DOI:10.19651/j.cnki.emt.2416367

基于注意力机制的多视图立体重建算法*

朱代先[1] 巩若琳[1] 孔浩然[1] 刘树林[2]
(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:针对多视图立体重建在光照不均匀、弱纹理、非朗伯表面等复杂场景中重建完整度差、泛化能力不足的问题,本文提出了一种基于注意力机制的多视图立体重建算法。在特征提取阶段,该算法采用基于深度可分离卷积和自注意力机制的多尺度特征提取模块,在扩大感受野的同时增强多视图间的空间特征关系,从而提升网络在复杂场景下特征的表征能力以实现更精确的特征匹配。在代价体正则化阶段,本文引入通道注意力机制来自适应调节不同通道的权重,从而减少无关信息对模型的干扰并过滤背景噪声,以提升模型的泛化能力。在 DTU 数据集上,本文算法的完整度和整体度分别为 0.286 和 0.334,与基准算法 CasMVSNet 相比,分别提升了 25.71%和 5.92%,与其他的 state-of-the-art (SOTA)算法相比,在复杂场景中重建点云的结构也更加完整。在 Tanks and Temples 中级数据集上,重建点云综合指标 F-score 为 61.49,这表明本文算法具有更好的鲁棒性和泛化能力。

关键词:三维重建;多视图立体;注意力机制
中图分类号:TP391.4;TN911.73 **文献标识码**:A **国家标准学科分类代码**:520.6030

Multi-view stereo reconstruction algorithm based on attention mechanism

Zhu Daixian[1] Gong Ruolin[1] Kong Haoran[1] Liu Shulin[2]
(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 problems of poor reconstruction completeness and insufficient generalization ability of multi-view stereo reconstruction in complex scenes such as uneven illumination, weak texture, and non-Lambertian surfaces, this paper proposes a multi-view stereo reconstruction algorithm based on the attention mechanism. In the feature extraction stage, the algorithm adopts a multi-scale feature extraction module based on depth-separable convolution and self-attention mechanism, which enhances the spatial feature relationships among multiple views while expanding the sensory field, thus improving the network's ability to characterize features in complex scenes to achieve more accurate feature matching. In the cost volume regularization stage, this paper introduces the channel attention mechanism to adaptively adjust the weights of different channels, so as to reduce the interference of irrelevant information on the model and filter the background noise to improve the generalization ability of the model. On the DTU dataset, the completeness and overall metrics of this paper's algorithm are 0.286 and 0.334, respectively, which are improved by 25.71% and 5.92% compared to the benchmark algorithm CasMVSNet. The structure of the reconstructed point cloud is also more complete in complex scenes compared to other state-of-the-art (SOTA) algorithms. On the Tanks and Temples intermediate dataset, the reconstructed point cloud composite index F-score is 61.49, indicating that the algorithm in this paper has better robustness and generalization ability.

Keywords: 3D reconstruction; multi-view stereo; attention mechanism

0 引 言

随着计算机视觉的快速发展,多视图立体(Multi-view stereo,MVS)重建成为推动科技进步和技术创新的关键力量,在虚拟现实、文化遗产保护、自动驾驶等领域广泛应用。对多视图立体的研究旨在从一系列已知相机固有参数的图像序列中恢复出图像的三维场景。在过去几十年中,传统 MVS 方法[1-2]受到广泛研究,但仍存在一些待解决的问题,

收稿日期:2024-07-04
* 基金项目:陕西省重点研发计划项目(2021GY-338)、西安市碑林区 2023 年应用技术研发储备工程项目(GX2333)资助

如重建场景不完整、泛化能力不足等。

近年来,基于深度学习的MVS重建算法发展迅速,其核心思想是利用卷积神经网络(convolutional neural networks,CNN)处理特征提取和代价体正则化等任务。Yao等[3]首次提出一种端到端的深度估计框架MVSNet。该算法通过2D U-Net网络提取图像特征,并引入方差度量构建三维代价体,再利用3D U-Net网络对三维代价体正则化以预测高精度的深度图,但三维卷积在正则化过程中会消耗大量内存。为此,该团队提出一种基于递归神经网络的多视图立体框架R-MVSNet[4],其引入了门控循环单元(gated recurrent unit, GRU)进行正则化,避免一次性对整个代价体进行调整,内存消耗降低,但增加了重建时间。Gu等[5]提出一种级联结构的CasMVSNet网络,其利用金字塔网络提取多尺度图像特征,并通过迭代前一轮的深度估计来优化本轮的深度采样区间,实现由粗到细的深度估计策略。然而,该方法在重建纹理缺乏或光照变化大的场景完整度较差,难以精确捕捉场景的所有细节和几何结构。Zhang等[6]提出了多粒度特征融合网络MG-MVSNet,通过引入密集特征自适应连接模块来自适应融合全局和局部特征,生成更完整的特征图,进而推断出更详细的深度图,使最终重建点云更完整。叶春凯等[7]提出一种基于特征金字塔网络的多视图深度估计方法,有效采集并融合多尺度图像特征,以应对弱纹理区域的挑战。尽管这些算法利用CNN引入全局语义信息,但在处理弱纹理、非朗伯表面等复杂场景时无法有效地聚合不同尺度的图像特征,这导致网络在构建代价体时存在较大误差,进而影响深度图回归的准确性。

为了应对上述挑战,最新的工作引入注意力机制来更全面地捕获不同尺度下的图像信息。Yu等[8]引入了独立自注意力机制来提取特征以捕获重要信息,并使用相似度度量代替基于方差生成代价体的过程,但在弱纹理区域重建点云准确性不佳。Ding等[9]提出了端到端网络TransMVSNet,其首次将特征匹配Transformer引入到多视图立体任务中,通过自注意力和交叉注意力来聚合图像内与图像间的长距离上下文信息,显著提高了重建准确性,但重建效率明显下降。Zhu等[10]提出一种基于扩展卷积和注意力机制的多尺度特征提取模块,有助于准确地捕获像素间的依赖关系并提取关键特征信息,但其泛化能力仍有待加强。Zhang等[11]提出一个具有注意力感知的3D U-Net网络,该网络采用深度可分离卷积进行代价体正则化,使信息有效聚合并缓解特征不匹配问题,但未充分考虑到通道之间的关联性,导致重建效果不理想。

综上所述,本文提出一种端到端的基于注意力机制的MVS重建网络。针对网络在复杂场景中造成的特征提取困难及颜色匹配误差问题,本文在特征提取阶段采用深度可分离卷积和自注意力机制的并行多尺度特征提取模块。旨在扩大感受野并增强不同视图间的特征交互,减少源视图在单应性变换后的细节损失,使得网络能够更全面地捕获远距离的全局上下文信息。此外,现有的一些算法未能充分利用通道之间的相关性,这导致在重建点云的过程中存在噪声信息多、泛化能力差的问题,为了应对这一挑战,本文在3D U-Net的对称编码跳跃连接中引入通道注意力模块。旨在对同一尺度下的代价体进行通道级对比来获取信息差异,并将其应用于解码结构中,抑制并过滤部分噪声和无用信息。

1 网络结构

本节描述了所提出算法的整体结构,如图1所示,本文在多尺度特征提取阶段引入了注意力块,并在代价体正则化阶段的跳跃连接中引入通道注意力模块。整体结构以CasMVSNet作为骨干网络,按照由粗到细的三阶段级联方式。本文首先采用基于深度可分离卷积和注意力机制的多尺度特征提取模块来提取特征图$\{F_i \in R^{C'\times H'\times W'}\}_{i=1}^{N}$。其次,通过可微单应性变换和方差聚合的思想将2D变换后的特征映射到参考图像的3D视锥体中,再将其融合以构建三维代价体。为了进一步缓解网络因错误匹配构建的有噪声代价体的影响,本文在3D U-Net跳跃连接中引入通道注意力机制对三维代价体进行正则化,以动态调整通道权重,增强模型的泛化能力和准确度。最后,再使用soft-argmax方法将其转换为深度概率体以回归深度图。

1.1 多尺度特征提取模块

本文的多尺度特征提取模块在原有的特征金字塔网络结构的顶层跳跃连接中引入了注意力块。如图1中注意力块部分所示,该注意力块采用3个并行的结构,第1个支路采用1×1的标准卷积,目的是在保持原有感受野的基础上增强特征图的非线性表达能力。同时,为了扩大输入特征的感受野来捕获更广泛的上下文信息,第2个支路采用卷积核大小为3×3,扩展率为2的深度可分离卷积。特别地,在第3个支路引入带有自注意力层的注意力模块,旨在捕获特征之间的依赖关系来提取更丰富的全局特征信息。之后,将并行的3个分支的特征图沿通道维度进行堆叠,再经过1×1标准卷积融合不同尺度的信息。最后通过具有Sigmoid函数的残差网络结构,可以在一定程度上缓解梯度消失问题并加速网络收敛。

带有自注意力层的注意力模块架构如图2所示,将特征输入到一个卷积核为3×3的卷积层,并经过一个Leaky ReLU激活函数,其作为ReLU的变体,通过引入一个小的斜率来解决ReLU中出现的“神经元死亡”问题,即负数区域的输出不再为零,而是取一个非零值[12]。该激活函数增加网络的非线性能力,有助于网络学习复杂的特征表示,更加稳定地训练神经网络,从而避免梯度消失的问题。其中带有LayerScale[13]的自注意力层,具体结构如图3所示。其采用的注意力机制将查询向量和一组键值对映射到输出。其中,值向量的权重由查询向量与键向量之间的兼容

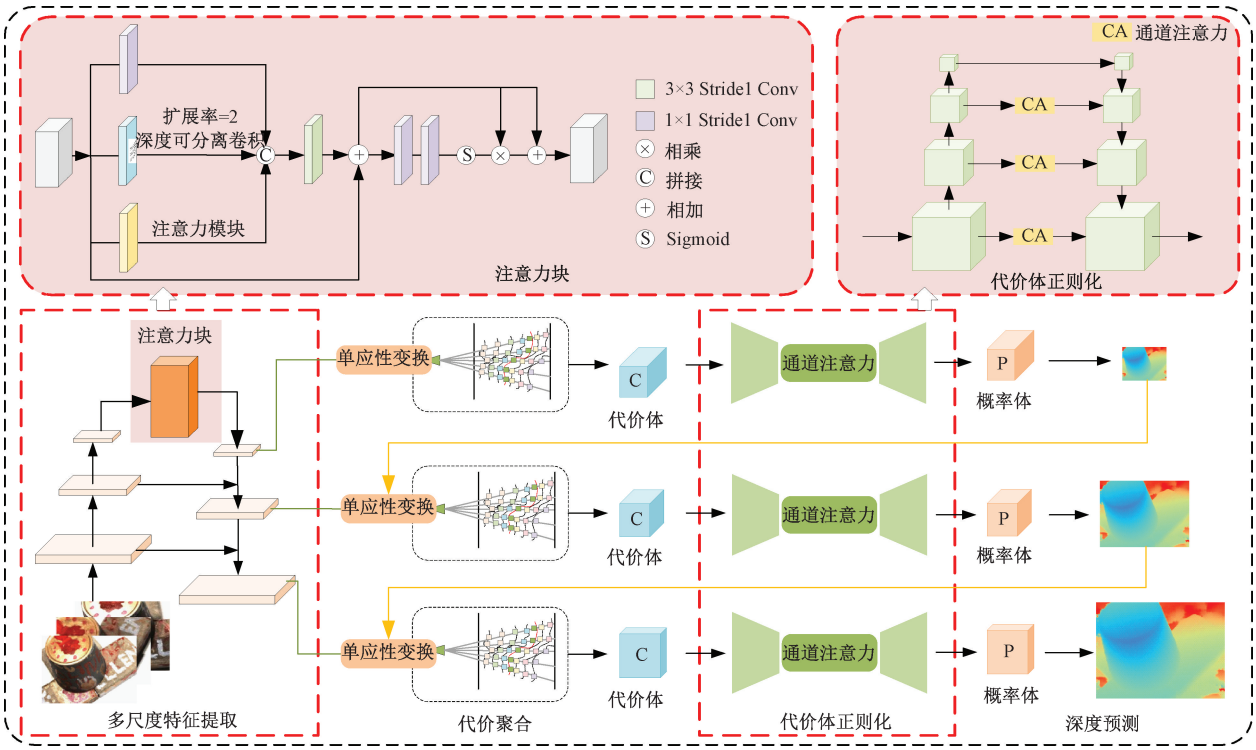

图 1　网络整体结构图

Fig. 1　Overall architecture of the proposed network

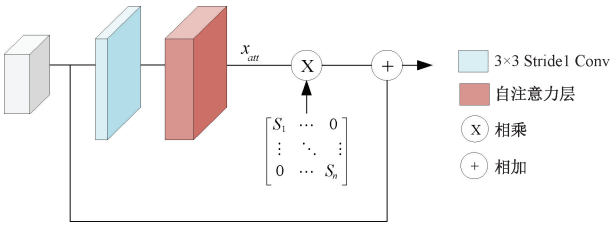

图 2　注意力模块的结构

Fig. 2　The architecture of the attention module

性函数计算，衡量查询和键之间的相似度，从而决定每个值向量在最终输出的重要程度。这类机制可以用于动态地调整模型对输入信息的关注度，使模型能够更好地适应不同的任务和输入数据，提高其适应性和鲁棒性。

为解决注意力机制中位置信息未被编码而导致的排列等价性问题，本文引入可学习的相对位置嵌入到键向量中，以捕获输入数据的空间结构和位置关系。从定义位置 ij 到每个位置 $ab \in R$ 的相对距离开始，将相对距离向量

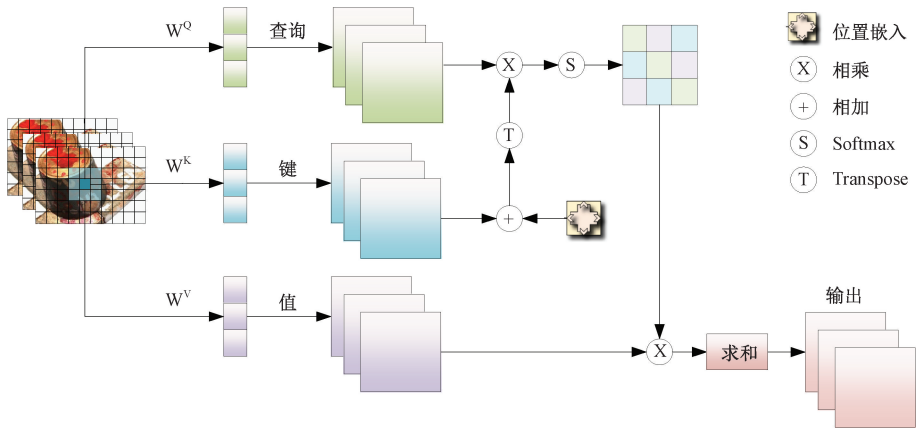

图 3　自注意力层的结构

Fig. 3　The architecture of the self-attention layer

$r_{a-i,b-i}$ 分解为行偏移量 $a-i$ 与列偏移量 $b-i$，与嵌入 r_{a-i} 和 r_{b-i} 相关联。具体实现如下：

$$\boldsymbol{y}_{ij}=\sum_{a,b\in R}Softmax(\boldsymbol{q}_{ij}^{\mathrm{T}}(\boldsymbol{k}_{ab}+\boldsymbol{r}_{a-i,b-i}))\boldsymbol{v}_{ab} \tag{1}$$

其中,$\boldsymbol{y}_{ij}$ 为输出向量,$\boldsymbol{q}_{ij}=\boldsymbol{W}_q\boldsymbol{x}_{ij}$、$\boldsymbol{k}_{ab}=\boldsymbol{W}_k\boldsymbol{x}_{ab}$、$\boldsymbol{v}_{ab}=\boldsymbol{W}_v\boldsymbol{x}_{ab}$ 分别表示查询、键和值,矩阵 $\boldsymbol{W}_p(p=q,k,v)$ 是一个学习的权重矩阵,由学习参数组成。R 表示输入大小为 3×3 的局部区域。

1.2 代价体构建

对于每个参考图像的匹配代价体构建,本文沿用 MVSNet[3] 中的平面扫描算法,将每张图像 $\{I_i\}_{i=1}^N$ 通过多尺度特征提取模块得到对应的特征图 $\{F_i\}_{i=1}^N$,以参考图像的主光轴为扫描方向,按照相同的深度间隔构建一个锥体。之后,将每个源图像通过可微单应性变换映射到参考图像视锥体下的每个深度层中,形成 N 个特征体 $\{V_i\}_{i=1}^N$。具体实现如下:

$$\boldsymbol{H}_i(d)=\boldsymbol{K}_i\cdot\boldsymbol{R}_i\cdot\left(\boldsymbol{E}-\frac{(t_1-t_i)\cdot\boldsymbol{n}_{ref}^{\mathrm{T}}}{d}\right)\cdot\boldsymbol{R}_1^{\mathrm{T}}\cdot\boldsymbol{K}_1^{\mathrm{T}} \tag{2}$$

其中,$\boldsymbol{H}_i(d)$ 为第 i 个源特征图变换到参考图像视锥体下深度值为 d 处的单应性变换矩阵;$\boldsymbol{K}$ 为相机的内部参数;$\boldsymbol{R}$,$\boldsymbol{t}$ 分别为旋转矩阵和平移矩阵;$\boldsymbol{E}$ 为单位矩阵;$\boldsymbol{n}_{ref}$ 为参考图像的平面法向量。

在此基础上,本文利用方差度量 M 将 N 个特征体 $\{V_i\}_{i=1}^N$ 构建成一个统一的代价体 C:

$$C=M(V_1,\cdots,V_N)=\frac{\sum_{i=1}^{N}(V_i-\bar{V})^2}{N} \tag{3}$$

其中,N 是输入图像的个数,V_i 是源图像的特征体,$\bar{V}$ 是所有特征体的均值。

1.3 代价体正则化

初步构建的代价体可能会因非朗伯表面而受到噪声的干扰,因此需结合平滑度约束推断深度图,将代价体经过一个类似 3D U-Net 的正则化网络,以生成更为精确的概率体进行深度图推断。本文考虑到通道之间的关联性,在每两个相同尺度的代价体之间引入挤压聚合激励(Squeeze Aggregated Excitation,SaE)模块[14],以学习通道之间的相关性。SaE 模块将多分支密集层和挤压激励模块相结合,使网络能够根据每个通道的重要性实现自适应调整,并同时过滤部分噪声和无用的信息,有助于网络更有效地捕捉目标区域的关键信息。特别地,在跳跃连接后的上采样过程中,SaE 模块能够使网络更关注全局信息,从而提高特征的辨别能力。

SaE 的内部结构如图 4 所示。输入特征图 X 的每一层都经过卷积算子 F_{tr} 得到 C 个输出的特征图组成 U。之后经过全局平均池化 F_{sq} 对输入模型沿空间维度压缩得到全局特征描述。每个通道的具体实现如下:

$$z_c=F_{sq}(u_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_c(i,j) \tag{4}$$

其中,z_c 表示通道级统计数据 z 的第 c 个元素,H、W 表示输入特征在空间维度的高和宽,$u_c(i,j)$ 表示对应点的像素值。

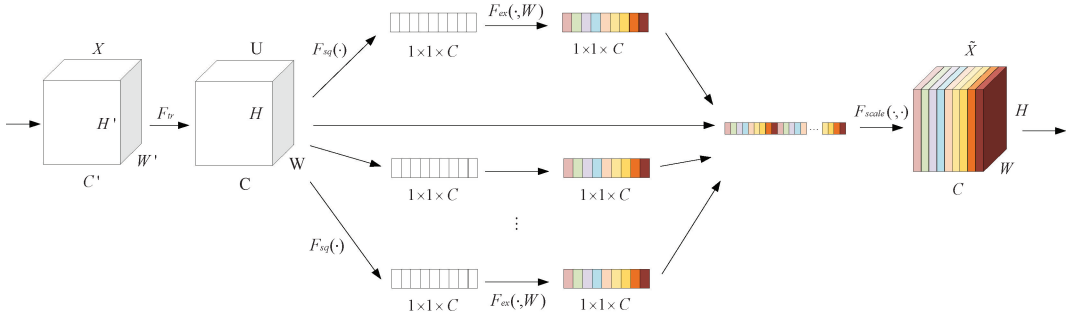

图 4 SaE 模块

Fig. 4 SaE module

经全局平均池化层后,增加了全连接层之间的基数,用来优化并增强网络自适应地关注重要部分的过程。之后,对全连接层进行激励 F_{ex},将中间表示映射回原始通道数,再应用 Sigmod 函数将注意力权重归一化,用来表示每个通道的重要性。具体实现如下:

$$s=F_{ex}(z,W)=\sigma(g(z,W))=\sigma(W_2\delta(W_1\sum_{i=1}^{n}z)) \tag{5}$$

其中,s 表示权重,$W_1\in\mathbb{R}^{\frac{C}{r}\times C}$、$W_2\in\mathbb{R}^{C\times\frac{C}{r}}$,$r$ 是降维超参数,为了减少参数量和计算量;n 表示全连接层的基数,δ 表示 ReLU 函数,σ 表示 Sigmod 函数。

最后,还需进行缩放 F_{scale},使特征图 U 中的每个特征乘以对应权重,从而得到最终的输出 $\tilde{X}_C$。具体实现如下:

$$\tilde{X}_C=F_{scale}(u_c,s_c)=s_cu_c \tag{6}$$

其中,特征映射 $u_c\in\mathbb{R}^{H\times W}$,$s_c$ 表示对应的标量。

1.4 损失函数

本文采用 CasMVSNet[5] 的三层级联结构。为了全面评估计算损失并获得更准确的深度图,本文考虑了每个阶段的深度图损失并采用 L1 损失函数来计算地面真实深度值和深度估计值之间有效区域的绝对距离,具体实现

如下：

$$Loss = \sum_{k=1}^{3}\sum_{p=\Omega}\lambda^k \| D_{GT}^k(p) - D^k(p) \|_1 \quad (7)$$

其中，Ω 表示地面真实有效像素点的集合，k 表示第 k 阶段，$D_{GT}^k(p)$ 表示像素 p 在第 k 阶段对应的地面真实深度值，$D^k(p)$ 表示像素 p 在第 k 阶段对应的深度估计值，λ^k 表示 3 个不同阶段对应的损失权重。

2　实验分析

2.1　数据集

本文采用公开的 DTU 数据集[15] 和 Tanks and Temples 数据集[16]对模型进行训练和评估。DTU 数据集包含 124 个不同的场景，涵盖 49 个不同的视角和 7 种光照条件。Tanks and Temples 数据集是真实的大型室外场景，包含 6 组中级数据集和 8 组高级数据集。

2.2　实验细节

在 DTU 数据集中进行训练得到预训练模型，并将初始的输入图像分辨率固定为 512×640，输入图像的数量设置为 $N=3$。本文将网络分为 3 个阶段，每个阶段的特征图分辨率分别采用原始图像分辨率的 1/16、1/4 和 1 倍来实现，深度假设数量设置为 48、32、8，对应的深度间隔设置为 4、2、1，深度间隔系数设置为 1.06。网络通过 PyTorch 实现，在训练阶段使用 $\beta_1=0.9$ 和 $\beta_2=0.999$ 的 Adam 优化器。总共训练 16 个阶段，初始学习率调整为 0.001，在训练第 10、12、14 个阶段后将学习率降低 2 倍。实验使用单张 NVIDIAGTX 3090 GPU 来训练网络，batch size 大小设为 2。

在测试阶段，对训练得到的预训练模型使用 DTU 测试数据集进行测试。设置输入图像的数量为 $N=5$。输入图像的分辨率设置为 864×1 152，深度假设数量、深度间隔及深度间隔系数与训练阶段保持一致。

2.3　实验结果

1）DTU 数据集

为了验证本文算法的有效性，将本文算法与传统算法 Gipuma[1]、Colmap[2]，以及最近基于深度学习的方法 MVSNet[3]、R-MVSNet[4]、CasMVSNet[5]、MG-MVSNet[6]、DSC-MVSNet[11]、P-MVSNet[17]、PVA-MVSNet[18]、CVP-MVSNet[19]、UCS-Net[20]、AA-RMVSNet[21]进行定量结果对比。实验使用 DTU 数据集官方提供的 MATLAB 评估代码来计算重建点云的准确度（Acc.）、完整度（Comp.）和整体度（Overall），其中整体度是准确度和完整度的均值，该指标综合反映重建效果的优劣。具体评估结果如表 1 所示，其中引用的其他研究结果均取自之前已发布的研究。本文算法在完整度和整体度方面的表现优于目前大多数现有的算法。与基准网络 CasMVSNet[5]相比，点云重建的完整度和整体度方面分别提升了 25.71% 与 5.92%，与最近 SOTA 方法 MG-MVSNet[6]相比，点云重建的完整度提升了 15.38%，整体度提升了 4%。此外，本文选择基准网络 CasMVSNet 与最近的 SOTA 网络 CVP-MVSNet 作为对比模型，并使用地面真实值作为基准，与本文算法在 DTU 测试数据集重建点云结果进行对比，得到如图 5 所示可视化结果。在 Scan11、Scan15 及 Scan29 这 3 个场景中 CasMVSNet 与 CVP-MVSNet 的重建点云都呈现出不同程度的细节结构不清晰与点云重建不完整现象，尤其是在弱纹理、非朗伯表面等复杂场景中。具体来说，如 Scan11 场景中的侧壁及标语，Scan15 场景中的门及树后的墙壁，Scan29 场景中的高楼，见图 5 矩形框中突出显示的细节部分所示。与 CasMVSNet、CVP-MVSNet 相比，本文模型在 Scan11 场景和 Scan15 场景中点云的噪声信息更少，字母显示更加清晰；在 Scan29 场景中成功补全了真实点云未能涵盖的部分，这表明在复杂场景中本文模型生成了更完整的、更精细的点云，有效保留了场景的纹理细节。

表 1　DTU 数据集定量结果

Table 1　Quantitative results for the DTU dataset are presented

方法	Acc./mm	Comp./mm	Overall/mm
Gipuma	**0.283**	0.873	0.578
Colmap	0.400	0.644	0.532
MVSNet	0.396	0.527	0.462
R-MVSNet	0.385	0.459	0.422
CasMVSNet	0.325	0.385	0.355
AA-RMVSNet	0.376	0.399	0.357
P-MVSNet	0.406	0.434	0.420
PVA-MVSNet	0.379	0.336	0.357
CVP-MVSNet	0.296	0.406	0.351
UCS-Net	0.338	0.349	0.344
MG-MVSNet	0.358	0.338	0.348
DSC-MVSNet	0.316	0.372	0.344
本文方法	0.382	**0.286**	**0.334**

注：加粗数值表示该列最优值

2）Tanks and Temples 数据集

本文使用在 DTU 数据集上训练，且经过 BlendedMVS 数据集[22]微调的权重文件对 Tanks and Temples 数据集进行测试，以评估本文算法的泛化能力。该数据集的评估是将重建点云提交到官方评估网站来获得综合性指标 F-score，该值愈高表示立体重建效果愈理想。定量结果如表 2 所示，本文的平均 F-score 指标高于其他最新的 SOTA 方法，如 CVP-MVSNet、DSC-MVSNet、MG-MVSNet，相较于基准网络 CasMVSNet，平均 F-score 提升了 8.18%。实现了与最先进的方法相当的泛化性能，在 Family、Francis、Horse、M60 和 Train 场景都达到了较高

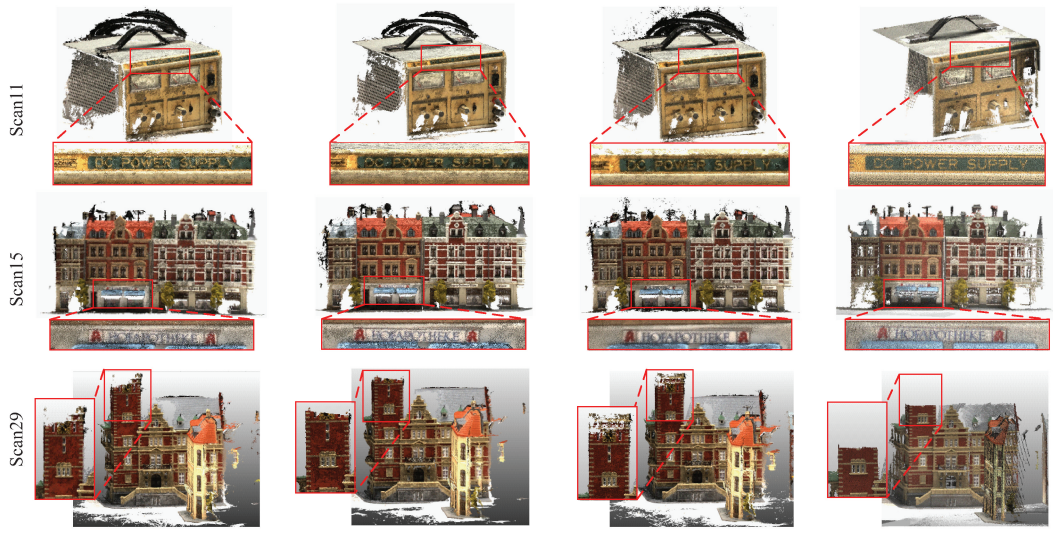

图 5 部分模型在 DTU 数据集重建结果对比

Fig. 5 Comparison of reconstruction results of some models on the DTU dataset

的精度。图 6 与图 7 展示了 Tanks and Temples 中级数据集和部分高级数据集的重建点云结果,矩形框突出显示了 Tanks and Temples 数据集的局部细节信息,如图 6(f) Train 场景完整地重建了火车的整体结构,且点云纹理细节更加清晰,在更复杂的图 7(a)Courtroom 场景清晰地重建出法院顶部的天窗细节。图 8 所示列举了 Lighthouse、Horse 和 Train 场景下根据相应的地面真实点云计算出的误差可视化,其中颜色较深的区域表示误差更高。如图 8 中矩形框所示区域,CVP-MVSNet 在 3 个场景中均有更多的错误点和噪声,相较于 CasMVSNet,本文方法显著提升整体的重建质量,特别是在弱纹理区域,能够在降低噪声的同时获得更准确的位置信息。

表 2 Tanks and Temples 中级数据集定量结果

Table 2 Quantitative results for the Tanks and Temples intermediate dataset are presented

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
R-MVSNet	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
P-MVSNet	55.62	70.04	44.64	40.22	**65.20**	55.08	55.17	**60.37**	54.29
MG-MVSNet	57.58	73.82	52.46	45.53	57.57	61.00	59.77	58.18	52.13
PVA-MVSNet	54.46	69.36	46.80	46.01	55.74	57.23	54.75	56.70	49.06
CVP-MVSNet	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54
DSC-MVSNet	53.48	68.06	47.43	41.60	54.96	56.73	53.86	53.46	51.71
UCS-Net	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89
CasMVSNet	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51
本文方法	**61.49**	**78.54**	**63.91**	**50.64**	60.45	**62.36**	59.51	58.30	**58.19**

注:加粗数值表示该列最优值

2.4 消融实验

为了进一步验证本文所提出的多尺度特征提取模块与通道注意力模块的有效性,本文进行了相关消融实验以及定量分析,具体结果如表 3 所示。该实验在 DTU 测试集上进行了四组对比实验,实验参数的设置与 2.2 节保持一致。其中 Baseline 表示 CasMVSNet 网络,FM 表示多尺度特征提取模块,SaE 表示通道注意力模块。

从表 3 的定量结果对比可知:多尺度特征提取模块和通道注意力模块都在不同程度上提高了重建点云的完整度和整体度。当这两个模块与基准网络 CasMVSNet 结合时,重建点云的完整度和整体度评估结果最佳。如图 9 所示,本文对以上模块的重建点云结果进行了可视化,结果表明在特征提取阶段利用深度可分离卷积和自注意力机制的多尺度图像特征提取方法,有效提取了复杂场景的特

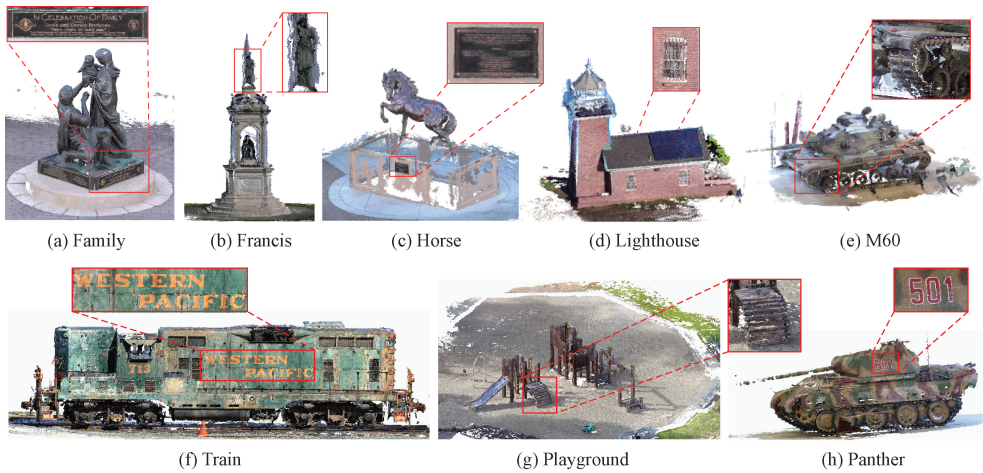

图 6　Tanks and Temples 中级数据集重建点云结果

Fig. 6　Point cloud reconstruction results on the Tanks and Temples intermediate dataset

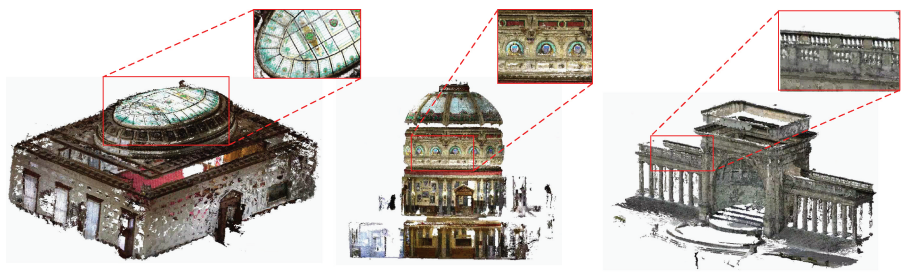

图 7　Tanks and Temples 高级数据集重建点云结果

Fig. 7　Point cloud reconstruction results on the Tanks and Temples advanced dataset

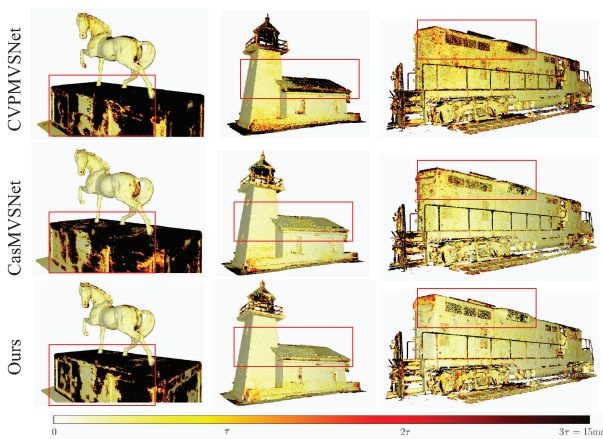

图 8　Tanks and Temples 中级数据集点云误差可视化比较

Fig. 8　Comparison of point cloud errors visualization on the Tanks and Temples intermediate dataset

征，避免了 CasMVSNet 因特征提取不足导致的特征匹配错误问题。此外，在代价体正则化阶段利用通道注意力模块使网络自适应关注重要通道信息，以捕获图像的目标区域，缓解了最近方法因未能充分利用通道间的相关性而导致重建点云噪声信息多的问题。这两种方式都进一步改善了光照不均匀、弱纹理和非朗伯表面等复杂场景的重建点云结果。

表 3　消融实验定量结果对比/mm

Table 3　Comparative quantitative results of ablation experiments

图 9	方法	Acc.	Comp.	Overall
(a)	Baseline	**0.325**	0.385	0.355
(b)	Baseline+FM	0.402	0.288	0.345
(c)	Baseline+SaE	0.393	0.299	0.346
(d)	Baseline+FM+SaE	0.382	**0.286**	**0.334**

注：加粗数值表示该列最优值

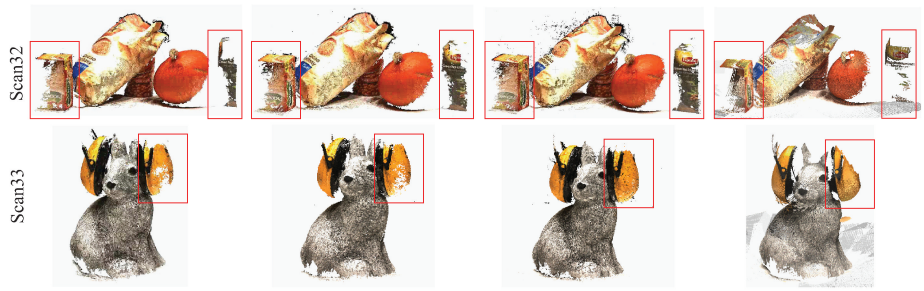

图 9 DTU 数据集的消融实验定性结果对比

Fig. 9 Qualitative results comparison of ablation experiments on the DTU dataset

3 结 论

本文提出一种基于注意力机制的多视图立体重建算法。本文将自注意力机制与深度可分离卷积相结合构成并行的多尺度特征提取模块,以充分提取图像深层特征。该模块旨在通过扩大感受野并建立特征间的空间相关性,以捕获像素间的依赖关系,从而有效缓解光照不均匀、弱纹理、非朗伯表面等复杂场景造成的特征提取困难、颜色匹配误差等问题。为了进一步加强通道信息的聚合能力和过滤噪声的能力,本文在代价体正则化网络的跳跃连接中引入通道注意力模块来自适应调节不同通道的权重。在 DTU 数据集上实验结果表明,本文算法优于现有的 SOTA 算法,其完整度和整体度分别为 0.286 和 0.334,同时生成更完整的稠密点云和更清晰的细节部分。另外在 Tanks and Temples 数据集上的实验结果表明,本文算法具有良好的泛化能力。

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 EN L, FRAHM J M, et al. Pixelwise view selection for unstructured multi-view stereo[C]. Computer Vision-ECCV 2016: 14th European Conference, 2016: 501-518.

[3] YAO Y, LUO Z X, LI SH W, et al. MVSNet: Depth inference for unstructured multi-view stereo [C]. European Conference on Computer Vision (ECCV), 2018: 767-783.

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

[5] GU X D, FAN ZH W, ZHU S Y, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2495-2504.

[6] ZHANG X D, YANG F ZH, CHANG M, et al. MG-MVSNet: Multiple granularities feature fusion network for multi-view stereo[J]. Neurocomputing, 2023, 528: 35-47.

[7] 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. Multi-view depth estimation based on feature pyramid networks [J]. Electronic Measurement Technology, 2020, 43(11): 91-95.

[8] YU AN ZH, GUO W Y, 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.

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

[10] ZHU D X, KONG H R, QIU Q, et al. Multi-view stereo network based on attention mechanism and neural volume rendering [J]. Electronics, 2023, 12(22): 4603.

[11] ZHANG S, WEI ZH W, XU W J, et al. DSC-MVSNet: Attention aware cost volume regularization based on depthwise separable convolution for multi-view stereo [J]. Complex & Intelligent Systems, 2023, 9(6): 6953-6969.

[12] 陈暄, 吴吉义. 基于优化卷积神经网络的车辆特征识别算法研究[J]. 电信科学, 2023, 39(10): 101-111.
CHEN X, WU J Y. Research on vehicle feature recognition algorithm based on optimized convolutional neural network [J]. Telecommunications Science, 2023, 39(10): 101-111.

[13] TOUVRON H, CORD M, SABLAYROLLES A, et al.

Going deeper with image transformers[C]. IEEE/CVF International Conference on Computer Vision, 2021: 32-42.

[14] NARAYANAN M. SENetV2: Aggregated dense layer for channelwise and global representations[J]. ArXiv preprint arXiv:2311.10807, 2023.

[15] 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.

[16] 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.

[17] LUO K Y, GUAN T, JU L L, et al. P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo[C]. IEEE/CVF International Conference on Computer Vision, 2019: 10452-10461.

[18] YI H W, WEI Z ZH, DING M Y, et al. Pyramid multi-view stereo net with self-adaptive view aggregation[C]. Computer Vision-ECCV 2020: 16th European Conference, 2020: 766-782.

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

[20] CHENG SH, XU Z X, ZHU SH L, et al. Deep stereo using adaptive thin volume representation with uncertainty awareness[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2524-2534.

[21] WEI Z ZH, ZHU Q T, MIN CH, et al. AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network [C]. IEEE/CVF International Conference on Computer Vision, 2021: 6187-6196.

[22] YAO Y, LUO Z X, LI SH W, et al. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1790-1799.

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，三维重建技术等。

E-mail：zhudaixian@xust.edu.cn

巩若琳（通信作者），硕士研究生，主要研究方向为三维重建技术。

E-mail：22207223096@stu.xust.edu.cn

孔浩然，硕士研究生，主要研究方向为三维重建技术。

E-mail：21207223089@stu.xust.edu.cn

刘树林，博士，教授，主要研究方向为计算机视觉。

E-mail：lsigma@163.com