DOI:10.19651/j.cnki.emt.2314738

# 注意力机制与神经渲染的多视图三维重建算法*

朱代先[1] 孔浩然[1] 秋 强[1] 刘树林[2] 张亚莉[1]
(1.西安科技大学通信与信息工程学院 西安 710054;2.西安科技大学电气与控制工程学院 西安 710054)

**摘 要**:针对多视图立体网络在弱纹理或非朗伯曲面等挑战性区域重建效果差的问题,首先提出一个基于 3 个并行扩展卷积和注意力机制的多尺度特征提取模块,在增加感受野的同时捕获特征之间的依赖关系以获取全局上下文信息,从而提升多视图立体网络在挑战性区域特征的表征能力以进行鲁棒的特征匹配。其次在代价体正则化 3D CNN 部分引入注意力机制,使网络注意于代价体中的重要区域以进行平滑处理。另外建立一个神经渲染网络,该网络利用渲染参考损失精确地解析辐射场景表达的几何外观信息,并引入深度一致性损失保持多视图立体网络与神经渲染网络之间的几何一致性,有效地缓解有噪声代价体对多视图立体网络的不利影响。该算法在室内 DTU 数据集中测试,点云重建的完整性和整体性指标分别为 0.289 和 0.326,与基准方法 CasMVSNet 相比,分别提升 24.9%和 8.2%,即使在挑战性区域也得到高质量的重建效果;在室外 Tanks and Temples 中级数据集中,点云重建的平均 F-score 为 60.31,与方法 UCS-Net 相比提升 9.9%,体现出较强的泛化能力。

**关键词**:多视图立体网络;三维重建;注意力机制;神经渲染
**中图分类号**:TP391.4 **文献标识码**:A **国家标准学科分类代码**:520.6030

# Attention mechanism and neural rendering for Multi-View 3D reconstruction algorithm

Zhu Daixian[1] Kong Haoran[1] Qiu Qiang[1] Liu Shulin[2] Zhang Yali[1]
(1. School of Communication and Information Engineering, Xi′an University of Science and Technology, Xi′an 710054, China;
2. School of Electrical and Control Engineering, Xi′an University of Science and Technology, Xi′an 710054, China)

**Abstract**: Aiming at the problem of poor reconstruction of Multi-View Stereo Networks in challenging regions such as weak textures or non-Lambertian surfaces, this paper first proposes a multi-scale feature extraction module based on three parallel dilated convolution and attention mechanism, which enables the network to capture the dependencies between features while increasing the sensory field to obtain global context information, thus enhancing the multi-view stereo network′s ability to characterize features in challenging regions for robust feature matching. Secondly, an attention mechanism is introduced in the 3D CNN part of the cost volume regularization so that the network pays attention to the important regions in the cost volume for smoothing. Additionally, a neural rendering network is built, which utilizes the rendering reference loss to accurately resolve the geometric appearance information expressed by the radiance field and introduces the depth consistency loss to maintain the geometric consistency between the multi-view stereo network and the neural rendering network, which effectively mitigates the detrimental effect of the noisy cost volume on the multi-view stereo network. The algorithm is tested in the indoor DTU dataset, achieving completeness and overall metrics of 0.289 and 0.326, respectively. Compared to the benchmark method CasMVSNet, there is an improvement of 24.9% and 8.2% in the two metrics, demonstrating high-quality reconstruction even in challenging regions. In the outdoor Tanks and Temples intermediate dataset, the average F-score for point cloud reconstruction is 60.31, showing a 9.9% improvement over the UCS-Net method. This reflects the algorithm′s strong generalization capability.

**Keywords**: multi-view stereo network;3D reconstruction;attention mechanism;neural rendering

## 0 引 言

随着计算机视觉技术的快速发展,多视图立体(multi-view stereo,MVS)重建成为了一个备受关注的领域。多视图立体的研究旨在从已知相机参数的多个视角图像中重建出场景的三维信息,在虚拟现实、增强现实和电影特效等领

收稿日期:2023-10-11
* 基金项目:国家自然科学基金(51774235)、陕西省重点研发计划项目(2021GY-338)、西安市碑林区科技计划项目(GX2333)资助

域中发挥着重要作用。几十年来传统 MVS 方法[1-2]得到了广泛的研究,但仍存在一些难以解决的问题,如重建不完全、可扩展性有限等。

随着深度学习技术的兴起,许多研究人员利用卷积神经网络(convolutional neural networks, CNN)处理 MVS 任务。作为代表性工作,Yao 等[3]提出了 MVSNet,一个基于深度学习的 MVS 网络,该网络通过可微单应性变换从不同视图的特征构建 3D 代价体,并利用 3D CNN 正则化代价体进行深度回归。然而 MVSNet 消耗了大量的内存,后续该团队成员寻求了更轻量化的方法,使用 Recurrent MVSNet 架构[4],通过 GRU(Gated Recurrent Unit,GRU)沿着深度方向顺序调整二维特征,从而避免因为一次性调整整个代价体而造成内存的消耗,使高分辨率重建成为可能。Yang 等[5]提出一个由粗到细的轻量化方法。该方法首先预测具有大范围深度间隔的低分辨率深度图,并且随着深度范围的减小而迭代的增加深度图分辨率,其算法有效的减少了代价体过大而引起的内存消耗。叶春凯等[6]提出了一种基于特征金字塔的多视图重建网络,该网络采用金字塔网络采集并融合多尺度图像特征,从而对 3D 场景进行准确的重建。但由于传统 CNN 卷积无法获得在挑战性区域(例如弱纹理和非朗伯表面区域)的特征信息,同一 3D 位置的特征在不同视图之间可能存在较大的差异,错误的特征匹配导致网络构建出有误差的代价体,即使利用具有泛化能力的 3D CNN 网络也无法回归出准确的深度图。因此后续工作将注意力机制引入 MVS 网络中,以提高特征对图像的表示能力。Yu 等[7]在特征提取阶段引入独立自注意力机制使网络更聚焦于重要信息,捕获像素之间的相互依赖关系。但是这种方法在挑战性区域重建的质量还有很大的发展空间。由于 Transformer[8]模型利用自注意力机制可以捕捉全局上下文信息的天然优势,后续许多工作[9-10]将其引入 MVS 中,TransMVSNet[9]引入了特征匹配 Transformer,利用内部和交叉注意力来加强图像内和图像间的远程全局上下文信息聚合。MVSTR[10]设计了全局上下文 Transformer 和三维几何 Transformer 模块,以便提取具有全局上下文的密集特征,实现特征的三维一致性,促进视图间信息交互。虽然 Tranformer 可以提供对 MVS 模型的全局理解,但这往往增加了运行时间和内存占比,在高分辨率的大范围场景中重建成本消耗大。

最近神经辐射场渲染技术(neural radiance fields, NeRF)[11]在计算机视觉和图形学研究领域取得了显著发展,其通过可微积分神经渲染技术将 3D 场景建模成连续的辐射场并渲染出新颖视图。后续 Yu 等[12]和 Chen 等[13]结合 MVS,即使是在弱纹理或非朗伯表面区域,NeRF 也可以输入少量的视图隐式学习 3D 场景的几何,合成真实新颖的视图。因此本文利用神经辐射场构建的场景 3D 几何信息,使 MVS 网络在挑战性区域中学习到除了表示场景几何信息的代价体之外的场景几何信息,可以有效缓解网络因错误匹配构建的有噪声代价体的影响。

综上所述,本文提出了一种端到端的基于注意力机制与神经渲染网络的 MVS 重建网络。针对网络在挑战性区域产生的特征错误匹配的问题,本文在特征提取部分提出了注意力机制的多尺度特征提取模块。该模块利用 3 个并行扩展卷积在扩大感受野的同时减少参数量,并且在特征提取的每个级别中加入注意力层使网络捕获像素之间的相互依赖关系从而聚焦于重要信息,即使在弱纹理或非朗伯表面区域也能提取环境中丰富的信息。为了缓解网络因错误匹配构建的有噪声代价体的影响,首先在 3D CNN 部分引入注意力机制平滑代价体,其次建立一个采用神经编码体的神经渲染网络。该网络采取置信度和深度引导的采样策略,高效率地集中物体表面上采取采样点,并应用渲染参考视图损失函数不断优化神经辐射场景,精确地解析辐射场景表达的几何外观信息。最后引入深度一致性损失函数保持 MVS 网络与神经渲染网络之间的几何一致性。

## 1 网络模型

本节阐述了所提出方法的总体架构,如图 1 所示。该架构主要由 MVS 网络和神经渲染网络构成。首先多张视图输入到注意力机制的多尺度特征提取模块中,提取到丰富的特征信息之后, MVS 网络采用由粗到细的方式构建概率体从而估计出深度图和置信度图。然后本文设计了一个全新的神经渲染网络,MLP(multi-layer-perceptron, MLP)网络以具有几何感知信息的神经编码体作为映射条件,同时采用以深度和置信度为引导的均匀采样策略,将场景采样集中在估计的深度表面区域。最后应用渲染参考视图损失函数 $\lambda_{RRV}$ 和深度一致性损失函数 $\lambda_{DC}$ 不断地训练整个网络。

### 1.1 注意力机制的多尺度特征提取模块

如图 2 所示,注意力机制的多尺度特征提取模块由包含注意力块的编码器和带跳跃连接的解码器构成,注意力块是由扩展卷积层与注意力模块构成的网络。如图 3 所示,在注意力块中每一层的特征送入到卷积核大小 $3\times3$,步长为 2 的卷积层进行下采样操作,随后利用 3 个并行且不同扩展率的扩展卷积层扩大输入特征的感受野。为了避免采用 3 个并行的扩展卷积层导致信息关联性降低的问题,特征经过扩展卷积层之后都会送入到一个带有 Sigmoid 函数的残差网络结构。最后将 3 个特征拼接送入带有注意力层的注意力模块中生成最终的特征图。当输入分辨率为 $H\times W$ 的参考图像 $I_1$ 和从不同视点拍摄的源图像 $\{I_i\}_{i=2}^{N}$,注意力机制的多尺度特征提取模块输出 3 个不同尺度的特征如式(1)所示,其中 $k$ 代表第 $k$ 个阶段。

$$\{F_{i,k=1}\in R^{\frac{H}{4}\times\frac{W}{4}\times32},F_{i,k=2}\in R^{\frac{H}{2}\times\frac{W}{2}\times16},F_{i,k=3}\in R^{H\times W\times8}\}_{i=1}^{N} \tag{1}$$

如图 4 所示的注意力模块体系结构中,首先将经过扩

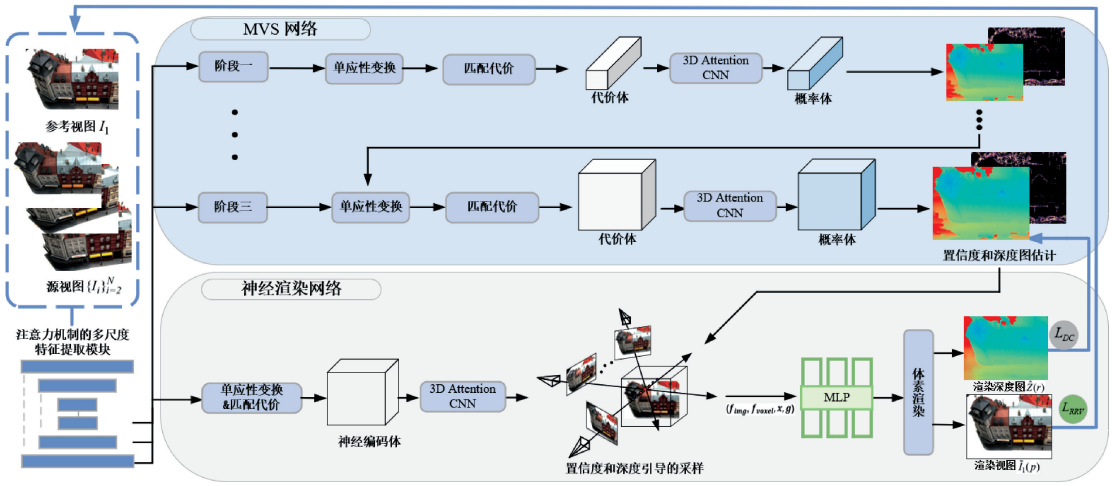



图 1　网络整体结构图

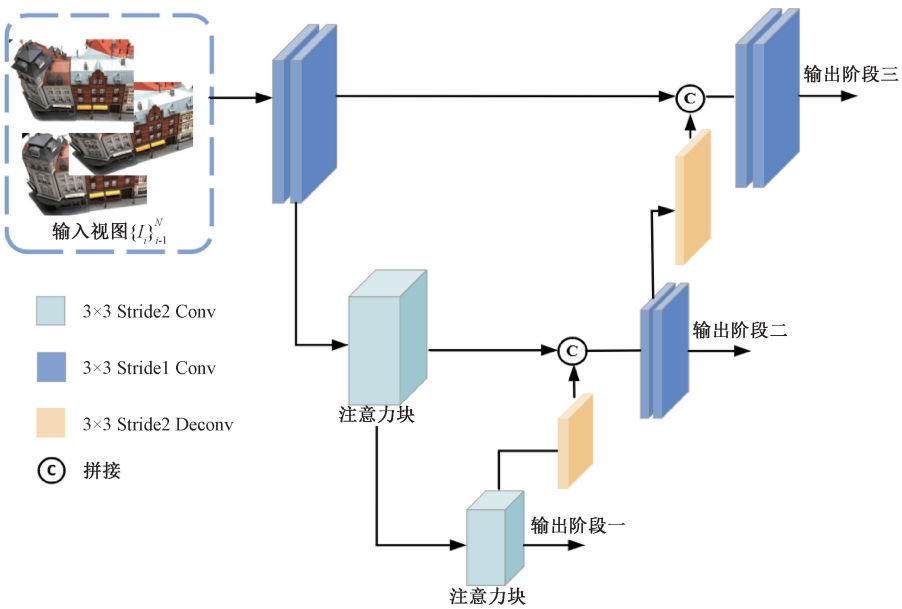



图 2　注意力机制的多尺度特征提取模块体系结构

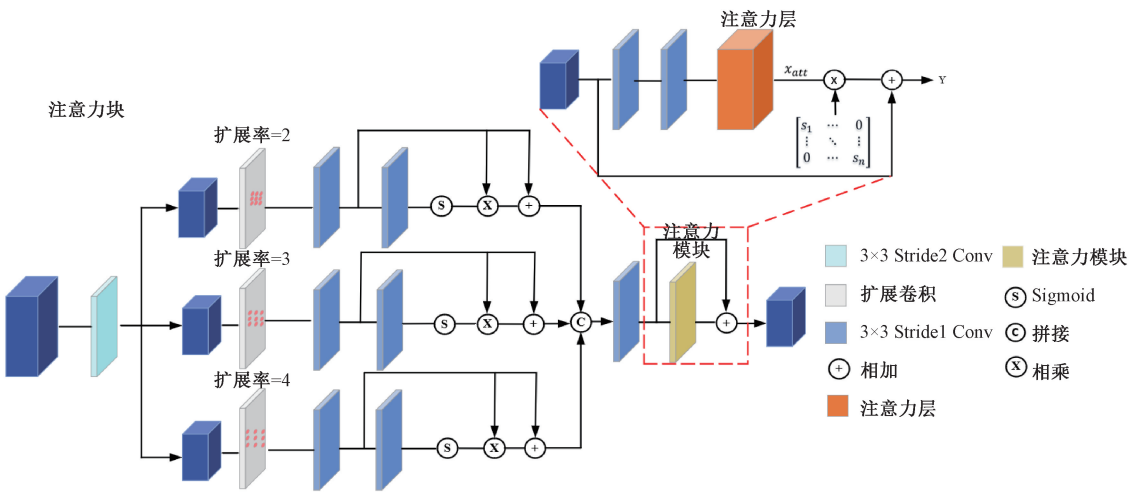



图 3　注意力块的体系结构

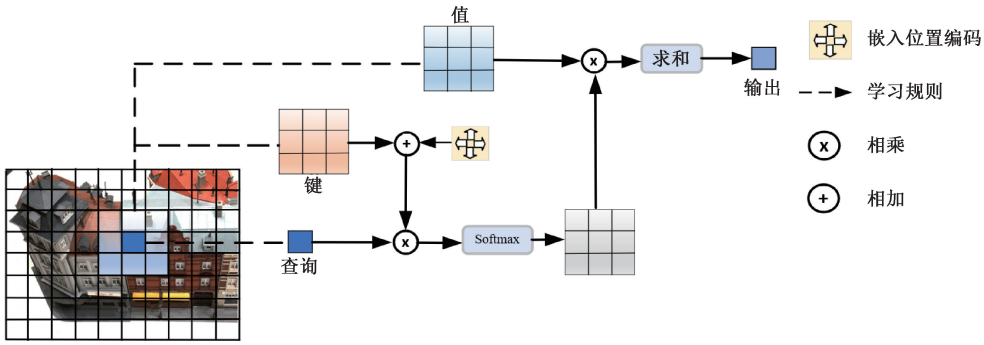



图 4 注意力层的体系结构

展卷积层后的特征输入到两个卷积核为 3×3 的卷积层，每个卷积层都会经过 BN(batch normalization，BN)和 ReLU(rectified linear unit，ReLU)层。然后输入到带有 LayerScale[14]的注意力层。注意力层的实现如图 4 所示，其将查询和一组键值对映射到输出，然后通过 softmax 函数计算像素输出：

$$y_{ij}=\sum_{a,b\in R}w_{ab}(q_{ij}^{T}(k_{ab}+\boldsymbol{r}_{ab}))v_{ab} \tag{2}$$

其中，$q_{ij}=\boldsymbol{W}_q x_{ij}$、$k_{ab}=\boldsymbol{W}_k x_{ab}$、$v_{ab}=\boldsymbol{W}_v x_{ab}$ 分别表示查询、键和值，矩阵 $\boldsymbol{W}_g(g=q,k,v)$ 由学习参数组成。$R$ 表示输入大小为 3×3 的局部区域。为了避免位置信息没有被编码导致排列等价的问题，本文通过向键添加可学习的参数来引入相对位置嵌入[15]。相对距离向量 $\boldsymbol{r}_{ab}$ 在维度上被分解，是使用输出向量的一半维度对行方向进行编码，另一半维度对列方向进行编码而分解的。注意力层输出的特征 $x_{att}$ 还需要乘上网络学习的对角矩阵权重。

$$Y=diag(s_1,\cdots,s_n)\times x_{att}+x \tag{3}$$

其中，$s_1\sim s_n$ 是可学习的权重。

### 1.2 构建代价体

将 2D 源图像特征可微分地扭曲(warp)到参考视图，构建了一组特征体 $\{V_i\}_{i=1}^{N}$。在深度平面假设 $d$ 下，参考视图处的像素 $p$ 与其在源视图处对应的像素 $p_i'$ 之间的扭曲被定义为：

$$p_i'=\boldsymbol{K}_i[(R_i(dK^{-1}p)+t_i] \tag{4}$$

其中，$R_i$ 和 $t_i$ 分别表示两个视图之间的旋转和平移。$\boldsymbol{K}_i$ 和 $\boldsymbol{K}$ 是源相机的内参矩阵和参考相机的内参矩阵。

之后采用基于方差的聚合策略将多个特征体 $\{V_i\}_{i=1}^{N}$ 聚合成一个 3D 代价体 $V$。接下来，利用 3D CNN 正则化代价体处理成概率体，本文将 3D 注意力层嵌入到 3D CNN 中，如图 5 所示的 3D Attention CNN 正则化体系结构。为了减少网络模型的参数量，仅在下采样最后一层的跳跃连接之间引入 3D 注意力层，这样使 3D CNN 集中注意力于代价体的重要区域，从而忽略不重要的特征或噪声信息，对代价体进行平滑处理。

最后，对于参考视图的一个像素 $p$，可以获得它在某一深度平面 $D_j(p)$ 的概率 $P_j(p)$，则像素 $p$ 的深度值定义为：

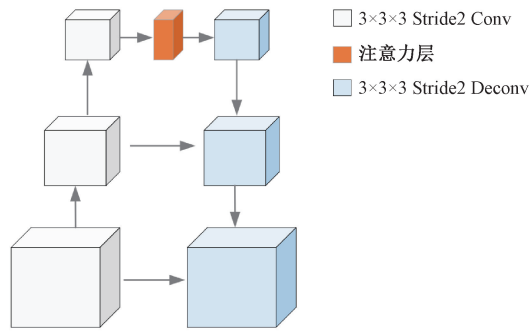



图 5 3D Attention CNN 正则化体系结构

$$\hat{D}(p)=\sum_{j=1}^{J}P_j(p)D_j(p) \tag{5}$$

### 1.3 神经渲染网络

1) 基于神经编码体的场景表示

对于沿着从参考相机中心发射到参考视图方向的射线上每个点 $x$，通过双线性采样的方式查询源视图对应的 RGB 值 $f_{img}$，除此之外，本文与 RC-MVSNet[16]做法相同，对利用 MVS 方法构造的 3D 神经编码体进行三线性插值以获得体素对齐的三维特征体素 $f_{voxel}$。最后将 $x$ 对应到各源视图的 RGB 值 $f_{img}$ 和三维特征体素 $f_{voxel}$ 传递到 MLP 网络中获得在参考视图方向的 3D 采样点的颜色 $c$ 和体密度 $\sigma$：

$$[c,\sigma]=\varphi(f_{img},f_{voxel},\gamma_x(x),\gamma_g(g)) \tag{6}$$

其中，$g$ 表示 3D 点 $x$ 的参考视图方向，$\gamma$ 是位置编码函数，有助于网络恢复高频深度。

2) 基于置信度和深度引导的采样策略

参考视图 $I_1$ 中每个像素点 $p$ 确定世界坐标系中的一条射线，其原点是摄像机投影中心 $o=e$。沿着与距离原点 $e$ 处的 $p$ 相关联的 3D 点表示为 $r_p=o+eg$。为了在像素 $p$ 处渲染颜色 $\tilde{I}_1(p)$，原始 NeRF 在近平面和远平面 $[e_n,e_f]$ 内的射线中对 $M$ 个离散样本距离 $e_m$ 进行均匀采样，并查询 3D 点处的辐射场 $\varphi$：

$$e_m\sim\mu\left[e_n+\frac{m-1}{M}(e_f-e_n),t_n+\frac{m}{M}(e_f-e_n)\right] \tag{7}$$

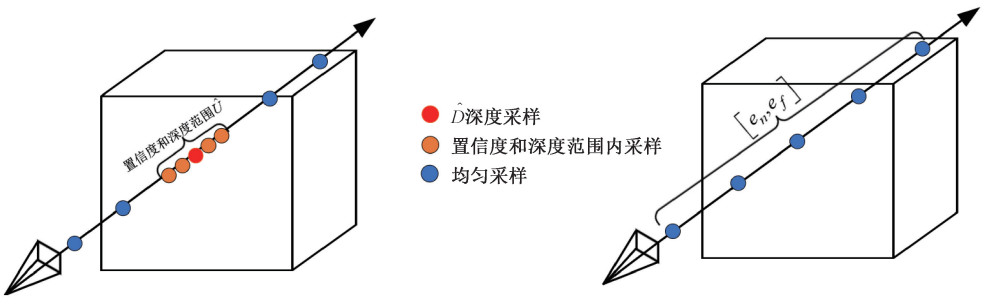
由于原始 NeRF 采用的均匀采样策略不能集中在物体表面上，这将降低辐射场构建 3D 几何的能力导致渲染参考视图的质量较差。因此对于像素 $p$，本文从 MVS 网络估计的深度值和置信度的先验范围指导下，对候选点进行采样。对于深度估计值为 $\hat{D}(p)$ 的像素 $p$，将其置信度定义为标准差 $\hat{S}$：

$$\hat{S}(p)=\sqrt{\sum_{j=1}^{J}P_j(p)(D_j(p)-\hat{D}(p))^2} \tag{8}$$

每个像素对应的物体表面位置可能位于由深度估计值 $\hat{D}(p)$ 和标准差 $\hat{S}(p)$ 定义的范围 $\hat{U}(p)$ 内：

$$\hat{U}(p)=[\hat{D}(p)-\hat{S}(p),\hat{D}(p)+\hat{S}(p)] \tag{9}$$

为了渲染 3D 点 $x$ 的颜色，本文替换了原始 NeRF 用于分层采样的粗网络。图 6 展示出了两种采样方法的比较。如图 6(a)所示，本文将采样样本数量的一半分布在近平面 $e_n$ 和远平面 $e_f$ 之间进行，采样样本的另一半由深度和置信度先验范围 $\hat{U}(p)$ 内提取。与图 6(b)所示的原始 NeRF 所采用的均匀采样策略相比，置信度和深度引导的采样策略更集中于物体表面。




(a) 基于置信度和深度引导的采样策略　　(b) 原始NeRF的均匀采样策略

图 6　两种采样方法对比

然后将各个采样点预测的颜色和体积密度值 $\{(c_m,\sigma_m)\}_{m=1}^{M}$ 渲染为参考视图的像素：

$$\tilde{I}_1(p)=\sum_{m=1}^{M}\alpha_m c_m \tag{10}$$

$$\alpha_m=E_m(1-\exp(-\sigma_m\delta_m)) \tag{11}$$

$$E_m=\exp\Big(-\sum_{m'=1}^{m}\sigma_{m'}\delta_{m'}\Big) \tag{12}$$

其中，$E_m$ 表示沿着 $e_m$ 的射线累计透射率，$\delta_m=e_{m+1}-e_m$ 是相邻样本之间的距离。

最后沿着参考视角的射线，将其密度积分渲染为该像素 $p$ 的深度值：

$$\hat{Z}(r)=\sum_{m=1}^{M}\alpha_m e_m \tag{13}$$

### 1.4　损失函数

在神经渲染网络中，本文引入了渲染参考视图损失函数，该函数使用均方误差测量体素渲染后的参考视图颜色和地面真实参考视图颜色之间的差异，从而优化网络对三维场景几何的表达能力。

$$L_{RRV}=\sum_{p}\|\tilde{I}_1(p)-I_1(p)\|_2^2 \tag{14}$$

为了保证两个网络之间几何的一致性，本文提出深度一致性损失函数，其使用 $L_1$ 损失函数最小化渲染参考视图深度和地面真实参考视图深度之间的差异并且最小化 MVS 网络估计的参考视图深度和神经渲染网络渲染的参考视图深度之间的差异。

$$L_{DC_1}=\sum_{p}\sum_{r}[\lambda_{DC_1}\|D_{GT}(p)-\hat{Z}(r)\|_1] \tag{15}$$

$$L_{DC_2}=\sum_{p}\sum_{r}[\lambda_{DC_2}\|\hat{D}(p)-\hat{Z}(r)\|_1] \tag{16}$$

$$L_{DC}=L_{DC_1}+L_{DC_2} \tag{17}$$

在 MVS 网络中，本文使用 $L_1$ 损失函数来测量地面真实深度值和网络估计深度值之间的绝对差。

$$L_{MVS}=\sum_{p}\|D_{GT}(p)-\hat{D}(p)\|_1 \tag{18}$$

因此本文设计的网络整体训练损失函数为：

$$L=\lambda_{RRV}L_{RRV}+\lambda_{DC}L_{DC}+\lambda_{MVS}L_{MVS} \tag{19}$$

## 2　实验结果与分析

本文根据 DTU[17] 和 Tanks and Temples[18] 数据集对网络进行训练和评估。DTU 数据集通过机械臂在 7 种不同的光照条件下采集室内的 124 个场景中 49 张不同视角下的视图。而 Tanks and Temples 数据集是在规模更大更复杂的现实场景中采集了 8 个中级数据集和 6 个高级数据集。

### 2.1　实验细节

在训练阶段，设置输入图像个数 $N=4$，将输入的原始图像分辨率剪裁到 512×640。在 MVS 网络中分为 3 个阶段，每个阶段分别输入图像为原始分辨率的 1/16、1/4 和 1。从最粗阶段到最细阶段本文分别假设了 48、32 和 8 个平面扫描深度数量，而且深度区间设为 4、2、1。在神经渲染网络中，射线采样数量设置为 1 024。$\lambda_{DC_1}$ 和 $\lambda_{DC_2}$ 分别设置为 0.8 和 0.2，$\lambda_{RRV}$、$\lambda_{DC}$ 和 $\lambda_{MVS}$ 分别设置为 1、0.01 和 1。

本实验是基于 Ubuntu20.04，CUDA11.3，pytorch1.10.0，Python3.8 搭建的深度学习环境，使用 2 个 Nvidia GTX 3090ti GPUs 上训练本文的方法，batch size 大小设置为 2。在训练过程中使用 Adam 优化器进行训练。

训练进行了 16 个阶段,初始学习率为 0.000 1,在 10、12 和 14 个 epoch 降低 2 倍学习率。

### 2.2 实验结果

1) DTU 数据集的结果

本文使用准确性、完整性和整体性这三种度量指标在 DTU 数据集进行评估网络模型。这 3 个度量指标数值越低表示重建效果越好。在评估时使用 5 个相邻视图,并输入分辨率为 1 600×1 184 的图像。使用 DTU 数据集提供的 MATLAB 评估脚本,用来测量重建的点云与地面真实点云之间的距离。本文方法与传统方法 Gipuma[1]、Colmap[2],以及最近先进的学习方法 MVSNet[3]、R-MVSNet[4]、CasMVSNet[5]、CVP-MVSNet[19]、UCS-Net[20]、P-MVSNet[21]、Point-MVSNet[22]、Fast-MVSNet[23]、PVA-MVSNet[24]、EPP-MVSNet[25]、PatchmatchNet[26]、AA-RMVSNet[27]进行定量评估对比。如表 1 所示,Gipuma 在准确性(Acc.)上表现最好,PatchmatchNet 在完整性(Comp.)上表现最好,而本文的方法在点云重建整体性指标(Overall)上优于所有方法,相较于基准网络 CasMVSNet,在点云重建完整性指标方面提升了 24.9%,在点云重建整体性指标方面提升了 8.2%。与最近 SOTA 方法 CVP-MVSNet 相比,点云重建完整性指标提高了 28.8%,并在点云重建整体性指标方面提高了 7.1%。在定量分析的基础上,本文增加了重建点云的可视化定性结果。如图 7 所示,在 scan9、scan32 和 scan49 中 CasMVSNet 和本文方法相比。CasMVSNet 在这些场景中都存在不同程度的点云空洞和细节缺失,尤其是在具有弱纹理和光照反射的挑战性区域,比如 scan9 的门和门窗都出现了边缘缺失,scan32 的饮料包装的细节缺失而且没有完整的重建效果,scan49 的靴子和标签出现了空洞,相比之下,本文的方法重建的点云在这些挑战性区域的重建结果最为完整,并且对表面的纹理细节恢复的更加精细。

**表 1 DTU 数据集上定量结果** mm

| 类别 | 方法 | Acc. | Comp. | Overall |
|---|---|---|---|---|
| 传统方法 | Gipuma | **0.283** | 0.873 | 0.578 |
| | Colmap | 0.400 | 0.644 | 0.532 |
| 基于学习的方法 | MVSNet | 0.396 | 0.527 | 0.462 |
| | R-MVSNet | 0.383 | 0.452 | 0.417 |
| | Fast-MVSNet | 0.336 | 0.403 | 0.370 |
| | AA-RMVSNet | 0.376 | 0.399 | 0.357 |
| | P-MVSNet | 0.406 | 0.434 | 0.420 |
| | Point-MVSNet | 0.342 | 0.411 | 0.376 |
| | PVA-MVSNet | 0.379 | 0.336 | 0.357 |
| | EPP-MVSNet | 0.413 | 0.296 | 0.355 |
| | PatchmatchNet | 0.427 | **0.277** | 0.352 |
| | UCS-Net | 0.338 | 0.349 | 0.344 |
| | CasMVSNet | 0.325 | 0.385 | 0.355 |
| | CVP-MVSNet | 0.296 | 0.406 | 0.351 |
| | 本文方法 | 0.363 | 0.289 | **0.326** |

注:加粗的数值是该列的最优值


(a) CasMVSNet (b) 本文方法 (c) 真实点云

图 7 DTU 数据集上与多种方法比较重建结果

2) Tanks and Temples 数据集的结果

本文采用在 DTU 数据集上训练且没有微调的权重模型进行测试 Tanks and Temples 数据集，从而评估网络的泛化能力。输入原始图像分辨率为 1 920×1 080，相邻视图数量为 5。中级数据集的定量结果如表 2 所示，该数据集的评估是通过将生成的点云提交到官方网站得到的 F-score 数值参数，其是衡量重建结果的准确性和完整性，数值越高代表重建效果越好。本文方法在中级数据集下的大多数场景中平均 F-score 高于其他基于学习的方法，如 CasMVSNet、UCS-Net 和 PatchmatchNet 等，其中比表现最好的 MVSTR 平均 F-score 还要高 6.5%，这证明所提方法具有较好的泛化能力，以及在室外大场景中重建效果的有效性。如图 8 所示本文方法在中级数据集上的可视化重建效果，选择了 Family、Horse、Lighthouse、Playground 场景进行重建结果显示。在图 8(a) Family 场景中，完成了对大理石雕塑的重建，重建点云的人像轮廓和纹理清晰、衣服凹凸有致；在图 8(b) Horse 场景中，完成对马雕塑的重建，重建点云的马身轮廓较为清楚，较完整且清晰的重建出马座中间标签里的字体。在光照强的室外图 8(c) Lighthouse 场景中，重建点云的建筑物楼顶、窗户区域较完整，在墙壁等纹理丰富区域更为清晰；在更加复杂的图 8(d) Playground 场景中完成了对儿童公园的重建，重建点云的滑梯的台阶部分清晰，草坪区域白色噪声较少且较完整，视觉效果较好。

**表 2　Tanks and Temples 中级数据集上定量结果**

| 方法 | Mean | Family | Francis | Horse | Lighthouse | M60 | Panther | Playground | Train |
|---|---|---|---|---|---|---|---|---|---|
| MVSNet | 43.48 | 55.99 | 28.55 | 25.07 | 50.79 | 53.96 | 50.86 | 47.90 | 34.69 |
| Point-MVSNet | 48.27 | 61.79 | 41.15 | 34.20 | 50.79 | 51.97 | 50.85 | 52.38 | 43.06 |
| Fast-MVSNet | 47.39 | 65.18 | 39.59 | 34.98 | 47.81 | 49.16 | 46.20 | 53.27 | 42.91 |
| CVP-MVSNet | 54.03 | 76.50 | 47.74 | 36.34 | 55.12 | 57.28 | 54.28 | 57.43 | 47.54 |
| R-MVSNet | 48.40 | 69.96 | 46.65 | 32.59 | 42.95 | 51.88 | 48.80 | 52.00 | 42.38 |
| PatchmatchNet | 53.15 | 66.99 | 52.64 | 43.24 | 54.87 | 52.87 | 49.54 | 54.21 | 50.81 |
| CasMVSNet | 56.42 | 76.36 | 58.45 | 46.20 | 55.53 | 56.11 | 54.02 | **58.17** | 46.56 |
| PVA-MVSNet | 54.46 | 69.36 | 46.80 | 46.01 | 55.74 | 57.23 | 54.75 | 56.70 | 49.06 |
| UCS-Net | 54.83 | 76.09 | 53.16 | 43.03 | 54.00 | 55.60 | 51.49 | 57.38 | 47.89 |
| MVSTR | 56.61 | 77.85 | 57.34 | 48.25 | 54.76 | 59.25 | 52.47 | 55.74 | 47.19 |
| 本文方法 | **60.31** | **79.68** | **62.33** | **53.60** | **56.86** | **61.06** | **57.66** | 56.48 | **54.85** |

注：加粗的数值是该列的最优值

(a) Family　(b) Horse　(c) Lighthouse　(d) Playground

图 8　Tanks and Temples 中级数据集上可视化本文方法的重建效果

### 2.3 消融实验

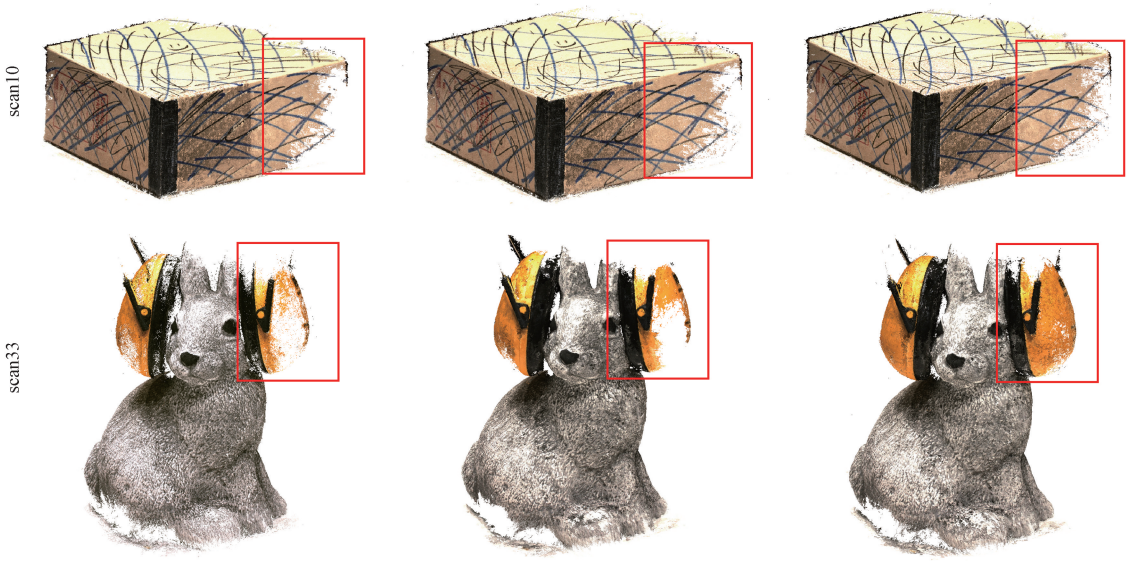
1) 改进模块和损失函数对重建结果的影响

为了进一步验证本文方法的有效性,对 MVS 网络进行了消融实验和定量定性结果实验分析,消融实验定量结果如表 3 所示,在 DTU 数据集上分别进行了 5 组对比实验,AM 代表注意力机制的多尺度特征提取模块,3DA 代表 3D Attention CNN 正则化部分,RRV 代表渲染参考视图损失函数,DC 代表深度一致性损失函数。单独使用 AM 模块后网络主要在完整性方面优于基础网络,这是因为 AM 模块能够更好地提取图像特征的细节和结构,从而提升了点云重建的完整性,在 AM 模块基础上增添了 3DA 模块后,将有噪声的代价体平滑,减少了无关噪声的影响,同时引入了神经渲染网络和相关的 RRV 与 DC 损失函数后,进一步减少了因错误匹配产生的有噪声的代价体对网络的影响,重建的完整性指标数值提升最大,同时保持了较高的整体性指标数值,达到 0.326。本文将不同构件对重建结果的影响可视化显现出来,如图 9 所示,在基线 CasMVSNet 上添加了 RRV 和 DC 损失函数后,网络学习到了额外场景几何信息,从而提升了重建的完整性。之后进一步加入 AM 模块和 3DA 模块,网络提取了丰富的特征信息,减少了传统 CNN 特征提取网络导致的特征匹配误差,又将 3D CNN 注意于代价体的重要信息区域,因此在弱纹理和非朗伯表面区域产生了更好的重建结果。

**表 3 DTU 数据集上消融实验定量结果** mm

| 方法 | Acc. | Comp. | Overall |
|---|---|---|---|
| Baseline | **0.325** | 0.385 | 0.355 |
| Baseline+AM | 0.357 | 0.324 | 0.340 |
| Baseline+AM+3DA | 0.375 | 0.297 | 0.336 |
| Baseline+RRV+DC | 0.369 | 0.310 | 0.339 |
| Baseline+AM+3DA+RRV+DC | 0.363 | **0.289** | **0.326** |

注:加粗的数值是该列的最优值




(a) Baseline (b) Baseline+RRV+DC (c) Baseline+RRV+DC+AM+3DA

图 9 DTU 数据集上消融实验定性结果对比

2) 在不同视图数量下置信度和深度引导的采样策略对重建结果的影响

消融实验结果如表 4 所示,CDG 代表置信度与深度引导的采样策略。当视图数量被设置为 4 时重建的整体性结果最好。因此本文在其他消融实验分析的视图数量设置为 4。

**表 4 DTU 数据集上不同视图数量下置信度与深度引导的采样策略的消融实验结果** mm

| 视图数量 | CDG | Acc. | Comp. | Overall |
|---|---|---|---|---|
| 3 | | 0.376 | 0.334 | 0.355 |
| 3 | √ | 0.370 | 0.320 | 0.345 |
| 4 | | **0.359** | 0.311 | 0.335 |
| 4 | √ | 0.363 | **0.289** | **0.326** |
| 5 | | 0.367 | 0.309 | 0.338 |
| 5 | √ | 0.367 | 0.297 | 0.332 |

注:加粗的数值是该列的最优值

此外,由于应用置信度与深度采样策略,采样的点集中于物体表面,使网络准确构造出神经辐射场景几何形状以缓解因匹配错误产生的有噪声代价体对网络的影响,因此重建的整体性结果会有所提升。

## 3 结 论

本文提出了一个基于注意力机制和神经渲染网络进行三维重建的端到端深度学习架构。注意力机制用于包含扩展卷积层的多尺度特征提取模块。扩展卷积在扩大感受野的同时减少参数量,注意力机制的引入使网络捕获

像素之间的相互依赖关系而聚焦于重要信息，充分提取原始视图的语义信息。此外，建立了神经渲染网络，应用渲染参考视图损失重建 3D 场景几何结构，引入深度一致性损失保持场景几何一致性，充分缓解了多视图立体网络在弱纹理或非朗伯曲面区域中特征发生错误匹配的影响。在 DTU 数据集上的实验结果表明，本文的方法优于现有的方法，在完整性和整体性方面达到了 0.289 和 0.326，同时得到了更高质量的重建效果。另外在 Tanks and Temples 中级数据集上表现出了较强的泛化能力，以及在室外大场景中重建效果的有效性。

## 参考文献


[1] GALLIANI S, LASINGER K, SCHINDLER K. Massively parallel multiview stereopsis by surface normal diffusion[C]. IEEE International Conference on Computer Vision, 2015: 873-881.

[2] SCHÖNBERGER J L, ZHENG E, FRAHM J M, et al. Pixelwise view selection for unstructured multi-view stereo[C]. European Conference on Computer Vision, 2016: 501-518.

[3] YAO Y, LUO Z, LI S, ET AL. MVSNet: Depth inference for unstructured multi-view stereo [C]. European Conference on Computer Vision, 2018: 767-783.

[4] YAO Y, LUO Z, LI S, et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 5525-5534.

[5] YANG J Y, MAO W, LIU M M, et al. Cost volume pyramid based depth inference for multi-view stereo[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 4876-4885.

[6] 叶春凯，万旺根. 基于特征金字塔网络的多视图深度估计[J]. 电子测量技术, 2020, 43(11): 91-95.
YE CH K, WAN W G. Feature pyramid network for multi-view depth estimation [J]. Electronic Measurement Technology, 2020, 43(11): 91-95.

[7] YU A, GUO W, LIU B, et al. Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 175: 448-460.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017, 30.

[9] DING Y, YUAN W, ZHU Q, et al. Transmvsnet: Global context-aware multi-view stereo network with transformers [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2022: 8585-8594.

[10] ZHU J, PENG B, LI W, et al. Modeling long-range dependencies and epipolar geometry for multi-view stereo [C]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023.

[11] 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.

[12] YU A, YE V, TANCIK M, et al. Pixelnerf: Neural radiance fields from one or few images[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2021: 4578-4587.

[13] CHEN A, XU Z, ZHAO F, et al. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo[C]. IEEE International Conference on Computer Vision, 2021: 14124-14133.

[14] TOUVRON H, CORD M, SABLAYROLLES A, et al. Going deeper with image transformers[C]. IEEE International Conference on Computer Vision, 2021: 32-42.

[15] SHAW P, USZKOREIT J, VASWANI A. Self-attention with relative position representations[J]. ArXiv Preprint, 2018, ArXiv: 1803.02155.

[16] CHANG D, BOŽIČ A, ZHANG T, et al. RC-MVSNet: Unsupervised multi-view stereo with neural rendering[C]. European Conference on Computer Vision, 2022: 665-680.

[17] AANÆS H, JENSEN R R, VOGIATZIS G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120: 153-168.

[18] KNAPITSCH A, PARK J, ZHOU Q Y, et al. Tanks and temples: Benchmarking large-scale scene reconstruction[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-13.

[19] YANG J, MAO W, ALVAREZ J M, et al. Cost volume pyramid based depth inference for multi-view stereo[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 4877-4886.

[20] CHENG S, XU Z, ZHU S, et al. Deep stereo using adaptive thin volume representation with uncertainty awareness [C]. IEEE International Conference on Computer Vision, 2020: 2524-2534.

[21] LUO K, GUAN T, JU L, et al. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 10452-10461.

[22] CHEN R, HAN S, XU J, et al. Point-based multi-view stereo network [C]. IEEE International Conference on Computer Vision, 2019: 1538-1547.

[23] YU Z, GAO S. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1949-1958.

[24] YI H, WEI Z, DING M, et al. Pyramid multi-view stereo net with self-adaptive view aggregation [C]. European Conference on Computer Vision, 2020: 766-782.

[25] MA X, GONG Y, WANG Q, et al. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo [C]. IEEE International Conference on Computer Vision, 2021: 5732-5740.

[26] WANG F, GALLIANI S, VOGEL C, et al. Patchmatchnet: Learned multi-view patchmatch stereo[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2021: 14194-14203.

[27] WEI Z, ZHU Q, MIN C, et al. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network [C]. IEEE International Conference on Computer Vision, 2021: 6187-6196.


## 作者简介


**朱代先**，博士，副教授，硕士生导师，主要研究方向为机器人 SLAM 技术、三维重建技术。
E-mail: zhudaixian@xust.edu.cn

**孔浩然**（通信作者），硕士研究生，主要研究方向为机器人 SLAM 技术、三维重建技术。
E-mail: 21207223089@stu.xust.edu.cn