基于自适应空间稀疏化的高效多视图立体匹配

周晓清[1,2,3],王 翔[1,2,3],郑 锦[1],百 晓[1,2,3]

(1. 北京航空航天大学计算机学院,北京 100191;2. 北京航空航天大学软件开发环境国家重点实验室,北京 100191;
3. 北京航空航天大学江西研究院,江西南昌 330000)

摘 要: 针对多视图立体匹配中构建和聚合匹配代价体时计算复杂度高的问题,现有研究通常采用级联架构或迭代优化方法. 然而这些方法仍面临两个亟待解决的挑战:级联架构在精细阶段缩小了深度采样范围,导致深度不连续区域可能陷入低分辨率的错误估计;而迭代优化网络的推理时间随迭代次数线性增长,难以满足实时系统需求. 为此,本文提出一种基于自适应空间稀疏化的高效多视图立体匹配网络. 我们提出一种稀疏匹配代价体构建方法,通过在完整深度范围内稀疏采样,在降低计算复杂度的同时保持了网络对深度不连续区域的建模能力. 同时,我们提出一种稀疏迭代优化方法,在迭代中通过自适应变分 Dropout 逐步剪枝深度值已收敛的区域,使推理时间随迭代次数亚线性增长. 在 DTU 和 Tanks & Temples 公共数据集上的实验结果表明,本文方法的推理速度相比 CasMVSNet 和 PatchmatchNet 分别快 1.2 倍和 0.35 倍,同时点云重建效果优异,边缘伪影显著减少,且泛化能力表现出色.

关键词: 多视图立体;三维重建;深度估计;稀疏神经网络;循环神经网络;Transformer

基金项目: 国家自然科学基金(No.62276016,No.62372029)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)11-3079-13

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230353

Adaptive Spatial Sparsification for Efficient Multi-View Stereo Matching

ZHOU Xiao-qing[1,2,3], WANG Xiang[1,2,3], ZHENG Jin[1], BAI Xiao[1,2,3]

(1. *School of Computer Science and Engineering*, *Beihang University*, *Beijing* 100191, *China*;
2. *State Key Laboratory of Software Development Environment*, *Beihang University*, *Beijing* 100191, *China*;
3. *Jiangxi Research Institute of Beihang University*, *Nanchang*, *Jiangxi* 330000, *China*)

Abstract: To reduce the high computational complexity in constructing and aggregating cost volumes for multi-view stereo matching, existing methods commonly employ cascaded architectures or iterative optimization. However, these approaches still face two main challenges. The cascaded architectures narrow down the depth sampling range during the refinement stage, which may lead to erroneous estimation of depth discontinuities. While the inference time of iterative optimization networks linearly increases with the number of iterations, making it difficult to meet the requirements of real-time systems. To address these challenges, this paper proposes an efficient multi-view stereo matching network via adaptive spatial sparsification. We introduce a sparse matching cost volume that sparsely samples within the complete depth range, reducing computational complexity while maintaining the network's ability to model depth-discontinuous regions. Meanwhile, we propose a sparse iterative optimization method that progressively prunes regions with converged depth values during iterations using adaptive variational Dropout, resulting in sub-linear growth in inference time with iteration count. Experimental results on the public datasets, DTU and Tanks & Temples, demonstrate that the proposed method achieves 1.2× and 0.35× improvements of inference speed compared to CasMVSNet and PatchmatchNet, respectively. Moreover, it exhibits excellent performance in point cloud reconstruction, effectively handles details in depth-discontinuous regions, and demonstrates outstanding generalization capability.

Key words: multi-view stereo; 3D reconstruction; depth estimation; sparse neural networks; recurrent neural networks; Transformer

Foundation Item(s): National Natural Science Foundation of China (No.62276016, No.62372029)

收稿日期:2023-04-18;修回日期:2023-09-27;责任编辑:朱梅玉

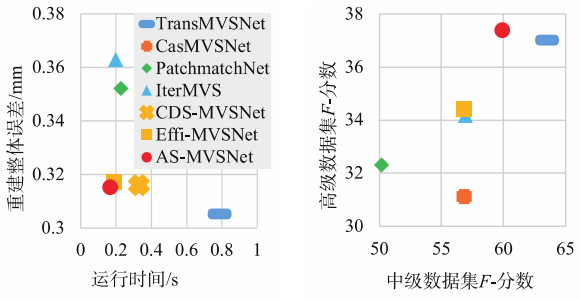
1 引言

获取有关环境的稠密3D信息是诸如自主导航[1]、智能机器人[2]和工业控制[3]等应用的关键. 与主动深度传感相比,基于摄像机的被动传感具有成本效益高、能源消耗低、尺寸紧凑和适用于广泛条件下运行等优势. 实践中由于增加视点能够提供更鲁棒的重建效果,多视图立体(Multi-View Stereo,MVS)通常是首选方法. 假设已知摄像机内参和采集图像间的姿态变换,MVS根据光度一致性假设和多视图几何约束从一组相邻视图的图像中恢复深度图,并将这些深度图融合为三维点云. 作为计算机视觉的核心问题之一,MVS已被广泛研究数十年[4~7]. 尽管传统MVS方法可以实现朗伯表面的鲁棒重建,但由于人工设计特征的限制,这些方法在弱纹理和非朗伯表面等不适定区域中重建完整性不足.

近年,深度神经网络在提升多视图立体效果方面展现出强大的能力,但可扩展性仍然面临挑战. 大多数基于学习的MVS方法[8,9]根据一组假设深度平面和扭曲特征构建3D匹配代价体,并使用3D卷积层聚合匹配代价. 由于计算复杂度随输入图像分辨率的增加呈立方增长,这些方法不能很好地扩展到数百万像素的图像. 最近,一些工作研究通过依序聚合匹配代价体切片[10]、生成降采样深度图[11,12]、构建级联匹配代价体[13~15]、迭代优化深度图[16,17]等方法减少多视图立体的显存占用. Gu等[13,14]在级联架构中由粗到细(coarse-to-fine)地聚合匹配代价,Wang等[15,16]采用2D门控循环单元(Gated Recurrent Unit,GRU)来迭代优化深度图估计,同时改善了显存占用和运行时间. 然而,这些方法仍面临两个亟待解决的挑战:一方面,级联架构在精细阶段缩小了深度采样范围,在深度不连续区域可能陷入低分辨率的错误估计;另一方面,迭代优化方法的推理时间随迭代次数增加呈线性增长,难以满足实时系统的要求. 此外,Jiang等人[18]使用k-近邻搜索构建稀疏相关体,进一步减少了显存占用,但精度欠佳.

针对上述问题,本文提出一种基于自适应空间稀疏化的高效多视图立体网络(Adaptive Spatial Sparsification MVS Network,AS-MVSNet). 与现有高效多视图立体方法不同,AS-MVSNet从立体匹配的空间稀疏性出发,分别构造深度空间和图像空间的稀疏计算,在提高多视图立体推理效率的同时保持了优异的重建效果. 网络首先使用Transformer提取位置感知和视图一致的多视图图像特征. 接着,提出一种稀疏匹配代价体构建方法,通过在完整深度范围内选取匹配相似性最大的K个的深度值作为稀疏假设深度,所构建的稀疏匹配代价体在特征上采样时覆盖完整深度范围,使网络在深度不连续区域具有精确建模能力. 此外,为减少深度图优化阶段图像空间的冗余计算,本文提出一种样本自适应的稀疏迭代优化方法,通过在GRU隐状态更新量上施以样本自适应的Dropout稀疏化,逐步剪枝深度值已收敛区域,使推理时间随不同图像动态变化.

在公共数据集上的实验表明,AS-MVSNet的推理速度相比CasMVSNet[13]和PatchmatchNet[17]分别快1.2倍和0.35倍,同时重建效果优异,边缘伪影显著减少,且泛化能力表现出色. 如图1(a)所示,本文方法在DTU评估集[19]上的运行速度相较于比较方法更快,同时重建效果优异. 图1(b)进一步展示了本文方法在Tanks & Temples评估集[20]上的泛化能力,在中级数据集和高级数据集上与主流MVS方法相比分别取得了第二和最优的F-分数.



(a) DTU (b) Tanks & Temples

图1 DTU[19]和Tanks & Temples[20]评估集上与先进MVS方法的比较

2 相关工作

本节首先简要概述基于学习的多视图立体方法,然后介绍稀疏变分Dropout理论和相关工作.

2.1 基于学习的多视图立体

一类典型的基于学习的多视图立体网络包括特征提取、匹配代价体构建、代价聚合和深度图优化四个部分[8]. 特征提取部分使用孪生网络(Siamese network)提取逐视图图像的稠密特征. 接着,在参考视图中采样一组正面平行(fronto-parallel)假设平面,基于假设平面诱导的单应性变换将每个源视图特征扭曲到参考视图,并计算不同视图特征方差得到匹配代价体. 其中,假设参考视图$\boldsymbol{I}_0$中像素$\boldsymbol{p}$的一个假设深度值为d_j,则单应性变换表示为

$$\boldsymbol{p}_{i,j}:=\boldsymbol{p}_i(d_j)=\boldsymbol{K}_i\cdot(\boldsymbol{R}_{0,i}\cdot(\boldsymbol{K}_0^{-1}\cdot\boldsymbol{p}\cdot d_j)+\boldsymbol{t}_{0,i}) \tag{1}$$

其中,$\{\boldsymbol{K}_i\}_{i=0}^{N-1}$表示摄像机内参,$\{[\boldsymbol{R}_{0,i}|\boldsymbol{t}_{0,i}]\}_{i=1}^{N-1}$表示摄像机相对位姿变换,$\boldsymbol{p}_{i,j}$表示像素$\boldsymbol{p}$重投影到源视图$\boldsymbol{I}_i$中的位置. 最后,使用3D卷积层将匹配代价体聚合为概率体,计算期望得到深度图估计$\boldsymbol{D}$,并使用残差网络优化深度图. 虽然可以实现更稳健的匹配,但由于这类方法计算复杂度与图像分辨率呈立方关系,方法扩展性不足.

2.2 稀疏变分 Dropout

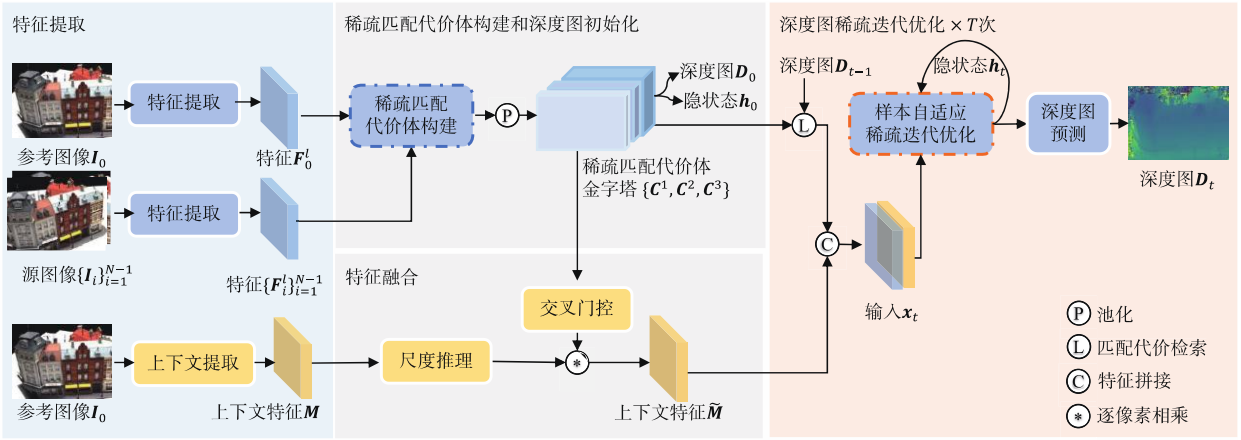
深度神经网络中的稀疏化最初作为一种正则化器,但近年来也被视为提高模型推理效率的方法之一[21~26]. 稀疏变分 Dropout[24]假设网络参数 Φ 服从对数均匀先验 p 和高斯近似后验 q_Φ,其中具有高后验方差的参数对最终网络性能的贡献很小,因此删除它们不影响网络效果. 参数 Φ 的最优值通过最大化证据下界(the Evidence Lower BOund,ELBO)得到,表示为

$$L(\Phi)=L_{\mathrm{d}}(\Phi)-D_{\mathrm{KL}}(q_\Phi \| p)\rightarrow \max_{\phi\in\Phi} \tag{2}$$

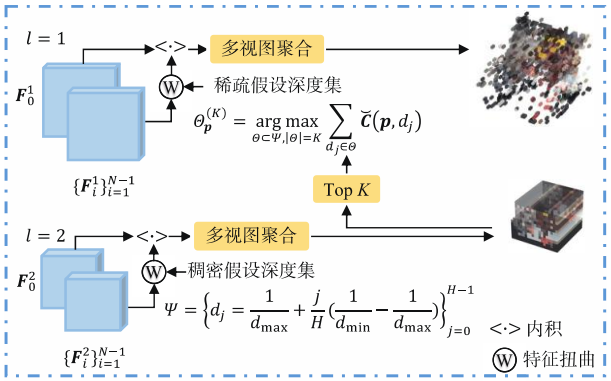
其中,$L_{\mathrm{d}}(\Phi)$表示模型期望,$-D_{\mathrm{KL}}(q_\Phi \| p)$表示对数均匀先验 p 和高斯近似后验 q_Φ 的 KL 散度. Lobacheva 等人[27]使用稀疏变分 Dropout 来稀疏化循环神经网络(Recurrent Neural Network,RNN)门控的预激活值,简化了网络结构,但由于 Dropout 概率对所有样本共享,网络灵活性和扩展性受限.

3 基于自适应空间稀疏化的高效多视图立体

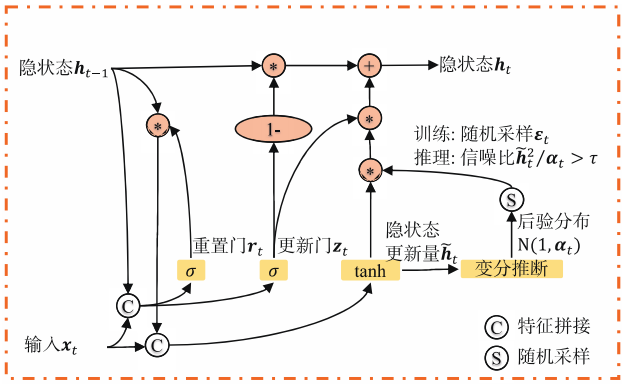
图 2(a)展示了基于自适应空间稀疏化的高效多视图立体网络的整体结构. 首先,特征提取部分提取逐视图的多尺度特征 $\{\boldsymbol{F}_i^l\}_{i=0}^{N-1}$ 和参考视图的上下文特征 $\boldsymbol{M}$;接着,通过计算参考视图特征 $\boldsymbol{F}_0^l$ 和源视图特征 $\{\boldsymbol{F}_i^l\}_{i=1}^{N-1}$ 的相似性,构建稀疏匹配代价体金字塔,并从中恢复深度图初始值 $\boldsymbol{D}_0$;最后,在深度图稀疏迭代优化部分,从稀疏匹配代价体金字塔检索几何特征,并与上下文特征 $\boldsymbol{M}$ 融合后,共同作为引导特征优化深度图估计.



(a) 网络整体结构



(b) 稀疏匹配代价体构建



(c) 样本自适应稀疏迭代优化

图 2　基于自适应空间稀疏化的高效多视图立体匹配

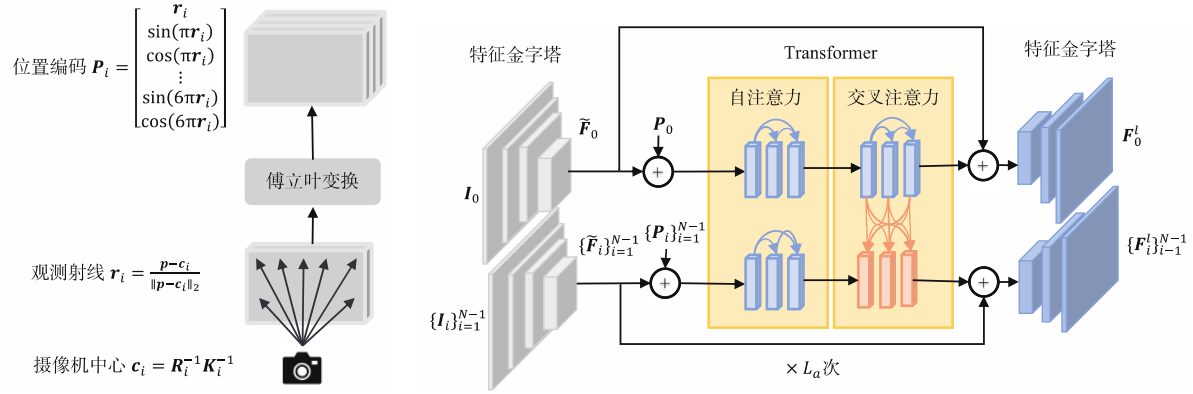
3.1 多尺度全局特征提取

多视图立体方法通常使用多尺度特征以减少不适定区域的匹配歧义[7]. 为进一步提取位置感知和视图一致的图像表征,本节提出一种使用 Transformer 增强特征金字塔网络(Feature Pyramid Network,FPN)的特征提取方法.

具体地,如图 3(a)所示,首先将逐视图的视线方向 $\boldsymbol{r}_i=\boldsymbol{R}_{0,i}^{-1}\boldsymbol{K}_i^{-1}[\boldsymbol{p}^{\mathrm{T}}\ 1]^{\mathrm{T}}$ 嵌入高维傅立叶空间作为位置编码 $\boldsymbol{P}_i$. 然后,如图 3(b)所示,将位置编码 $\boldsymbol{P}_i$ 与 FPN 瓶颈层特征 $\tilde{\boldsymbol{F}}_i$ 相加后,输入 Transformer 中交替执行 $L_a=4$ 次视图内

自注意力和视图间交叉注意力. 其中视图间交叉注意力使源视图从参考视图提取视图一致表征. 注意力使用Linear Transformer计算[28],使本文方法适用于高分辨率图像. 最后,将Transformer输出的特征上采样后与FPN对应层特征相加,输出2个尺度的特征$\boldsymbol{F}_i^l$,其中$l\in\{1,2\}$,每个尺度的特征分辨率为输入图像分辨率的$1/2^l$.

此外,我们使用一个独立的上下文提取器提取参考图像$\boldsymbol{I}_0$的上下文特征$\boldsymbol{M}$. 该上下文提取器的结构与图3的特征提取器相同,但不包含Transformer的交叉注意力层.



(a) 摄像机几何位置编码　　(b) Transformer增强的特征金字塔网络

图3　多尺度全局特征提取

3.2　稀疏匹配代价体构建和深度图初始化

沿用平面扫描立体[29],本节在参考视图采样一组正面平行假设平面,经过假设平面诱导的单应性变换将源视图特征$\{\boldsymbol{F}_i^l\}_{i=1}^{N-1}$扭曲到参考视图,然后基于分组相关性度量构建匹配代价体,并计算深度图初始估计. 为降低计算复杂度,我们参照文献[13,14]构建由粗到细的匹配代价体. 然而,与先前方法在精细阶段缩小深度采样范围不同,本节方法在完整深度范围内选取匹配相似性最大的K个的深度值,作为精细阶段的稀疏假设深度集. 所构建的稀疏匹配代价体使网络在特征上采样过程中保持了深度不连续区域的建模能力,有助于减少边缘伪影现象.

3.2.1　稀疏假设深度和稀疏匹配代价体金字塔

图2(b)展示了稀疏匹配代价体的构建过程. 给定深度范围$[d_{\min},d_{\max}]$,首先沿逆深度区间等间隔采样,得到稠密假设深度集,表示为

$$\Psi=\left\{d_j|d_j=\frac{1}{d_{\max}}+\frac{j}{H}\left(\frac{1}{d_{\min}}-\frac{1}{d_{\max}}\right)\right\}_{j=0}^{H-1}\tag{3}$$

其中,H是深度值稠密采样数量,d_j是深度值. 由于逆深度区间上的等间隔采样对应图像对极线上的等间隔采样,本节方法适用于重建宽深度范围的大规模场景[10,30]. 确定稠密深度假设集Ψ后,按3.2.2节方法使用特征$\{\boldsymbol{F}_i^2\}_{i=0}^{N-1}$构建1/8分辨率的稠密匹配代价体$\overline{\boldsymbol{C}}(\boldsymbol{p},d_j)$.

随后,从$\overline{\boldsymbol{C}}(\boldsymbol{p},d_j)$中逐像素选取匹配相似性最大的$K$个深度值,组成稀疏假设深度集$\Theta_{\boldsymbol{p}}^{(K)}$,表示为

$$\Theta_{\boldsymbol{p}}^{(K)}=\underset{\Theta\subset\Psi,|\Theta|=K}{\operatorname{argmax}}\sum_{d_j\in\Theta}\overline{\boldsymbol{C}}(\boldsymbol{p},d_j)\tag{4}$$

其中,K是深度值稀疏采样数量,d_j是深度值,$\boldsymbol{p}$是参考视图的像素. 使用最近邻插值将$\Theta_{\boldsymbol{p}}^{(K)}$上采样,并按3.2.2节方法使用特征$\{\boldsymbol{F}_i^1\}_{i=0}^{N-1}$构建1/4分辨率的稀疏匹配代价体$\boldsymbol{C}\left(\boldsymbol{p},d_j\right)$. 与构建1/4分辨率的稠密匹配代价体相比,构建稀疏匹配代价体的计算复杂度从$O(nH/4^2)$减少至$O(nH/8^2+nK/4^2)$,其中n表示图像像素数量,H和K分别是深度值稠密采样和稀疏采样数量,$K\ll H$.

最后,将稀疏匹配代价体$\boldsymbol{C}\left(\boldsymbol{p},d_j\right)$在深度维度进行步长分别为1、2、4的平均池化,得到稀疏匹配代价体金字塔$\{\boldsymbol{C}^1,\boldsymbol{C}^2,\boldsymbol{C}^3\}$. 金字塔的每一级有增加的感受野,但由于池化操作仅在深度维度进行,保留了高分辨率图像中的高频细节信息,从而能够恢复视差较大的近距离精细结构. 结合式(3),稀疏匹配代价体金字塔第l级中,逆深度最小采样间隔为

$$R^l=\left(\frac{1}{d_{\min}}-\frac{1}{d_{\max}}\right)\frac{2^{l-1}}{H},l=1,2,3\tag{5}$$

其中,H是深度值稠密采样数量.

3.2.2　匹配代价体的构建

给定图像特征$\{\boldsymbol{F}_i^l\}_{i=0}^{N-1}$和假设深度集,本节基于分组相关[15]的特征一致性度量构建多视图匹配代价体. 下文将省略$\boldsymbol{F}_i^l$的尺度标记l以简化符号.

已知参考视图$\boldsymbol{I}_0$中像素$\boldsymbol{p}$的假设深度值d_j,使用式(1)中单应性变换得到该像素在源视图$\boldsymbol{I}_i$中的位置$\boldsymbol{p}_{i,j}$. 使

用该重投影位置$\boldsymbol{p}_{i,j}$在源视图特征中双线性插值,得到源视图$\boldsymbol{I}_i$扭曲到参考视图$\boldsymbol{I}_0$的特征,记为$\tilde{\boldsymbol{F}}_i(\boldsymbol{p}_{i,j})$. 随后,计算扭曲特征$\tilde{\boldsymbol{F}}_i(\boldsymbol{p}_{i,j})$和参考视图特征$\boldsymbol{F}_0(\boldsymbol{p})$的分组相关[15]得到源试图$\boldsymbol{I}_i$的匹配代价$\boldsymbol{C}_i(\boldsymbol{p},d_j)$.

多视图立体将来自任意数量源视图的匹配代价聚合到单个匹配代价体. 考虑到可能存在遮挡和视场外等违反多视图光度一致性假设的区域,参照文献[31]学习每个源视图像素级的可见性,即沿匹配代价体$\boldsymbol{C}_i(\boldsymbol{p},d_j)$的深度维度使用softmax函数将其归一化为概率体$\boldsymbol{P}_i(\boldsymbol{p},d_j)$,从中逐像素取最大概率值作为可见性权重,表示为

$$\boldsymbol{w}_i(\boldsymbol{p})=\max_{d_j}\left\{\boldsymbol{P}_i(\boldsymbol{p},d_j)\right\} \tag{6}$$

聚合后的匹配代价体表示为

$$C(\boldsymbol{p},d_j)=\frac{1}{\sum_{i=1}^{N-1}\boldsymbol{w}_i(\boldsymbol{p})}\sum_{i=1}^{N-1}\boldsymbol{w}_i(\boldsymbol{p})\cdot\boldsymbol{C}_i(\boldsymbol{p},d_j) \tag{7}$$

3.2.3　深度图初始化

为得到深度图初始值$\boldsymbol{D}_0$,将稀疏匹配代价体$\boldsymbol{C}^1$归一化为概率体$\boldsymbol{P}^1$,并计算深度值期望. 由于深度值在不连续区域呈多峰分布[32],为了避免对多峰分布求期望导致的过度平滑(over-smooth)问题,我们仅在最大概率值局部邻域取加权平均,表示如下:

$$\boldsymbol{D}_0(\boldsymbol{p})=\sum_{d_j\in\Theta_{\boldsymbol{p}}^{(K)}}d_j\cdot\hat{\boldsymbol{P}}(\boldsymbol{p},d_j) \tag{8}$$

$$\hat{\boldsymbol{P}}(\boldsymbol{p},d_j)=\frac{\boldsymbol{P}^1(\boldsymbol{p},d_j)}{\sum_{d\in\Theta_{\boldsymbol{p}}^{(K)},\,1/d\in[1/d_j-rR^1,\,1/d_j+rR^1]}\boldsymbol{P}^1(\boldsymbol{p},d)} \tag{9}$$

其中,$\boldsymbol{P}^1(\boldsymbol{p},d_j)$表示像素$\boldsymbol{p}$处深度值为$d_j$的概率,$R^1$表示式(5)中逆深度区间的最小采样间隔,$r$指定了最大概率值的邻域区间,实验中设置$r=4$.

3.3　深度图稀疏迭代优化

本节对深度图初始值$\boldsymbol{D}_0$进行迭代优化,输出一系列深度图预测值$\{\boldsymbol{D}_1,\cdots,\boldsymbol{D}_t,\cdots,\boldsymbol{D}_T\}$. 3.3.1节介绍迭代优化模块的输入;3.3.2节介绍迭代优化方法;3.3.3节介绍样本自适应稀疏迭代优化方法.

3.3.1　几何特征与上下文特征融合

如图2(*a*)所示,每个迭代步中,使用当前深度图预测值$\boldsymbol{D}_{t-1}$检索稀疏匹配代价体金字塔得到几何特征,与上下文特征融合后输入稀疏迭代优化模块. 由于上下文特征和几何特征的异构性质,我们使用一个尺度推理网络学习上下文特征局部变化的感受野,并使用交叉门控有选择地控制特征传播.

具体地,迭代步t中,使用$\boldsymbol{D}_{t-1}$检索稀疏匹配代价体金字塔$\{\boldsymbol{C}^1,\boldsymbol{C}^2,\boldsymbol{C}^3\}$每一级$l$逆深度范围的$[-4R^l,4R^l]$邻域,将结果拼接为几何特征$\boldsymbol{G}_t$,其中,$R^l$是式(5)中的逆深度最小采样间隔. 然后,从上下文特征$\boldsymbol{M}$中学习局部动态尺度,参照文献[33]使用该动态尺度构建采样网格,并插值得到尺度变换的上下文特征$\tilde{\boldsymbol{M}}$. 最后,从匹配代价体$\boldsymbol{C}^1$中学习限制上下文特征传播的交叉门控$\boldsymbol{A}$. 最终迭代优化模块的输入$\boldsymbol{x}_t$表示如下:

$$\boldsymbol{x}_t=\mathrm{cat}(\boldsymbol{G}_t,\boldsymbol{A}\odot\tilde{\boldsymbol{M}}) \tag{10}$$

其中,$\odot$表示逐像素相乘,cat(·,·)表示拼接运算.

3.3.2　基于GRU的深度图迭代优化

深度图迭代优化模块基于GRU实现. GRU隐状态的更新操作本质上基于平滑先验传播空间邻域信息,为了缓解跨深度值边缘传播导致的边缘模糊问题,我们使用逐像素采样偏移和调制系数的动态卷积更新GRU隐状态.

具体地,使用2个卷积层和tanh函数对匹配代价体$\boldsymbol{C}^1$进行编码,得到GRU隐状态初始值$\boldsymbol{h}_0$. 然后,每个迭代步t中,在输入$\boldsymbol{x}_t$的引导下更新GRU隐状态,表示为

$$\boldsymbol{z}_t=\sigma(\mathrm{conv}(\mathrm{cat}(\boldsymbol{h}_{t-1},\boldsymbol{x}_t),\boldsymbol{W}_{\mathrm{z}})) \tag{11}$$

$$\boldsymbol{r}_t=\sigma(\mathrm{conv}(\mathrm{cat}(\boldsymbol{h}_{t-1},\boldsymbol{x}_t),\boldsymbol{W}_{\mathrm{r}})) \tag{12}$$

$$\tilde{\boldsymbol{h}}_t=\tanh\left(\sum_{k=1}^{N_k}\boldsymbol{m}^k\cdot\boldsymbol{W}_{\mathrm{h}}^k\cdot\mathrm{cat}(\boldsymbol{r}_t\odot\boldsymbol{h}_{t-1},\boldsymbol{x}_t)(\boldsymbol{p}+\boldsymbol{p}^k+\Delta\boldsymbol{p}^k)\right) \tag{13}$$

$$\boldsymbol{h}_t=(1-\boldsymbol{z}_t)\odot\boldsymbol{h}_{t-1}+\boldsymbol{z}_t\odot\tilde{\boldsymbol{h}}_t \tag{14}$$

其中,$\boldsymbol{z}_t$表示更新门,$\boldsymbol{r}_t$表示重置门,$\sigma(\cdot)$表示Sigmoid函数,N_k表示卷积核大小,$\Delta\boldsymbol{p}^k$和$\boldsymbol{m}^k$分别表示动态卷积的逐像素采样偏移和调制系数. $\Delta\boldsymbol{p}^k$和$\boldsymbol{m}^k$由2个卷积层从上下文特征$\tilde{\boldsymbol{M}}$和深度图初始值$\boldsymbol{D}_0$中学习,并在不同迭代步中共享.

每个迭代步t后,一个深度图预测头从隐状态$\boldsymbol{h}_t$中学习深度图残差. 深度预测头包含2个卷积层,并使用tanh函数返回逆深度区间内的残差百分比.

最后一次迭代后,参考RAFT[34]使用3×3邻域网格的加权凸组合将1/4分辨率的深度图$\boldsymbol{D}_T$上采样到全分辨率$\hat{\boldsymbol{D}}_T$. 加权凸组合的掩膜由2个卷积层从隐状态$\boldsymbol{h}_t$中学习,训练阶段在所有迭代步中进行监督.

3.3.3　样本自适应稀疏迭代优化

为减少迭代优化中图像空间的冗余计算,受文献[25]启发,我们扩展了稀疏变分Dropout[24],使其支持样本自适应的Dropout分布,并应用于GRU隐状态更新量作为一种样本自适应的迭代式剪枝方法. 图2(*c*)展示了相应的计算过程. 具体而言,训练阶段假设式(14)中GRU隐状态更新量$\tilde{\boldsymbol{h}}_t$服从对数均匀先验$p(\|\tilde{\boldsymbol{h}}_t\|)\propto 1/\|\tilde{\boldsymbol{h}}_t\|$和可完全分解的高斯近似后验$q_{\Phi}(\tilde{\boldsymbol{h}}_t|1,\boldsymbol{\alpha}_t)=\mathrm{N}(\tilde{\boldsymbol{h}}_t|\tilde{\boldsymbol{h}}_t,\boldsymbol{\alpha}_t\tilde{\boldsymbol{h}}_t^2)$,其中,后验方差$\boldsymbol{\alpha}_t$由参数为$\Phi$的卷积网络从隐状态更新操作的输入cat($\boldsymbol{r}_t\odot\boldsymbol{h}_{t-1}$, $\boldsymbol{x}_t$)中学习. 为提高网络参数利用率,实现中参照文献[25]

使用两个堆叠卷积层$g_{\Phi}(\cdot)$学习$\tilde{\boldsymbol{h}}_t$的后验方差. 然后，从高斯分布$\mathrm{N}(1,\boldsymbol{\alpha}_t)$中采样随机噪声$\boldsymbol{\varepsilon}_t$乘在更新量$\tilde{\boldsymbol{h}}_t$上，即将式(14)替换为

$$\boldsymbol{\gamma}_t=g_{\Phi}(\tilde{\boldsymbol{h}}_t),\boldsymbol{\alpha}_t=\boldsymbol{\gamma}_t^2/\tilde{\boldsymbol{h}}_t^2 \tag{15}$$

$$\boldsymbol{h}_t=(1-\boldsymbol{z}_t)\odot\boldsymbol{h}_{t-1}+\boldsymbol{z}_t\odot\tilde{\boldsymbol{h}}_t\odot\boldsymbol{\varepsilon}_t,\boldsymbol{\varepsilon}_t\sim\mathrm{N}(1,\boldsymbol{\alpha}_t) \tag{16}$$

其中，式(15)使用加性重参数化[24]以支持无界的Dropout概率. 参数$\boldsymbol{\Phi}$的最优值通过式(2)最大化证据下界得到.

推理阶段，由于更新量$\tilde{\boldsymbol{h}}_t$服从对数均匀先验，其中大多数元素的信噪比$\tilde{\boldsymbol{h}}_t^2/\boldsymbol{\alpha}_t$趋于0. 实验中设置信噪比$\tilde{\boldsymbol{h}}_t^2/\boldsymbol{\alpha}_t$小于0.02时，停止相应位置的更新量计算，同时激活像素占比小于0.01时结束迭代过程. 稀疏卷积运算采用基于标题的稀疏卷积[35]. 由于本方法仅将Dropout作用在隐状态的候选更新量上，GRU保留了长期记忆能力[26].

3.4 损失函数

本文方法在训练过程中使用3个监督项：初始分类损失L_{ce}、深度图回归损失L_{d}和稀疏正则化项L_{reg}.

初始深度图估计由匹配概率体$\boldsymbol{P}^1$和对应真值的交叉熵监督，表示为

$$L_{\mathrm{ce}}=\sum_{\boldsymbol{p}}\sum_{d_j\in\Theta_{\boldsymbol{p}}^{(K)}}-\boldsymbol{P}^1(\boldsymbol{p},d_j)\cdot\log\boldsymbol{Q}(\boldsymbol{p},d_j) \tag{17}$$

其中，$\boldsymbol{p}$是具有有效深度真值的像素，$\Theta_{\boldsymbol{p}}^{(K)}$是稀疏假设深度集，$\boldsymbol{P}^1(\boldsymbol{p},d_j)$是像素$\boldsymbol{p}$处深度值为$d_j$的概率，$\boldsymbol{Q}$是以高斯分布表示的匹配概率体真值，其均值为深度图真值，方差σ控制峰值锐度，实验中设置$\sigma=2$.

深度图回归损失监督深度图估计序列和真值之间的L1距离，并对早期迭代乘以指数衰减权重. 为解决透视投影中固有的深度分布不均衡问题，我们参照文献[36]引入深度感知注意力，使网络更加关注远距离区域，表示如下：

$$L_{\mathrm{d}}=\sum_{t=1}^{T}\gamma^{t-T}\boldsymbol{\beta}_D\odot\sum_{\boldsymbol{p}}\|\boldsymbol{D}_{\mathrm{gt}}-\hat{\boldsymbol{D}}_t\|_1 \tag{18}$$

其中，$\boldsymbol{D}_{\mathrm{gt}}$是深度图真值，$\hat{\boldsymbol{D}}_t$是深度图估计值，$\boldsymbol{\beta}_D=\dfrac{\max\boldsymbol{D}_{\mathrm{gt}}-\boldsymbol{D}_{\mathrm{gt}}}{\max\boldsymbol{D}_{\mathrm{gt}}-\min\boldsymbol{D}_{\mathrm{gt}}}$是深度感知注意力，$\gamma$是序列指数衰减系数，实验中设置$\gamma=0.9$.

稀疏正则化项约束GRU隐状态更新量的稀疏性，使其高斯近似后验与对数均匀先验一致，表示为

$$L_{\mathrm{reg}}=\sum_{t=1}^{T}\gamma^{t-T}\odot\sum_{\boldsymbol{p}}\kappa(\boldsymbol{\alpha}_t) \tag{19}$$

其中，$\kappa(\boldsymbol{\alpha}_t)$表示式(2)中KL散度$-D_{\mathrm{KL}}(\boldsymbol{q}_{\Phi}\|\boldsymbol{p})$的近似，通过从近似后验分布中采样$\boldsymbol{\varepsilon}$计算[24]，其结果仅与后验方差$\boldsymbol{\alpha}_t$有关.

最终损失函数如下：

$$L=\lambda_{\mathrm{ce}}L_{\mathrm{ce}}+\lambda_{\mathrm{d}}L_{\mathrm{d}}+\lambda_{\mathrm{reg}}L_{\mathrm{reg}} \tag{20}$$

其中，λ_{ce}、λ_{d}和λ_{reg}是平衡监督项的超参数，实验中设置$\lambda_{\mathrm{ce}}=10$，$\lambda_{\mathrm{d}}=1$，$\lambda_{\mathrm{reg}}=1\mathrm{e}^{-4}$.

4 实验与分析

本节在公开数据集上验证本文方法的有效性. 首先介绍数据集、评价标准和实现细节；然后介绍与主流MVS方法在重建效果、效率和泛化性方面的对比实验；最后进行消融实验研究.

4.1 数据集和评价标准

4.1.1 数据集

DTU数据集[19]是一个大规模室内MVS数据集. 所有场景由49个视图在7种照明条件下扫描. 点云真值通过结构光扫描仪获取，并经过泊松表面重建和渲染生成训练所需的深度图[8]. 参考MVSNet[8]将数据集划分为79个训练场景、18个验证场景和22个评估场景，共有27 097张图像用于训练.

BlendedMVS数据集[37]是用于训练MVS网络的大规模合成数据集. 数据集划分为106个训练场景和7个验证场景，提供了超过17 000张训练图像.

Tanks & Temples数据集[20]包含真实照明条件下的室内和室外场景，分为包含8个场景的中级数据集和包含6个场景的高级数据集. 在Tanks & Temples上评估模型时，参照文献[9,15,17]使用BlendedMVS数据集对模型进行微调.

4.1.2 评价标准

为评估DTU数据集[19]上的重建效果，输入结构光点云真值和MVS重建点云，按照文献[19]计算精度、完整度和总体误差. 其中，精度定义为重建点云到点云真值的平均距离，完整度定义为点云真值到重建点云的平均距离，总体误差定义为精度和完整度的平均值. 为计算一个三维点到点云的距离，使用k-近邻算法搜索点云中与该三维点距离最小的一个点，并剔除最小距离大于20 mm的异常点.

本文还评估了DTU数据集[19]中深度不连续区域的重建精度. 深度不连续区域定义为图像中水平或垂直相邻像素深度值变化大于2 mm的像素. 评估指标采用柔性边缘误差(Soft Edge Error，SEE_k)[32,38]，其定义为预测深度值和局部$k\times k$邻域内深度真值的最小绝对误差，实验中取k=5.

为评估Tanks & Temples数据集[20]的重建效果，我们将MVS重建点云提交至在线排行榜，使用网站计算的平均F-分数作为评价指标，该指标反映了重建精度和完整度的平均效果.

4.2 实现细节

本文模型使用PyTorch深度学习框架实现，在DTU

训练集[19]上训练,相关代码将公开发布在 https://github.com/colorfulgreen/ASMVSNet. 训练时输入多视图图像数量 N=5,图像分辨率640×512,深度采样范围为[425 mm,935 mm],深度值稠密采样和稀疏采样数量分别为 H=192 和 K=16. 网络使用 AdamW 优化器和 OneCycleLR 策略训练,最大学习率为0.000 2. 首先使用 $\lambda_{ce}L_{ce}$ 训练5个轮次,预热特征提取和稀疏匹配代价体构建模块;随后使用 $\lambda_{ce}L_{ce}+\lambda_{d}L_{d}$ 联合训练27个轮次,得到基于GRU的稠密迭代优化模型;最后使用 $\lambda_{ce}L_{ce}+\lambda_{d}L_{d}+\lambda_{reg}L_{reg}$ 微调5个轮次,得到稀疏迭代优化模型. 训练过程使用2块Nvidia RTX 3090 GPU,设置批大小为8. 评估时,对输出深度图应用几何一致性检查[39]以滤除异常值,然后将其融合为三维点云.

4.3　性能基准测试

4.3.1　DTU数据集

本节在DTU评估集[19]上对比本文方法与现有主流MVS方法的重建效率. 评估时输入多视图图像数量 N=5,图像分辨率1 600×1 184,深度采样范围为[425 mm,935 mm],不特别说明时迭代优化次数 T=4. 所有方法的运行时间和显存占用在一块NVIDIA RTX 3090 GPU上测量. 统计运行时间时,首先执行200次预热运行并测量接下来200次运行的平均延时;显存占用使用nvidia-smi工具统计. 表1总结了定量对比结果,其中的粗体标出的数值表示最优结果,下划线表示次优结果,TransMVSNet[9]和CasMVSNet[13]评估时输入图像分辨率为1 152×864. 本文方法的运行速度相较对比方法更快,例如,迭代优化4次(T=4)时,运行速度比TransMVSNet[9]快3.6倍、比CasMVSNet[13]快1.2倍、比PatchmatchNet[17]快0.35倍、比IterMVS[15]快0.18倍、比Effi-MVSNet[16]快0.12倍,同时,显存占用仅次于PatchmatchNet[17],重建总体效果仅次于TransMVSNet[9]. 值得注意的是,在仅迭代优化一次(T=1)的快速版本中,本文方法仍然取得了优于CasMVSNet[13]和PatchmatchNet[17]等高效MVS方法的重建总体效果.

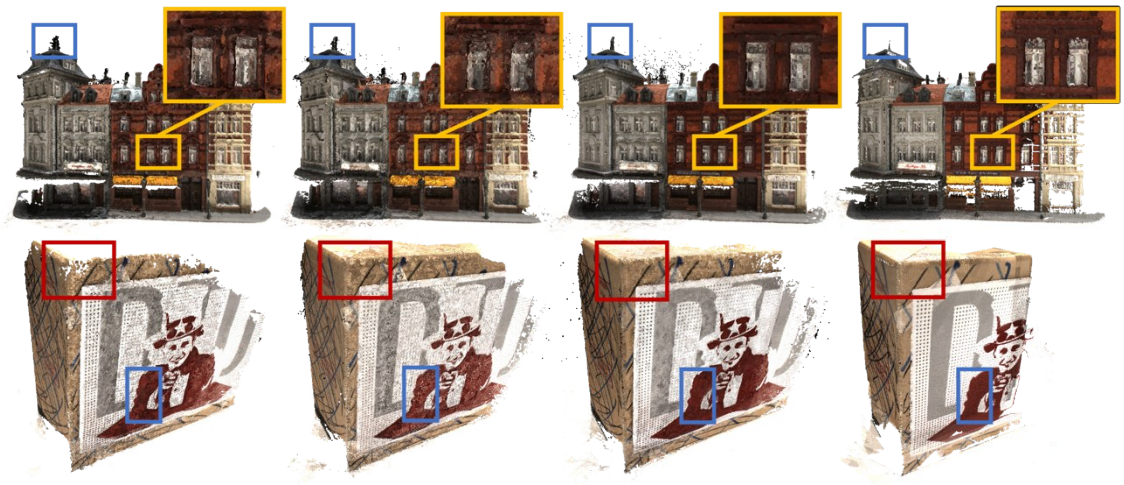
表1　在DTU评估集[19]上与现有主流MVS方法的定量对比结果

方法	精度/mm ↓	完整度/mm ↓	总体/mm ↓	时间/s ↓	显存/GB ↓
COLMAP[40]	0.400	0.667	0.532	—	—
MVSNet[8]	0.396	0.527	0.562	—	—
TransMVSNet[9]	<u>0.321</u>	0.289	**0.305**	0.79	4.24
CVP-MVSNet[41]	**0.296**	0.406	0.351	1.88	8.65
CasMVSNet[13]	0.325	0.385	0.355	0.37	4.46
UCS-Net[14]	0.338	0.349	0.344	0.71	5.79
CDS-MVSNet[42]	0.352	<u>0.280</u>	0.316	0.33	4.86
PatchmatchNet[17]	0.427	**0.277**	0.352	0.23	**3.59**
IterMVS[15]	0.373	0.354	0.363	0.20	4.68
Effi-MVSNet[16]	<u>0.321</u>	0.313	0.317	0.19	4.55
AS-MVSNet(T=1)	0.343	0.342	0.342	**0.14**	<u>4.15</u>
AS-MVSNet(T=4)	0.328	0.302	<u>0.315</u>	<u>0.17</u>	<u>4.15</u>

图4展示了本文方法与高效MVS方法中级联方法CasMVSNet[13]和块匹配方法PatchmatchNet[17]的重建点云比较. 如第一行的蓝色和黄色矩形框部分所示,在建筑物顶部的精细结构和物体边缘区域,本文方法的重建点云比CasMVSNet[13]和PatchmatchNet[17]包含更多精确细节. 第二行的红色矩形框部分进一步展示了在弱纹理等不适定区域,本文方法的重建点云具有更高的完整性. 表2展示了本文方法与CasMVSNet[13]和PatchmatchNet[17]在深度不连续区域的定量对比结果,表中可见,本文方法的平均柔性边缘误差较低,并且在所有误差阈值上占比更优. 产生上述结果的原因如下:

(1) 在深度不连续区域,级联方法CasMVSNet[13]在特征降采样后,单个像素处可能对应多个深度值真值,如果这些真值的跨度超过了后续精细阶段的深度采样范围,则导致模型陷入局部错误估计;随后尽管PatchmatchNet[17]在精细阶段引入了深度采样的随机扰动,但该扰动仍受局部采样范围限制. 相比之下,本文3.2节所构建的稀疏匹配代价体覆盖了完整的深度范围,其允许单个像素存在多个跨度较大的深度值估计,从而避免了采样范围受限引起的边缘伪影问题;此外在3.2.3节中,我们仅在最大匹配相似性的局部邻域计算深度值期望,进一步避免了由于深度在不连续区域呈多峰分布而导致的过度平滑问题[32].

(2) 在弱纹理等不适定区域,本文在3.1节根据摄像机先验设计了位置编码,并利用Transformer中视图内自注意力层和跨视图交叉注意力层增强特征,提升了网络对位置感知表征和视图间依赖关系的建模能力;此外,本文3.3.2节在门控循环单元中引入了逐像素采样偏移和调制系数的动态卷积,使其能够自适应聚合具有不同尺度和纹理丰富性的表征,从而具有更高



(a) CasMVSNet[13] (b) PatchmatchNet[17] (c) 本文方法 (d) 点云真值

图4 DTU评估集[19]场景13和场景15的重建点云比较

表2 DTU评估集[19]上深度不连续区域重建精度的定量对比结果

方法	柔性边缘误差(SEE_5)				
	平均/mm ↓	<1 mm /% ↑	<2 mm /% ↑	<4 mm /% ↑	<8 mm /% ↑
CasMVSNet[13]	19.65	49.02	60.49	69.41	75.88
PatchmatchNet[17]	10.50	54.59	66.74	76.01	83.07
AS-MVSNet	**8.63**	**60.75**	**72.23**	**80.10**	**85.61**

的重建完整度.

4.3.2 Tanks & Temples数据集

为验证模型的泛化能力,我们在Tanks & Temples数据集[20]榜单上提交重建点云验证效果. 评估时输入多视图图像数量N=9,图像分辨率为1 920×1 024,迭代次数为T=6,相机参数、深度采样范围和视图选择与R-MVSNet[10]一致. 表3展示了与主流MVS方法的比较结果,本文方法在中级和高级数据集上分别取得第二和最优的结果. 与CasMVSNet[13]和PatchmatchNet[17]相比,本文方法在中级数据集上的平均F-分数分别提高了3.1%和6.8%,在高级数据集的平均F-分数分别提高了6.3%和5.1%. 特别地,中级数据集上的最优方法TransMVSNet[9]使用了高计算复杂度的3D卷积,而本文仅使用2D卷积. 与TransMVSNet[9]相比,本文方法在中等数据集上平均F-分数降低了3.6%,但实验中测量的平均运行时间减少了83%.

图5进一步展示了本文方法在Tanks & Temples[20]中级数据集和高级数据集上的点云重建可视化效果. 即使在镜面反射、重复结构和光照变化等具有挑战性的场景中,本文方法仍然呈现出高质量的重建效果,验证了其在真实复杂场景中的有效泛化性.

表3 在Tanks & Temples[20]评估集的定量结果

方法	中级数据集F-分数/% ↑									高级数据集F-分数/% ↑						
	平均值	Fam.	Fra.	Hor.	L.H.	M60	Pan.	P.G.	Tra.	平均值	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP[40]	42.14	50.41	22.25	26.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
MVSNet[8]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69	—	—	—	—	—	—	—
R-MVSNet[10]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38	24.91	12.55	29.09	25.06	38.68	19.14	24.96
CasMVSNet[13]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
TransMVSNet[9]	**63.52**	**80.92**	**65.83**	**56.94**	**62.54**	**63.06**	**60.00**	**60.20**	**58.67**	37.00	24.84	**44.59**	34.77	46.49	**34.69**	36.62
PatchmatchNet[17]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
Effi-MVSNet[16]	56.88	72.21	51.02	51.78	58.63	58.71	56.21	57.07	49.38	34.39	20.22	42.39	33.73	45.08	29.81	35.09
AS-MVSNet	59.94	76.44	58.11	55.06	61.14	58.45	58.99	56.9	54.44	**37.37**	**28.07**	41.93	**35.76**	**47.08**	33.26	**38.14**

4.4 消融实验与分析

本节通过DTU评估集[19]上的消融实验验证AS-MVSNet各个模块的有效性,并分析样本自适应的稀疏迭代优化方法. 模型实现细节与4.2节相同,不特殊说明时,评估时输入多视图图像数量N=5,图像分辨率为1 600×1 184,深度采样范围为[425 mm,935 mm],深度

(*a*) 中级数据集

(*b*) 高级数据集

图5　Tanks & Temples基准数据集[20]上的点云重建可视化效果

图迭代优化次数T=4.

4.4.1　消融实验

消融实验结果如表4所示,其中每个实验独立测试了特定模块. 最终采用的方案用下划线标出. 下文将详细描述每个实验.

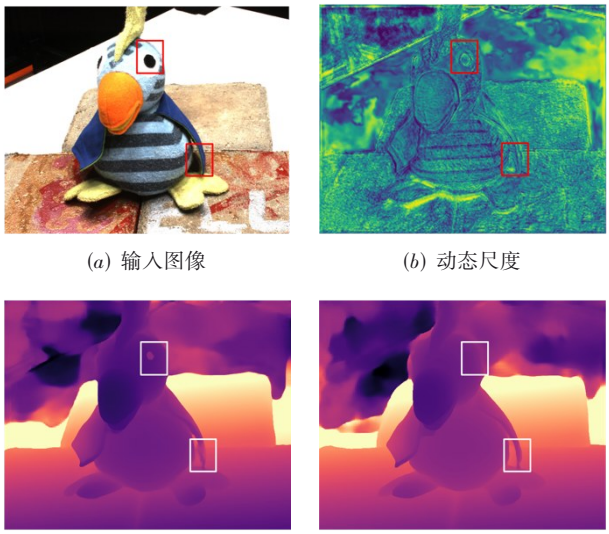
(1)特征提取　消融版本仅使用FPN提取多尺度匹配特征. 与消融版本相比,使用Transformer增强的FPN提取多尺度全局特征使重建总体误差从0.348 mm降低至0.315 mm. Transformer的全局感受野和摄像机几何位置编码将特征转换为更易于匹配的表示形式.

(2)稀疏匹配代价体　消融版本中使用$\{\boldsymbol{F}_i^1\}_{i=0}^{N-1}$直接构建输入图像1/4分辨率的稠密匹配代价体,其中深度采样数量设置为192. 如表4所示,使用稀疏匹配代价体使模型推理时间从0.25 s减少至0.17 s,减少了32%,同时点云重建效果与稠密匹配代价体相当.

表4　消融实验

实验	方法	精度/mm ↓	完整度/mm ↓	总体/mm ↓	时间/s ↓
特征提取	多尺度全局特征	**0.328**	**0.302**	**0.315**	0.17
	FPN	0.379	0.317	0.348	**0.12**
匹配代价体	稀疏匹配代价体	**0.328**	**0.302**	**0.315**	0.17
	常规匹配代价体	0.338	0.301	0.319	0.25
上下文特征尺度变换	尺度推理	**0.328**	**0.302**	**0.315**	0.17
	常规卷积	0.331	0.311	0.321	0.17
特征融合	交叉门控	**0.328**	**0.302**	**0.315**	0.17
	直接拼接	0.337	0.309	0.323	0.17
GRU隐状态更新	动态卷积	**0.328**	**0.302**	**0.315**	0.17
	常规卷积	0.332	0.315	0.324	0.17

(3)上下文特征尺度变换　消融版本中将上下文特征尺度推理模块替换为参数量相同的2个标准卷积层. 与消融版本相比,使用尺度推理将重建完整度的平均误差从0.311 mm降低至0.302 mm. 图6展示了上下文特征动态尺度推理的可视化效果,网络在无纹理区域和精细结构区域学习到内容自适应的最佳局部尺度,从而输出更加精细和准确的深度图.



(a) 输入图像　(b) 动态尺度

(c) 深度图(常规卷积)　(d) 深度图(尺度推理)

图6　上下文特征尺度推理变换

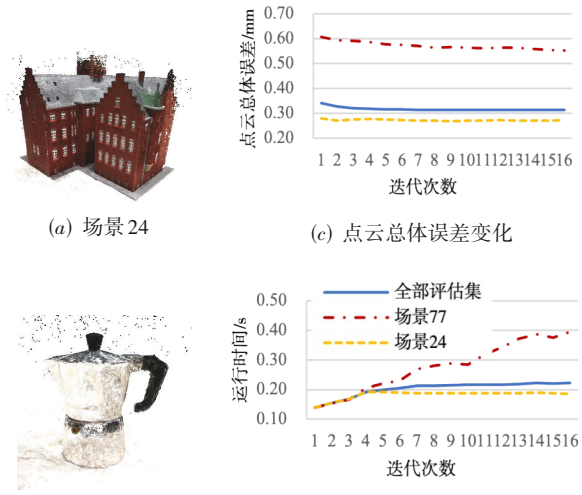
(4)基于交叉门控的特征融合　本文使用交叉门控融合上下文特征和几何特征,使得上下文特征提取器能够学习与多视图立体匹配任务相关的特定信息. 相比直接拼接特征的消融版本,使用交叉门控使总体误差从0.323 mm降低至0.315 mm.

(5)基于动态卷积的GRU隐状态更新　为了缓解深度值边缘的模糊问题,本文将GRU隐状态更新操作中的标准卷积替换为动态卷积. 相比标准卷积,使用动态卷积使总体误差从0.324 mm降低至0.315 mm.

4.4.2　样本自适应稀疏迭代优化分析

为验证样本自适应稀疏迭代优化模块的有效性,本节分析迭代中点云总体误差和模型运行时间的变化. 图7展示了分析结果. 随着迭代次数的增加,图7(*c*)中总体误差逐渐减小,图7(*d*)中运行时间以亚线性增长. 此外,场景77由于存在弱纹理和反光等不适定区域,总体误差需要经过多次迭代逐渐减小,运行时间相应增加;相比之下,场景24在4次迭代后,总体误差和运行时间趋于稳定. 通过对比场景77和场景24,可见本文方法能够自适应学习输入图像的重建难度,并动态调整运行时间.

图8展示了稀疏迭代优化中的深度图估计、误差图和Dropout掩膜的变化. 误差图中白色区域误差较大,Dropout掩膜中白色区域为Dropout区域. 第1行展示了随着迭代次数的增加,深度图精度逐步提高. 第2行的误差图显示,平滑区域内部的误差在第2次迭代后趋于稳定,而深度边缘和精细结构需要更多次迭代达到稳定. 第3行的Dropout掩膜展示了每个迭代步中网络学习到的剪枝区域和误差图中误差较小的区域一致,并且剪枝区域逐步扩大,验证了稀疏迭代优化的有效性.



(a) 场景24　(c) 点云总体误差变化

(b) 场景77　(d) 运行时间变化

图7　不同稀疏迭代次数的点云总体误差和运行时间变化

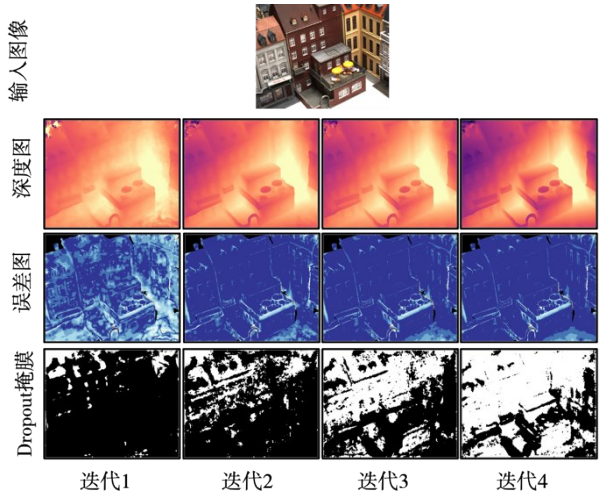


图8　稀疏迭代优化中深度图、误差图和Dropout掩膜变化

5　结论

本文针对现有多视图立体匹配网络计算复杂度高的问题,提出了一种基于自适应空间稀疏化的高效多视图立体匹配网络. 网络从立体匹配的空间稀疏性出发,分别构造深度空间和图像空间的稀疏计算,从而提高了推理效率. 在公开数据集上的实验结果表明,所提

方法有效减少了模型推理时间,同时重建效果和泛化性能表现优异. 本文方法适用于计算资源有限的嵌入式平台上的三维重建任务. 本文方法可进一步与其他现代的结构化稀疏神经网络技术相结合,以利用矢量处理架构进行硬件加速.

参考文献

[1] 李博洋, 刘思健, 崔明月, 等. 基于最小回环检测的多车协同SLAM框架[J]. 电子学报, 2021, 49(11): 2241-2250.
LI B Y, LIU S J, CUI M Y, et al. Multi-vehicle collaborative SLAM framework for minimum loop detection[J]. Acta Electronica Sinica, 2021, 49(11): 2241-2250. (in Chinese)

[2] 金紫凤, 潘思聪, 危辉. 可变环境下基于位姿变换矩阵的机器人无标定手眼协调方法[J]. 电子学报, 2022, 50(10): 2318-2328.
JIN Z F, PAN S C, WEI H. Uncalibrated hand eye coordination method for robot based on pose transformation matrix in variable environment[J]. Acta Electronica Sinica, 2022, 50(10): 2318-2328. (in Chinese)

[3] 樊亚红, 刘宾, 陈平, 等. 基于轮廓先验约束的复杂异形工件CT成像方法研究[J]. 电子学报, 2020, 48(10): 1976-1982.
FAN Y H, LIU B, CHEN P, et al. Research on CT imaging method of complex shaped workpiece based on contour prior constraint[J]. Acta Electronica Sinica, 2020, 48(10): 1976-1982. (in Chinese)

[4] SEITZ S M, DYER C R. Photorealistic scene reconstruction by voxel coloring[C]//Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2002: 1067-1073.

[5] FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362-1376.

[6] SCHÖNBERGER J L, FRAHM J M. Structure-from-motion revisited[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 4104-4113.

[7] XU Q S, TAO W B. Multi-scale geometric consistency guided multi-view stereo[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 5478-5487.

[8] YAO Y, LUO Z X, LI S W, et al. MVSNet: Depth inference for unstructured multi-view stereo[C]//Computer Vision—ECCV 2018. Cham: Springer International Publishing, 2018: 785-801.

[9] DING Y K, YUAN W T, ZHU Q T, et al. TransMVSNet: Global context-aware multi-view stereo network with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8575-8584.

[10] YAO Y, LUO Z X, LI S W, et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 5520-5529.

[11] CHEN R, HAN S F, XU J, et al. Point-based multi-view stereo network[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 1538-1547.

[12] YU Z H, GAO S H. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1946-1955.

[13] GU X D, FAN Z W, ZHU S Y, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2492-2501.

[14] CHENG S, XU Z X, ZHU S L, et al. Deep stereo using adaptive thin volume representation with uncertainty awareness[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2521-2531.

[15] WANG F, GALLIANI S, VOGEL C, et al. IterMVS: Iterative probability estimation for efficient multi-view stereo [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8596-8605.

[16] WANG S Q, LI B, DAI Y C. Efficient multi-view stereo by iterative dynamic cost volume[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8645-8654.

[17] WANG F, GALLIANI S, VOGEL C, et al. PatchmatchNet: Learned multi-view patchmatch stereo[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 14189-14198.

[18] JIANG S H, LU Y, LI H D, et al. Learning optical flow from a few matches[C]//2021 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 16587-16595.

[19] AANÆS H, JENSEN R R, VOGIATZIS G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.

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

[21] TORSTEN H, DAN A, TAL B N, et al. Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks[J]. Journal of Machine Learning Research, 2021, 22(241): 1-124.

[22] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15: 1929-1958.

[23] KINGMA D P, SALIMANS T, WELLING M. Variational dropout and the local reparameterization trick[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems — Volume 2. New York: ACM, 2015: 2575-2583.

[24] MOLCHANOV D, ASHUKHA A, VETROV D. Variational dropout sparsifies deep neural networks[C]//Proceedings of the 34th International Conference on Machine Learning - Volume 70. New York: ACM, 2017: 2498-2507.

[25] FAN, X J, ZHANG, S J, TANWISUTH, K, et al. Contextual dropout: An efficient sample-dependent dropout module[C]//Proceedings of the 9th International Conference on Learning Representations. Appleton: ICLR, 2021: 1-12.

[26] SEMENIUTA S, SEVERYN A, BARTH E. Recurrent dropout without memory loss[C]//Proceedings of the 26th International Conference on Computational Linguistics. Stroudsburg: ACL, 2016: 175-1766.

[27] LOBACHEVA E, CHIRKOVA N, MARKOVICH A, et al. Structured sparsification of gated recurrent neural networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 4989-4996.

[28] KATHAROPOULOS A, VYAS A, PAPPAS N, et al. Transformers are RNNs: Fast autoregressive transformers with linear attention[C]//Proceedings of the 37th International Conference on Machine Learning - Volume 119. Cambridge: JMLR, 2020: 5156-5165.

[29] COLLINS R T. A space-sweep approach to true multi-image matching[C]//Proceedings of 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2002: 358-363.

[30] XU Q S, TAO W B. Learning inverse depth regression for multi-view stereo with correlation cost volume[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12508-12515.

[31] CHEN R, HAN S F, XU J, et al. Visibility-aware point-based multi-view stereo network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43 (10): 3695-3708.

[32] CHEN C R, CHEN X Z, CHENG H. On the over-smoothing problem of CNN based disparity estimation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 8996-9004.

[33] KIM S, KIM S, MIN D B, et al. LAF-net: Locally adaptive fusion networks for stereo confidence estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 205-214.

[34] TEED Z, DENG J A. RAFT: Recurrent all-pairs field transforms for optical flow[C]//Computer Vision—ECCV 2020. Cham: Springer International Publishing, 2020: 402-419.

[35] LI M Y, LIN J, MENG C L, et al. Efficient spatially sparse inference for conditional GANs and diffusion models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14465-14480.

[36] JIAO J B, CAO Y, SONG Y B, et al. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss[C]//Computer Vision—ECCV 2018. Cham: Springer International Publishing, 2018: 55-71.

[37] YAO Y, LUO Z X, LI S W, et al. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1787-1796.

[38] TOSI F, LIAO Y Y, SCHMITT C, et al. SMD-Nets: Stereo mixture density networks[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 8938-8948.

[39] YAN J F, WEI Z Z, YI H W, et al. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking[C]//Computer Vision—ECCV 2020. Cham: Springer International Publishing, 2020: 674-689.

[40] SCHÖNBERGER J L, ZHENG E L, FRAHM J M, et al.

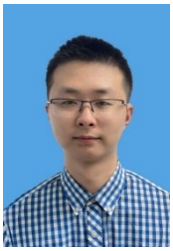
Pixelwise view selection for unstructured multi-view stereo[C]//Computer Vision—ECCV 2016. Cham: Springer International Publishing, 2016: 501-518.

[41] YANG J Y, MAO W, ALVAREZ J M, et al. Cost volume pyramid based depth inference for multi-view stereo[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4876-4885.

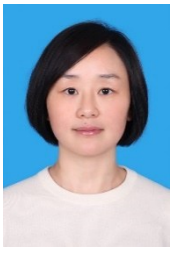
[42] GIANG K T, SONG S, JO S. Curvature-guided dynamic scale networks for multi-view stereo[C]//Proceedings of the 10th International Conference on Learning Representations. Appleton: ICLR, 2022: 1-16.

作者简介

周晓清　女,1989年3月出生于河南省洛阳市. 现为北京航空航天大学计算机学院硕士研究生. 主要研究方向为三维视觉.
E-mail: zhouxiaoqing@buaa.edu.cn



王　翔　男,1995年2月出生于浙江省宁波市. 现为北京航空航天大学计算机学院博士研究生. 主要研究方向为三维视觉. 中国电子学会会员编号:E190002331S.
E-mail: vixwang@buaa.edu.cn



郑　锦　女,1978年10月出生于四川省乐山市. 现为北京航空航天大学计算机学院副教授、博士生导师. 主要研究方向为计算机视觉、视频图像处理等.
E-mail: JinZheng@buaa.edu.cn

百　晓(通讯作者)　男,1979年3月出生于甘肃省兰州市. 现为北京航空航天大学计算机学院教授、博士生导师. 主要研究方向为计算机视觉、模式识别等. 中国电子学会会员编号:E190010986M.
E-mail: baixiao@buaa.edu.cn