Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 15, No. 3, 2024

Mixer-NeRF: Research on 3D Reconstruction Methods of Neural Radiance Fields Based on Hybrid Spatial Feature Information

Linhua Jiang [1, 2, 3], Ruiye Che[1], Lingxi Hu [1, 2, 3], Xinfeng Chen [1], Yuanyuan Yang [1], Lei Chen [1], Wentong Yang [1], Peng Liu [1], Wei Long[1, 2, 3, *]

[1] School of Information Engineering, Huzhou University, Huzhou 313000, China
[2] Huzhou Key Laboratory of Waters Robotics Technology, Huzhou University, Huzhou, 313000, China
[3] Zhejiang-France Joint Laboratory for Digital Monitoring of Aquatic Resources and the Environment, Huzhou, 313000, China
[*] **Corresponding author**: Wei Long

Abstract: Compared to traditional 3D reconstruction techniques, Neural Radiance Fields (NeRF) have demonstrated significant performance advantages in implicit 3D reconstruction. However, simple multi-layer perceptron (MLP) models often result in blurry details in the reconstructed 3D scenes due to their lack of necessary local information capture during the sampling process. To address this issue, this paper proposes a joint learning method based on hybrid spatial feature information to enhance detail reconstruction. First, after the position encoding output, a hybrid spatial feature information learning module (MLCA) is added before entering the MLP network and before the input direction information, capturing spatial feature information and mixing the channels of these features to obtain the features of different channels at each spatial location. Then, a squeeze-and-excitation network module is introduced between the NeRF sampling and inference layers to filter out high-weight information. Experimental results show that the proposed 3D reconstruction method based on hybrid spatial feature information can effectively improve NeRF's reconstruction performance. It is applicable to complex, real, multi-view scenarios, especially showing the best performance under high-texture and complex lighting conditions. Compared to NeRF, this method improves the learning of perceptual image block similarity by more than 30%.

Keywords: Neural Radiance Fields; Joint Learning; 3D Reconstruction; Feature Extraction.

1. Introduction

The research on 3D reconstruction can be traced back to the 1960s and is one of the core topics in the field of computer vision. 3D reconstruction refers to the process of recovering the three-dimensional shape from two-dimensional images and interpreting the spatial relationships between objects[1]. Although 3D scanners based on structured light principles can provide high-precision 3D reconstruction[2], their high costs and complex operation procedures remain major bottlenecks in practical applications[3]. In recent years, researchers have combined traditional physical multi-view geometry techniques with deep learning[4].This innovative approach leverages the powerful learning capabilities of neural networks to directly infer the structural information of 3D scenes from 2D image data[5]. This advancement enables neural networks to efficiently infer 3D scene information directly from 2D images[6].

Currently, 3D reconstruction methods can be categorized into geometric explicit methods and neural implicit methods based on how they process 3D data[7]. Common explicit representations include point clouds[8], meshes[9], and voxels[10]. While these methods can successfully capture the 3D features of objects, they typically focus on densely sampling and reconstructing local regions of objects. This means that a large amount of image data is required to obtain a complete 3D model. Furthermore, these methods often query 3D geometric prior information during reconstruction, which increases the memory resources needed for view reconstruction[11].In contrast, view reconstruction methods based on implicit representations offer broader application prospects due to their compact scene representations that occupy less storage space and their independence from view resolution[12]. One of the most effective implicit methods in recent years is Neural Radiance Fields (NeRF)[13]. NeRF's sampling strategy involves inputting a set of sparse views into a Multi-Layer Perceptron (MLP) model, implicitly mapping the color and density of objects from 3D spatial positions to 2D pixels. This achieves high-quality novel view synthesis. However, NeRF's simple linear fully connected sampling method often results in the loss of local information, leading to blurry and aliased reconstructed views[14].To address these issues, Zhang K., Riegler G., and Snavely N., among others, analyzed NeRF's potential problems in their work NeRF++ and proposed a novel spatial parameterization scheme—Inverse Spherical Parameterization—which can handle a new class of unbounded scenes[15]. Arandjelović R. and Zisserman A. proposed a training method called NeRF-ID, which introduced smaller MLP modules to achieve superior view synthesis quality compared to NeRF on benchmarks. However, the issues of view blurring and aliasing remain[16]. Yang G. W., Zhou W. Y., and Peng H. Y. proposed the Recursive Neural Fields, which apply different numbers of linear layers in the MLP structure at different stages recursively, enabling adaptive reconstruction across different scenes[17].

Although there have been various improvements to the MLP structure of NeRF in the literature, research on the extraction and integration of hybrid spatial information remains limited, making it difficult to address issues such as feature sparsity or missing details when local details appear infrequently. To address this, this paper proposes a joint

learning method based on hybrid spatial feature information to enhance the ability to capture mixed information in the spatial dimension, making it more flexible when processing large-scale scenes or integrating multi-modal data. During the processing of multi-view information, this paper designs a new hybrid spatial feature learning module (MLCA) that adapts to the input information of NeRF. It can effectively integrate the input hybrid spatial features and enhance the ability to capture global spatial information. Additionally, during feature inference, a squeeze-and-excitation network module (SE-Net)[18]is incorporated, which uses a channel attention mechanism to dynamically weight the features for the MLP, suppressing redundant features and highlighting important information.

2. Neural Radiance Fields (NeRF)

The NeRF structure is divided into a cosine-sine position encoding network, a feature extraction network, and a volume rendering network, with the overall network structure shown in Figure 1. The feature extraction network consists of two MLP models: one is an 8-layer fuly connected sampling MLP model, and the other is a 3-layer fully connected inference MLP model. The core idea of NeRF is to use a neural network to represent the radiance field in 3D space. The radiance field mainly contains two types of information: first, volume density, which represents the geometric information of each point in the scene (e.g., whether the point is occupied by an object); and second, color values, which represent the color emitted by each point in the scene in a specific direction. NeRF uses multi-layer perceptrons (MLPs) as functions to map 3D coordinates and viewing directions to volume density (σ) and color values (c):

$$\mathrm{F}(\theta):(X,d)\rightarrow(c,\sigma) \tag{1}$$

NeRF integrates the camera rays into pixels and uses the volume rendering equation to independently compute the color of each sampling point along the ray, ultimately reconstructing the color of the pixel C(r)in the image:

$$C(r)=\int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t),d)dt \tag{2}$$

Where:

$$\mathrm{T}(t)=\exp(-\int_{t_n}^{t}\sigma(r(s))ds) \tag{3}$$

It is the accumulated opacity, representing the probability that the ray is not occluded from the starting point to t. The input to NeRF is a 3D spatial point X and a direction d. To enhance the model's ability to capture high-frequency details and allow the MLP to better fit high-frequency information to compensate for the network's spectral bias, the input is mapped to a high-dimensional space through Fourier position encoding:

$$\gamma(x)=(\sin(2^0\pi x),\cos(2^0\pi x),\sin(2^1\pi x),\cos(2^1\pi x),\cdots,\sin(2^L\pi x),\cos(2^L\pi x)) \tag{4}$$

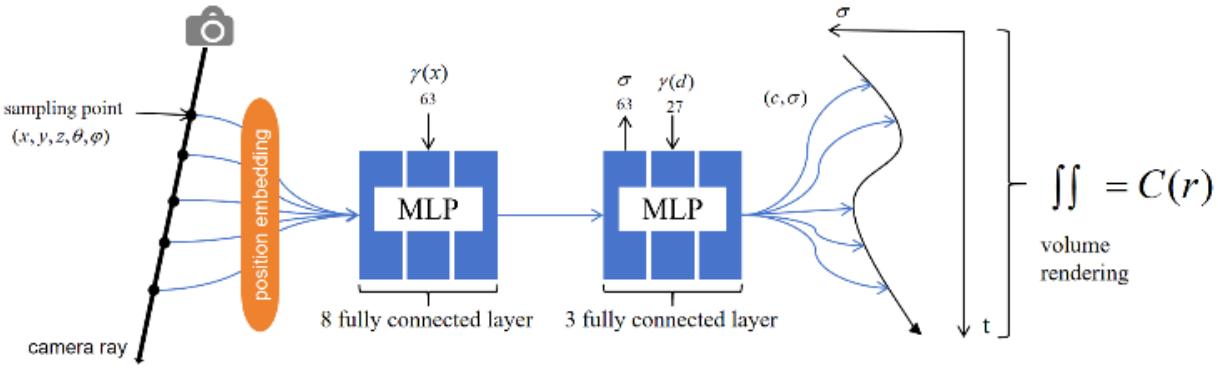

Figure 1. The neural network structure used in NeRF to represent the radiance field.

In the equation, $\gamma(x)$ represents the position encoding of the 3D location X and the orientation d; L denotes the dimensionality of the position encoding. In NeRF, the position encoding for the 3D location X is set with $L=10$, while for the camera orientation d, the position encoding is set with $L=4$. After this position encoding, X and d are mapped to high-dimensional features.

To improve the sampling frequency and enhance rendering efficiency and detail quality, NeRF incorporates stratified sampling in its sampling strategy. It is mainly divided into two stages: coarse sampling and fine sampling. The coarse sampling provides a rough estimate of the ray color while generating the weight distribution of the volume density. The fine sampling stage uses these weights to guide the distribution of new sampling points along the ray, more accurately capturing regions that significantly contribute to the result. This strategy optimizes the distribution of sampling points in stages, allowing the network to focus more on regions that make a larger contribution to the rendering outcome.

From the description of the neural radiance field above, it is clear that the core idea of NeRF is to use a neural network to fit the radiance field in 3D space and generate high-quality images through volume rendering. Its network structure includes a cosine-sine position encoding network, a feature extraction network (which contains sampling and inference MLPs), and a volume rendering network. Through the stratified sampling strategy, NeRF improves sampling efficiency and rendering detail quality. However, there are still some issues in calculating the sparsity of viewpoints, which affect the view reconstruction results. Therefore, this paper proposes a neural radiance field based on hybrid spatial feature information, Mixer-NeRF, to address these issues in NeRF.

3. Construction of the Mixer-NeRF Network

Mixer-NeRF consists of an improved backbone, a hybrid spatial feature learning module, and a squeeze-and-excitation

network module, with the network structure shown in Figure 2. After the position encoding output and before entering the MLP network, as well as before the input direction information, the hybrid spatial feature learning module (MLCA) is added. The position encoding generates high-frequency features, which are the core of the continuous coordinate mapping. Through the MLCA module, mixed associations can be established in the feature dimension space. Additionally, a squeeze-and-excitation network module is constructed between the sampling and inference layers of the feature extraction network to filter out high-weight information.

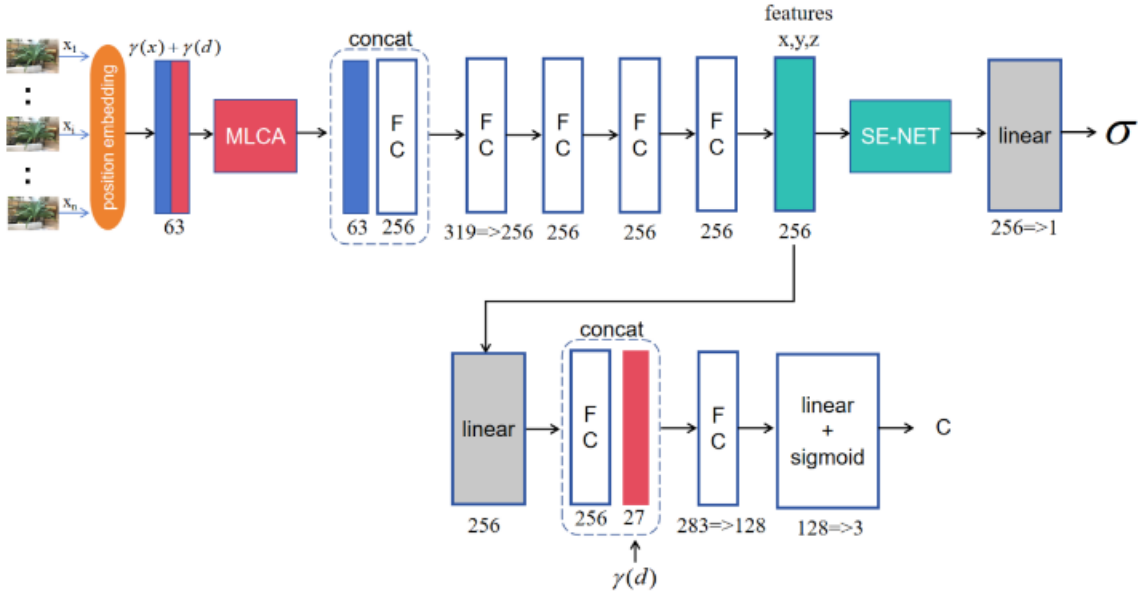

Figure 2. The network structure of Mixer-NeRF

3.1. Hybrid Spatial Information Learning Module

The structure of the hybrid spatial feature learning module is shown in Figure 3. NeRF typically takes 3D spatial positions and viewing angles as inputs. To accommodate these inputs, the hybrid spatial feature learning module requires some transformations. After the spatial position is encoded through an encoding layer, the position encoding is fed into the fully connected layer of the module for sine and cosine function encoding, which increases high-frequency information. The position encoding process can be represented as:

$$PE(x) = [\sin(2^i \pi x), \cos(2^i \pi x)]_i^L \quad (5)$$

Here, L represents the number of encoding layers, which adds high-frequency information at multiple scales to help the network learn more details. Similarly, the viewing angle information is processed in a similar encoding manner to generate the corresponding embedding.

$$PE(d) = [\sin(2^i \pi x), \cos(2^i \pi x)]_i^L \quad (6)$$

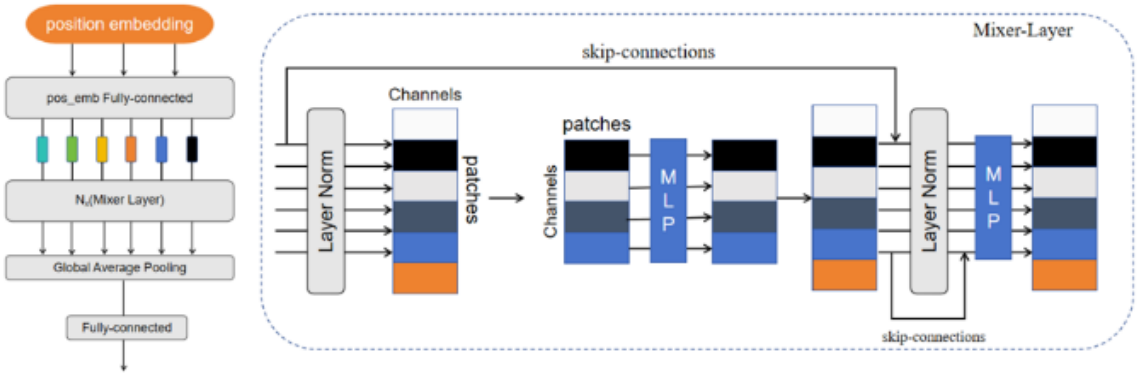

Figure 3. Structure of the Hybrid Spatial Information Learning Module

The structure of the Mixer Layer can be divided into two parts: the spatial feature capture layer and the mixed feature channel layer. In NeRF, each input spatial position X and view direction d are encoded and then passed into the first layer of the Hybrid Spatial Information Learning Module. The spatial features of the position are then mixed through the spatial feature capture layer. The encoding of the input position and view direction can be represented as:

$$X_{input} = Concat(PE(x), PE(d)) \quad (7)$$

Next, the spatial feature capturing layer will pass this input through fully connected layers to mix the spatial features:

$$X_{token} = GELU(W_{token} X_{input} + b_{token}) \quad (8)$$

The hybrid feature channel layer mixes the feature channels to capture the relationships between different channel features (such as color and density) at each spatial location. This

operation is performed through a fully connected layer and is typically applied after the spatial feature capturing layer:

$$X_{channel} = GELU\left(W_{channel} X_{token} + b_{channel}\right) \quad (9)$$

The final output comes from the features passed through the mixed feature channel layer, which are then used to generate color c and density σ.

3.2. Squeeze-and-Excitation Network Module

Cheng et al.[19]demonstrated that the Squeeze-and-Excitation Network (SE-net) module performs well in feature extraction for images with complex textures and intensity distributions. Therefore, in this paper, the SE-net module is placed between the sampling layer and the inference layer, where it can directly affect the prediction of density and color, particularly helping with detail restoration and light color in complex scenes. The structure of the Squeeze-and-Excitation Network is shown in Figure 4. Given an input x, with c_1 feature channels, a series of convolutions and other transformations are applied to obtain a feature with c_2 feature channels.

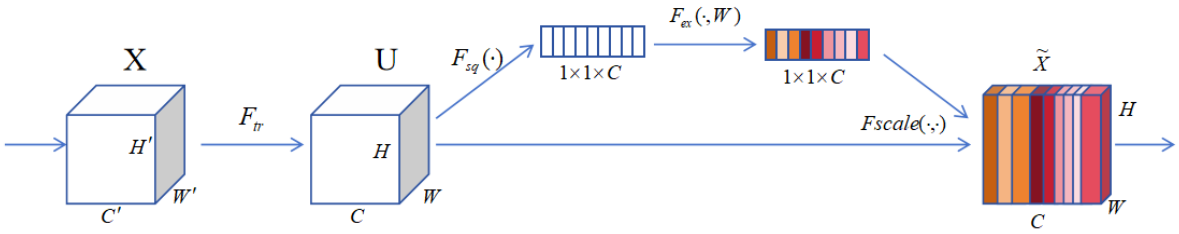

Figure 4. Squeeze-and-Excitation Network Structure

The squeeze-and-excitation network module (SE-net) consists of two main operations: Squeeze (global information extraction) and Excitation (channel relationship modeling). The Squeeze operation performs global average pooling on the input features F, resulting in global features:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{H} F_c(i, j) \quad (10)$$

Here, H and W represent the dimensions of the feature space. The Excitation operation processes the global feature z through a two-layer perceptron to generate the channel weights:

$$s = \sigma(W_2 \cdot \mathrm{Re}\,LU(W_1 \cdot z + b_1) + b_2) \quad (11)$$

Here $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$, r is the compression ratio, and σ is the sigmoid activation function.

Then, the features F of each channel are weighted:

$$F'_c = s_c \cdot F_c \quad (12)$$

The weighted features F' are then output and passed separately into NeRF's density and color branches:

$$\sigma = MLP_\sigma(F'), c = MLP_c(F', view) \quad (13)$$

4. Experimental Results and Analysis

4.1. Datasets and Experimental Parameters

Two public datasets were used for the experiments, namely the nerf_Synthetic dataset and the nerf_llff_data dataset. The nerf_Synthetic dataset contains path-tracked images of eight objects with complex geometric structures and realistic non-Lambertian materials. It includes six viewpoints sampled from a hemisphere and two viewpoints sampled from a full sphere. Each scene has 100 views as input, with 200 views reserved for testing. All views are 800x800 pixels. The nerf_llff_data dataset consists of scenes captured with a handheld mobile phone, with each scene containing 20 to 62 images, and one-eighth of the images are reserved for the test set. All images are 1008x756 pixels.

The experiments were conducted on a Linux server running Ubuntu 22.04.1. The CPU used is an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz, and the GPU is an NVIDIA A100-PCIE-40GB. The CUDA 12.4 software toolkit was used to accelerate parallel computing applications on the GPU. The algorithm was implemented using the PyTorch deep learning framework based on Python 3.8. To evaluate the quality of the new view reconstruction, image quality metrics were employed. The Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) were used as performance metrics to assess the quality of the new view reconstruction. A higher PSNR and SSIM value, along with a lower LPIPS value, indicate better reconstruction quality.

4.2. Ablation Experiment

To verify the effectiveness of the proposed method, ablation experiments were conducted by sequentially setting up the following groups: Experiment Group A used only the NeRF backbone network; Experiment Group B added only the SE-net module; Experiment Group C added only the MLCA module; and Experiment Group D incorporated both the MLCA and SE-net modules. The performance of novel view reconstruction was evaluated on the Chair scene from the nerf_Synthetic dataset and the Horns scene from the nerf_llff_data dataset. Tables 1 and 2 present the results of the ablation experiments, with the optimal metrics highlighted in bold, and the percentages in parentheses indicating the improvement relative to NeRF.

Table 1. Ablation Experiment on Novel View Reconstruction in the Chair Scene

Ablation experiment	PSNR/dB	SSIM	LPIPS
A	33.00	0.967	0.019
B	34.21(+3.67%)	0.973(+0.62%)	0.017 (-10.5%)
C	35.23(+6.76%)	0.981(+1.45%)	0.013 (-31.6%)
D	36.78(+11.5%)	0.986(+1.96%)	0.010 (-47.3%)

Table 2. Ablation Experiment on Novel View Reconstruction in the Horns Scene

Ablation experiment	PSNR/dB	SSIM	LPIPS
A	27.45	0.828	0.068
B	27.56(+0.4%)	0.834(+0.72%)	0.068 (-0.00%)
C	27.94(+1.79%)	0.845(+2.05%)	0.065 (-4.41%)
D	28.31(+3.13%)	0.861(+3.99%)	0.062 (-8.82%)

From the data in Tables 1 and 2, the following observations can be made: Compared to Experiment Group A, Experiment Group B, which only added the SE-net module, improved the average PSNR, SSIM, and LPIPS values of reconstructed views in the Chair scene by 1.21 dB (3.67%), 0.06 (0.62%), and −0.002 (−10.5%), respectively. In the Horns scene, the average PSNR and SSIM values increased by 0.11 dB (0.4%) and 0.006 (0.72%), respectively, indicating a moderate enhancement in reconstruction performance. Compared to Experiment Group A, Experiment Group C, which only added the MLCA module, improved the average PSNR, SSIM, and LPIPS values of reconstructed views in the Chair scene by 2.23 dB (6.76%), 0.014 (1.45%), and −0.006 (−31.6%), respectively. In the Horns scene, the average PSNR, SSIM, and LPIPS values increased by 0.49 dB (1.79%), 0.017 (2.05%), and −0.003 (−4.41%), respectively, without compromising reconstruction performance. Compared to Experiment Group A, Experiment Group D, which incorporated both the MLCA and SE-net modules, achieved the most significant improvements in the Chair scene, with increases of 3.78 dB (11.5%), 0.019 (1.96%), and −0.009 (−47.3%) in average PSNR, SSIM, and LPIPS values, respectively. In the Horns scene, the average PSNR, SSIM, and LPIPS values increased by 0.86 dB (3.13%), 0.033 (3.99%), and −0.006 (−8.82%), respectively. These results demonstrate that the combination of the two modules enables the network to simultaneously capture spatial hybrid information and channel attention, thereby enhancing its ability to process overall global information. This validates the effectiveness of the combined use of the two proposed modules.

4.3. Comparative Experiments

To verify the performance of the proposed method in novel view reconstruction, training and testing were conducted on a total of 16 scenes from the nerf_Synthetic and nerf_llff_data datasets. Additionally, experiments were carried out for comparison with NeRF and NeRF-ID, using consistent experimental parameters. The results of the comparative experiments are shown in Tables 1 and 2, with the optimal results in the comparison experiments highlighted in bold. The views related to the experiments are shown in Figure 5.

The comparison experiment data on the nerf_Synthetic dataset is shown in Table 3. The proposed method improves the average PSNR, SSIM, and LPIPS values compared to NeRF-ID by 0.15 dB (0.46%), 0.001 (0.1%), and −0.001 (3.44%), respectively. Compared to NeRF, it shows improvements of 1.48 dB (4.77%), 0.011 (1.16%), and −0.013 (31.7%). The proposed method outperforms the NeRF comparison group in all eight scenes of the dataset, with certain improvements compared to the NeRF-ID comparison group in the reconstructed views.

The comparison experiment data on the nerf_llff_data dataset is shown in Table 4. The proposed method improves the average PSNR, SSIM, and LPIPS values compared to NeRF-ID by 0.69 dB (2.59%), 0.017 (2.06%), and −0.003 (4.29%), respectively. Compared to NeRF, it shows improvements of 0.95 dB (3.58%), 0.028 (3.45%), and −0.003 (4.29%). The proposed method performs better than the other comparison groups in most scenes of the nerf_llff_data dataset.

From the data comparison above, the proposed method demonstrates a significant advantage over the other two methods in terms of PSNR. The improvement is even more pronounced on the nerf_llff_data dataset, indicating its better performance in handling complex textures and lighting conditions.

According to the comparison experiment results in Table 3 (nerf_Synthetic dataset) and Table 4 (nerf_llff_data dataset), Mixer-NeRF shows significantly lower LPIPS values on both datasets compared to NeRF and NeRF-ID, indicating its more effective optimization of visual perceptual details. For example, on the nerf_Synthetic dataset, LPIPS is reduced by 31.7% compared to NeRF, and on the nerf_llff_data dataset, it is reduced by 4.29%. This improvement can be attributed to the Mixed Space Information Learning Module (MLCA), which enhances the ability to capture spatial and channel features by combining high-frequency features with position encoding, thus effectively reducing reconstruction blur and modal aliasing. Meanwhile, the Squeeze-and-Excitation Network (SE-Net) module highlights key features through channel attention, demonstrating significant restoration ability, especially for complex lighting conditions and texture details.

The performance on the real-world dataset (nerf_llff_data) surpasses that on the synthetic scene dataset (nerf_Synthetic), with Mixer-NeRF showing a greater improvement on the nerf_llff_data dataset. For example, in the Fern scene, Mixer-NeRF improves the PSNR by 3.49% compared to NeRF. The nerf_llff_data dataset contains more complex geometry and lighting characteristics, which may explain why the MLCA module is better suited to handle these intricate features, enabling more accurate capture of spatial correlations and variations in light ray color. By combining global and local features, Mixer-NeRF is able to better handle irregular lighting and multi-view information in the scene, thereby improving reconstruction quality.

At the same time, in simple texture scenes (such as Flower, Hotdog, Lego, etc.), the performance improvement of Mixer-NeRF is smaller, sometimes slightly worse than NeRF-ID. For example, in the Flower scene, the PSNR of Mixer-NeRF is 27.82, slightly lower than NeRF-ID's 27.85. This may be because simple texture scenes require less complex feature extraction, and the global feature capture in Mixer-NeRF could introduce feature interference, adversely affecting the color and density near object surfaces in featureless backgrounds. In some scenes, there is an issue with view blur. For example, in the Lego scene of the nerf_Synthetic dataset, the LPIPS of Mixer-NeRF is 0.017, higher than NeRF-ID's 0.015. This could be because the feature mixing in the MLCA module may negatively impact detail restoration in uniform texture backgrounds, leading to local blur.

Overall, the superior performance of Mixer-NeRF compared to other methods can be attributed to the complementary and synergistic nature of its modular design. First, the Mixed Space Information Learning Module (MLCA) enhances the ability to capture spatial information, effectively

improving the expression accuracy of complex geometry and high-frequency features. Second, the Squeeze-and-Excitation Network (SE-Net) optimizes the channel attention mechanism, dynamically selecting key features and suppressing redundant information interference, which allows for more precise restoration of scene details. Furthermore, the improved backbone architecture works in tandem with the hierarchical sampling strategy, achieving a better balance between sampling efficiency and reconstruction quality. Thanks to the overall design and adaptability of these modules, Mixer-NeRF demonstrates significant advantages in complex and real-world scenes. Its consistent and coherent improvements across various performance metrics, such as PSNR, SSIM, and LPIPS, fully demonstrate the superiority of this method in new view reconstruction tasks.

Table 3. Comparison of Parameters for Different Methods on the nerf_Synthetic Dataset

Scene	NeRF			NeRF-ID			Mixer-NeRF		
	PSNR/dB	SSIM	LPIPS	PSNR/dB	SSIM	LPIPS	PSNR/dB	SSIM	LPIPS
Chair	33.00	0.967	0.019	34.54	0.978	0.014	36.78	0.986	0.010
Drums	25.01	0.925	0.058	25.15	0.926	0.057	26.47	0.932	0.052
Lego	32.54	0.961	0.024	34.73	0.974	0.015	32.62	0.968	0.017
Ship	28.65	0.856	0.119	29.75	0.876	0.081	29.35	0.873	0.091
Mic	32.91	0.980	0.023	34.71	0.988	0.009	35.10	0.988	0.009
Ficus	30.13	0.964	0.022	32.24	0.976	0.015	31.14	0.973	0.016
Materials	29.62	0.949	0.029	30.37	0.956	0.024	31.29	0.963	0.017
Hotdog	36.18	0.974	0.016	37.26	0.981	0.013	37.21	0.978	0.015
Mean	31.01	0.947	0.041	32.34	0.957	0.029	32.49	0.958	0.028

Table 4. Comparison of Parameters for Different Methods on the nerf_llff_data Dataset

Scene	NeRF			NeRF-ID			Mixer-NeRF		
	PSNR/dB	SSIM	LPIPS	PSNR/dB	SSIM	LPIPS	PSNR/dB	SSIM	LPIPS
Fern	25.20	0.792	0.092	25.01	0.800	0.089	26.08	0.825	0.082
Flower	27.40	0.827	0.061	27.85	0.842	0.058	27.82	0.839	0.061
Fortress	31.16	0.881	0.030	31.51	0.888	0.028	31.84	0.903	0.026
Horns	27.45	0.828	0.068	27.88	0.843	0.065	28.31	0.861	0.062
Leaves	20.92	0.690	0.111	21.09	0.708	0.108	21.53	0.755	0.105
Orchids	20.36	0.641	0.121	20.38	0.643	0.120	20.44	0.674	0.116
Room	32.70	0.948	0.041	32.93	0.954	0.039	32.86	0.941	0.042
Trex	26.80	0.880	0.057	27.45	0.897	0.051	27.73	0.917	0.045
Mean	26.50	0.811	0.073	26.76	0.822	0.070	27.45	0.839	0.067

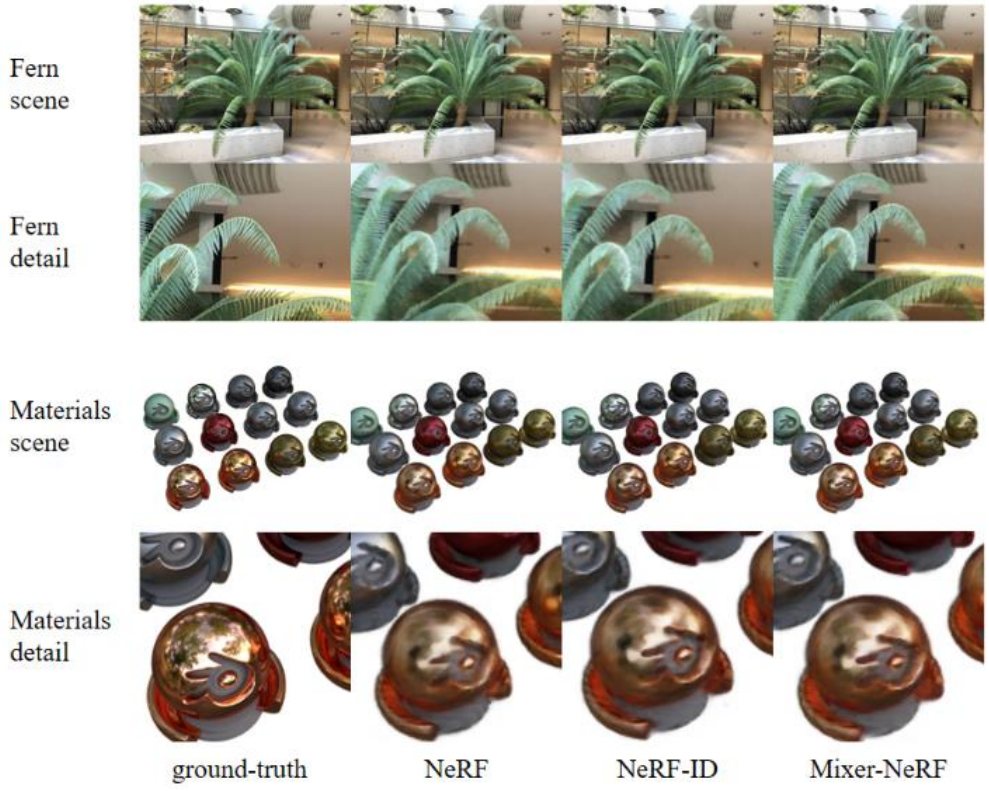

Figure 5. The performance of different methods on two datasets

Figure 5 shows the view results obtained by different reconstruction methods for the Fern scene in the nerf_llff_data dataset and the Materials scene in the nerf_Synthetic dataset. The first column of Figure 5 shows

the ground truth images that were not used for training, while the second to fourth columns show the reconstructed views from NeRF, NeRF-ID, and Mixer-NeRF, respectively. By comparing the reconstructed views with the ground truth images in the first column, we can observe the local variations in the new view reconstruction results of different methods. Qualitative results show that the proposed method demonstrates excellent view reconstruction ability, avoiding rendering blur and aliasing, and provides richer details in the reconstruction, such as the branching of leaves and the bump texture on the sphere.

5. Conclusion

Neural Radiance Fields (NeRF) is an important representation of implicit neural expressions. To address the issue of NeRF's limited ability to capture local and mixed dependencies in space, this paper proposes a Mixed Space Information Learning method, Mixer-NeRF. The method is based on an improved backbone structure, where the Mixed Space Information Learning Module (MLCA) is added after the positional encoding output, before entering the MLP network, and before the input direction information, to enhance the ability to capture mixed space information. Additionally, a Squeeze-and-Excitation Network (SE-Net) module is introduced between the sampling and inference layers of the feature extraction network to effectively filter high-weight information.

Mixer-NeRF, through training and testing on two public datasets, shows significant performance improvement over NeRF. On the nerf_Synthetic dataset, the proposed method improves PSNR, SSIM, and LPIPS by 1.48 dB (4.77%), 0.011 (1.16%), and -0.013 (31.7%) respectively. On the nerf_llff_data dataset, the proposed method improves PSNR, SSIM, and LPIPS compared to NeRF by 0.95 dB (3.58%), 0.028 (3.45%), and -0.003 (4.29%). By introducing the mixed space information learning method, this paper not only effectively enhances the performance of the Mixer-NeRF model but also provides valuable insights for the further improvement of NeRF.

Acknowledgements

Funding: This work was supported in part by the Zhejiang-French Digital Monitoring Lab for Aquatic Resources and Environment, Department of Science and Technology of Zhejiang Province and in part by the Huzhou Key Laboratory of Waters Robotics Technology (2022-3), Huzhou Science and Technology Bureau.

References

[1] Zhang R, Tsai P S, Cryer J E, et al. Shape-from-shading: a survey[J]. IEEE transactions on pattern analysis and machine intelligence, 1999, 21(8): 690-706.

[2] Yin Y K, Yu K, Yu C Z, et al. 3D imaging using geometric light field: a review[J]. Chinese Journal of Lasers, 2021, 48(12): 1209001.

[3] Zhang B,Yu J,Jiao X,et al.Line Structured Light Binocular Fusion Filling and Reconstruction Technology[J].Laser Optoelectron.Prog,2023, 60: 1611001.

[4] Shi S F, Ye N, Zhang L Y. Calibration of two-camera vision system with far and near sight distance[J]. Acta Optica Sinica, 2021, 41(24): 2415001.

[5] Liu H X, Zhao Y M, Zhang C L. Study on tooth cone beam CT image reconstruction based on improved U-net network[J]. Chinese Journal of Lasers, 2022, 49(24): 2407207.

[6] Wang M J, Li L, Yi F. Three-dimensional reconstruction and analysis of target laser point cloud data under simulated real water environment[J]. Chinese Journal of Lasers, 2022, 49(3): 0309001.

[7] Mingyang L I, Wei C, Shanshan W, et al. Survey on 3D reconstruction methods based on visual deep learning[J]. Journal of Frontiers of Computer Science & Technology, 2023, 17(2): 279.

[8] Bui G, Le T, Morago B, et al. Point-based rendering enhancement via deep learning[J]. The Visual Computer, 2018, 34: 829-841.

[9] Riegler G, Koltun V. Free view synthesis[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. Springer International Publishing, 2020: 623-640.

[10] Sitzmann V, Thies J, Heide F, et al. Deepvoxels: Learning persistent 3d feature embeddings[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2437-2446.

[11] Srinivasan P P, Tucker R, Barron J T, et al. Pushing the boundaries of view extrapolation with multiplane images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 175-184.

[12] Schirmer L, Schardong G, da Silva V, et al. Neural networks for implicit representations of 3D scenes[C]//2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2021: 17-24.

[13] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[14] Zhu Z, Peng S, Larsson V, et al. Nice-slam: Neural implicit scalable encoding for slam[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12786-12796.

[15] Zhang K, Riegler G, Snavely N, et al. Nerf++: Analyzing and improving neural radiance fields[J]. arXiv preprint arXiv:2010.07492, 2020.

[16] Arandjelović R, Zisserman A. Nerf in detail: Learning to sample for view synthesis[J]. arXiv preprint arXiv:2106.05264, 2021.

[17] Yang G W, Zhou W Y, Peng H Y, et al. Recursive-nerf: An efficient and dynamically growing nerf[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(12): 5124-5136.

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[19] Cheng D, Meng G, Cheng G, et al. SeNet: Structured edge network for sea–land segmentation[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 14(2): 247-251.