SCNeRF: Feature-Guided Neural Radiance Field from Sparse Inputs

Junting Li[1], Yanghong Zhou[1,2] and P. Y. Mok[1,2,*]
[1] The Hong Kong Polytechnic University [2]Research Centre of Textiles for Future Fashion

Abstract

Reconstructing neural radiance fields from limited or sparse views has given very promising potential for this field of research and its future development. Previous methods usually constrain the reconstruction process with additional priors, e.g. semantic information from observed views were exploited as a regularization for unseen views. Nevertheless, patch-based regularization only gives rough supervision to the field and it lacks an additional guide for training views. Instead, we propose a Self-Conditioned NeRF (SCNeRF) in this paper that can learn extra information from features extracted from pre-trained neural networks for the sparse training views, and use as extra guide for the training of the radian field. With such extra feature guides, SCNeRF predicts more accurate color and density when synthesizing novel views. Experimental results have shown that SCNeRF can effectively improve the quality of the synthesized novel views with only limited or sparse inputs.

1. Introduction

Neural radiance field (NeRF), with an impressive ability of novel-view synthesis, and its related studies [1, 7] have grown rapidly in recent years. Nevertheless, NeRF is trained on a large number of views, typically requiring tens of views, which has imposed various restrictions on its potential applications. It is indeed difficult, if not impractical, to collect such large numbers of views for training purpose in some real-life scenarios.

Learning a NeRF from sparse inputs, therefore, has attracted a great deal of research attentions. However, training a NeRF with input views as few as three is very challenging, and the resulted novel views have significantly degraded quality. One approach of existing methods is to predict a new scene by learning some knowledge from similar scenes. For example, PixelNeRF [13] proposed to condition a NeRF on convolutional feature maps projected from each view. Another approach is to provide regularization for the unobserved views by introducing different priors. RegNeRF [8] regularizes unobserved views using geometry priorities and appearance estimation from a 2D model. DietNeRF [3] constrains the learning process by measuring the high-level semantic similarity between the unobserved view and the input views. Although these methods can improve the result in a sparse setting, patch-based regularization can only provide local supervision.

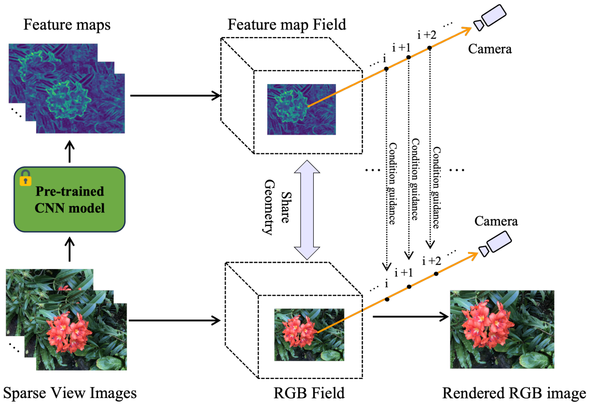

Figure 1. Feature guidance in SCNeRF: we learn the feature field and RGB radiance field simultaneously and the abstract feature is employed to help the learning of RGB field. The feature field and RGB radiance field share the same geometry.

We argue that the RGB representation learned from the NeRF training process often lacks a holistic understanding of the image content. The novel view rendered from this representation is highly sensitive to changes in lighting and viewing angles, making the task even more challenging. Therefore, we propose to improve NeRF by taking advantage of the abstract representations extracted from images, which have a view-independent feature field and can effectively help novel view synthesis. More specifically, we first extract feature maps from a pre-trained CNN model as an abstract representation and learn a feature field. For each point in the field, the feature is utilized as a condition to predict RGB values. The main contributions of the present work are as follows:

- We propose a novel approach that exploits feature representation to enhance the scene representation of NeRF for sparse view inputs.
- Experiments have shown the effectiveness of our proposed method. Compared with the method that uses the

[*]Corresponding author: tracy.mok@polyu.edu.hk

semantic feature for sparse view inputs, SCNeRF can effectively improve performance.

2. Related Work

Novel-view Synthesis with Sparse Inputs by Regularizing Appearance and Geometry. Diet-NeRF [3] regularizes the field by comparing the semantic embedding of unseen viewpoints to that of known viewpoints. RegNeRF [8] regularizes unobserved views using patch-based depth and color constraints. Instead of constraining the field from unobserved views, other methods explored to improve the result from the limited training views. SparseNeRF [9] distills depth information predicted from a prior model to constrain the geometry. FreeNeRF [12] explores frequency of position encoding and trains the field in a coarse-to-fine process. ReconFusion [11] exploits a diffusion prior for novel view synthesis. Although regularization from unobserved views is useful for sparse input, existing methods do not fully explore using prior information as a guidance for training.

Semantic Decomposition of Neural Scene Representations Many methods explore detailed semantic information to understand the neural radiance field. Zhi et al. [16] used semantic labels as additional supervision to learn a field for segmentation. DFF [5] uses semantic feature as a supervision to learn a semantic field that can decomposite different parts of a scene. LeRF [4] distinguishes different parts of a scene through a language model. Furthermore, Latent-NeRF [6] generates 3D scenes represented by latent 3D representations. These works show that a semantic field can also be learned from 2D supervision. Nevertheless, the detailed information of each 3D point is not known yet, in particular in field learning under sparse inputs.

3. Method

3.1. Preliminary

Figure 2(a) shows the typical network structure to learn the RGB radiance field. Given a coordinate point $p_i(x, y, z)$, a network $MLP_\sigma(p)$ is used to predict bottleneck feature f_b and density σ. The feature f_b is then fed into an MLP $MLP_c(p)$ to predict the corresponding color c_i conditioned on the view direction d.

$$(\sigma_i, f_b) = MLP_\sigma(p_i) \tag{1}$$

$$c_i = MLP_c((f_b, d)) \tag{2}$$

With density σ_i, the corresponding weight w_i is then computed by

$$w_i = T_i(1 - exp(-\sigma_i\delta_i)), \tag{3}$$

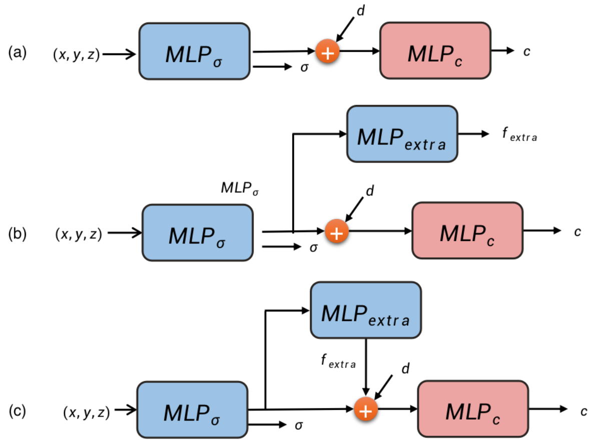

Figure 2. Network structure comparison of different methods: (a) original structure of NeRF [7]. (b) Other methods [5, 15] learned a network to predict feature representation and color of a point simultaneously. (c) our network predicts the color of a point with the guidance of its feature prediction.

where δ_i is the distance to the adjacent sample point, and $T_i = exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j)$ is the transmittance of the ray which presents the probability that information from the 1^{st} sampled point can pass through to the $i-1^{th}$ sampled point. The pixel color can be rendered by

$$C(r) = \sum_{i=1}^{N} w_i(r)c_i(r), \tag{4}$$

where N is the number of sampled points along r between the predefined near and far planes. For model training, Mean Square Error (MSE) is exploited to minimize the distance between rendered colors and ground truth colors:

$$L_c = ||C(r) - C^{gt}(r)||_2^2 \tag{5}$$

3.2. SCNeRF

Different from the typical NeRF, we propose to predict both color and feature fields, and the knowledge of the predicted feature field can explicitly guide the color prediction. The network structure of our proposed method is shown in Figure 2(c).

Feature Supervision For the input view images, we first extract the corresponding feature maps F_f^{gt} from the ReLU3-1 layer of the pre-trained VGG model, which are trained on the ImageNet dataset for classification. Based on the shared feature f_b, an MPL layer is exploited to predict the features of input images:

$$f_{extra} = MLP_{extra}(p_i) \tag{6}$$

where f_{extra} represents the share feature at the point p_i. Similar to the rendering of pixel color, we render the feature representation with eqs (3) and (4), and optimize the model parameters by minimizing the distance between the

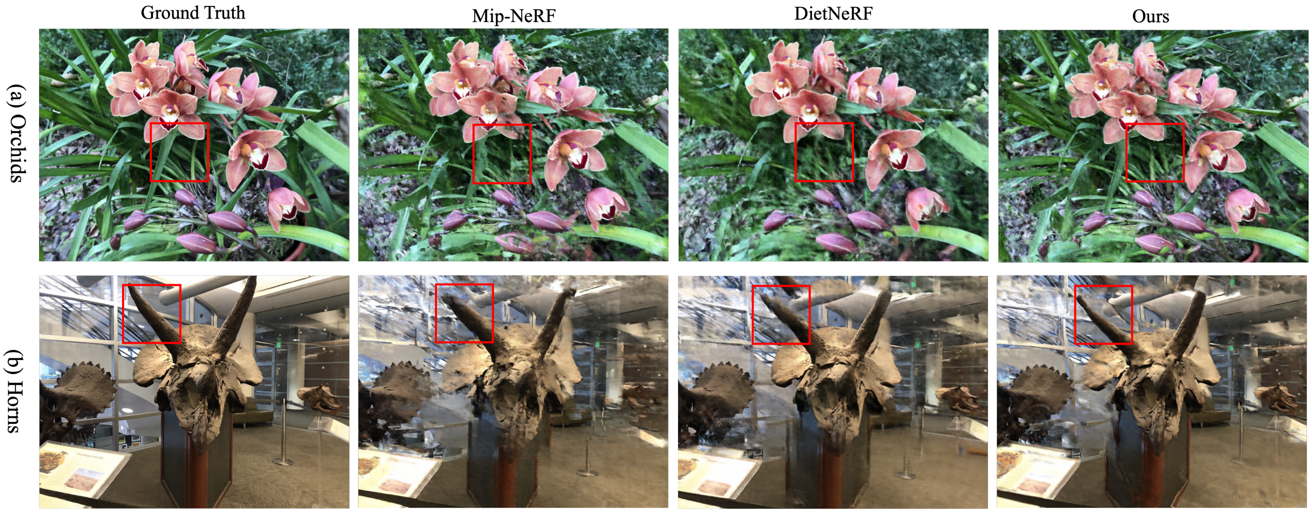

Figure 3. Qualitative comparison results: with extra feature guides, our method can reduce floaters and reconstruct better quality objects.

rendered feature representations $F_{extra}(r)$ and the extracted features from pre-trained model with L_2 loss:

$$L_{infor} = ||F_{extra}(r) - F^{gt}_{extra}(r)||_2^2 \tag{7}$$

Feature guidance In order to exploit the knowledge of feature field to improve the RGB field, we concatenate the shared feature f_b and the predicted feature representation f_{extra}, and use an MLP layer MLP_c to enhance the scene representation and predict the pixel color as follows:

$$c_i = MLP_c(f_b, d; f_{extra}) \tag{8}$$

The overall loss function is:

$$L = \lambda L_{infor} + L_c, \tag{9}$$

where λ is a weight balances the constraint of extra information.

4. Experimental Results

Dataset. We conducted experiments on the LLFF dataset. LLFF consists of 8 forward-facing scenes. Following [7], we kept every 8^th image as the hold-out test set and selected the training views evenly from the remaining images. We report our result of setting in 3 views.

Baseline and Evaluation Metrics. DietNeRF [3] is chosen as our baseline, because it synthesizes well results given spare views. Different from DietNeRF [3] that extracts semantic representation from CLIP model to guarantee semantic consistency between different views, we extract the feature from CNN model which not only contains the abstract representations of the image but also the local spatial information, and use it to guide the learning of color prediction of 3D scene. Based on the public code of DietNeRF [3], we trained our baseline model in 3 views. To evaluate the view synthesis performance, we used PSNR, SSIM [10] and LPIPS [14] as evaluation metrics. In the ablation study, we conducted the experiments on flower scene and compared these metrics within this context as shown in Tables 2–3. For fair comparison, we conducted the experiments on all scenes and calculate the mean of these metrics as shown in Table 1.

Implementation Details. We implement our framework based on the network structure of Figure 2(c). We used Adam optimizer to optimize models, with the exponential learning rate decreasing from 5×10^{-4} to 5×10^{-5}. Each scene was trained for 200k iteration. We used an ImageNet pre-trained VGG encoder [2] to extract image feature and set λ to 0.01 in our experiment. The model can be replaced with more powerful CNN models, like ResNet.

4.1. Comparison with SOAT Methods

The qualitative and quantitative comparison results with the SOAT methods are shown in Table 1 and Figure 3, respectively. We compared the SOAT methods that do not use depth information, Mip-NeRF [1] and DietNeRF [3]. As shown in Table 1, our method surpasses both Mip-NeRF [1] and DietNeRF [3] in terms of all the metrics. As shown in Figure 3, our method can also reconstruct some more reasonable information of scene. It's important to note that unlike DietNeRF [3], our method is supervised solely with sparse input views, without the need for a sampling strategy to acquire unobserved views during training.

4.2. Ablation Studies

To show the effectiveness of our framework, we compare our method with the NeRF [7] structure Figure 2 (a) and the DFF [5] structure Figure 2 (b) in a sparse view setting. DFF

Table 1. Quantitative Comparison with SOAT methods.

Method	PSNR↑	SSIM↑	LPIPS↓
Mip-NeRF	14.62	0.351	0.495
DietNeRF	14.94	0.370	0.496
Ours	**17.80**	**0.568**	**0.432**

Table 2. Ablation study evaluating the impact of the proposed modules on the flower scene of LLFF dataset.

Method	PSNR↑	SSIM ↑	LPIPS↓
w/o FS	17.78	0.540	0.434
w/ FS	16.01	0.461	0.534
Full	18.65	0.567	0.428

also learns feature representtion and color from the train view simultaneously. However, feature and color are predicted parallel from a bottleneck in DFF.

We conducted the experiments on the flower scene of LLFF dataset to evaluate the effect of different modules of our method in a setting of 3 views. We trained a feature field and a radiance field without the feature guidance similar to the DFF [5]. The qualitative and quantitative comparison results are shown in Figure 4 and Table 2, respectively. The feature guidance can reduce floaters [1] in the scene. As shown, in a sparse setting, the prediction of color without feature representation as a condition (e.g., [5]) leads to an under-constrained reconstruction field (Figure 4(a)). Furthermore, simultaneous learning of feature representation exacerbates the issue (Figure 4(b)). However, this observation conversely highlights that additional feature supervision has an impact on the geometry of the field. By incorporating our feature guidance, the field is capable of reconstructing a more meaningful scene. Moreover, compared to solely feature supervision, our method excels in capturing finer details of the leaves (Figure 4(c)). In addition, the feature L_{infor} loss can also constrain the geometry of the scene and lead to fewer artifacts.

4.3. Discussion

We explore the effect of loss weight λ in eq. (9) and the different layers for extracting the features as guidance for color prediction. It can be seen from Table 3 that the model obtains the best result when the λ is set as 0.1 in terms of PSNR and SSIM. As shown in Table 4, when using the feature from Relu1-1 layer to guide color prediction, our method achieves the best results. This is because that the feature maps from lower layers have more low-level information such as color, edge and corners, which provide more detail information to guide the novel view image synthesis.

5. Conclusions

In this paper, we propose a novel framework SCNeRF that synthesizes novel views with sparse view inputs. To tackle the under-constrained few-shot NeRF problem, our proposed SCNeRF learns the feature field to help the learning of the color field. The experimental results have also shown that our method significantly improves the synthesis performance in a sparse setting and is complementary to the previous methods.

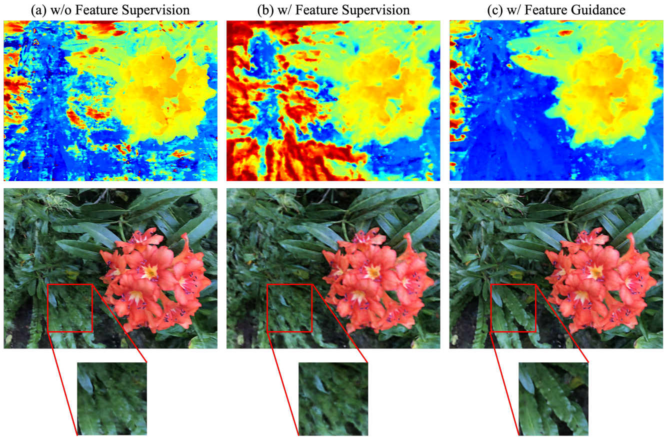

Figure 4. Ablation study of our proposed different modules: the result with feature supervision is better than that without feature supervision and the results with feature guidance are better than that with feature supervision. The first row is the depth map and the second row is the synthesized novel view image.

Table 3. Comparison of different loss weights for the flower scene of the LLFF dataset

λ	PSNR↑	SSIM↑	LIPIPS↓
0.1	18.86	0.589	0.408
0.01	18.65	0.567	0.428
0.001	18.29	0.583	0.400

Table 4. Effectiveness comparison of using features extracted from different layers as guidance for color prediction of novel view images on the flower scene of the LLFF dataset.

Layers	PSNR↑	SSIM↑	LIPIPS↓
Relu1-1	18.93	0.594	0.378
Relu2-1	18.61	0.593	0.391
Relu3-1	18.65	0.567	0.428

Acknowledgments

We sincerely thank Prof. J.T. Fan, Dr D.H. Shou and Dr S. Xu for constructive feedback and guidance. The work described in this paper is supported, in part, by The Hong Kong Polytechnic University (Grant No. Q-8883), National Natural Science Foundation of China (Grant No. 62202241), and Jiangsu Province Natural Science Foundation for Young Scholars (Grant No. BK20210586).

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1, 3, 4

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[3] 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. 1, 2, 3

[4] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 2

[5] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. 2, 3, 4

[6] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 2

[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. In *ECCV*, 2020. 1, 2, 3

[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

[9] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 2

[10] 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. 3

[11] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. *arXiv preprint arXiv:2312.02981*, 2023. 2

[12] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 2

[13] 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. 1

[14] 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. 3

[15] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[16] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 2