendering results of our method with NeRF and baseline InfoNeRF on NeRF llff dataset, demonstrating our method's superior performance.

Methods	PSNR↑	SSIM↑	LPIPS↓
NeRF [1]	18.74	0.569	0.282
InfoNeRF [24]	16.35	0.446	0.437
Our-method	**18.87**	**0.582**	**0.263**

Table 3: The table shows average PSNR, SSIM, and LPIPS values for different methods on NeRF llff dataset, highlighting our method's superior performance.

NeRF real 360. In contrast to NeRF synthetic and NeRF llff datasets, NeRF real 360 dataset contains 360-degree scene images with intricate details and environmental factors. We compared our method with baseline InfoNeRF on this dataset, using four training images. Table 5 in the appendix shows our method's superior PSNR, SSIM, and LPIPS metrics, confirming its effectiveness in complex environments.

4.2. Analysis

Method	$\mathcal{L}_{KL}$	$\mathcal{S}$	$\mathcal{RM}$	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline				18.65±0.18	0.811±0.008	0.230±0.008
Our-method w/o $\mathcal{S}+\mathcal{RM}$	✓			19.06±0.42	0.815±0.004	0.223±0.011
Our-method w/o $\mathcal{RM}$	✓	✓		19.07±0.44	0.817±0.005	**0.212±0.013**
Our-method	✓	✓	✓	**19.08±0.59**	**0.818±0.006**	0.214±0.016

Table 4: The table compares our method's module performance on NeRF synthetic dataset ($\mathcal{S}$ indicates selector, $\mathcal{RM}$ indicates residual module).

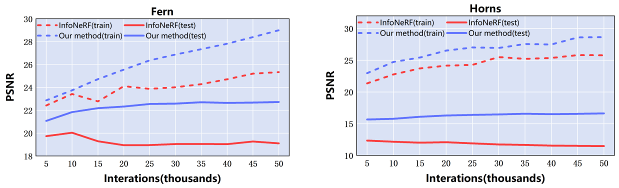

Fig. 6: The figures compare PSNR for baseline InfoNeRF and our method on NeRF llff's fern and horns, demonstrating our method's better overfitting prevention compared to InfoNeRF.

Module Impact Analysis. Table 4 shows the results of ablation experiments of our method on NeRF synthetic dataset. The experimental results show that the KL divergence of depth difference and weight greatly improves the model performance compared to the baseline. Moreover, the combined use of the selector and residual module results in a synergistic effect, not just additive, leading to a marked increase in our method's overall efficiency and accuracy.

Overfitting Assessment. Our analysis highlights the overfitting issue in baseline InfoNeRF. Fig. 6 show that baseline InfoNeRF peaks at 10000 iterations for the 'fern' dataset and 5000 for 'horns', but then performance declines, indicating overfitting. In contrast, our method consistently performs well throughout training, as seen in the same figure. This consistency demonstrates our method's robustness and superior generalization ability compared to baseline InfoNeRF, particularly in its resistance to overfitting challenges.

5. CONCLUSION

This paper introduces Sid-NeRF, a few-shot novel view synthesis method based on scene information distribution. Specifically, Sid-NeRF combines color and geometric supervision, using a selector during training for data accuracy and a residual module to counteract optimization bias. Efficient in generating high-quality views from limited inputs, Sid-NeRF performs well in various settings from synthetic to real-world scenes, adeptly managing challenges like lighting and obstructions. Sid-NeRF does not rely on any prior knowledge and therefore does not cause additional consumption of storage and computational resources, increasing its utility in various applications of novel view synthesis.

6. REFERENCES

[1] 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*, vol. 65, no. 1, pp. 99–106, 2021.

[2] X. Miao, Y. Bai, H. Duan, F. Wan, Y. Huang, Y. Long, and Y. Zheng, "Conrf: Zero-shot stylization of 3d scenes with conditioned radiation fields," *arXiv preprint arXiv:2402.01950*, 2024.

[3] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, "Depth-supervised nerf: Fewer views and faster training for free," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12882–12891, 2022.

[4] Y.-J. Yuan, Y.-K. Lai, Y.-H. Huang, L. Kobbelt, and L. Gao, "Neural radiance fields from sparse rgb-d images for high-quality view synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[5] J. Wang, Z. Sun, Z. Tan, X. Chen, W. Chen, H. Li, C. Zhang, and Y. Song, "Towards effective usage of human-centric priors in diffusion models for text-based human image generation," 2024.

[6] F. Wan, J. Wang, H. Duan, Y. Song, M. Pagnucco, and Y. Long, "Community-aware federated video summarization," in *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2023.

[7] H. Duan, Y. Long, S. Wang, H. Zhang, C. G. Willcocks, and L. Shao, "Dynamic unary convolution in transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12747–12759, 2023.

[8] Z. Wang, X. Li, H. Duan, and X. Zhang, "A self-supervised residual feature learning model for multifocus image fusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 4527–4542, 2022.

[9] W. Tong, X. Guan, J. Kang, P. Z. H. Sun, R. Law, P. Ghamisi, and E. Q. Wu, "Normal assisted pixel-visibility learning with cost aggregation for multiview stereo," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24686–24697, 2022.

[10] A. Criminisi, A. Blake, C. Rother, J. Shotton, and P. H. Torr, "Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming," *International Journal of Computer Vision*, vol. 71, pp. 89–110, 2007.

[11] X. Miao, Y. Bai, H. Duan, Y. Huang, F. Wan, Y. Long, and Y. Zheng, "Ctnerf: Cross-time transformer for dynamic neural radiance field from monocular video," 2024.

[12] F. Shao, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Depth map coding for view synthesis based on distortion analyses," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 4, no. 1, pp. 106–117, 2014.

[13] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1831–1838, IEEE, 2010.

[14] R. Gao, F. Wan, D. Organisciak, J. Pu, H. Duan, P. Zhang, X. Hou, and Y. Long, "Privacy-enhanced zero-shot learning via data-free knowledge transfer," in *ICME 2023*, pp. 432–437, 2023.

[15] F. Wan, X. Miao, H. Duan, J. Deng, R. Gao, and Y. Long, "Sentinel-guided zero-shot learning: A collaborative paradigm without real data exposure," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[16] S. Pang, X. He, W. Hao, and Y. Long, "Feature fine-tuning and attribute representation transformation for zero-shot learning," *Computer Vision and Image Understanding*, vol. 236, p. 103811, 2023.

[17] J. Wang, Y. Jiang, Y. Long, X. Sun, M. Pagnucco, and Y. Song, "Deconfounding causal inference for zero-shot action recognition," *IEEE Transactions on Multimedia*, 2023.

[18] M. M. Johari, Y. Lepoittevin, and F. Fleuret, "Geonerf: Generalizing nerf with geometry priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18365–18375, 2022.

[19] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann, "Point-nerf: Point-based neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5438–5448, 2022.

[20] P. Truong, M.-J. Rakotosaona, F. Manhardt, and F. Tombari, "Sparf: Neural radiance fields from sparse and noisy poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4190–4200, 2023.

[21] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5610–5619, 2021.

[22] X. Miao, Y. Bai, H. Duan, Y. Huang, F. Wan, X. Xu, Y. Long, and Y. Zheng, "Ds-depth: Dynamic and static depth estimation via a fusion cost volume," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[23] C. Deng, C. Jiang, C. R. Qi, X. Yan, Y. Zhou, L. Guibas, D. Anguelov, *et al.*, "Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors," in *CVPR*, pp. 20637–20647, 2023.

[24] M. Kim, S. Seo, and B. Han, "Infonerf: Ray entropy minimization for few-shot neural volume rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12912–12921, 2022.

[25] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. Nießner, "Dense depth priors for neural radiance fields from sparse input views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12892–12901, 2022.