计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):259−267 [doi: 10.15888/j.cnki.csa.009824] [CSTR: 32024.14.csa.009824]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合上下文增强与图像频率引导的 MVS 方法①

陈　曦[1,2], 刘　美[1], 陈嘉升[1]

[1](广东石油化工学院 自动化学院, 茂名 525000)
[2](广东工业大学 自动化学院, 广州 510006)
通信作者: 刘　美, E-mail: liumei_gdpt.edu.cn@hotmail.com

摘　要: 基于学习的多视图立体匹配算法目前成果显著, 但是仍然存在的卷积感受野受限以及忽略图像频率信息导致在低纹理、重复和非兰伯曲面匹配性能不足的问题, 针对以上问题提出了上下文增强与图像频率引导的多视图立体匹配网络 CAF-MVSNet. 首先, 在特征提取阶段, 将上下文增强模块融合到特征金字塔网络中, 有效地扩大网络的感受野. 然后引入了图像频率引导注意力模块, 通过编码图像的不同频率获取图像的线条、形状、纹理和颜色等信息, 增强图像的远程上下文联系的同时进一步解决低纹理、重复和非兰伯曲面的精确匹配问题, 以实现可靠的特征匹配. 在 DTU 数据集上的实验结果显示, 与经典的级联模型 CasMVSNet 相比综合误差 (overall) 提升了 12.3%, 展现了优秀的性能. 此外, 在 Tanks and Temples 数据集上也取得了不错的效果, 展现了良好的泛化性能.

关键词: 多视图立体匹配; 三维重建; 上下文增强; 图像频率引导; 深度学习

引用格式: 陈曦,刘美,陈嘉升.融合上下文增强与图像频率引导的 MVS 方法.计算机系统应用,2025,34(3):259–267. http://www.c-s-a.org.cn/1003-3254/9824.html

MVS Method Combining Context-enhanced and Image-frequency-guide

CHEN Xi[1,2], LIU Mei[1], CHEN Jia-Sheng[1]

[1](School of Automation, Guangdong University of Petrochemical Technology, Maoming 525000, China)
[2](School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Learning-based multi-view stereo matching algorithms have achieved remarkable results, but still have the problems of limited convolutional receptive field and ignoration of image frequency information, which lead to insufficient matching performance on low-texture, repetitive, and non-Lambertian surfaces. To address these problems, this study proposes CAF-MVSNet, a context-enhanced and image-frequency-guided multi-view stereo matching network. First, the context enhancement module is fused into the feature pyramid network in the feature extraction stage to effectively expand the receptive field of the network. Then the image-frequency-guided attention module is introduced to obtain the information of lines, shapes, textures, and colors of the images by encoding different frequencies of the images, which enhances the remote contextual connection of the images and further solves the problem of accurate matching of low-texture, repetitive, and non-Lambertian surfaces for reliable feature matching. Experimental results on the DTU dataset show that CAF-MVSNet has a 12.3% improvement in the combined error compared to the classical cascade model CasMVSNet, demonstrating excellent performance. In addition, good results are achieved on the Tanks and Temples dataset, reflecting the good generalization performance of CAF-MVSNet.

Key words: multi-view stereo (MVS); 3D reconstruction; context enhancement; image-frequency-guide; deep learning

① 基金项目: 国家自然科学基金 (62073091)
收稿时间: 2024-09-10; 修改时间: 2024-09-30; 采用时间: 2024-11-18; csa 在线出版时间: 2025-01-21
CNKI 网络首发时间: 2025-01-22

多视图立体匹配 (multi view stereo, MVS) 就是从一系列图像中重建场景的三维结构[1]. 由于 MVS 在自动驾驶[2]、机器人技术[3]和虚拟现实[4]中的广泛应用, 近年越来越受到关注. 尽管传统方法如 OpenMVS[5]、COLMAP[6]、Gipuma[7]等都取得了良好的结果, 但它们仍然存在一些局限性, 例如由于遮挡、光照变化、无纹理区域和非兰伯特表面等问题导致重建完整度不高.

为了克服上述挑战, 基于学习的多视图立体匹配方法得到了提出. Yao 等首先提出了一种基于深度学习的端到端的多视图立体匹配方法 MVSNet[8], MVSNet 使用卷积神经网络 (CNN) 提取特征, 有效地获取图像高级特征. 在代价体正则化阶段, MVSNet 引入了基于方差的多视图聚合方法, 以适应任意视图的输入. 此外, MVSNet 还采用 3DCNN 结构对代价体进行过滤, 然而该结构占用内存较高, 实际应用中成本较大. 为解决内存占用较高的问题, Yao 等继续提出了 R-MVSNet[9], 该网络使用循环神经网络中的门控单元 GRU 来替代 3DCNN, 虽然效率有所降低, 但是有效降低了内存占用. Gu 等提出 Cascade-MVSNet[10], 在下采样低分辨率图像上采用大的深度间隔和较少的深度区间进行深度预测, 然后将预测结果应用于未下采样的高分辨率图像上, 显著降低了模型的内存需求. Ding 等提出了 TransMVSNet[11], 将 Transformer[12]引入到 MVSNet 网络中, 有效聚合了图像的全局上下文信息, 提高了重建精度和完整度. Peng 等提出了 UniMVSNet[13], 在深度估计阶段结合回归和分类的优点, 既能像分类方法一样直接约束成本量, 也可以像回归方法一样实现亚像素深度预测, 使深度估计更有效. Li 等提出了 NR-MVSNet[14], 通过正态一致性深度假设模块和深度细化的可靠注意力机制, 提高了深度估计的准确性. Zhang 等提出 RA-MVSNet[15], 通过点对面距离使每个假设平面与更宽的表面相关联, 在无纹理区域和物体边界推断周围表面信息, 使得 RA-MVSNet 获得了不俗的性能. Ye 等提出 DMVSNet[16], 引入了由鞍形 cell 组成的理想深度几何形状, 其预测深度图围绕真实表面向上和向下振荡, DMVSNet 同样取得不错的性能. Vats 等提出了 GC-MVSNet[17], 首次在训练期间使用跨多个源视图的几何一致性检查为模型提供明确的多视图几何线索, 显著提高了准确性, 同时显著降低了训练迭代要求.

尽管基于学习的 MVS 方法相比传统方法在低纹理、重复和非兰伯曲面等问题上展现了更好的性能, 但这些问题仍然是 MVS 的重要挑战. 根据以上方法, 我们发现目前的基于学习的 MVS 方法存在两个主要问题: (1) 卷积在获取局部特征上拥有优秀的性能, 同时卷积在获取特征时有限的感受野也影响了上下文信息的感知, 这使得在 MVS 任务中, 对低纹理、重复图案和非兰伯曲面等具有挑战性的区域难以进行鲁棒的深度估计. (2) 尽管已有基于学习的 MVS 方法通过引入 Transformer, 增强图像的远程上下文联系, 使得在低纹理、重复和非兰伯曲面等存在的精确匹配问题得到一定的解决, 但是其忽略了自然图像存在的丰富频率, 通过编码不同的频率对获取图像不同信息 (如线条、形状、纹理和颜色等) 有着非常重要的作用, 这些信息对 MVS 任务中低纹理、重复和非兰伯曲面的精确匹配同样有着不可替代的作用.

为了解决以上问题, 我们设计了一个端到端的 MVS 网络 CAF-MVSNet, 该网络利用基于空洞卷积[18]的上下文增强模块 (context enhancement module, CEM) 来增强图像的上下文信息, 同时利用基于 Transformer 的图像频率引导注意力模块 (image frequency-guide attention module, IFAM), 通过编码图像的不同频率获取图像的线条、形状、纹理和颜色等信息, 增强图像的远程上下文联系的同时使得低纹理、重复和非兰伯曲面的精确匹配问题得到进一步的解决. CAF-MVSNet 在 DTU 数据集上实现了较好的结果, 精确度和完整度都达到了较高的水平.

1 CAF-MVSNet 网络原理

1.1 CAF-MVSNet 网络结构

CAF-MVSNet 采用与 CasMVSNet 相同的从粗到细的级联结构, 级联结构如图 1 由上中下 3 个阶段构成, 每个阶段的深度假设都是从深度范围中均匀采样的. 上层的阶段以低分辨率获取图像特征, 并构建具有预定深度范围但具有较大深度间隔的成本体, 中下层阶段则以更高的分辨率、更窄的深度范围和更小的深度间隔构建成本体.

CAF-MVSNet 网络从左到右首先由特征金字塔网络 (feature pyramid network, FPN)[19]提取 N 张图像在不同的 3 个分辨率下的多尺度特征, 同时, 为了让中下层阶段的特征拥有与之匹配得感受野, 通过上下文增强模块扩大中下层阶段的感受野; 然后通过图像频率引导注意力模块进一步的解决低纹理、重复、镜面和

反射区域的精确匹配问题, 以捕获图像的局部细节和全局结构等信息; 再利用可微的单应性变换 (differentiable warping) 将源视图的特征扭曲到参考相机坐标系, 并根据特征图的相似性来生成代价体 (cost volume), 之后通过 3DCNN 对代价体进行正则化生成概率体 (probability volume) 用于深度图的预测; CAF-MVSNet 网络使用焦点损失函数进行端到端的训练. 焦点损失函数将深度估计视为一种分类任务, 基于预测准确度对样本权重调整, 使模型更关注分类错误和分类困难的样本, 这样保证了模型预测深度的稳定性和准确性. 最后对所获得的图像深度图进行融合操作, 得到该场景的点云.

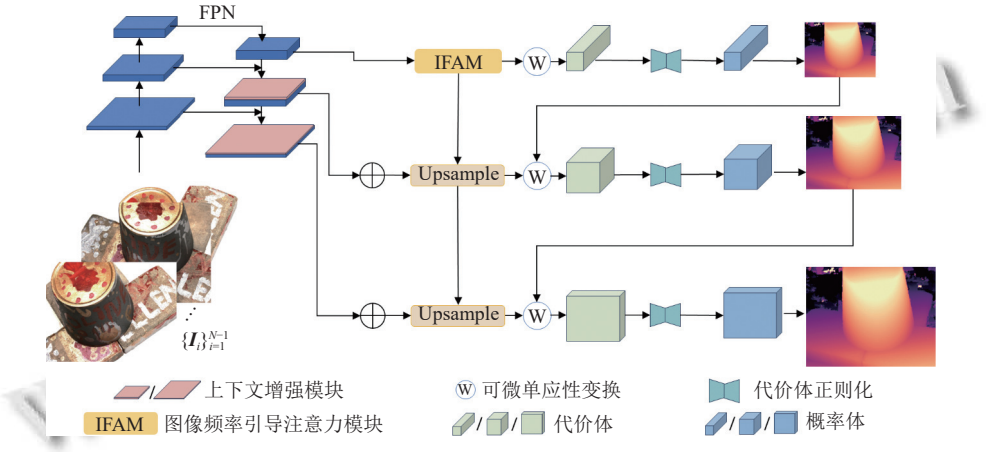

图 1 CAF-MVSNet 网络结构

1.2 特征提取网络

基于深度学习的 MVS 方法中, 通常采用特征金字塔网络来对图像进行多尺度的特征提取. 特征金字塔网络通过引入自上而下的方式, 从输入图像中提取出 3 个不同尺度的特征图, 并通过自上向下的方式将高层特征图与低层特征图进行融合, 以获得更加丰富和具有多尺度信息的特征表示. 通过特征金字塔网络中获得 3 个特征图的分辨率分别是 $H/4\times W/4$、$H/2\times W/2$、$H\times W$, 然而特征金字塔网络每个尺度的感受野并不能完全适应这 3 个尺度下的分辨率, 更高分辨率的图像需要更大的感受野才能获得有效的语义[20], 这使得特征金字塔网络性能受到限制.

因此, 我们设计了一个上下文增强模块来增强高分辨率下不同感受野的上下文信息. 具体而言, 如图 1 所示, 对于分辨率更高的中下两层, 在获取其特征映射之后, 为了获取丰富的上下文信息, 我们将其输入到上下文增强模块中. 上下文增强模块如图 2 所示, 其包括了 4 个不同膨胀率的空洞卷积层[18], 膨胀率分别设为 $rate$、$2\times rate$、$3\times rate$、$4\times rate$, 具体的 $rate$ 数值可以根据分辨率自行设置, 这些分离的卷积层可以在不同的感受野中获取多个特征映射. 此外, 为了精细地合并多尺度信息, 我们将每个空洞卷积的输出拼接在一起, 并通过一个 1×1 的卷积层以融合不同感受野下的上下文信息. 最后, 为了保持初始输入的粗粒度信息, 我们将空洞卷积的输出与输入相加进行粗粒度和细粒度特征的融合.

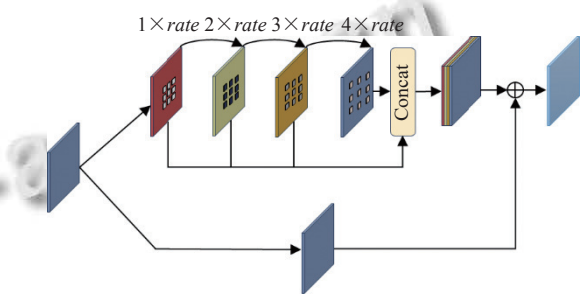

图 2 上下文增强模块

1.3 图像频率引导注意力模块

自然图像包含丰富的频率, 高频捕获对象的局部细节 (例如线条和形状), 而低频编码全局结构 (例如纹理和颜色)[21,22], 而基于学习的 MVS 方法少有能考虑到对图像高低频信息的利用, 导致对于提取到的特征图, 其丰富的上下文信息没有得到充分地利用, 最终使得预测质量不够理想, 特别是对于低纹理区域和重复区域影响更加明显. 考虑到 Transformer 的能够有效地捕获图像局部细节和全局结构, 我们设计了一个基于 Transformer 的图像频率引导注意力模块, 用于捕获图

像中的高低频信息.

为了聚合图像不同频率的信息, 图像频率引导注意力模块由高频路径和低频路径组成, 如图 3 所示. 在聚合图像频率信息之前, 对提取到的特征进行位置编码, 以补充窗口的位置信息, 提高网络鲁棒性. 高频路径为图 3 中的上层路径, 通过局部自注意力编码高频信息, 分配 $(1-\alpha)Nh$ 个注意力头, 低频路径为图 3 中的下层路径, 通过全局注意力编码低频信息, 分配 αNh 个注意力头. 注意力头的分配只需根据任务对高低频信息的敏感程度的不同直接更改 α 值, 既简单直接, 又能避免了对两条路径同时分配 Nh 个注意力头造成的计算资源的消耗.

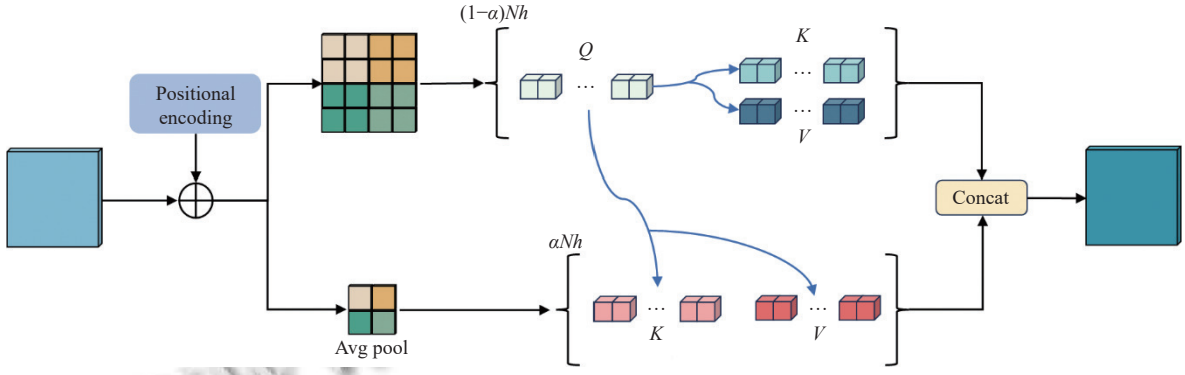

图 3 图像频率引导注意力模块

高频路径的主要任务是编码对象的局部细节, 考虑到局部窗口自注意力对图像局部特征的有效聚合, 将图像划分为若干个 $n\times n$ 的局部窗口 (图中 n=2), 通过计算局部窗口的自注意力捕获图像的高频信息. 同时考虑到计算资源有限, 对窗口进行不重叠地划分, 相比于窗口多尺度划分[23]的操作, 节省了计算资源. 低频路径的主要任务是编码图象的全局结构, 由于平均池化是一种低通滤波器[24], 低频路径采取对高频路径的每个窗口进行平均池化的操作, 池化后的窗口保留了丰富的低频信息, 通过计算池化后的窗口注意力对这些低频信息进行捕获.

图像频率引导注意力模块的注意力使用了缩放点积注意力进行计算, 缩放点积结构图如图 4 所示. 根据图 4, 计算注意力时需要使用 Q、K、V 矩阵, Q、K、V 分别表示查询、键、值, 通过计算查询和键之间的相似性, 再根据相似性对值进行加权求和, 得到最终注意力, 计算公式如式 (1). 计算高频路径注意力时 Q、K、V 由划分窗口后的特征图获得, 低频路径的 KV 由窗口平均池化后的特征图得到, 与高频路径的 Q 进行注意计算. 在计算完高频路径和低频路径的注意力输出之后, 将两个路径的输出拼接在一起, 作为最终输出.

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

此外, 由于高分辨率下的图像特征计算缩放点积注意力公式消耗大量的计算成本, 如何将特征从低分辨率有效地传递到高分辨率仍然是一个问题, 因此我们加入了一个传递路径, 具体来说, 首先将图 1 中最上层特征图经过图像频率引导注意力模块处理后的输出通过一层卷积神经网络对齐下一层高分辨率特征的通道, 再对其进行上采样 (Upsample) 操作后和高分辨率特征相加, 最后通过一层卷积神经网络对相加后的结果进行融合.

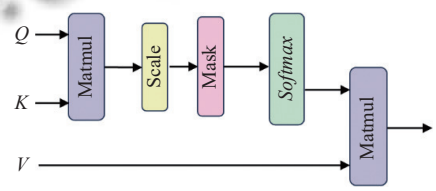

图 4 缩放点积注意力

1.4 代价体的构建与正则化

我们应用可微分单应性变换来将所有的源图像对齐到参考视图. 在深度假设 d 下, 参考视图上的像素 p 与源视图上对应的像素 $\hat{p}$ 之间的单应变换定义如式 (2).

$$\hat{p}=K\left[R(K_0^{-1}pd)+t\right] \tag{2}$$

其中, R 和 t 表示两个视图之间的旋转和平移. K_0 和 K 是参考相机和源相机的本质矩阵. 单应变换后的特征图通过双线性插值保持原始的分辨率. 由于将已知深

度空间离散为 D 个深度值, 参考视图和 N–1 个源视图将获得 D 个深度的单应变换后的特征映射. 同时, 在同一深度 d 下相同位置 p 处的特征映射, 将源视图一一和参考视图计算相关度, 相关度关系式如式 (3) 所示.

$$C_i^d(p) = < \mathcal{F}_0(p), \hat{\mathcal{F}}_i^{(d)}(p) > \tag{3}$$

其中, $\hat{\mathcal{F}}_i^{(d)}(p)$表示深度 d 处第 i 个源图像单应变换的特征映射, 此时得到了 D 个深度下 N–1 个特征相关体. 同时, 为了聚合 N–1 个特征相关体, 我们为最大深度维度相关性分配一个逐像素权重. 聚合的相关性体积定义如式 (4).

$$C^{(d)}(p) = \sum_{i=1}^{N-1} \max_d \left\{ c_i^{(d)}(p) \right\} \cdot c_i^{(d)}(p) \tag{4}$$

此时得到了一个深度为 D 的代价体, 然后使用一个多尺度的 3DCNN 网络对代价体进行正则化, 优化代价体可能包含的噪声.

1.5 损失函数

先前由粗到精的级联结构多采用基于 $l1$ 的损失函数回归损失, 这种方法由于间接学习代价体和深度之间的关系, 表现出了过拟合的问题. 我们则应用焦点损失函数[25], 将深度估计视为一种分类任务, 加强模糊区域的单热点监督. 焦点损失函数公式如下:

$$L = \sum_{p \in \{p_v\}} -\left(1 - P^{(\tilde{d})}(p)\right)^{\gamma} \log\left(P^{(\tilde{d})}(p)\right) \tag{5}$$

其中, $P^{(d)}(p)$表示深度假设 d 在像素 p 处的预测概率, $\tilde{d}$表示在所有假设中最接近真实值的深度值. $\{p_v\}$表示具有有效真值的像素子集. 当聚焦参数$\gamma = 0$时, 焦点损失会退化为交叉熵损失.

2 实验结果

2.1 实验数据集

实验的数据集主要有 3 个, 分别为 DTU[26]、Tanks and Temples[27]以及 BlendedMVS[28]. DTU 数据集是一个包含 128 个场景的室内 MVS 数据集, 每个场景在 7 种不同的照明条件下有 49 或 64 个视图. 我们按照 MVSNet[8]的方式将 DTU 数据集分为训练集和测试集. BlendedMVS 数据集是用于多视图立体匹配训练的大规模合成数据集, 包含多种对象和场景. 我们通常将 BlendedMVS 用于对模型的微调, 以确保 Tanks and Temples 数据集的测试达到最好的结果. Tanks and Temples 数据集包含了真实光照条件下的室内和室外场景, 且有较大的尺度变化, 它包括了一个拥有 8 个场景的中间子集和一个拥有 6 个场景的高级子集, 为了和其他方法比较, 我们主要测试中间子集的结果.

2.2 实验细节

CAF-MVSNet 通过 PyTorch 框架实现, 同时可在 4 个英伟达 RTX3090 上进行训练. 实验训练过程包括了 DTU 训练、BlendedMVS 微调 DTU 训练后的模型. 在 DTU 的训练阶段, 我们设置图像输入数量为 N=5, 图像原始分辨为 640×512, 同时深度范围从 425 mm 到 935 mm. 对于由粗到细的过程, 我们分别对 8、32、48 这 3 个数量的假设深度进行采样, 深度间隔分别设置为 1、2、4. 训练通过 Adam 优化器迭代 16 个 epoch, 初始学习率为 0.001, 并且于第 6、第 8、第 10 个 epoch 后减半, batchsize 设置为 1. BlendedMVS 微调阶段则设置图像输入数量 N=7, 图像原始分辨率为 768×576, 其余参数设置与 DTU 训练阶段相同.

2.3 实验

2.3.1 对比实验

为了验证 CAF-MVSNet 的有效性, 首先在 DTU 测试集上进行了实验. 为了定量分析 DTU 数据集的实验结果, DTU 数据集中提出了两种评价指标, 分别是精确度 (Acc) 和完整度 (Comp), 这两种评判指标由 Seitz 等[29]提出. 精确度计算了从三维重建的结果到真实场景的距离, 描述的是重建的三维点云结果的质量. 完整度的计算采用标准点云中每一个点到重建点云的最近距离的平均值来表示, 描述的是真实点云在重建出的点云上的覆盖率. 将上述两种指标取平均值, 得到模型的综合性能 (Overall). 上述的指标均用于衡量重建的模型和真实模型的偏差程度, 数值越小代表重建出的效果越好.

CAF-MVSNet 本身只产生深度图, 要想生成三维点云, 需要对深度图进行滤波融合处理. MVSNet 常用的滤波融合方法有 fusibile 和 normal 两种, 其中 fusibile 为 Gipuma[7]中的滤波融合方法, 通过对得到的深度图进行光度一致性滤波和几何一致性滤波剔除部分不可信的点, 然后进行深度融合获得点云; normal 则是在 PyTorch 框架下实现光度一致滤波、几何一致性滤波、深度融合的方法. 由于这两种方法生成的三维点云性能上存在些许差别, 为了更综合地评估 CAF-MVSNet, 我们将两种方法的结果与其他方法在数据集 DTU 上

的结果进行定量对比, 如表 1 所示, 其中 CAF-MVSNet* 为使用 fusibile 方法进行深度融合的结果.

表 1 DTU 数据集上的定量评估结果

Method	Acc↓	Comp↓	Overall↓
Gipuma[7]	**0.283**	0.873	0.578
COLMAP[6]	0.400	0.664	0.532
MVSNet[8]	0.396	0.527	0.426
R-MVSNet[9]	0.389	0.459	0.422
AA-RMVSNet[30]	0.376	0.339	0.357
CasMVSNet[10]	0.325	0.385	0.355
EPP-MVSNet[31]	0.413	0.296	0.355
PatchmatchNet[32]	0.427	0.277	0.352
CVP-MVSNet[33]	0.296	0.406	0.351
UCS-Net[34]	0.338	0.349	0.344
RayMVSNet[35]	0.341	0.319	0.330
UniMVSNet[13]	0.352	0.278	0.315
NR-MVSNet[14]	0.331	0.285	0.308
TransMVSNet[11]	0.321	0.289	0.305
CAF-MVSNet	0.325	0.297	0.311
CAF-MVSNet*	0.342	**0.262**	**0.302**

注: 加粗数值为该列最优值

根据表 1, CAF-MVSNet 使用 fusibile 深度融合时综合性能和完整度最好, 另外, 使用 normal 深度融合时精度和完整度虽然不是最佳, 但是都处于较高水平, 且其综合性能优于除 NR-MVSNet 和 TransMVSNet 外的其他方法, 总体而言, CAF-MVSNet 在 DTU 数据集上展现出了较好的性能. 同时, 图 5 展示了 DTU 数据集的点云对比结果, 为了方便比较, 图 5 中各方法的点云均为选择使用 fusibile 进行深度融合得到的点云. 根据图 5, 其他方法的重建点云都有不同程度的细节结构不清晰和点云重建不完整的现象, CAF-MVSNet 得益于上下文增强模块和图像频率引导注意力模块获得了全面的上下文信息和丰富的频率信息, 获得的点云结果细节结构更清晰、重建完整度更高. 具体来说, CAF-MVSNet 在第 1 行结果中与其他方法相比, 人物及人物背后的图案都更清晰, 噪点更少, 同时 CAF-MVSNet 在第 2、3 行结果中与其他方法相比, 红框部分的点云更完整.

此外, 表 2 展示了推理成本对比, 由于 CAF-MVSNet 采用了级联结构, 所以在成本推理对比选择了同为级联结构的 CasMVSNet 和 TransMVSNet 进行对比.

根据表 2, CAF-MVSNet 由于上下文增强模块和图像频率引导模块的加入, 其推理消耗的内存最多, 推理时间略优于 TransMVSNet, 点云重建的综合性能相对较好.

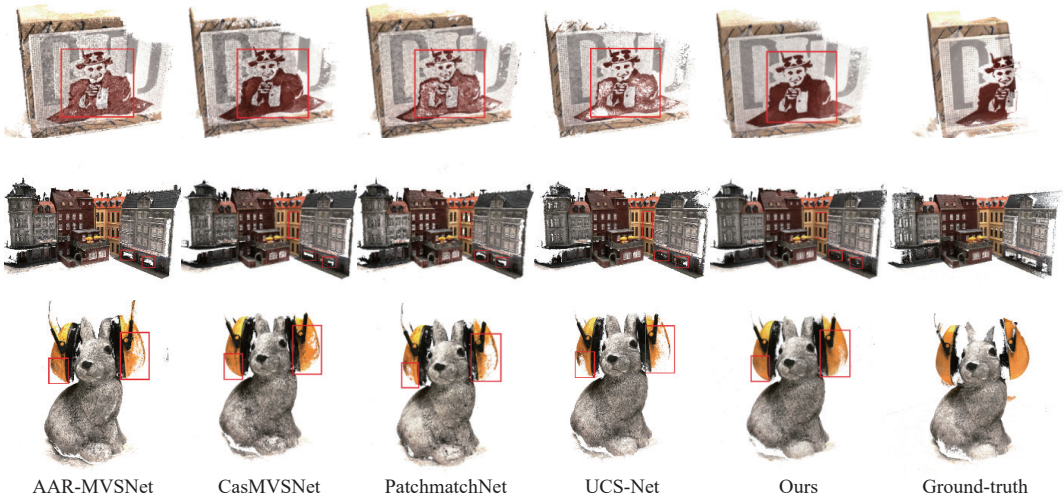

图 5 DTU 数据集点云对比

表 2 推理成本对比

Method	Mem (MB)	Time (s)	Overall (mm)
CasMVSNet[10]	3 803	0.20	0.355
TransMVSNet[11]	7 459	0.71	0.305
CAF-MVSNet*	12 597	0.49	0.302

2.3.2 泛化实验

我们通过在 Tanks and Temples 测试集上进行评估来验证 CAF-MVSNet 的泛化能力. 为了与其他方法的 Tanks and Temples 实验部分保持一致, 我们首先通过 BlendedMVS 训练集微调已有的模型, 在 Tanks and Temples 测试时则调用微调后的模型. 为了定量分析测试结果, 需要将重建结果上传至 Tanks and Temples 官网, 然后获得每个场景的 *F*-score, *F*-score 代表了精确

率 (precision) 和召回率 (recall) 的调和平均值, 性能越好得到的 *F*-score 值越高, 我们最终根据 *F*-score 判断 CAF-MVSNet 的泛化性能.

CAF-MVSNet 与其他先进的方法在数据集 Tanks and Temples 上的重建结果如表 3 所示. 根据表 3, CAF-MVSNet 在 Family 和 Train 场景得到了最好 *F*-score 值, 同时在 Francis 场景取得了第 2 好的得分, 且 CAF-MVSNet 在 *F*-score 上的平均得分仅次于 AA-RMVSNet 和 EPP-MVSNet, 优于其他大部分算法, 表现出了良好的性能. 另外我们从 Tanks and Temples 官网下载了 AA-RMVSNet、CasMVSNet、CVP-MVSNet、EPP-MVSNet 等方法的部分场景的召回图, 并与本文方法进行比较, 如图 6 所示. τ 是 Tanks and Temples 官方确定的与场景相关的距离阈值, 颜色较暗的区域表示遇到的 τ 误差较大. 图 6 中 AA-RMVSNet、CasMVSNet、CVP-MVSNet、EPP-MVSNet 等方法的召回图在复杂场景处颜色较暗, 误差较大, CAF-MVSNet 召回图在相同位置相对颜色较浅, 误差更小. 这表明 CAF-MVSNet 得到的点云误差更小更可靠, 显示出了较好的性能.

表 3 Tanks and Temples 数据集上定量测试结果

Method	Mean↑	Family	Francis	Horse	LightHouse	M60	Panther	PlayGround	Train
COLMAP[6]	42.14	50.41	22.25	26.63	56.43	44.83	46.97	48.53	42.04
MVSNet[8]	43.48	55.59	28.55	25.07	50.79	53.96	50.86	47.9	34.69
R-MVSNNet[9]	50.55	73.01	54.46	43.42	43.88	46.80	46.69	50.87	42.25
PatchmatchNet[32]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81
CVP-MVSNet[33]	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54
CasMVSNet[10]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51
D2HC-RMVSNet[36]	59.20	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92
NP-CVP-MVSNet[37]	59.64	78.93	**64.09**	51.82	59.42	58.39	55.71	56.07	52.71
AA-RMVSNet[30]	61.51	77.77	59.53	51.53	**64.02**	**64.05**	59.47	60.85	55.50
EPP-MVSNet[31]	**61.68**	77.86	60.54	**52.96**	62.33	61.69	**60.34**	**62.44**	55.30
Ours	60.32	**80.30**	63.12	49.80	58.06	59.87	57.85	55.89	**57.69**

注: 加粗数值为该列最优值

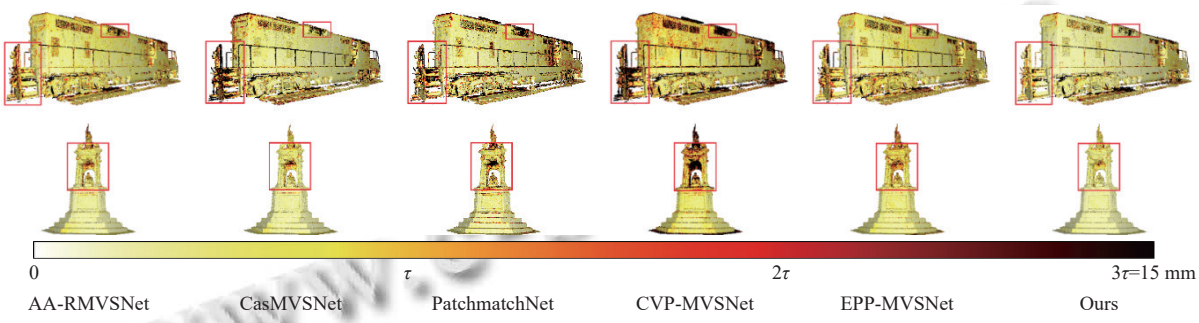

图 6 Tanks and Temples 结果对比

2.3.3 消融实验

为了验证 CAF-MVSNet 中上下文增强模块和图像频率引导注意力模块的有效性, 我们进行了消融实验. 我们选取 DTU 测试集作为消融实验的数据集, 为了方便比较, 消融实验均采用 normal 方法进行深度融合, 其余参数设置与第 2.2 节相同的情况. 实验分成了 4 组, 分别是 Baseline、Baseline+CEM 模块、Baseline+IFAM 模块和 Baseline+CEM 模块+IFAM 模块 4 种组合, 其中 Baseline 为 CAF-MVSNet 去除 CEM 和 IFAM 两个模块后的算法.

4 组的消融实验结果如表 4 所示. 根据表 4, 单独使用 CEM 模块时, 点云重建的精确度和完整度都得到明显的提高, 证明了 CEM 模块扩大对高分率特征图的感受野有利于重建的结果. 而单独使用 IFAM 模块时, 虽然完整度改变不明显, 但是精度的明显提升, 证明了 IFAM 模块针对图像高低频信息的提取, 对深度估计的精确性具有充分的影响. CEM 模块和 IFAM 这两个模块同时使用时, 网络的性能相比单独使用某一模块也

得到提升, 使得 CAF-MVSNet 在 DTU 测试集上实现了较高的性能.

表 4 消融实验定量结果对比

Method	Acc↓	Comp↓	Overall↓
Baseline	0.357	0.311	0.334
Baseline+CEM	0.345	**0.295**	0.320
Baseline+IFAM	0.334	0.318	0..326
CAF-MVSNet	**0.325**	0.297	**0.311**

注: 加粗数值为该列最优值

3 结论与展望

本文提出了上下文增强与图像频率引导的多视图立体网络 CAF-MVSNet. 具体而言, 使用融合空洞卷积的特征提取模块自适应扩大感受野, 同时通过图像引导模块, 聚合图像中不同频率的信息, 使聚合的特征得到更可靠的匹配. 在 DTU、Tanks and Temples 数据集上的实验表明, 相比于其他基于学习的多视图立体匹配方法, 本文提出的方法具有优秀性能, 同时拥有良好的泛化能力.

参考文献

1 Zhu QT, Min C, Wei ZZ, *et al*. Deep learning for multi-view stereo via plane sweep: A survey. arXiv:2106.15328, 2021.

2 Chen XZ, Ma HM, Wan J, *et al*. Multi-view 3D object detection network for autonomous driving. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6526–6534.

3 Schmid K, Hirschmüller H, Dömel A, *et al*. View planning for multi-view stereo 3D reconstruction using an autonomous multicopter. Journal of Intelligent & Robotic Systems, 2012, 65(1): 309–323.

4 Muzzupappa M, Gallo A, Spadafora F, *et al*. 3D reconstruction of an outdoor archaeological site through a multi-view stereo technique. Proceedings of the 2013 Digital Heritage International Congress (DigitalHeritage). Marseille: IEEE, 2013. 169–176.

5 Cernea D. OpenMVS: Multi-view stereo reconstruction library. https://cdcseacave.github.io/openMVS. [2024-08-25].

6 Zheng EL, Dunn E, Jojic V, *et al*. Patchmatch based joint view selection and depthmap estimation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1510–1517.

7 Galliani S, Lasinger K, Schindler K. Massively parallel multiview stereopsis by surface normal diffusion. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 873–881.

8 Yao Y, Luo ZX, Li SW, *et al*. MVSNet: Depth inference for unstructured multi-view stereo. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 785–801.

9 Yao Y, Luo ZX, Li SW, *et al*. Recurrent MVSNet for high-resolution multi-view stereo depth inference. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5520–5529.

10 Gu XD, Fan ZW, Zhu SY, *et al*. Cascade cost volume for high-resolution multi-view stereo and stereo matching. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2492–2501.

11 Ding YK, Yuan WT, Zhu QT, *et al*. TransMVSNet: Global context-aware multi-view stereo network with Transformers. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8575–8584.

12 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

13 Peng R, Wang RJ, Wang ZY, *et al*. Rethinking depth estimation for multi-view stereo: A unified representation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8635–8644.

14 Li JL, Lu ZD, Wang YQ, *et al*. NR-MVSNet: Learning multi-view stereo based on normal consistency and depth refinement. IEEE Transactions on Image Processing, 2023, 32: 2649–2662. [doi: 10.1109/TIP.2023.3272170]

15 Zhang YS, Zhu JK, Lin LX. Multi-view stereo representation revist: Region-aware mvsnet. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 17376–17385.

16 Ye XY, Zhao WY, Liu TQ, *et al*. Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 17615–17624.

17 Vats VK, Joshi S, Crandall DJ, *et al*. GC-MVSNet: Multi-view, multi-scale, geometrically-consistent multi-view stereo. Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2024. 3230–3240.

18 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Proceedings of the 4th International Conference on Learning Representations. San Juan: OpenReview.net, 2016.
19 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.
20 Cao JX, Chen Q, Guo J, *et al*. Attention-guided context feature pyramid network for object detection. arXiv:2005.11475, 2020.
21 Cooley JW, Lewis PAW, Welch PD. The fast Fourier transform and its applications. IEEE Transactions on Education, 1969, 12(1): 27–34. [doi: 10.1109/TE.1969.4320436]
22 Deng G, Cahill LW. An adaptive Gaussian filter for noise reduction and edge detection. Proceedings of the 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference. San Francisco: IEEE, 1993. 1615–1619.
23 Yang JW, Li CY, Zhang PC, *et al*. Focal self-attention for local-global interactions in vision Transformers. arXiv:2107.00641, 2021.
24 Voigtman E, Winefordner JD. Low-pass filters for signal averaging. Review of Scientific Instruments, 1986, 57(5): 957–966. [doi: 10.1063/1.1138645]
25 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2999–3007.
26 Aanæs H, Jensen RR, Vogiatzis G, *et al*. Large-scale data for multiple-view stereopsis. International Journal of Computer Vision, 2016, 120(2): 153–168. [doi: 10.1007/s11263-016-0902-9]
27 Knapitsch A, Park J, Zhou QY, *et al*. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG), 2017, 36(4): 78.
28 Yao Y, Luo ZX, Li SW, *et al*. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1787–1796.
29 Seitz SM, Curless B, Diebel J, *et al*. A comparison and evaluation of multi-view stereo reconstruction algorithms. Poceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE. [doi: 10.1109/CVPR.2006.19]
30 Wei ZZ, Zhu QT, Min C, *et al*. AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 6167–6176.
31 Ma XJ, Gong Y, Wang QR, *et al*. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 5712–5720.
32 Wang FJH, Galliani S, Vogel C, *et al*. PatchmatchNet: Learned multi-view patchmatch stereo. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14189–14198.
33 Yang JY, Mao W, Alvarez JM, *et al*. Cost volume pyramid based depth inference for multi-view stereo. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4876–4885.
34 Cheng S, Xu ZX, Zhu SL, *et al*. Deep stereo using adaptive thin volume representation with uncertainty awareness. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2521–2531.
35 Xi JH, Shi YF, Wang YJ, *et al*. RayMVSNet: Learning ray-based 1D implicit fields for accurate multi-view stereo. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8585–8595.
36 Yan JF, Wei ZZ, Yi HW, *et al*. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 674–689.
37 Yang JY, Alvarez JM, Liu MM. Non-parametric depth distribution modelling based depth inference for multi-view stereo. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8616–8624.

(校对责编: 张重毅)