E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS


**中图法分类号**:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)10-3346-15

**论文引用格式**:Jia D, Liu Y, Li W, Han X F, Song H L, Meng X H and Liu Y Q. 2025. Novel-view synthesis method integrating local spatial information. Journal of Image and Graphics, 30(10):3346-3360(贾迪, 刘洋, 李维, 韩雪峰, 宋慧伦, 孟晓华, 刘宇琪. 2025. 融合局部空间信息的新视角合成方法. 中国图象图形学报, 30(10):3346-3360)[DOI:10.11834/jig.240673]

# 融合局部空间信息的新视角合成方法

贾迪[1,2],刘洋[2*],李维[2],韩雪峰[1],宋慧伦[2],孟晓华[2],刘宇琪[2]

1. 辽宁工程技术大学鄂尔多斯研究院,鄂尔多斯　017000; 2. 辽宁工程技术大学电子与信息工程学院,葫芦岛　125105

**摘　要**:**目的**　基于点云的神经渲染方法受点云质量及特征提取的影响,易导致新视角合成图像渲染质量下降,为此提出一种融合局部空间信息的新视角合成方法。**方法**　针对点云质量及提取特征不足的问题,首先,设计一种神经点云特征对齐模块,将点云与图像匹配区域的特征进行对齐,融合后构成神经点云,提升其特征的局部表达能力;其次,提出一种神经点云Transformer模块,用于融合局部神经点云的上下文信息,在点云质量不佳的情况下仍能提取可靠的局部空间信息,有效增强了点云神经渲染方法的合成质量。**结果**　实验结果表明,在真实场景数据集中,对于只包含单一物品的数据集Tanks and Temples,本文方法在峰值信噪比(peak signal to noise ratio, PSNR)指标上与NeRF(neural radiance field)方法相比提升19.2%,相较于使用点云输入的方法Tetra-NeRF和Point-NeRF分别提升了6.4%和3.8%,即使在场景更为复杂的ScanNet数据集中,与NeRF方法及Point-NeRF相比分别提升了34.6%和2.1%。**结论**　本文方法能够更好地利用点云的局部空间信息,有效改善了稀疏视角图像输入下因点云质量和提取特征导致的渲染质量下降,实验结果验证了本文方法的有效性。

**关键词**:神经辐射场(NeRF);点云;神经渲染;三维重建;体积密度

# Novel-view synthesis method integrating local spatial information

Jia Di[1,2], Liu Yang[2*], Li Wei[2], Han Xuefeng[1], Song Huilun[2], Meng Xiaohua[2], Liu Yuqi[2]

*1. Ordos Institute of Liaoning Technical University, Ordos　017000, China;*

*2. School of Electronic and Information Engineering, Liaoning Technical University, Huludao　125105, China*

**Abstract**: **Objective** Modeling real-world scenes from image data and generating photorealistic novel views introduce great challenges within the fields of computer vision and graphics. Neural radiance field (NeRF) and its extensions have emerged as highly successful approaches for addressing this challenge by leveraging neural radiance fields. However, these methods frequently reconstruct radiance fields across the entire space using global multi-layer perceptions (MLPs) through ray marching, which results in prolonged reconstruction times. This delay is primarily attributed to the slow fitting of per-scene networks and the excessive sampling of extensive empty spaces. To solve these problems, neural radiance field representation based on point clouds is proposed, which uses 3D points to model the scene. Unlike NeRF, which relies purely on per-scene fitting, this method can be efficiently initialized by a feed-forward deep neural network pretrained across scenes. Furthermore, ray sampling in an empty scene space is avoided by utilizing a classical point cloud that approximates

**收稿日期**:2024-11-15;**修回日期**:2025-02-07;**预印本日期**:2025-02-14

*通信作者:刘洋　liuyang2254399194@163.com

**基金项目**:国家自然科学基金项目(61601213);辽宁省教育厅重点项目(LJ212410147003);辽宁工程技术大学鄂尔多斯研究院校地科技合作培育项目(YJY-XD-2023-003)

**Supported by**: National Natural Science Foundation of China (61601213); Key Project of the Education Department of Liaoning Province (LJ212410147003)

the actual scene geometry. However, the neural radiance field representation based on point clouds is affected by the quality of the point cloud, and the influence of extracted image features leads to a decrease in the rendering quality of new perspective images. To this end, a novel-view synthesis method integrating local spatial information is proposed from the two key points of aligning point cloud features and fusing local point cloud context information field. **Method** The network architecture in this study comprises neural networks for point cloud generation and neural radiance fields based on point clouds. The point cloud and confidence are produced by the depth prediction network in the neural network component for point cloud generation. The image is processed using the feature pyramid network to extract features at different scales. The neural alignment module for point cloud features is subsequently employed to integrate the feature derived from the point cloud and the image. The features are aligned to extract the semantic information of the image. This step enables the network to more effectively adjust to the structural and textural characteristics of various scene images. Neural point clouds are created by mixing points, confidence levels, and image features. In the neural radiance field network structure based on point clouds, the RGB value and volume density of the sampling point are predicted by aggregating the neural point cloud features near the sampling point. This experiment utilizes the Transformer layer combined with the contextual information of the local neural point cloud to better capture the spatial and geometric shape details, and it outputs high-quality synthetic images through volume rendering. **Result** This experiment establishes the environment on the Ubuntu18. 04 system to assure the reliability of the training and testing procedure. The CPU is an Intel Core i9-10900, the memory capacity is 32 GB, and the graphics card is an RTX 3090. The experiment primarily uses the peak signal-to-noise ratio (PSNR) as the metric for evaluating the test results. It also utilizes the structural similarity index measure and the learned perceptual image patch similarity. Network training uses the Adam algorithm for adaptive learning rate optimization. By dynamically adjusting the learning rate, the network can more effectively balance the convergence speed and stability during the training process. The initial learning rate is set to 0. 000 5, and the decay rate parameters are set to 0. 9 and 0. 99. Four widely utilized datasets(DTU, NeRF Synthetic, Tanks and Temples, and ScanNet)will be utilized in the experiment. DTU is a dataset comprising indoor scenes. Each scene is composed of 49 object photography angles, with 7 brightness levels per angle, and has a resolution of 512 pixels by 640 pixels. NeRF Synthetic is a synthetic dataset containing eight scenes, each with 100 training images and 200 test images. These images are fully rendered and synthesized by Blender. ScanNet is an interior scanning dataset. Scenes 241 and 101 will be applied to the evaluation. A total of 20% of the total image count will be allocated for training objectives (1 463 images for Scene-241 and 1 000 images for Scene-101), with the remaining images being utilized for evaluation purposes. The Tanks and Temples dataset is an extensive collection of indoor scene data and comprises 14 distinct scenes. Experimental results show that, for the Tanks and Temples datasets containing only a single object, the PSNR of this method is improved by 19. 2% compared with the NeRF method. The ratios are enhanced by 6. 4% and 3. 8% compared with those obtained by Tetra-NeRF and Point-NeRF using point cloud input, respectively. Even in the ScanNet dataset with more complex scenes, the ratios are improved by 34. 6% and 2. 1% compared with the NeRF method and Point-NeRF, respectively. **Conclusion** This study presents a novel-view synthesis method integrating local spatial information that, in conjunction with a neural alignment module for point cloud features, dynamically modifies the alignment characteristics of a neural point cloud. When the points correspond to aligned features, our approach can enhance the precision of this procedure through the extraction of features from images at various dimensions along with the semantic information they encompass. The neural Transformer module based on point clouds enhances the capability of the network to extract spatial position and geometry information from the neural point cloud by incorporating context information from nearby sampling points. This improved efficiency is particularly useful when dealing with points of different qualities and shapes. The experimental results for the Tanks and Temples, Synthetic Blender, and ScanNet datasets show that this method outperforms existing advanced neural radiation field representations based on point clouds in terms of visual effects and assessment indicators. Overall, the method outlined in this document improves the combination of point clouds and image characteristics. It utilizes the contextual information found in local point cloud features to assist the network in merging sparse point cloud features. This process leads to more lifelike and unique details in the resulting image. Moreover, high-quality scene images are produced from input images that contain only a small number of shots.

**Key words**: neural radiance field (NeRF); point cloud; neural rendering; three dimensional reconstruction; volume density

## 0 引 言

从一组2D图像与其相关的摄影位姿合成新视角场景影像是计算机视觉领域的一项重要课题。Mildenhall等人(2020)提出一种神经渲染方法——神经辐射场(neural radiance field, NeRF),采用全连接的多层感知机(multi-layer perception, MLP)存储真实场景信息,并将场景表示为一个隐式神经网络。NeRF作为一种新颖的场景表示方法,在渲染任务中的视觉效果获得了较大提高。尽管NeRF合成图像的质量高,但该方法以牺牲训练时间为代价,在场景重建时,每个采样点都需要通过整个MLP模型。为缓解原始NeRF重建场景速度过慢的问题,Yu等人(2021a)结合体素引入一种PlenOctrees的分层数据结构。该结构通过八叉树避免过度采样,能够减少贡献度不多的采样点,与NeRF相比能够实现实时渲染图像。Fridovich-Keil等人(2022)完全抛弃神经网络,改用稀疏的体素网格存储球谐函数表示场景,并通过三线性插值获取采样点的RGB值和体积密度,训练速度提升近2个数量级。Müller等人(2022)提出一种多分辨率哈希编码,采用CUDA(compute unified device architecture)编程减少计算时间,以此提高场景的重建效率,训练一个高质量的场景只需要几秒钟。

另一方面,诸多学者开展了采用NeRF提高场景重建质量的研究工作。Verbin等人(2022)采用反射辐射(reflected radiance)取代了NeRF的参数化视相关的出射辐射(outgoing radiance),并通过一组空间变化场景属性构造该函数。与原始NeRF相比,该方法能够额外输出表面法向、漫反射颜色、镜面反射和粗糙度,根据视点方向与粗糙度计算镜面反射颜色,并利用漫反射颜色和镜面反射预测最终的颜色。Wang等人(2023)将表面反射建模为神经网络,将入射光线的方向和场景中的光照条件作为输入,输出为表面反射的颜色和强度。通过在训练过程中使用带有光照和材质信息的渲染图像优化网络参数,该网络能够学习到复杂材质的反射行为,包括金属、塑料以及皮肤等各种材质的光学特性。

尽管神经辐射场合成新视角图像的质量高,但在只有少数场景的视图可用的条件下仍存在较多问题(Chen等,2021;Yu等,2021b;Jain等,2021)。诸多学者试图通过多种方式解决该问题,包括基于内容正则化(Jain等,2021)、利用图像特征(Yu等,2021b)、基于深度监督(肖强 等,2024;刘晓楠 等,2024)以及基于高斯渲染(Gaussian splatting)(Chung等,2024)的方法。Yu等人(2021b)提出一种pixel-NeRF网络,使得NeRF可以在不同场景下进行训练,通过学习场景的先验知识,只需一幅或几幅图像重建场景。Chen等人(2021)使用预训练的卷积网络提取图像特征,并将图像特征映射到参考视图的扫描平面上,以此构建基于三维体素(three dimensional voxel)的代价体,并采用3D卷积网络重建神经编码体,利用MLP将神经编码体内的隐式特征解码成体积密度和颜色值。Xu等人(2022a)仅使用单幅视角图像对神经辐射场进行训练,重建出复杂的视觉场景。其构建出一个半监督学习的框架,通过设计基于几何及语义的监督信号完成场景的重建。几何监督通过图像变形和相机的位姿参数,以保证多视角下的几何一致性。而语义监督则是通过一个预训练的视觉Transformer(visual Transformer, ViT)对图像特征进行提取,再与整体结构进行比较,优化未知视角下的语义质量。肖强等人(2024)使用深度监督的策略,将更多的采样点分配到物体表面,并引入未知视角下的光线损失,使得最终结果更加精细化。刘晓楠等人(2024)则是利用深度预估网络求取预估深度值和稀疏深度值之间的标准差,对整个网络进行监督,从而解决稀疏视角输入的问题。Chung等人(2024)通过大量高斯渲染表示3D场景,使用深度图对高斯渲染的优化过程进行正则化,从而有效地减少了合成结果中的伪影。

解决稀疏视角问题的另一种方式可以利用传感器或摄影测量中的点云信息。Xu等人(2022b)利用输入的稀疏视角图像生成点云,再将点云特征结合采样点信息输入到神经辐射场中进行场景重建;Kulhanek和Sattler(2023)则对输入点云进行网格化,引入一种四面体结构辅助采样点进行插值;Zhang等人(2023)则通过傅立叶函数将点云转换到频率域进行优化,简化了场景表示的过程;Govindarajan等人(2025)提出拉格朗日散列法,即一种结合点云和分层哈希表的神经辐射场表示方法。

基于点云的神经辐射场方法是一种从稀疏视角图像重建场景的新方法,其利用点云场景的几何先验提高网络采样效率,有效改善了NeRF训练时间

过长的问题。但由于该方法需额外输入点云，最终合成的新视角图像质量受点云准确度、位置分布和疏密的限制。解决该问题的关键在于如何从有限的点云特征中提取出可靠的局部空间信息，本文通过提高点云与图像局部特征的相关性，并利用Transformer（Vaswani等，2017）的注意力机制捕获局部神经点云上下文信息，提出一种融合局部空间信息的新视角合成方法。

本文主要贡献如下：1）提出一种新的基于点云的神经渲染方法，通过将图像的语义信息融入到点云特征中，对局部神经点云特征进行关联性建模和信息提取，能更好地从稀疏图像输入中合成新视角场景；2）提出一种神经点云特征对齐模块，通过自适应调整卷积核大小更好地对齐点云与图像特征，提高神经点云对局部空间信息的表达能力，从而提升整个神经辐射场的合成质量；3）在估计体积密度前，充分利用采样点周围神经点云的局部空间信息，给出一种基于神经点云的Transformer模块结构，能够有效提升场景渲染图像的局部细节。

## 1 相关工作

### 1.1 神经辐射场

Mildenhall等人（2020）提出的神经辐射场，其核心思想在于通过一个深度神经网络对场景进行建模，其输入为一组3D坐标和视角方向，输出采样点的颜色和体积密度，最后通过体渲染公式生成高质量的图像，渲染过程本质上就是积分运算，通过对场景中每个采样点的颜色和体积密度进行加权求和，从而获取特定视角下的观察图像。

### 1.2 基于点云的新视角合成方法

基于点云的新视角合成方法主要分为两种。一种是Kerbl等人（2023）提出的三维高斯泼溅（3D Gaussian splatting）方法，与神经辐射场的场景表示方法不同，其选用可微的三维高斯函数表示场景。对于输入的场景点云，首先建立一个稠密的三维高斯合集，在训练过程中不断优化三维高斯的参数，即位置、协方差、不透明度和球谐函数，这种可见性的感知渲染方法不仅加快了训练速度还实现了实时渲染。

另一种方法是使用神经辐射场表示场景，利用点云的连续性和稀疏性，在多视图数据中有效捕捉场景的几何特征的同时，避免了在空白空间的不必要计算。例如，Xu等人（2022b）在选取采样点的过程中，借助于点云判断其是否接近表面，避免了采样点在空白区域出现。与三维高斯泼溅方法不同，基于点云的神经辐射场更注重于对点云特征的提取，将采样点的信息与点云特征融合后，再输入到神经网络中预测采样点的颜色的体积密度。Xu等人（2022b）在网络模型中直接对采样点附近的点云特征进行加权求和，扩充了采样点信息，而Kulhanek和Sattler（2023）则是通过将点云三角网格化，获取采样点特征只需要查询采样点所在的四面体，并利用四面体顶点特征进行插值，进一步压缩了查询点云特征的所需时间，将该特征通过MLP处理后获得该点体积密度和颜色值，最后通过体渲染获得合成图像中像素的颜色估计值。Zhang等人（2023）提出一种新的点云渲染方法FreqPCR（frequency-modulated point cloud rendering with easy editing），该方法的创新点在于提出一个自适应频率调制模块，将点云数据从空间域转换到频率域，通过傅立叶变换将点云数据在更高维度的频率域进行处理，对点云特征进行降噪和优化。这种表示方式不仅能够更紧凑地捕捉点云的几何细节，在频率域中还能简化某些复杂操作，从而提高渲染质量。Govindarajan等人（2025）提出一种拉格朗日散列法，这是一种新的神经辐射场表示方法，通过将点云合并到分层哈希表的高分辨率层，能够有效压缩场景表示的重建信号，且不会影响最终的结果质量。

## 2 本文方法

如图1所示，总体网络结构主要由神经点云生成网络和神经辐射场网络两个部分组成。神经点云生成网络对输入图像进行处理，通过对齐点云和提取的图像特征生成神经点云。在神经辐射场网络部分，通过聚合采样点附近神经点云特征和相机视角信息预测采样点的RGB值和体积密度，最终通过体渲染公式合成新视角场景图像。

### 2.1 神经点云生成网络

#### 2.1.1 深度预测网络

本文基于多视点立体视觉（multi-view stereo，MVS）深度估计方法，通过使用3D卷积网络和构建代价体（cost volume）预测出点云的位置（Yao等，2018；Cheng等，2020）。首先对输入的多视角图像

进行特征提取，将不同视角间的图像进行特征匹配后，利用不同视角特征点的方差值构建代价体，使用3D卷积网络对代价体进行正则化生成概率体（probability volume）用以预测深度图，最终对每个图像的深度图进行滤波和融合操作获得稠密点云。由于概率体能够表示点云在重建物体表面的概率，因此使用三线性插值能够获得点云对应的置信度。

### 2.1.2 神经点云特征对齐模块

图2为神经点云特征对齐模块，该模块的主要功能是实现点云和图像特征对齐。在图像特征提取方面采用多尺度特征融合的方法，借助特征金字塔网络（feature pyramid network，FPN）结构进行图像特征的提取，有助于处理不同分辨率图像的细节信息，从而提升网络模型的泛化能力（Lin等，2017）。然而，在使用固定的卷积核提取图像特征时，由于对图像不同位置区域的感受野相同，难以有效处理不同形状和纹理区域的特征。为了更好地对齐点云和图像特征，提高对齐特征的相关性，对FPN网络的输出特征进一步处理：将图像特征通过特征偏移网络改变卷积核的采样位置，动态调整所对齐特征的感受野，使点云的对齐特征聚焦于相关性强的图像区域，更好地适应不同场景结构和纹理特征，增强神经点云对周围空间信息的表达能力。特征偏移网络基于可变形卷积网络（Dai等，2017；Zhu等，2019）实现，通过学习采样位置的额外偏移量，改变卷积核的特征映射关系。

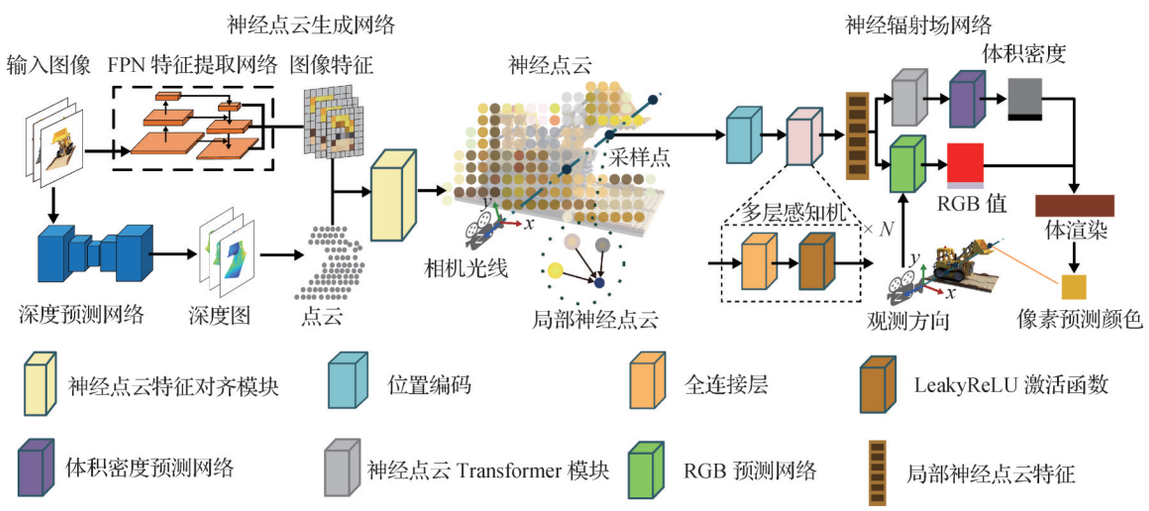



图1 总体网络架构

Fig. 1 Overall network structure

对于传统的标准卷积，给定一个3 × 3的卷积核$\boldsymbol{R} \in \{(-1,-1),(-1,0),\cdots,(1,1)\}$，令$x(p)$与$y(p)$代表输入特征图$\boldsymbol{x}$和输出特征图$\boldsymbol{y}$位置$p$的特征值，$\omega_n$为卷积核权重，$p_n$为$\boldsymbol{R}$中对于$p$的相对坐标，其特征映射具体为

$$\boldsymbol{y}(p)=\sum_{p_n\in\boldsymbol{R}}\omega_n\cdot\boldsymbol{x}(p+p_n) \tag{1}$$

相对于标准卷积，可变形卷积能够学习额外的偏移量，使采样位置向周围扩散，具体为

$$\boldsymbol{y}(p)=\sum_{p_n\in\boldsymbol{R}}\omega_n\cdot\boldsymbol{x}(p+p_n+\Delta p_n) \tag{2}$$

式中，$\Delta p_n$是位置的可学习的偏移量，一个无约束范围的实数。如图3所示，采用标准卷积会将不需要的图像信息融入提取特征中，而可变形卷积提取图像特征时使用一个单独的卷积层来学习获得偏移量，能根据物体不同位置和不同形状自适应地调整感受野的大小，在点云对齐图像特征时，着重对感兴趣的区域进行采样。

为了建立点云与图像特征之间的对齐关系，本文对图像特征进行网格化，将其坐标值限定在[−1,1]的范围内，其中，(−1,−1)表示左上角坐标，(1, 1)表示右下角坐标。在查询点云的特征时，根据相机的位姿参数，将点云空间坐标$(x,y,z)$投影到图像特征对应的像平面上，并将像平面的坐标映射到[−1,1]范围内，即可获得对应图像特征的相对位置坐标$(u,v)$，通过查询该位置坐标获得采样点的图像特征。将通过第1节中的方法获得点云置信度$\gamma_i$和位置坐标$p_i$，以及对齐采样点特征$\boldsymbol{f}_i$进行特征级联，获得最终的神经点云$\boldsymbol{P}=\{(p_i,\gamma_i,\boldsymbol{f}_i)|i=1,\cdots,N\}$。

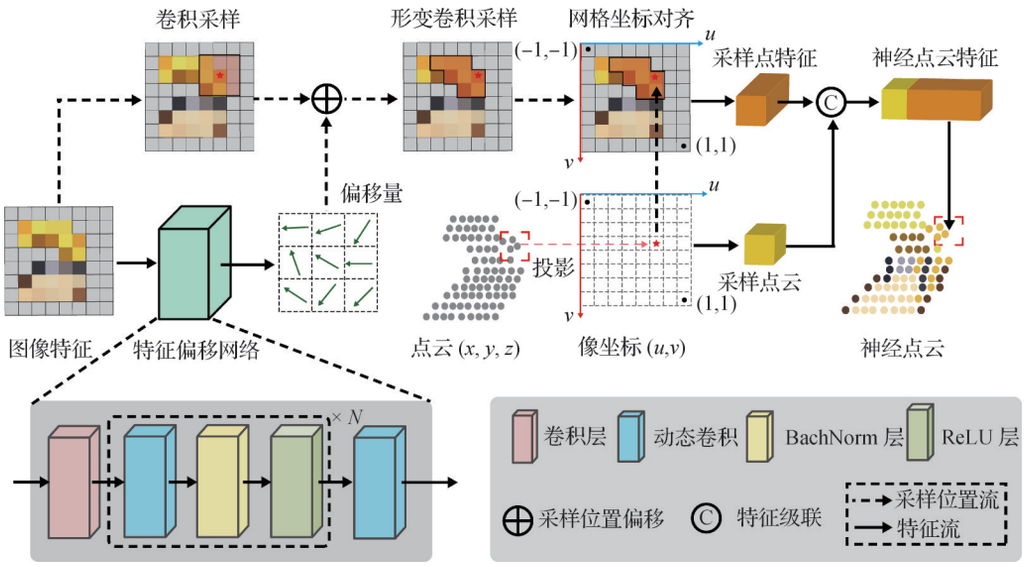



图 2　神经点云特征对齐模块

Fig. 2　Neural point cloud feature alignment module

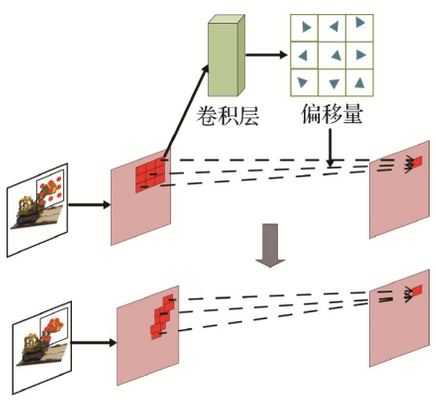



图 3　可变形卷积特征的提取过程

Fig. 3　Extraction process of deformable convolution feature

## 2.2　基于神经点云的神经辐射场网络

图 4 给出了神经辐射场网络对生成的神经点云进行处理的过程，通过聚合神经点云特征，该网络能够预测采样点的 RGB 值和体积密度，主要过程为：由相机光心向目标图像的每个像素发射光线，沿光线方向选取多个位置进行采样，在每个采样点处通过 K 最近邻（K-nearest neighbors，KNN）算法选择距离最近的 $K$ 个神经点云 $\{P_i \mid i = 1, \cdots, K\}$。输入采样点的位置 $x$、相机视角 $d$ 和选取的神经点云特征，输出采样点所对应的 RGB 值 $c$ 和体积密度 $\sigma$。具体映射为

$$(c, \sigma) = G(x, d, P_1, \cdots, P_K) \tag{3}$$

式中，$G$ 表示神经辐射场网络。通过该网络获取的 RGB 值和体积密度，能够利用体渲染公式预测出合成场景图像的颜色值。

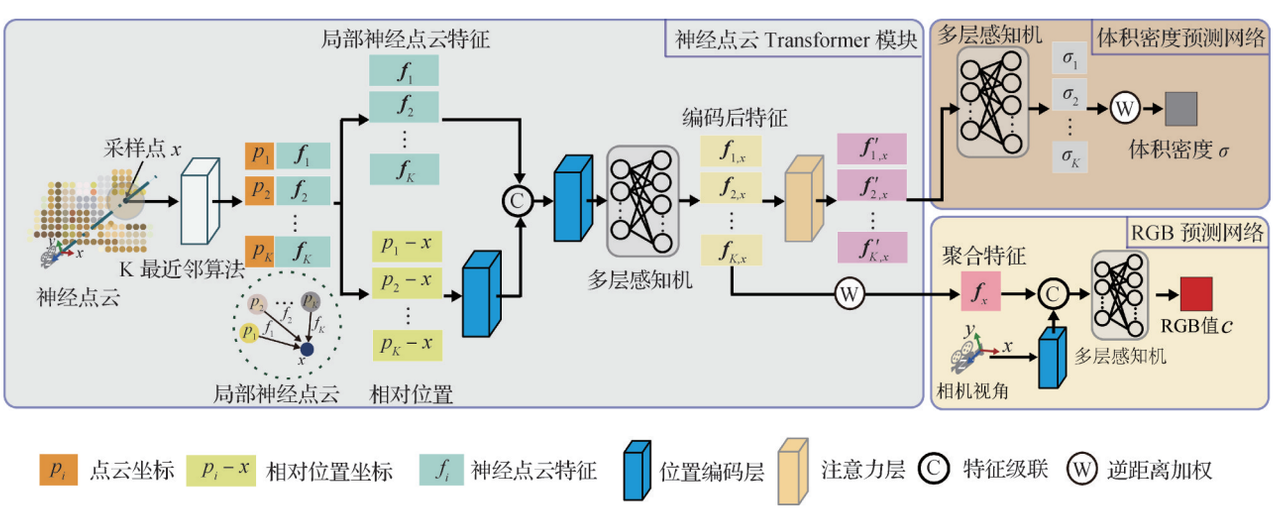



图 4　基于神经点云的神经辐射场网络结构

Fig. 4　Structure of neural radiation field network based on neural point cloud

2.2.1 神经点云Transformer模块

在预测采样点体积密度时，直接将神经点云的特征$f_i$输入到MLP网络中获取体积密度$\sigma$，该方法精确度不足，易导致输出结果质量下降，因此单独地处理神经点云特征不足以预测复杂场景的体积密度。为了解决该问题，在预测体积密度之前，采用神经点云Transformer模块来更好地捕捉局部神经点云间的相互关系和补充信息，使神经点云特征包含更多的细节线索，生成更加逼真的体积密度分布，在新视角图像合成时表现出更高的细节保真度。

神经点云Transformer模块主要由位置编码和多头注意力机制两部分组成。与传统Transformer的位置编码层不同，本文通过将神经点云的相对位置坐标映射到高维特征空间，将编码后的相对位置坐标与神经点云特征级联后输入到位置编码层，将其进一步编码升维，以此捕获神经点云的高频特征信息。采用MLP网络提取神经点云编码后的特征$\boldsymbol{f}_{i,x}$，由于神经点云的特征$f_i$包含局部空间几何信息，融合相对位置信息后获得特征向量$\boldsymbol{f}_{i,x}$，可以更好地拟合采样点周围场景的空间信息，提高网络对点云的平移不变性，增强网络提取场景局部特征的能力，并加强整个网络模型的泛化能力。具体为

$$\begin{aligned}&PE(x)=(\sin(2^0\pi x),\cos(2^0\pi x),\cdots,\\&\sin(2^{L-1}\pi x),\cos(2^{L-1}\pi x))\\&\boldsymbol{f}_e=f_{\text{Concat}}(\boldsymbol{f}_i,PE(p_i-x))\\&\boldsymbol{f}_{i,x}=f_{\text{MLP}}\left(f_{\text{Concat}}\left(\boldsymbol{f}_e,PE(\boldsymbol{f}_e)\right)\right)\end{aligned}\tag{4}$$

式中，$PE(\cdot)$为位置编码层的特征映射，$L$为编码维度控制变量，$f_{\text{Concat}}(\cdot)$为特征级联操作，$f_{\text{MLP}}(\cdot)$为MLP网络，$\boldsymbol{f}_e$表示神经点云的中间特征。对于编码后的神经点云特征序列$\boldsymbol{X}=\{\boldsymbol{f}_{i,x}|i=1,\cdots,K\}$，将其输入到注意力层中，其中注意力计算方法为

$$(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{X}\times(W_q,W_k,W_v)\tag{5}$$

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\tag{6}$$

式中，$W_q$、$W_k$和$W_v$分别为Query矩阵$(\boldsymbol{Q})$、Key矩阵$(\boldsymbol{K})$和Value矩阵$(\boldsymbol{V})$可学习的权重参数，$d_k$为$\boldsymbol{K}$的特征维度，softmax为归一化指数函数。

通过将每个注意力的权重与对应的值相乘并求和，得到多头注意力的输出表示，具体为

$$f_{\text{MultiHead}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{Concat}}(head_1,\cdots,head_n)\boldsymbol{W}^O\tag{7}$$

式中，$head_i=f_{\text{Attention}}(\boldsymbol{Q}W_q^{(i)},\boldsymbol{K}W_k^{(i)},\boldsymbol{V}W_v^{(i)})$，$\boldsymbol{W}^O$表示输出权重矩阵，$i$表示注意力头的数量。经过计算，获取采样点附近神经点云新的特征向量$\{\boldsymbol{f}'_{i,x}|i=1,\cdots,K\}$，将其输入到体积密度预测网络中，用于计算采样点的体积密度值。

2.2.2 体积密度预测网络

将通过神经点云Transformer模块得到的神经点云特征$\{\boldsymbol{f}'_{i,x}|i=1,\cdots,K\}$输入到MLP网络，预测每个神经点云的体积密度$\{\sigma_{i,x}|i=1,\cdots,K\}$，并通过神经点云的置信度$\gamma_i$进行逆距离加权处理，获得采样点$x$最终的体积密度$\sigma$。具体为

$$\begin{aligned}&\sigma_i=MLP(\boldsymbol{f}'_{i,x})\\&\sigma=\sum\sigma_i\gamma_i\frac{w_i}{\sum w_i}\\&w_i=\frac{1}{\|p_i-x\|}\end{aligned}\tag{8}$$

2.2.3 RGB预测网络

与体积密度预测网络不同，为了预测采样点的RGB值，对于神经点云Transformer模块中编码后的中间特征$\boldsymbol{f}_{i,x}$，先进行逆距离加权计算每个特征向量的权重，将采样点周围神经点云的特征进行聚合，具体为

$$\boldsymbol{f}_x=\sum_1^K\gamma_i\frac{w_i}{\sum w_i}\boldsymbol{f}_{i,x}\tag{9}$$

获取聚合特征$\boldsymbol{f}_x$后，通过MLP网络预测采样点的RGB值，该方法在保持预测结果足够精确的基础上，只需要对聚合后的特征进行解码，能够降低网络模型的计算量。由于采样点的RGB值与观测方向相关，将相机视角方向$d$的编码特征与聚合特征$\boldsymbol{f}_x$拼接，输入到MLP网络中获取采样点的RGB值$c$，具体为

$$c=f_{\text{MLP}}\left(f_{\text{Concat}}\left(\boldsymbol{f}_x,PE(d)\right)\right)\tag{10}$$

## 2.3 体渲染

对于得到的RGB值和体积密度，采用NeRF中的体渲染(volume rendering)方法(Kajiya和von Herzen，1984)合成场景的新视角图像。具体操作为：向目标图像的每个像素投射一条贯穿整个场景光线$r$，在光线上对训练好的网络模型进行采

样，得到采样位置的RGB值$c$和体积密度$\sigma$，最后沿光线方向进行积分得到该像素的颜色值，具体为

$$C(r)=\int_{t_n}^{t_f}T(t)\sigma\big(r(t)\big)c\big(r(t),d\big)\mathrm{d}t$$
$$T(t)=\mathrm{e}^{\int_{t_n}^{t_f}-\sigma(s)\mathrm{d}s} \tag{11}$$

式中，区间$\left[t_n,t_f\right]$表示相机光线的采样距离，$T(t)$表示累计投射率，即光线从$t_n$到$t$没有被吸收的概率。

由于连续积分的难度过大，该式在NeRF中给出一种简化的计算方法。采用分层采样的方法将区间$\left[t_n,t_f\right]$划分为$N$个相同的区间，具体为

$$t_i\sim\left[t_n+\frac{i-1}{N}\big(t_f-t_n\big),t_n+\frac{i}{N}\left(t_f-t_n\right)\right] \tag{12}$$

式中，$t_i$表示第$i$个划分区间，在划分后的每个区间中将随机抽取一个样本进行计算，即

$$\hat{C}(r)=\sum_{i=1}^{N}T_i\big(1-\mathrm{e}^{-\sigma_i\delta_i}\big)c_i$$
$$T_i=\mathrm{e}^{-\sum_{j=1}^{i-1}\sigma_i\delta_i} \tag{13}$$

式中，$\hat{C}(r)$表示合成新视角场景图像中像素的预测颜色，$i$表示沿光线方向的采样点，$\delta_i=t_i-t_{i-1}$表示当前采样点与上一个采样点间的距离。

### 2.4 损失函数

为度量输出图像与真值图像之间的误差，使用L2函数作为损失函数，在空间采样射线$R$中，通过计算合成图像的像素预测值$\hat{C}(r)$与真实的图像像素真值$C_{\mathrm{gt}}$的差值平方的均值作为渲染损失$L_{\mathrm{render}}$，具体为

$$L_{\mathrm{render}}=\sum_{r\in R}\left\|\hat{C}(r)-C_{\mathrm{gt}}\right\|_2^2 \tag{14}$$

置信度$\gamma$能表示点云在物体表面的概率，为了使模型在训练过程中更偏向于物体表面的神经点云特征，引入稀疏性损失函数(Lombardi等，2019)，具体为

$$L_{\mathrm{sparse}}=\frac{1}{\left|\gamma\right|}\sum_{\gamma_i}\Big[\log\big(\gamma_i\big)+\log\big(1-\gamma_i\big)\Big] \tag{15}$$

该损失函数可以使点云的置信度$\gamma$趋于0或1，提高对贡献度不高点云的约束。在逐场景优化时，将渲染损失和稀疏性损失的联合损失函数作为网络的整体损失，并使用固定值$a=2\mathrm{e}{-3}$平衡稀疏性损失函数对整体损失的影响，具体为

$$L_{\mathrm{opt}}=L_{\mathrm{render}}+aL_{\mathrm{sparse}} \tag{16}$$

## 3 实 验

### 3.1 实验环境

本文采用Ubuntu 18.04系统、CPU型号Intel Core i9-10900、内存32 GB、显卡型号为RTX 3090、编程语言使用Python 3.7，模型的搭建和测试均在Vscode编译器上进行。

### 3.2 数据集和评估指标

本文采用峰值信噪比(peak signal to noise ratio，PSNR)、结构相似度(structural similarity index measure，SSIM)和学习感知相似度(learned perceptual image patch similarity，LPIPS)评估图像合成质量(Mildenhall等，2020)。其中，PSNR指标更关注合成图像与真实图像间的像素值误差，该值越大表明误差越小；SSIM指标用于评价两幅图像间结构的相似性，范围为[0,1]，当两幅图像完全相同时，SSIM的值为1；LPIPS指标是一种用于评估图像中人类感知相似度的指标，旨在比较两幅图像在视觉上的差异，而不仅限于像素级别的不同，该值越小则视觉效果越好。

采用DTU、NeRF Synthetic、Tanks and Temples和ScanNet 4个常用数据集开展实验。DTU为室内场景数据集，每个场景包含49个物品拍摄视角，每个视角具有7种亮度，图像分辨率为512 × 640像素。NeRF Synthetic是一个合成数据集，包含8个场景，每个场景有100幅训练图像和200幅测试图像，这些图像完全由Blender渲染合成，图像分辨率为800 × 800像素。ScanNet是室内扫描数据集，实验将在其中两个场景Scene-101和Scene-241上进行评估。本文实验与PointNeRF保持一致，对20%的图像进行采样，即Scene-241的1 463幅图像，场景Scene-101的1 000幅图像用于训练，其余图像用于评估。Tanks and Temples是大型室外场景数据集，其中包含14个场景，本次实验将挑选其中的5个场景进行测试。

### 3.3 实验细节

首先在DTU数据集上对模型进行预训练，并按pixelNeRF和MVSNeRF的方法对训练集和测试集进行划分。在此阶段，基于真实图像的实际像素颜色对渲染式(11)的预测值进行监督，仅使用渲染损失对整个网络进行端到端训练。预训练结束后在NeRF Synthetic、Tanks and Temples和ScanNet数据

集的各场景上进行微调，为了更好地约束点云，采用联合损失函数$L_{opt}$进行训练。网络训练采用自适应学习率优化Adam算法，通过动态调整学习率，使网络在训练过程中更有效地平衡收敛速度和稳定性。初始学习率设置为5e−4，衰减率参数设置为0.9和0.99，在Synthetic Blender、Tanks and Temples数据集上的迭代轮数为200 k，在ScanNet数据集上的训练轮数设置为300 k。

### 3.4 实验结果

将本文方法与较为先进的方法（如Point-NeRF、Tetra-NeRF、FreqPCR等）在NeRF Synthetic、Tanks and Temples、ScanNet 3种数据集上进行测试评估。表1给出了不同方法在合成场景NeRF Synthetic数据集上的PSNR结果，该数据集的场景图像均为Blender合成，由于输入图像的光照环境和清晰度为理想条件下生成的结果，各种基于点云的模型在此数据集上均有较好的表现。尽管本文方法更加关注改善复杂环境下由于点云获取困难而导致的渲染质量下降问题，各场景的PSNR均值达到了33.45 dB，仍取得了几种方法中最好的结果。

图5给出了在Synthetic Blender数据集上进行对比实验的可视化结果，相较于其他同类方法，本文方法能够更好地处理因点云稀疏而产生的空洞和噪点问题。此外，由于神经点云的特征来源于不同视角图像，将点云与视角图像结合，可以帮助减少镜面反射的影响，在使用神经点云特征对齐模块后，每个神经点云所包含的不同视角光照信息将更为准确，在神经点云的Transformer模块中，对采样点周围的神经点云进行关联性建模，估计各视角特征对镜面反射的贡献率，因此在镜面反射区域的渲染结果有所改善，在图像纹理相近及物体边缘的区域部分过渡更为自然。

**表1 不同方法在NeRF Synthetic数据集上的PSNR**

**Table 1 PSNR of different methods on the NeRF Synthetic dataset**

| 方法 | PSNR/dB | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Chair | Drums | Ficus | Hotdog | Lego | Materials | Mic | Ship | 平均 |
| NPBG*(Aliev等，2020) | 26.47 | 21.53 | 24.60 | 29.01 | 24.84 | 21.58 | 26.62 | 21.83 | 24.56 |
| NPBG++*(Rakhimov等，2022) | 28.72 | 23.60 | 28.11 | 32.22 | 27.84 | 27.12 | 31.23 | 26.11 | 28.12 |
| NeRF(Mildenhall等，2020) | 33.00 | 25.01 | 30.13 | 36.18 | 32.54 | 29.62 | 32.91 | 28.65 | 31.01 |
| NSVF(Liu等，2020) | 33.19 | 25.18 | 31.23 | 37.14 | 32.54 | **32.68** | 34.27 | 27.93 | 31.77 |
| Instant-NGP(Müller等，2022) | 35.00 | 26.02 | 33.51 | 37.40 | **36.39** | 29.78 | 36.22 | 31.10 | 33.18 |
| Gauss. Splat.* (Kerbl等，2023) | **35.83** | 26.15 | 34.87 | **37.72** | 35.78 | 30.00 | 35.36 | 30.80 | 33.32 |
| FreqPCR*(Zhang等，2023) | 33.06 | 25.95 | 32.19 | 35.82 | 31.56 | 29.69 | 33.64 | 27.97 | 31.24 |
| Point-NeRF*(Xu等，2022b) | 35.40 | 26.06 | 36.13 | 37.30 | 35.04 | 29.61 | 35.95 | 30.97 | 33.31 |
| Tetra-NeRF*(Kulhanek和Sattler，2023) | 35.05 | 25.01 | 33.31 | 36.16 | 34.75 | 29.30 | 35.49 | **31.13** | 32.53 |
| LHCN*（Govindarajan等，2025） | 35.61 | 25.67 | 33.89 | 37.23 | 35.6 | 29.63 | **36.45** | 30.84 | 33.12 |
| 本文* | 35.47 | **26.17** | **36.27** | 37.22 | 35.56 | 30.07 | 36.40 | 30.46 | **33.45** |

注：加粗、下划线字体表示各列最优、次优结果。“*”表示基于点云的方法。

表2和表3给出不同方法在真实场景Tanks and Temples和ScanNet数据集的对比结果。由于真实场景图像受到拍摄设备性能、环境光照等因素的影响，合成点云的质量往往不佳。Tetra-NeRF通过将点云三角化获得一个用四面体表示的神经辐射场，以插值的方法获取采样点的特征，使图像结果更加平滑减少空洞的产生，但同时也会丢失局部纹理细节。本文方法的神经点云Transformer模块无需对点云进行额外操作，基于局部神经点云的局部性、位置和高维特征信息使合成图像的局部细节得到明显改善。如表2所示，本文方法在Tanks and Temples数据集上的PSNR和SSIM均取得了最好的结果，其中LPIPS仅低于Tetra-NeRF，因此与其他同类型先进方法相比，本文方法能够更有效地处理真实场景图像，具有良好的泛化性，合成的图像失真度低，与真值的图像相似度更高，图6为可视化对比结果。

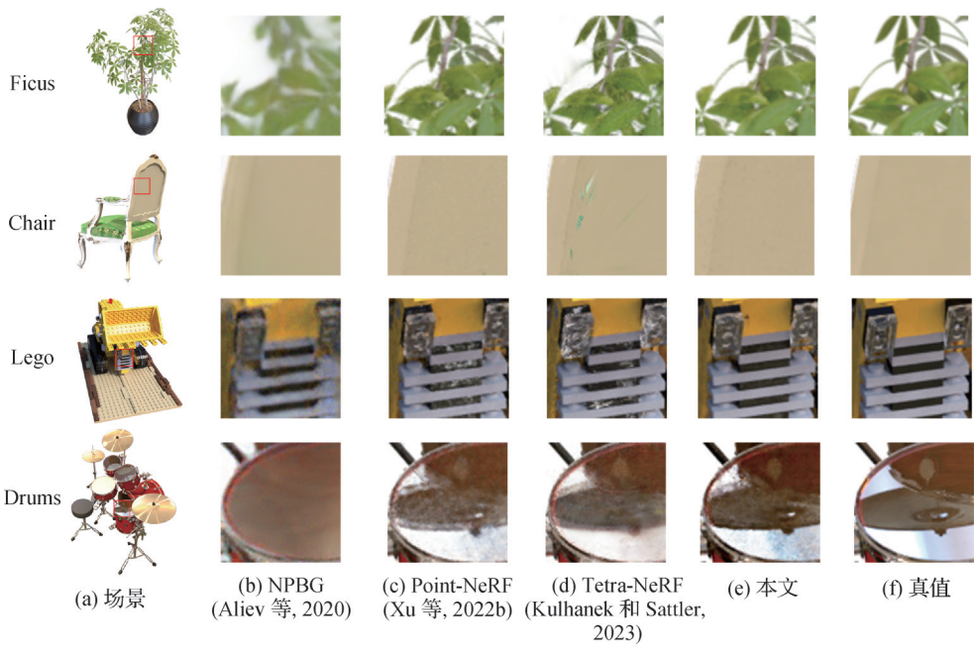



图 5　Synthetic Blender 可视化实验结果

Fig. 5　Visual experimental results on Synthetic Blender ((a) scenes; (b) NPBG (Aliev et al. , 2020); (c) Point-NeRF (Xu et al. , 2022b); (d) Tetra-NeRF (Kulhanek and Sattler, 2023); (e) ours; (f) ground truth)

**表 2　不同方法在 Tanks and Temples 数据集上的结果**

**Table 2　Results of different methods on the Tanks and Temples dataset**

| 方法 | PSNR/dB | SSIM | LPIPS |
|---|---|---|---|
| NPBG*(Aliev 等,2020) | 25.97 | 0.889 | 0.137 |
| NPBG++*(Rakhimov 等,2022) | 26.04 | 0.892 | 0.130 |
| NeRF(Mildenhall 等,2020) | 25.78 | 0.864 | 0.198 |
| Point-NeRF*(Xu 等,2022b) | 29.61 | 0.954 | 0.080 |
| Gauss. Splat.*(Kerbl 等,2023) | 23.14 | 0.841 | 0.183 |
| FreqPCR*(Zhang 等,2023) | 27.79 | 0.902 | 0.125 |
| Tetra-NeRF*(Kulhanek 和 Sattler,2023) | 28.90 | 0.957 | **0.059** |
| 本文* | **30.74** | **0.971** | 0.061 |

注:加粗字体表示各列最优结果。“*”表示基于点云的方法。

**表 3　不同方法在 ScanNet 数据集中两个场景上的结果**

**Table 3　Results of different methods on two ScanNet scenes**

| 方法 | 两个场景的平均指标 | | | 场景 PSNR/dB | |
|---|---|---|---|---|---|
| | PSNR/dB | SSIM | LPIPS | Scene-101 | Scene-204 |
| NeRF(Aliev 等,2020) | 22.99 | 0.620 | 0.369 | – | – |
| NSVF(Liu 等,2020) | 25.48 | 0.688 | 0.301 | – | – |
| Point-NeRF*(Xu 等,2022b) | 30.32 | 0.909 | 0.220 | 30.13 | 30.51 |
| 本文* | **30.95** | **0.926** | **0.218** | **30.74** | **31.16** |

注:加粗字体表示各列最优结果。“*”表示基于点云的方法,“–”表示数据未发布。

ScanNet为室内扫描数据集，是基于点云的神经辐射场的经典应用场景。将本文方法与NeRF、NSVF、Point-NeRF在此数据集上进行对比实验，如表3所示，由于该场景中的大部分区域缺少纹理并相对平滑，传统神经辐射场NeRF和基于体素的方法NSVF在该数据集的表现均不佳，而基于点云的方法在此场景下能够获得更为理想的结果。本文方法在Scene-101和Scene-204两个场景上的平均指标（PSNR、SSIM、LPIPS）均优于Point-NeRF，分别为30.95 dB、0.926、0.218，获得了最高的分数。

图7给出了本文方法与Point-NeRF在Synthetic Blender数据集中Lego、Ficus、Drums场景上PSNR值的可视化结果。结果表明，在进行场景微调时，与Point-NeRF相比，本文方法在不同场景下初始化的

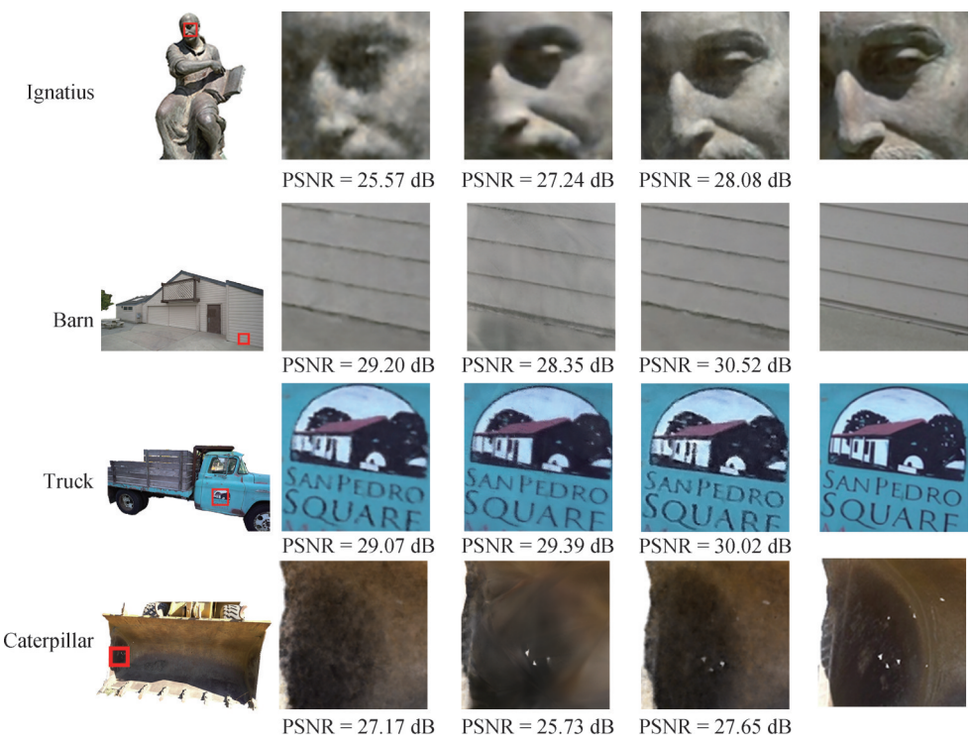



图6　Tanks and Temples可视化实验结果

Fig. 6　Visual experimental results on Tanks and Temples

((a) scenes; (b) Point-NeRF (Xu et al., 2022b); (c) Tetra-NeRF (Kulhanek and Sattler, 2023); (d) ours; (e) ground truth)

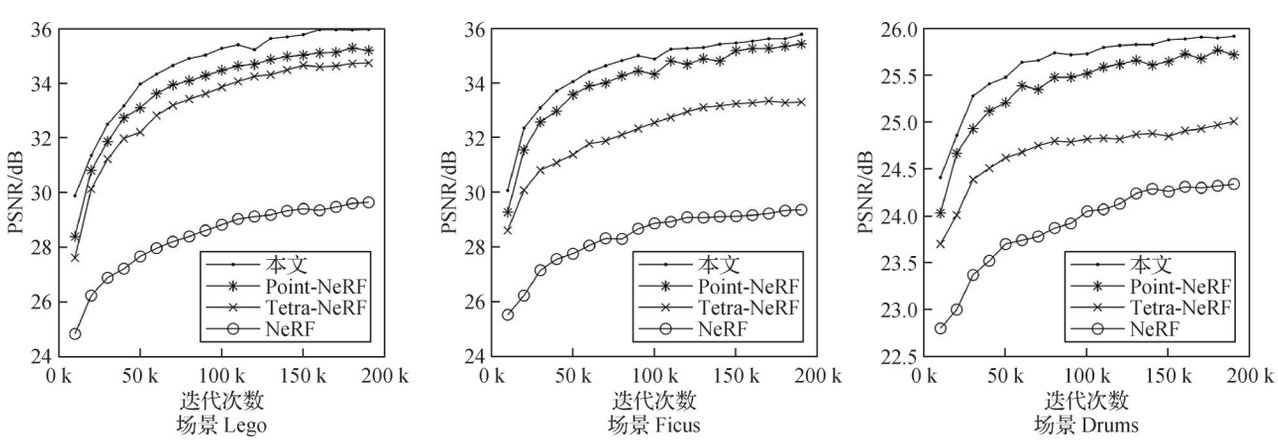



图7　验证集上的训练表现

Fig. 7　The training performance on validation set

效果更好。在较低的迭代轮次下，本文方法同样可以获得较好的实验结果。经相同轮次迭代后，本文方法的合成图像 PSNR 值更高，由此可见，在进行场景微调时，本文方法能更有效地拟合不同场景。

### 3.5 消融实验

为验证网络模型的神经点云特征对齐模块和神经点云 Transformer 模块的有效性，探究不同模块对合成图像质量的影响，在 Synthetic Blender、Tanks and Temples 数据集上开展消融实验。消融实验中采取控制变量的原则，分为添加神经点云特征对齐模块，添加神经点云 Transformer 模块和完整网络的测试，实验迭代次数设置为 200 k，实验过程中各参数设置保持一致，其中基准网络为去除上述模块后实验结果，实验结果如表 4 所示。

#### 3.5.1 神经点云特征对齐模块

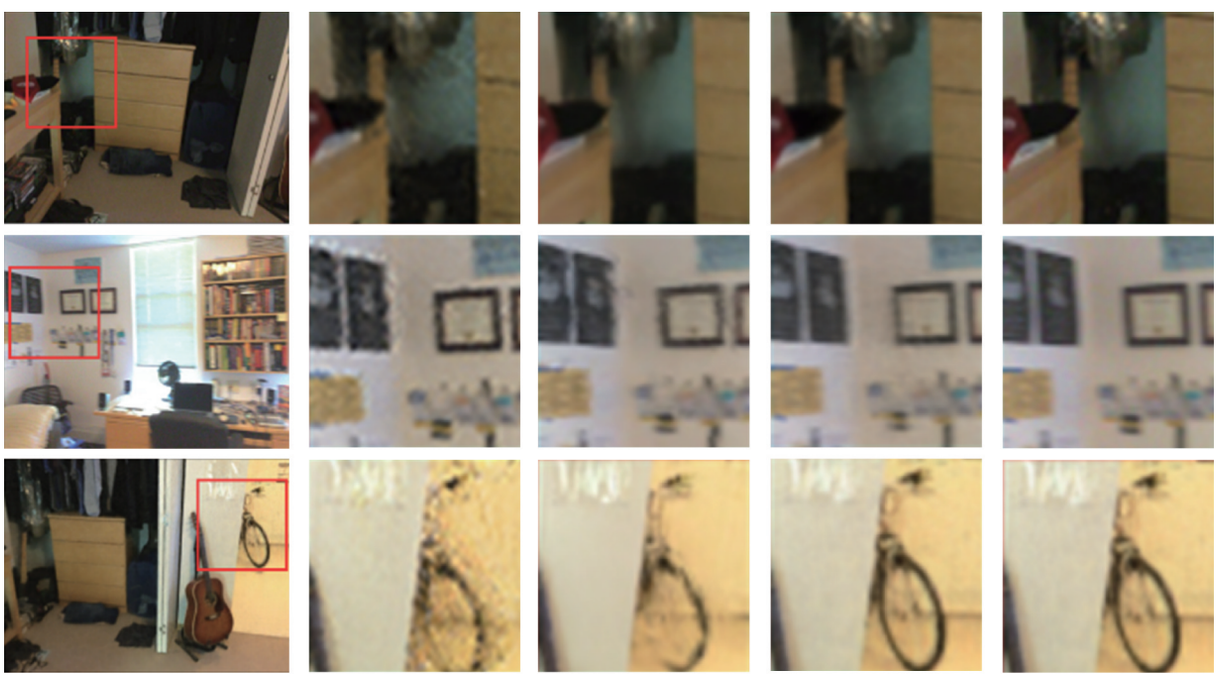
由表 4 可见，基准网络的实验结果指标值最低，单独添加神经点云特征对齐模块后，在两个数据体上的指标均高于基准网络。该模块通过动态调整点云所对齐特征的感受野，增强了神经点云对局部空间特征的适配度，使网络在处理不同形状、纹理的场景图像输入时能专注于相关区域，以此提升合成图像的真实度。图 8 给出了一组消融实验在 ScanNet

**表 4 消融实验：不同模块在数据集上的效果**

**Table 4 Ablation experiments: the effect of different modules on datasets**

| 方法 | Synthetic Blender | | | Tanks and Temples | | |
|---|---|---|---|---|---|---|
| | PSNR/dB | SSIM | LPIPS | PSNR/dB | SSIM | LPIPS |
| 基准网络 | 33.21 | 0.975 | 0.051 | 29.59 | 0.952 | 0.084 |
| 神经点云特征对齐模块 | 33.31 | 0.976 | 0.048 | 29.95 | 0.954 | 0.072 |
| 神经点云 Transformer 模块 | 33.42 | 0.977 | 0.047 | 30.13 | 0.959 | 0.063 |
| 完整网络 | **33.45** | **0.979** | **0.045** | **30.74** | **0.971** | **0.061** |

注：加粗字体表示各列最优结果。



(a) 场景　(b) 基准网络　(c) 神经点云特征对齐模块　(d) 完整网络　(e) 真值

图 8 消融实验的可视化结果

Fig. 8 Visualization results of ablation experiments

((a) scenes; (b) baseline network; (c) neural point feature alignment module; (d) ours; (e) ground truth)

数据集上的可视化结果。与NeRF Synthetic、Tanks and Temples数据集不同,ScanNet数据集包含多种不同形状物体的复杂场景,与单一物体的场景相比,重建场景难度大幅度增加。在形成神经点云的过程中,如果点云的对齐特征不够准确,可能会在神经点云特征中引入噪声,造成模糊和伪影,从而影响最终结果质量。此外,采用该模块形成神经点云特征,也能够提高后续网络在提取局部空间特征上的能力。

3.5.2 神经点云Transforme模块

神经点云Transformer模块在编码时,通过将神经点云的信息映射到高维空间中,获取神经点云的高频空间特征信息。利用该信息,网络能够更好地捕获采样点周围的上下文信息,并有效利用邻近神经点云之间的关联性,在点云稀疏的区域也能够提取出可靠的局部空间特征。运用该模块,模型在两个数据集上的指标均有所提高,在点云质量不佳的真实场景Tanks and Temples中,通过该模块合成出的场景图像提升更加明显。

根据消融实验的结果,在各个模块的共同作用下,完整网络在两个数据集上的PSNR、SSIM、LPIPS指标均表现出最佳效果,能够有效地利用神经点云的特征信息合成高质量场景图像,验证了各个模块对网络性能提升的有效性。

## 4 结 论

本文提出一种融合局部空间信息的新视角合成方法。在形成神经点云时,使用特征金字塔网络获取融合了场景不同尺度信息的图像特征,通过神经点云特征对齐模块调整局部特征的采样位置,查询每个点云所对应的图像特征区域并生成神经点云,与直接使用2D卷积网络进行下采样相比,能够更好地关注相关区域并捕获更多有用的特征,并将其融入到神经点云中。在体积密度预测网络前,通过神经点云Transformer模块捕获局部空间的上下文表示,局部神经点云特征通过自注意力机制对自身线索进行补全,能够有效处理点云的稀疏性和不规则性,提升最终合成图像的质量。实验结果表明,本文方法在真实场景和弱纹理区域的合成结果得到了显著提升。尽管本文方法能够借助点云完成场景重建的任务,但仍然存在局限性。由于本文方法采用了与NeRF相同的体渲染方法,导致渲染速度较慢,未来工作将着重提高模型的实时性。

## 参考文献(References)

Aliev K A, Sevastopolsky A, Kolos M, Ulyanov D and Lempitsky V. 2020. Neural point-based graphics//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 696-712 [DOI: 10.1007/978-3-030-58542-6_42]

Chen A P, Xu Z X, Zhao F Q, Zhang X S, Xiang F B, Yu J Y and Su H. 2021. MVSNeRF: fast generalizable radiance field reconstruction from multi-view stereo//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 14104-14113 [DOI: 10.1109/ICCV48922.2021.01386]

Cheng S, Xu Z X, Zhu S L, Li Z W, Li L E, Ramamoorthi R and Su H. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 2521-2531 [DOI: 10.1109/CVPR42600.2020.00260]

Chung J, Oh J and Lee K M. 2024. Depth-regularized optimization for 3D Gaussian splatting in few-shot images//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE: 811-820 [DOI: 10.1109/cvprw63382.2024.00086]

Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]

Fridovich-Keil S, Yu A, Tancik M, Chen Q H, Recht B and Kanazawa A. 2022. Plenoxels: radiance fields without neural networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5491-5500 [DOI: 10.1109/CVPR52688.2022.00542]

Govindarajan S, Sambugaro Z, Shabanov A, Takikawa T, Rebain D, Sun W W, Conci N, Yi K M and Tagliasacchi A. 2025. Lagrangian hashing for compressed neural field representations//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer-Verlag: 183-199 [DOI: 10.1007/978-3-031-73383-3_11]

Jain A, Tancik M and Abbeel P. 2021. Putting NeRF on a diet: semantically consistent few-shot view synthesis//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 5865-5874 [DOI: 10.1109/ICCV48922.2021.00583]

Kajiya J T and von Herzen B P. 1984. Ray tracing volume densities.

ACM SIGGRAPH Computer Graphics, 18(3): 165-174 [DOI: 10.1145/964965.808594]

Kerbl B, Kopanas G, Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG), 42(4): #139 [DOI: 10.1145/3592433]

Kulhanek J and Sattler T. 2023. Tetra-NeRF: representing neural radiance fields using tetrahedra//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 18412-18423 [DOI: 10.1109/ICCV51070.2023.01692]

Lin T Y, Dollar P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Liu L J, Gu J T, Lin K Z, Chua T S and Theobalt C. 2020. Neural sparse voxel fields//Proceedings of the 34th International Conference on Neural Information Processing System. Vancouver, Canada: Curran Associates Inc.: 15651-15663

Liu X N, Chen C Y, Hu X J and Yu H Y. 2024. Virtual viewpoint image synthesis using neural radiance fields with depth information supervision. Journal of Image and Graphics, 29(7): 2035-2045 (刘晓楠, 陈纯毅, 胡小娟, 于海洋. 2024. 带深度信息监督的神经辐射场虚拟视点画面合成. 中国图象图形学报, 29(7): 2035-2045) [DOI: 10.11834/jig.221188]

Lombardi S, Simon T, Saragih J, Schwartz G, Lehrmann A and Sheikh Y. 2019. Neural volumes: learning dynamic renderable volumes from images. ACM Transactions on Graphics (TOG), 38(4): #65 [DOI: 10.1145/3306346.3323020]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer-Verlag: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Müller T, Evans A, Schied C and Keller A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41(4): #102 [DOI: 10.1145/3528223.3530127]

Rakhimov R, Ardelean A T, Lempitsky V and Burnaev E. 2022. NPBG++: accelerating neural point-based graphics//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 15948-15958 [DOI: 10.1109/CVPR52688.2022.01550]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin L. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Verbin D, Hedman P, Mildenhall B, Zickler T, Barron J T and Srinivasan P P. 2022. Ref-NeRF: structured view-dependent appearance for neural radiance fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5481-5490 [DOI: 10.1109/CVPR52688.2022.00541]

Wang Z Y, Yang W, Cao J M, Hu Q, Xu L, Yu J Q and Yu J Y. 2023. NeReF: neural refractive field for fluid surface reconstruction and rendering//Proceedings of 2023 IEEE International Conference on Computational Photography (ICCP). Madison, USA: IEEE: 1-11 [DOI: 10.1109/ICCP56744.2023.10233838]

Xiao Q, Chen M L, Zhang H and Huang X H. 2024. Neural radiance field reconstruction for sparse indoor panoramas. Journal of Image and Graphics, 29(9): 2596-2609 (肖强, 陈铭林, 张晔, 黄小红. 2024. 室内稀疏全景图的神经辐射场重建. 中国图象图形学报, 29(9): 2596-2609) [DOI: 10.11834/jig.230643]

Xu D J, Jiang Y F, Wang P H, Fan Z W, Shi H and Wang Z Y. 2022a. SinNeRF: training neural radiance fields on complex scenes from a single image//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 736-753 [DOI: 10.1007/978-3-031-20047-2_42]

Xu Q G, Xu Z X, Philip J, Bi S, Shu Z X, Sunkavalli K and Neumann U. 2022b. Point-NeRF: point-based neural radiance fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5428-5438 [DOI: 10.1109/CVPR52688.2022.00536]

Yao Y, Luo Z X, Li S W, Fang T and Quan L. 2018. MVSNet: depth inference for unstructured multi-view stereo//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 785-801 [DOI: 10.1007/978-3-030-01237-3_47]

Yu A, Li R L, Tancik M, Li H, Ng R and Kanazawa A. 2021a. PlenOctrees for real-time rendering of neural radiance fields//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 5732-5741 [DOI: 10.1109/ICCV48922.2021.00570]

Yu A, Ye V, Tancik M and Kanazawa A. 2021b. pixelNeRF: neural radiance fields from one or few images//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 4576-4585 [DOI: 10.1109/CVPR46437.2021.00455]

Zhang Y, Huang X Y, Ni B B, Zhang W J and Li T. 2023. Frequency-modulated point cloud rendering with easy editing//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 119-129 [DOI: 10.1109/CVPR52729.2023.00020]

Zhu X Z, Hu H, Lin S and Dai J F. 2019. Deformable ConvNets V2: more deformable, better results//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 9300-9308 [DOI: 10.1109/cvpr.2019.00953]

## 作者简介

贾迪，男，教授，主要研究方向为立体匹配与三维重建、摄影测量、视觉空间定位和机器人。E-mail：lntu_jiadi@163.com
刘洋，通信作者，男，硕士研究生，主要研究方向为三维重建和新视角合成。E-mail：liuyang2254399194@163.com
李维，男，硕士研究生，主要研究方向为光场相机和深度估计。E-mail：15804293941@163.com
韩雪峰，男，研究员，主要研究方向为人工智能和智慧矿山。E-mail：173657676@qq.com
宋慧伦，女，硕士研究生，主要研究方向为单目深度估计。E-mail：lntu_songhuilun@163.com
孟晓华，女，硕士研究生，主要研究方向为三维重建。E-mail：mxh_wy2022@163.com
刘宇琪，男，硕士研究生，主要研究方向为单目深度估计。E-mail：yuqi030204@sina.com