第 47 卷第 1 期
Vol. 47 No. 1 2026

青岛理工大学学报
Journal of Qingdao University of Technology

基于轻量级 Transformer 和质数基混合编码的神经辐射场

万子寒[1]，李兆阳[2]，孔巍吉[3]，丁纪翔[4]

（1. 北京科技大学 计算机与通信工程学院，北京 100083；2. 东南大学 生物科学与医学工程学院，南京 211189；
3. 西南交通大学 计算机与人工智能学院，成都 611756；4. 北京理工大学 前沿交叉科学研究院，北京 100081）

摘　要：神经辐射场（NeRF）作为当前三维场景重建领域最热门的技术，存在着伪影和高频信号拟合能力不足的问题。为了解决伪影问题，提出了一种轻量级 Transformer 模型，通过 Attention 结构对输入的特征进行筛选。为了解决高频信号拟合能力不足的问题，提出了一种质数基混合编码方法，通过向原始位置编码引入高斯编码，并以多个质数为基底重构高频编码部分。将两种方法进行融合，并在 NeRF-synthetic 和 NeRF-LLFF 数据集上进行实验，实验结果表明，本文提出的方法在峰值信噪比（PSNR）和结构相似性指数（SSIM）两项指标上得到了提升，验证了本文方法的有效性。

关键词：神经辐射场；三维场景重建；混合编码
中图分类号：TP37　　**文献标志码**：A　　**文章编号**：1673-4602(2026)01-0082-07
DOI：10.3969/j.issn.1673-4602.2026.01.011

Neural radiance fields based on lightweight transformer and prime-based hybrid encoding

WAN Zihan[1], LI Zhaoyang[2], KONG Weiji[3], DING Jixiang[4]

(1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China; 2. School of Biological Science and Medical Engineering, Southeast University, Nanjing 211189, China; 3. School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China; 4. Advanced Research Institute of Multidisciplinary Sciences, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Neural radiance fields (NeRF) has emerged as a leading technique in 3D scene reconstruction. However, it suffers from artifacts and insufficient capacity for fitting high-frequency signals. To address these issues, a lightweight Transformer model was introduced to mitigate artifacts, in which the Attention mechanism was leveraged to filter input features. To enhance the fitting of high-frequency signals, a prime-based hybrid encoding method was designed, which incorporated Gaussian encoding into the original positional encoding and replaced standard high-frequency components with encodings based on multiple prime numbers. By integrating two approaches, experiments were conducted on the NeRF-synthetic and NeRF-LLFF datasets. The results demonstrate that the proposed method achieves improvements in both PSNR and SSIM metrics, validating its effectiveness.

Key words: neural radiance fields (NeRF); 3D reconstruction; hybrid encoding

收稿日期：2025-03-25
基金项目：福建省自然科学基金(2023J01135)
作者简介：万子寒(2002—　)，男，安徽芜湖人。硕士，研究方向为计算机视觉。E-mail：wanzihan@xs.ustb.edu.cn。

近年来，三维场景渲染和三维场景重建技术被广泛应用于人工智能、机器人、虚拟现实和自动驾驶领域[1]。传统的三维场景重建领域，通常采用运动恢复结构法或基于统计机器学习的方法[2]。随着神经网络的发展，传统方法在很多领域都被基于神经网络的方法超越，三维场景重建领域也不例外，比较有代表性的显示场景表示有基于体素表示的 3D-R2N2[3]、基于点云表示的 PointNet[4]、基于网格表示的 Pixel2Mesh[5]；具有代表性的隐式场景表示有 OccupancyNet[6]、DeepSDF[7]以及 Neural Radiance Fields (NeRF)[8]。

2020 年，MILDENHALL 等[8]采用神经网络拟合连续 5D 坐标到像素和体积密度的映射函数，再通过体渲染技术生成指定视角下的图片。其出色的性能以及较小的内存消耗，使得 NeRF 成为最为火热的研究方向，紧随其后有着一批出色的研究人员有针对性地对 NeRF 现存的问题做了一系列的工作。BARRON 等[9]通过采用圆锥体采样方式解决了严重的混叠伪影问题。BIAN 等[10]通过引入未失真的单目深度先验来约束连续帧之间的相对姿态，从而解决在复杂相机运动下的 NeRF 训练问题。DENG 等[11]提出使用运动恢复结构法的稀疏输出对 NeRF 进行监督，这样可以减少输入的视角并提高训练速度。NIEMEYER 等[12]通过引入多项损失对三维场景进行约束，以实现在稀疏视角下的新视角合成。METZER 等[13]将 NeRF 带入潜在空间，并通过得分蒸馏技术结合 Sketch-Shape 来引导 3D 形状生成，从而增加对生成过程的控制。LI 等[14]通过引入 Transformer 来联合聚合辐射场和语义嵌入场，以促进上下文感知的 3D 场景感知。

尽管 NeRF 已经在视图合成领域取得了令人瞩目的成就，但仍面临一些挑战，如图像细节不清晰的问题，这主要是由于高频分量的空间编码缺失造成的。同时，高频分量空间编码重叠也会导致伪影现象的出现。为了解决这些问题，本文提出了一种基于 Transformer 架构的神经网络解决方案，该方案有效地增强了模型的抗伪影能力，进而提升了图像的真实感和可信度。此外，本文还引入了一种质数基混合位置编码方法，通过引入新的高频编码函数来补充输入中的高频成分，从而有效弥补了原有模型在高频信息捕捉方面的不足。

这种结合轻量级 Transformer 结构与质数基混合位置编码的方法不仅强化了对复杂场景的处理能力，还在很大程度上解决了 NeRF 中存在的伪影问题，并且提高了图像细节的表现力，为三维场景重建和渲染技术的进步提供了新的视角和方法。这种方法的灵活性和有效性使其在不同的应用场景中都具有广阔的前景。

1　轻量级 Transformer 网络

TANCIK 等[15]指出，位置编码操作可以视为一种预训练的高频特征提取卷积核。基于这一观点，本文提出将位置编码的输出视作卷积神经网络中的通道，并采用基于通道的 Attention 机制对这些位置编码的输出进行筛选，以赋予重要通道更高的权重。这种方法能够在一定程度上缓解由于过高频率编码导致的高频混叠伪影问题，同时尽可能保留原始信息。

基于上述思路，本文引入了 Transformer 结构来替代传统的多层感知机(MLP)结构。考虑到 Transformer 最初是为自然语言处理任务设计的，在将其应用于当前场景时，对 Transformer 进行了适应性修改。具体而言，本文提出的轻量级 Transformer 架构仅包含一个编码器和一个解码器，并且在编码输入的过程中，仅在需要对输入进行特征筛选时使用 Attention 机制。这种设计不仅提高了模型的效率，还增强了其对于关键特征的选择性关注能力。图 1 展示了本文所提出的完整模型结构。

输入的 x、y、z 由质数基混合编码层编码后进入编码器进行编码，通过 Attention 结构对质数基混合编码的特征进行筛选，筛选后将特征编码到合适的维度，在中间引入残差连接[16]，将经过特征筛选后的原始输入信息传入中间层进行编码。参考 NeRF 中的结论，采用两种不同的解码器对编码信息进行解码。由于辐射场中点的体积密度 σ 只与该点的坐标相关，因此，设计一个独立的解码器对坐标 x、y、z 的编码结果进行解码得到体积密度σ。由于辐射场中点的颜色信息不仅和点的坐标 x、y、z 相关并且与其角度信

息 θ、φ 也相关，因此，同样对角度信息 θ、φ 进行位置编码，再采用同样的通道 Attention 机制对特征进行筛选，并将筛选后的结果与编码器的输出进行拼接，最后输入联合的解码器中对拼接后的信息进行解码，得到坐标点对应的 R、G、B 信息。

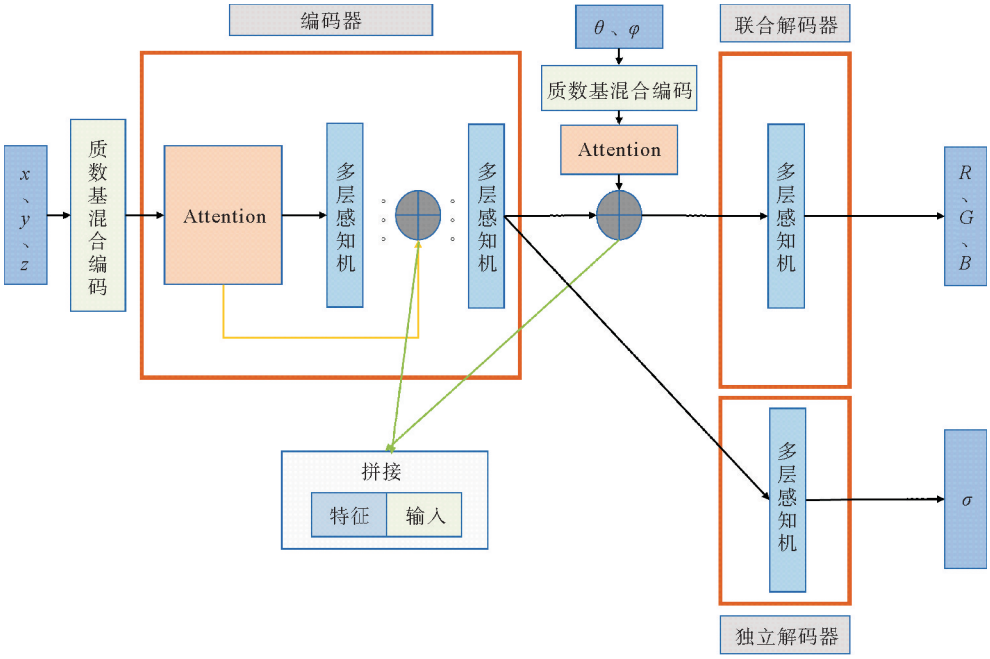


图 1 网络结构

2 质数基混合位置编码

ZHENG 等[17]提出采用非周期性的高斯函数作为位置编码中的编码函数。高斯函数如式(1)所示，其中 σ 是高斯函数的超参数。该方法不仅可以取得与傅里叶编码相同的性能，并且在编码高维度时更加有效。

$$G(x)=\exp\left(\frac{-x^2}{2\sigma}\right) \tag{1}$$

鉴于周期性编码函数可能限制采样效率，ZHENG 等认为使用非周期性函数可以解决这一问题。然而，本文认为尽管傅里叶编码引入了周期性，但其在频率域上提供了特定的特征和信息，因此，完全用高斯函数替代傅里叶编码可能会丢失这些重要的变换信息。基于此，本文提出一种新颖的混合位置编码方法，将非周期性的高频成分与周期性成分相结合，通过补充而非替换的方式利用非周期性高频信息，以提高采样效率并保留原有信息。该方法首先对输入进行高斯编码和傅里叶编码，然后将两者的编码结果拼接起来形成最终的位置编码 $\gamma'(x)$，如式(2)所示。其中，h 为高频函数，在本文中为高斯函数。

$$\gamma'(x)=\begin{bmatrix}\sin(2^0x) & \cdots & \sin(2^Nx)\\ \cos(2^0x) & \cdots & \cos(2^Nx)\\ \vdots & \vdots & \vdots\\ h(2^0x) & \cdots & h(2^Nx)\end{bmatrix} \tag{2}$$

由式(2)可以看出，当使用 2 的整数次幂作为频率基础时，sin 函数与 cos 函数易导致高频与低频成分的重叠。为了克服这个问题，提出了改进方案：利用互质的质数组合作为底数构造一组低频编码函数 $\gamma''(x)$，如式(3)所示，其中 $p_1,\cdots,p_n$ 为 n 个较小的质数，N 为编码的最大频率。

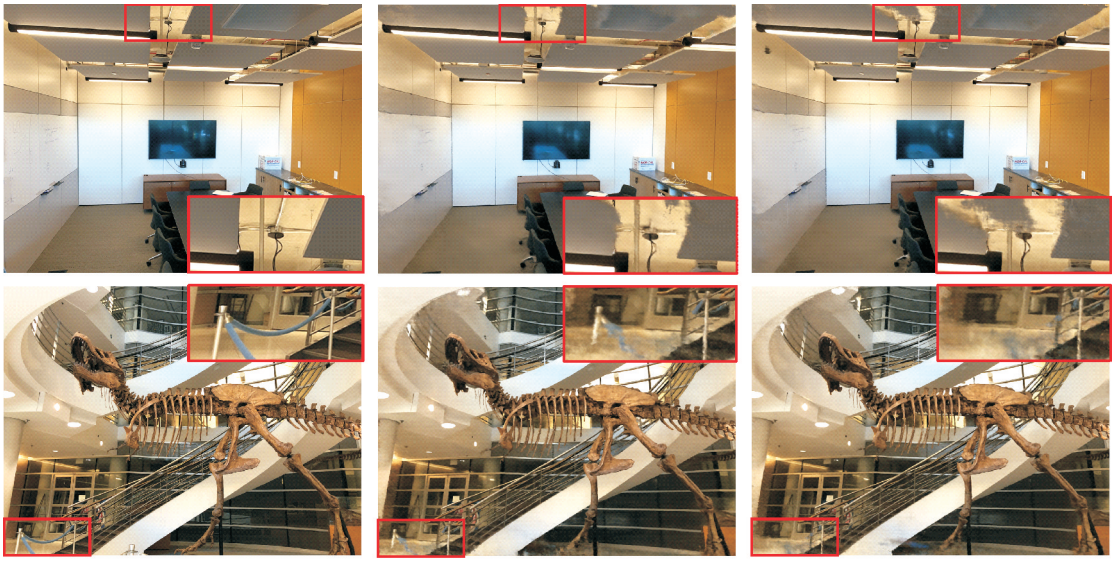
$$\gamma''(x)=\begin{pmatrix}\sin(p_1^0x) & \cdots & \sin(p_n^0x)\\ \vdots & & \vdots\\ \sin(p_1^Nx) & \cdots & \sin(p_n^Nx)\\ \cos(p_1^0x) & \cdots & \cos(p_n^0x)\\ \vdots & & \vdots\\ \cos(p_1^Nx) & \cdots & \cos(p_n^Nx)\\ G(p_1^0x) & \cdots & G(p_n^0x)\\ \vdots & & \vdots\\ G(p_1^Nx) & \cdots & G(p_n^Nx)\end{pmatrix} \tag{3}$$

如果将位置编码方法视为将原本的编码空间拓展到了频率域空间，那么质数编码的方法就是拓展了频率域空间的大小，用互质的若干簇数作为系数，可以在更小的编码频率下，取得与高编码频率近似的结果，同时解决混叠和重复的问题。扩大频率域空间也可以提升渲染的效果，在更大的空间中，神经网络可以更好地学习到数据中的信息，更便于拟合 NeRF 中的函数映射关系。

3　实验结果

本文实验数据采用 NeRF-synthetic[8] 和 NeRF-LLFF[18] 中的部分数据。NeRF-synthetic 数据集是由软件制作的合成图像组成，所有视图均为 800×800 像素。NeRF-LLFF 数据集由真实场景下拍摄的图片构成的，所有视图均为 1008×756 像素。实验在 RTX3090 上训练 200 000 轮。本文采用一个 batch size 4096 个 rays，粗采样 64 个点，精细采样 128 个点。采用 Adam 优化器，学习率设置为 5×10^{-4}，优化器一阶矩估计的衰减率 $\beta_1=0.9$，二阶矩估计的衰减率 $\beta_2=0.99$。

本次实验选取了渲染图像中的两组进行抗伪影能力的对比，实验结果如图 2 所示。最左侧列展示了作为模型渲染参考的原图，中间列显示了在 NeRF 基础上引入 Transformer 结构后渲染的结果图，最右侧列则是采用原始 NeRF 算法渲染的结果。每张图中红色框标记的区域是相同的观察点，用于直接对比不同方法下的渲染效果。从图中红色框标注的区域可以明显看出，原始 NeRF 方法仍存在较为显著的伪影现象，而本文提出的基于 Transformer 结构的改进版 NeRF 则展现了非常优异的抗伪影性能。具体来说，基于 Transformer 的NeRF在网络生成红色框内区域时产生的伪影面积显著少于原始NeRF，并且该区

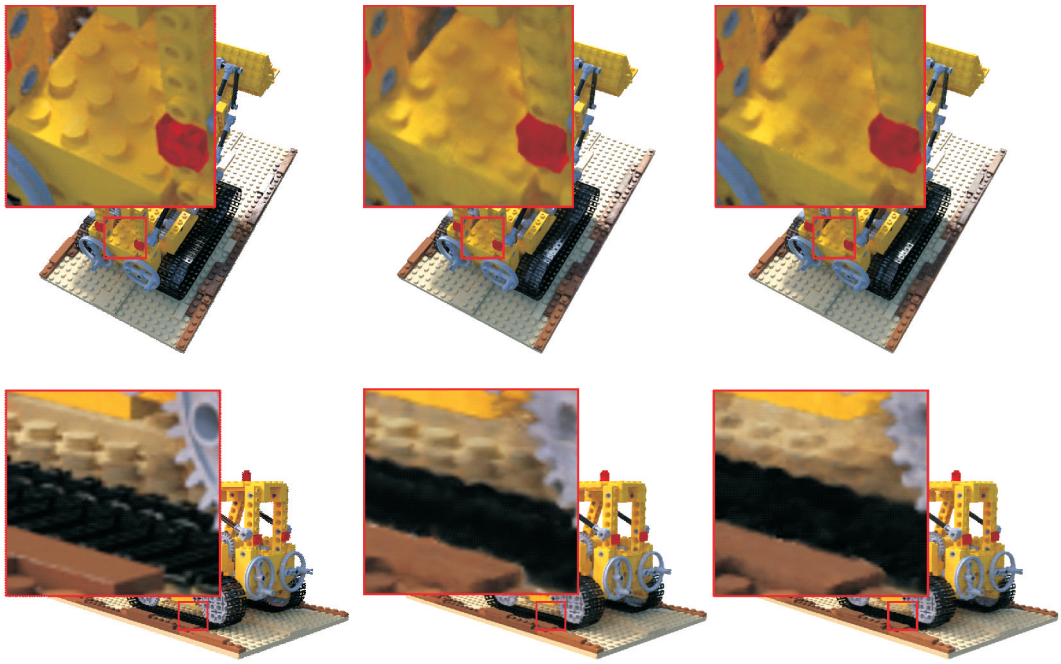


(a) 原图　　(b) NeRF+Transformer　　(c) NeRF

图 2　抗伪影能力对比

域内的图像清晰度也更高。因此，通过对渲染图像细节的仔细分析可以看出，本文所提出的结合 Transformer 结构的 NeRF 不仅有效地减轻了伪影问题，还提高了图像的整体质量与细节表现力，进一步证明了其更优的抗伪影能力和更高的视觉保真度。

本次实验通过对比引入质数基混合编码后的 NeRF 与原始 NeRF 渲染的图像，验证了质数基混合编码在高频信号重建能力上的优势，实验结果如图 3 所示。最左列展示了作为模型渲染参考的原图，中间列显示了基于 NeRF 并引入质数基混合编码后渲染的结果图，最右列则是采用原始 NeRF 算法渲染的结果。每张图中的红色框标记了同一个区域，以便于直接比较不同方法下的渲染效果。从图中红色框标注的区域可以看出，在细节处理上，引入质数基混合编码的 NeRF 表现出了显著的优势，尤其是在一些包含细节和高频信息的场景中，质数基混合编码能够有效地减少模糊现象，使得图像细节更加清晰。这表明，相较于原始 NeRF 使用的普通编码方式，质数基混合编码不仅能保留更多的细节信息，还能更准确地重建高频信号，从而提供更加细腻和真实的视觉效果。



(a) 原图 (b) NeRF+质数基混合编码 (c) NeRF

图 3 高频信号重建效果对比

本文提出将轻量级 Transformer 与质数基混合编码引入神经辐射场(NeRF)，构建了一个融合模型。实验结果显示，该模型在多个数据集上表现优异，具体结果如表 1 所示。表中以加粗字体标示了各数据集上的最佳表现。由表 1 可以看出，在大多数数据集中，融合模型均取得了最优的表现，仅在 Lego 数据集上，加入了 Transformer 结构的 NeRF 模型展现出了最佳性能。这表明所提出的融合方法对于提升 NeRF 模型的渲染质量具有显著效果。

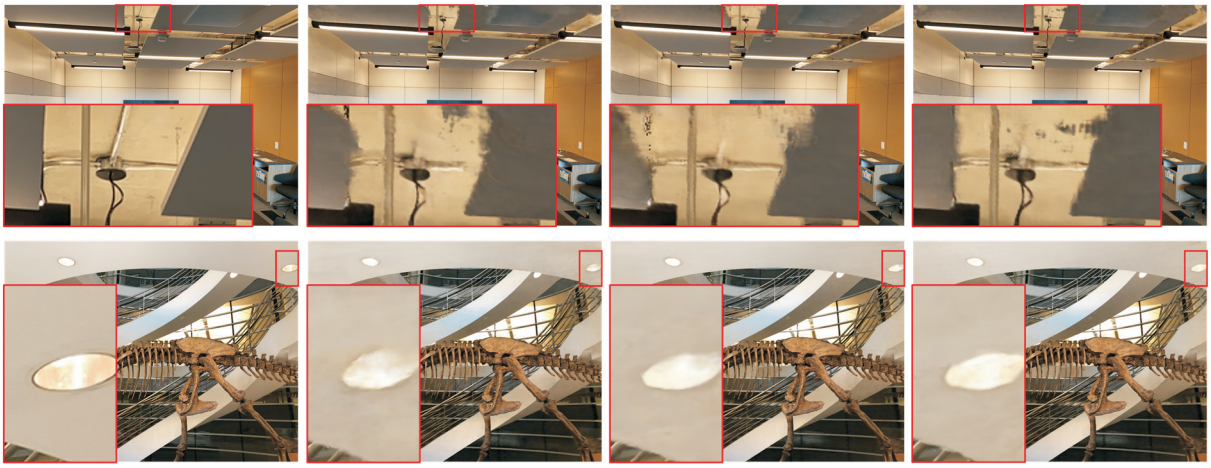
4 种模型的渲染结果对比如图 4 所示。从中可以明显看出融合模型相较于仅包含 Transformer 或者质数基混合编码的模型，在渲染效果上实现了显著提升。具体而言，融合模型不仅有效消除了大部分伪影，还在高频细节的渲染上表现更佳，成功结合了两种方法的优点。这种融合之所以能够实现上述优势，主要是因为质数基混合编码与 Transformer 结构在架构层面不存在冲突，质数基混合编码的结果可以直接作为输入传递给 Transformer 结构，从而避免了不同模块间的不兼容问题。此外，Transformer 中的 Attention 机制能够对质数基编码的结果进行特征筛选，这一过程实质上提升了质数基编码的质量。最终的实验结果表明，通过采用 Transformer 结构在一定程度上缓解了伪影问题，借助质数基混合编码增强了对高频信号的拟合能力，Attention 机制在此基础上进一步优化了由质数基引入的高频特征，确保了融合

模型能够实现更加精细和真实的渲染效果。

表 1　融合模型与其他方法的实验结果对比

数据集	NeRF		NeRF+Transformer		NeRF+质数基混合编码		融合模型	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Lego	31.1024	0.9681	**31.5802**	**0.9714**	31.1801	0.9694	31.2588	0.9700
Flower	27.2047	0.8918	27.0718	0.8894	27.2535	0.8948	**27.3563**	**0.8967**
Room	31.2377	0.9659	31.4137	0.9669	31.4091	0.9672	**31.5142**	**0.9673**
Horns	26.4006	0.8778	26.3494	0.8759	26.3856	0.8773	**26.4270**	**0.8782**
Trex	25.6112	0.9035	25.6925	0.9026	25.9176	0.9081	**25.9333**	**0.9083**
Avg	28.3113	0.9214	28.4215	0.9212	28.4292	0.9234	**28.4979**	**0.9241**

注：PSNR(Peak Signal-to-Noise Ratio)，峰值信噪比，用于衡量图像重建质量；SSIM(Structure Similarity Index Measure)，结构相似性指数，反映图像结构信息的相似程度。



(a) 原图　(b) NeRF+Transformer　(c) NeRF+质数基混合编码　(d) 融合模型

图 4　4 种模型渲染结果对比

4　结论

针对 NeRF 在渲染过程中出现的伪影现象以及对高频信号拟合不足的问题，本文提出了一种改进方案，包括引入轻量级 Transformer 结构和采用质数基混合编码的方法。这些改进旨在提升 NeRF 在新视角合成任务中的性能。实验基于 NeRF-synthetic 和 NeRF-LLFF 的部分数据集进行，结果表明，与原始 NeRF 模型相比，图像中伪影现象明显减少，且高频细节表现更为清晰。

本文所提出的轻量级 Transformer 架构展现了强大的抗伪影能力，有效缓解了空间中高频分量编码重叠的问题，在多个数据集上均表现出色。研究表明，通过适当调整网络结构并嵌入有效的 Attention 机制，可以筛选特征，从而在很大程度上解决高频分量的空间编码重叠问题。此外，本文提出的质数基混合编码方法可以很好地保留图像高频信息，尤其是在图像边缘处，能够更清晰地呈现细节。该方法具有广泛的适用性，通过向任何 NeRF 结构中的位置编码添加高频函数来补充高频成分，可以弥补原有编码方式缺失的信息，进而增强图像细节的渲染效果。使用质数基的位置编码不仅有助于降低图像编码维度，还能提高渲染质量，防止高频编码的重叠现象。

最后，本文探讨了将轻量级 Transformer 结构与质数基混合编码相结合的可能性，发现融合这两种方法的模型相较于单独应用其中任何一种方法，在 PSNR 和 SSIM 两项指标以及最后的图像渲染结果上都取得了更为优异的结果。这种结合不仅强化了模型处理复杂场景的能力，还进一步提升了图像渲染的质量，为未来视图合成技术的发展提供了新的方向。

参考文献(References):

[1] 李英群,胡啸,徐翔,等. 三维场景注视点渲染深度学习方法综述[J]. 中国图象图形学报,2024,29(10):2955-2978.
LI Yingqun,HU Xiao,XU Xiang,et al. Deep learning-based foveated rendering in 3D space:A review[J]. Journal of Image and Graphics,2024,29(10):2955-2978.

[2] 张彦雯,胡凯,王鹏盛. 三维重建算法研究综述[J]. 南京信息工程大学学报(自然科学版),2020,12(5):591-602.
ZHANG Yanwen,HU Kai,WANG Pengsheng. Review of 3D reconstruction algorithms[J]. Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2020,12(5):591-602.

[3] CHOY C B,XU D,GWAK J Y,et al. 3D-R2N2:A unified approach for single and multi-view 3D object reconstruction[C]//Proceedings of the European Conference on Computer Vision. Amsterdam,2016:628-644.

[4] QI C R,SU H,MO K,et al. PointNet:Deep learning on point sets for 3D classification and segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,2017:652-660.

[5] WANG N,ZHANG Y,LI Z,et al. Pixel2Mesh:Generating 3D mesh models from single RGB images[C]//Proceedings of the European Conference on Computer Vision. Munich,2018:52-67.

[6] MESCHEDER L,OECHSLE M,NIEMEYER M,et al. Occupancy networks:Learning 3D reconstruction in function space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,2019:4455-4465.

[7] PARK J J,FLORENCE P,STRAUB J,et al. DeepSDF:Learning continuous signed distance functions for shape representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,2019:165-174.

[8] 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.

[9] BARRON J T,MILDENHALL B,VERBIN D,et al. Mip-NeRF 360:Unbounded anti-aliased neural radiance fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans,2022:5470-5479.

[10] BIAN W J,WANG Z R,LI K J,et al. NoPe-NeRF: Optimising neural radiance field with no pose prior[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023:4160-4169.

[11] DENG K,LIU A,ZHU J Y,et al. Depth-supervised NeRF:Fewer views and faster training for free[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022:12882-12891.

[12] NIEMEYER M,BARRON J T,MILDENHALL B,et al. RegNeRF:Regularizing neural radiance fields for view synthesis from sparse inputs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans,2022:5480-5490.

[13] METZER G,RICHARDSON E,PATASHNIK O,et al. Latent-NeRF for shape-guided generation of 3D shapes and textures[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver,2023:12663-12673.

[14] LI H,ZHANG D,DAI Y,et al. Gp-NeRF:Generalized perception nerf for context-aware 3D scene understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,2024:21708-21718.

[15] TANCIK M,SRINIVASAN P,MILDENHALL B,et al. Fourier features let networks learn high frequency functions in low dimensional domains[J]. Advances in Neural Information Processing Systems,2020,33:7537-7547.

[16] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,2016:770-778.

[17] ZHENG J,RAMASINGHE S,LUCEY S. Rethinking positional encoding[J]. arXiv Preprint,2021. DOI:10.48550/arXiv.2107.02561.

[18] MILDENHALL B,SRINIVASAN P P,ORTIZ-CAYON R,et al. Local light field fusion:Practical view synthesis with prescriptive sampling guidelines[J]. ACM Transactions on Graphics (ToG),2019,38(4):1-14.

(责任编辑 张晓靖;英文校审 徐 飞)