ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 4,Jul. 2025,pp. 1023－1034
DOI:10. 16337/j. 1004-9037. 2025. 04. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于扩张注意力与深度最优化校正的多视图三维重建网络

徐　蕾[1]，雷有元[1]，朱　军[1]，周　杰[1]，邵根富[2]，张家铭[1]

(1. 南京信息工程大学电子与信息工程学院,南京 210044；2. 杭州电子科技大学自动化学院,杭州 310018)

摘　要：与CVP-MVSNet网络和CasMVSNet网络相比,MVSNet重建网络存在的内存消耗量问题降低了模型处理高分辨率图像时的内存消耗量以及重建点云的准确性误差,但是两者点云的完整性误差却很大。针对此问题,本文提出了基于扩张注意力与深度最优化校正的多视图三维重建网络DA-MVSNet。DA-MVSNet是以CasMVSNet作为基准网络,额外引入一个融合了深度可分离卷积的并行空洞卷积与注意力模块构成的特征增强网络,增强了重建网络对输入视图的全局特征捕获能力,提升了重建点云的完整度。为进一步提升输出深度图的精度,防止特征增强网络提取过多的视图非相关背景信息导致重建点云准确度的下降,在网络的输出部分还引入了一个基于非线性最小二乘的最优化校正机制模块。结果表明,DA-MVSNet重建网络在室内场景数据集DTU上运行得到的重建点云的准确性误差与完整性误差分别降低了2.5%和4.7%,具有较好的综合性能。但也由于额外引入了增强网络和校正机制,其内存和时间消耗均约高于CVP-MVSNet与CasMVSNet网络。

关键词：深度学习；三维重建；注意力机制；空洞卷积；最优化校正

中图分类号：TP391.41　　**文献标志码**：A

Multi-view 3D Reconstruction Network Based on Dilated Attention and Depth Optimal Correction

XU Lei[1], LEI Youyuan[1], ZHU Jun[1], ZHOU Jie[1], SHAO Genfu[2], ZHANG Jiaming[1]

(1. School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China;2. School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: The memory consumption issue in MVSNet reconstruction networks, compared with CVP-MVSNet and CasMVSNet networks, reduces memory usage when processing high-resolution images and improving the accuracy of reconstructed point clouds. However, both networks still exhibit significant errors in point cloud completeness. To address this issue, this paper proposes DA-MVSNet, a multi-view 3D reconstruction network based on dilated attention and depth optimal correction. DA-MVSNet uses CasMVSNet as the baseline network, with an additional feature enhancement network that integrates a parallel dilated convolution and attention module, incorporating the concept of depth-wise separable convolutions. This enhancement strengthens the network's ability to capture global features of input views, improving point cloud completeness. To further enhance the accuracy of output depth maps and prevent the feature enhancement network from extracting irrelevant background information, which can degrade the accuracy of the reconstructed point cloud, an optimization correction mechanism based on

基金项目:国家自然科学基金青年项目(62101275);国家自然科学基金面上项目(61971167)。
收稿日期:2024-10-07;**修订日期**:2025-03-29

nonlinear least squares is introduced at the output stage of the network. The results show DA-MVSNet reduces the accuracy and completeness errors of the reconstructed point cloud by 2.5% and 4.7%, respectively, on the indoor scene DTU dataset, achieving better overall performance. However, due to the additional feature enhancement network and correction mechanism, the memory and time consumption of DA-MVSNet are not very higher than those of CVP-MVSNet and CasMVSNet.

Key words: deep learning; 3D reconstruction; attention mechanisms; void convolution; optimal correction

引　言

近年来深度学习网络的发展推动了计算机视觉诸多领域的进步。研究人员为克服传统多视图三维重建方法存在的局限性,将深度学习网络引入到三维视图重建领域,取得了丰硕的成果。2017年Ji等[1]提出了基于体素的多视图三维重建网络SurfaceNet模型,利用深度学习网络学习场景的表面结构特征实现三维视图重建。同年Charles等[2]提出了基于点云的三维重建网络,同样基于深度学习的方法可对重建网络进行训练以输出理想的视图重建结果。Yao等[3]在2018年提出了多视图三维重建网络MVSNet,该网络利用可微单应性变换将相机几何耦合在深度学习网络中,实现了端到端的深度估计,以及实现了基于融合算法生成三维点云。2019年Yao等为进一步改善MVSNet内存消耗过大的问题,提出了R-MVSNet网络[4]。2020年Luo等[5]提出的P-MVSNet网络在先前模型的基础上额外引入基于各向同性和各向异性的三维卷积对输入视图进行特征聚合与深度预测,旨在进一步改进重建点云质量。其后Yi等[6]提出的PVA-MVSNet网络引入自适应视图聚合的代价体构建方案,旨在提高重建点云的完整度。Yang等[7]在提出的CVP-MVSNet网络中引入“由粗到细”的深度评估结构,并进行了基于多尺度特征图进行深度图的迭代计算,能有效地提升重建点云的准确度,但其重建点云的完整度表现较差。在Gu等[8]提出的CasMVSNet与CVP-MVSNet网络中也拥有相似的网络结构,因此该网络在重建点云的完整度表现上同样存在不足。Zhang等[9]提出的Vis-MVSNet网络主要针对视图中遮挡像素的重建完整度进行了改善,在代价体构建的部分增强了对被遮挡像素的特征表达。因此该网络为提升重建点云的整体质量提供一个优秀的方案,但与同时代提出的其他方法相比均未取得更好的表现。2021年Wang等[10]提出了PatchmatchNet网络,它将传统的Patchmatch算法引入到重建网络中,使得PatchmatchNet网络拥有出色的重建效率及重建点云完整度,但是准确性误差却显著增加。后来Wei等[11]提出了AA-RMVSNet网络,该网络主要为一种自适应的代价体构建方式,目标旨在提升模型对输入视图弱纹理区域的重建质量。针对重建点云完整度的提升,在2022年Gao等[12]以牺牲重建点云的准确度为代价,提出了MSCVP-MVSNet网络,它使得重建点云在完整度上的表现尤为出色。同年Cao等[13]也针对重建点云的完整度,提出了MVSFormer网络。在该网络模型中引入了预训练的特征提取网络,以其牺牲重建的准确度生成了完整度更高的三维点云模型。在2023年Liu等[14]提出的ET-MVSNet网络基于线对点的非局部增强策略,能有效提升重建点云的完整度,并通过实验数据证明了其引入策略的有效性。

虽然近年来基于深度学习的多视图三维重建方法取得了显著进步,但其中设计方案在提升某些方面的视图表现,例如完整度与准确度时,都无可避免地牺牲了其他方面的质量要求,常常导致重建点云的整体质量较差以及重建网络的综合性能较低。由此本文提出了基于扩张注意力(Dilated attention, DA)与深度最优化校正相结合的多视图三维重建网络DA-MVSNet。DA-MVSNet是以CasMVSNet作为基准网络,沿用了其级联形式的分阶段深度评估结构。为增强重建网络对输入视图的全局特征捕获能力和提升重建点云的完整度,网络中额外引入一个融合了深度可分离卷积的并行空洞卷积与注意

力模块构成的特征增强网络。为防止特征增强网络提取过多的视图非相关背景信息导致重建点云准确度的下降,在网络的输出部分还引入了一个基于非线性最小二乘可对深度图进行最优化的校正机制模块,能提升输出深度图的精度且避免了深度图融合后重建点云准确性误差的增大。

1　MVSNet与CasMVSNet网络

MVSNet是把深度学习与基于深度图方法的多视图三维重建技术相结合的方法,它实现了一种由端到端的深度估计算法网络,通过深度图融合实现了视图的三维重建[15]。MVSNet的基准流程分别为:特征提取、代价体构建、代价体正则化以及深度回归流程。在特征提取模块中以多张源视图及参考视图作为输入,利用卷积神经网络进行细微特征提取。在代价体构建模块中采用了基于平面扫描算法为每个像素有效生成均匀分布的深度假设,再利用单应性变换将源视图特征图投影至参考视图的坐标空间上,实现了在每个深度假设下计算视图特征图之间的相似度进而构建成代价体。针对代价体进行正则化的目的是将代价体中离散的相似度信息转换为连续的概率值信息,其过程主要使用基于Softmax激活函数构建和实现三维卷积网络层。此代价体也被称为概率体,其体素值表示各深度假设接近深度真值的概率。如果对每个像素沿概率体深度假设方向求其期望值,即可得网络预测的深度图,此过程也被称为深度回归。在深度图输出后还需进行一个深度残差网络对深度图优化,使其具备更好的精度。但是研究发现MVSNet内存消耗较大,因此其改进型CasMVSNet网络是在基础流程中引入了多尺度特征提取网络和级联的深度评估结构,但逐个阶段仍将遵循MVSNet的流程规律。CasMVSNet视图重建网络模型如图1所示,在CasMVSNet网络结构中的多尺度特征提取网络可以对特征图进行跨尺度输出,进而可以提升重建网络对视图局部特征的捕获能力[16]。另外级联的深度评估结构用于降低模型在处理高分辨率输入视图时的内存消耗量,为每个阶段输出的深度图上采样后,作为后续阶段的深度先验信息配合多尺度特征图对深度假设的采样范围及采样间隔进行收敛。因此,收敛的采样范围就可生成较少的深度假设数量,降低代价体构建中所消耗的内存量,从而提升CasMVSNet的视图重建效率。

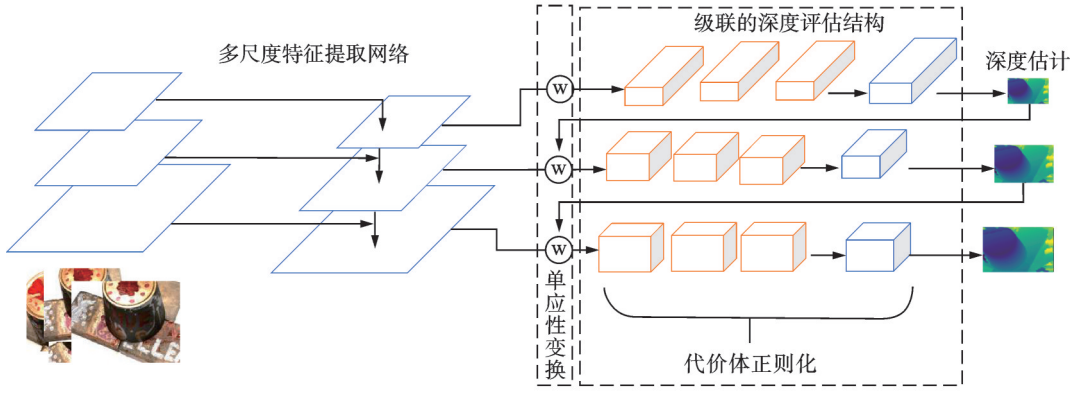

图1　CasMVSNet视图重建网络模型

Fig.1　CasMVSNet multi-view reconstruction network model

2　算法改进

如图1所示,CasMVSNet的特征提取网络是基于FPN(Feature pyramid network)[17]自上向下的金字塔结构对输入的源视图和参考视图进行多尺度特征提取。在模型结构中,低分辨率的高层特征图包含了更丰富的对象类别与结构等语义信息,高分辨率的低层特征图则拥有更丰富的目标位置、空间布

局等定位信息。基于FPN结构的特征提取网络虽然可以将低分辨率高层特征图的强语义信息通过上采样的方式进行传递，以增强各尺度特征图的语义信息表达，但该结构并未对逐个特征图层的全局定位信息进行增强，所以上述特征提取网络只注重高层特征对低层特征的语义信息加强，却忽略了低层特征对高层特征的定位信息补充，并且高层的语义特征在传递过程中也会被损耗。同时受限于二维卷积的感受野，特征图表达更多的是局部语义信息。此特征提取网络提取到的全局特征信息越多，网络对输入视图的结构信息则能越充分地表达，进而重建出更多的场景几何表面结构。因此，基于FPN结构的特征提取网络对输入视图的全局特征信息捕获能力相对有限，会导致重建点云丢失较多的表面信息。

综上分析，DA-MVSNet网络中融合了深度可分离卷积的并行空洞卷积与注意力模块，搭建了其特征增强网络。它对特征提取网络输出的多尺度特征图作了进一步的全局特征补充，为后续模块提供更丰富的场景结构信息，正向地增加重建点云的表面数量以降低完整性误差，然而也会增加重建网络对视图中背景信息的特征提取。对于重建三维点云的输入视图而言，其物体以外的背景特征信息是无用的，这会降低深度图的精度导致重建点云准确性误差增加，因此在DA-MVSNet重建网络的输出部分，引入了一个基于非线性最小二乘可对深度图进行最优化的校正机制模块，由此可获得深度值校正的最优解。

3 DA-MVSNet网络

3.1 网络模型

DA-MVSNet的整体网络模型如图2所示，其主要分为4个部分：特征提取网络、特征增强网络、级联的深度评估结构以及深度图校正网络。

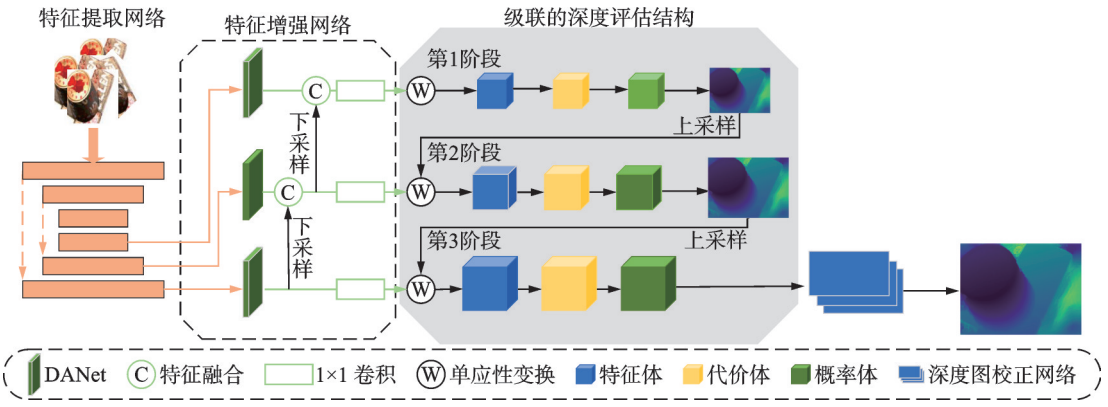

图2 DA-MVSNet整体网络模型

Fig.2 DA-MVSNet reconstruction network model

DA-MVSNet网络各部分基本组成结构及主要任务为：特征提取网络依然基于FPN的金字塔结构，对输入参考视图和源视图进行多尺度特征提取，并将各尺度的初始特征图输入到特征增强网络；特征增强网络内部包括3个相同的网络层，每个网络层即由融合了深度可分离卷积的并行空洞卷积与注意力模块组建而成，因此将该网络层简称为DANet。每个DANet网络层对输入的初始特征图进行处理，增强其全局特征信息的表达能力，并且以自下向上的特征融合方式将处理后的特征图输入到级联的深度评估结构；级联的深度评估结构沿用了CasMVSNet的网络结构设计，其每个阶段均包括代价体构建、代价体正则化和深度估计，各阶段基于输入特征图以及深度先验信息（除第1阶段）对深度图进行迭代计算；深度图校正网络对第3阶段深度评估结构输出的深度图进行最优化校正后，再作为重建网络

的最终输出。由于本文提出的端到端重建网络是基于深度图的多视图立体匹配网络,因此还需通过后处理对深度图进行融合,生成三维视图重建点云。

3.2　基于深度可分离空洞卷积与注意力机制的特征增强网络

在DA-MVSNet引入的特征增强网络旨在捕获更多输入视图的全局特征信息。特征增强网络内部由3个并行的DANet网络层组成,每个DANet的输出部分又以自下向上的方式进行特征融合,作为后续级联深度评估结构各阶段的输入信息。DANet内部网络层结构如图3所示,其中C、H、W分别表示通道数,以及通道的高度和宽度。

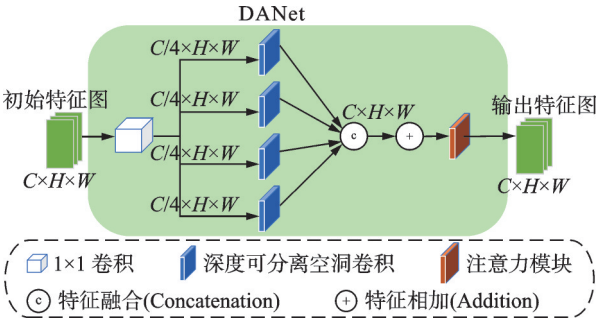

图3　DANet网络层内部结构

Fig.3　Internal structure of DANet reconstruction network

在DANet中主要由一组并行的空洞卷积与注意力模块组成,并且在初始特征图输入的部分,利用1×1卷积将输入特征图的特征通道降维成原来的1/4。在如此4个并行的空洞卷积输出的特征图进行特征融合后,无需再进行通道降维即可输入到注意力模块中进行处理。在其内部引入了4个3×3的并行空洞卷积,通过设置不同的空洞率增加卷积核的感受野进而可为初始特征图捕获更多的跨尺度特征信息。为保留初始特征图自身存在的局部特征信息,还可将其中一个空洞卷积的空洞率设置为1,即退化为普通的3×3卷积。因此在每个DANet网络层中,各空洞卷积由上至下的空洞率分别设置为1、4、6和8。

考虑到并行空洞卷积带来的额外计算量与参数量会增加网络的冗余度,将深度可分离卷积与空洞卷积相融合,将其分解为逐个通道上的空洞卷积操作与逐点卷积操作。将标准卷积过程进行分解,可以显著减少网络的参数量与计算量。如图4所示,DANet网络层内部的每个深度可分离空洞卷积网络结构。融合深度可分离卷积的空洞卷积简称为深度可分离空洞卷积。

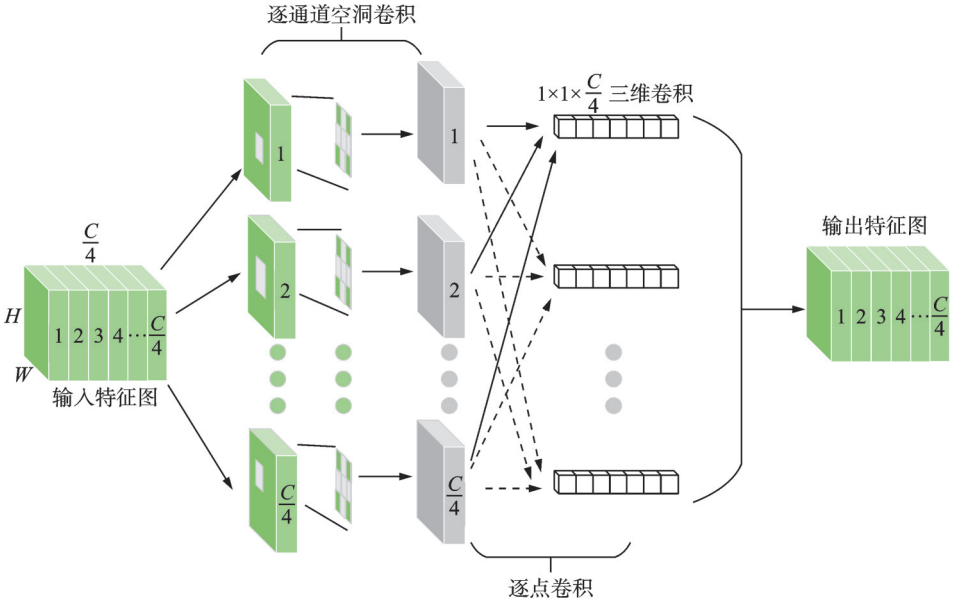

图4　深度可分离空洞卷积网络结构

Fig.4　Depth-wise separable convolutions network structure

结合图3和图4,假设初始特征图为F,其尺度信息为[C, H, W],DANet在并行的深度可分离空洞卷积部分对初始特征图的处理流程可表示为

$$F'=\mathrm{Conv}_{1\times1}(F) \tag{1}$$

$$\{F_k\}_{k=1}^{4}=\{\mathrm{DWDConv}_{3\times3}^{r_k}(F')\}_{k=1}^{4} \tag{2}$$

$$F_{\mathrm{out}}=F\oplus\mathrm{concat}(F_1,F_2,F_3,F_4) \tag{3}$$

式中:F'表示初始特征图F经1×1卷积处理后,其通道降维成原来的1/4;F_k表示F'输入到空洞率分别为$r_1=1$、$r_2=4$、$r_3=6$和$r_4=8$的并行深度可分离空洞卷积后输出的特征图;DWD表示深度可分离空洞卷积模块;concat(F_1, F_2, F_3, F_4)表示对各输出特征图进行特征融合;“⊕”表示特征相加,可将深度可分离空洞卷积特征融合后的结果与F进行特征相加即得到空洞卷积部分输出的特征图F_{out}。

空洞卷积通过设置空洞率来扩大感受野,可以为输入的初始深度图捕获更多的全局特征信息,然而特征融合后得到的特征图可能会存在一些冗余信息或非相关特征,这会降低特征图的质量。综上所述,可额外引入一个模块,对深度可分离空洞卷积处理后的特征图在通道维度上进行特征的权重处理,弱化非相关特征,最大程度地保留重要信息。因此DANet网络层在并行的深度可分离空洞卷积后耦合了一个注意力模块,可用于处理特征通道。若耦合的注意力模块选择的复杂度过高,依然会为重建网络带来较大的开销,因此研究中选择了ECA-Net[18]中的轻量级通道注意力模块对特征图通道进行处理。

3.3 级联的深度评估结构

DA-MVSNet级联的深度评估结构沿用了CasMVSNet的网络设计,对深度图进行由粗到细的迭代计算。每个阶段的深度评估结构均包括代价体构建、代价体正则化以及深度估计,交替阶段中上阶段输出的深度图上采样后作为先验信息对后续阶段的深度假设区间进行收敛,以便实现更准确的深度估计[19]。因此DA-MVSNet沿用了CasMVSNet级联组合的深度评估结构,旨在减少高分辨率代价体构建带来的复杂计算量,并且促进深度图正向收敛,实现逐阶段“由粗到细”地提升深度估计值的准确度。

以上实现的过程主要体现在深度评估结构交替的阶段,具体执行过程为:先前阶段输出深度图通过双线性插值进行上采样,作为当前阶段的深度先验信息;同时当前阶段依据深度先验信息对深度假设采样区间进行收敛,并缩小深度采样间隔,旨在保证深度估计精度的同时降低代价体构建的成本量[20-21]。对于当前阶段的深度评估结构而言,其深度采样区间以及采样间隔的计算方法为

$$\begin{cases}R_{k+1}=R_k\cdot\omega\\ \Delta r_{k+1}=\Delta r_k\cdot\varepsilon\end{cases} \tag{4}$$

式中:R_{k+1}和Δr_{k+1}分别表示当前阶段的深度假设采样区间以及采样间隔;ω和ε分别表示0到1区间的比例因子。另外在每个阶段的深度评估结构均执行相同的代价体构建、代价体正则化以及深度估计,最后输出重建网络预测的深度图。

3.4 基于非线性最小二乘思想的深度校正网络

如前所述,在DANet中引入特征增强网络虽然可以增强输入视图特征图的全局特征表达能力,但是也会导致非相关区域特征的增强。为避免非相关区域对输出深度图精度的影响,在DA-MVSNet的输出部分增加引入了一个基于非线性最小二乘算法的最优化深度校正模块。首先假设在输出的深度图已经具有了较高的精度,可将参考特征图中某一个像素点p利用深度图信息投影到第i张源视图特征图得到投影像素点,其像素点为

$$p_i=K_i\cdot(\boldsymbol{R}_{\mathrm{ref},i}\cdot(K_{\mathrm{ref}}^{-1}\cdot p\cdot D(p))+\boldsymbol{t}_{\mathrm{ref},i}) \tag{5}$$

式中:K_i和K_{ref}分别表示第i张源视图与参考视图的相机内参;$\boldsymbol{R}_{\mathrm{ref},i}$表示参考视图相对于第$i$张源视图的旋转矩阵;$\boldsymbol{t}_{\mathrm{ref},i}$表示平移矩阵,$D(p)$表示像素点$p$的深度估计值。若预测深度图具有较高的精度,则参考

视图上的像素点p与其在源视图特征图上的投影像素点的特征差值应该是无限小的。基于以上讨论,可令参考视图特征图将像素点p在所有源视图特征图上进行投影,并对所有的特征差值进行求和。对于像素点p而言,输出深度图的精度越高,其求和特征差值就应越小。反而言之,若想对输出的深度图精度进行校正,则是对参考视图特征图上每个像素点的投影差值进行最优化处理,即求最小值。因此将投影差值之和视作求解非线性最小二乘算法的最优化问题,可以将参考视图特征图上像素点p的特征信息$F_{\mathrm{ref}}(p)$视作观测数据,其在源视图特征图上的投影点特征信息$F_i(p_i)$视作基于深度图生成的预测数据。在所有源视图特征图投影点上求得二者差值平方和的最小值,即可实现对像素点p的深度值校正。对参考视图特征图的所有像素点均进行以上操作,则可实现对预测深度图的精度校正,这也是最优化校正网络的核心思想。以参考视图特征图上的像素点p为例,结合式(5)和以上所述,其基于预测深度图在所有源视图上的投影点特征差值平方和表示为

$$F(p)=\sum_{i=1}^{N-1}\left(F_{\mathrm{ref}}(p)-F_i\left(K_i\cdot\left(\boldsymbol{R}_{\mathrm{ref},i}\cdot\left(K_{\mathrm{ref}}^{-1}\cdot p\cdot D(p)\right)+\boldsymbol{t}_{\mathrm{ref},i}\right)\right)\right)^2 \tag{6}$$

将参考视图特征图上像素点p的初始深度值定义为$D(p)+\Delta D$,利用非线性最小二乘优化算法迭代计算求得ΔD的最优解使得式(6)达到最小值。由于ΔD的初始值设置为0,通过迭代优化法求得ΔD最优解即为深度值的最优校正值,其与初始深度值相加可实现对深度值的校正。

求解可借助经典的高斯-牛顿(Gauss-Newton)迭代优化算法,定义一个残差向量$\boldsymbol{\theta}_i(p)$,目标是找到一组残差向量求得ΔD解使得式(6)达到最小值,其中残差向量$\boldsymbol{\theta}_i(p)$为

$$\boldsymbol{\theta}_i(p)=F_{\mathrm{ref}}(p)-F_i(p_i) \tag{7}$$

基于$\boldsymbol{\theta}_i(p)$对$D(p)$求一阶导数,表示为

$$J_i(p)=\frac{\partial(F_i(p_i))}{\partial p_i}\cdot\frac{\partial p_i}{\partial D(p)} \tag{8}$$

由式(8)可求得ΔD最优解,即参考视图像素点p深度值校正的优化偏移值,表示为

$$D_{\mathrm{out}}(p)=D(p)-\left[(\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J})^{-1}\boldsymbol{J}^{\mathrm{T}}\boldsymbol{\theta}\right]_k \tag{9}$$

式中:$\boldsymbol{\theta}\sim\{\theta_1(p),\theta_2(p),\cdots,\theta_{N\text{-}1}(p)\}$表示像素点$p$在各源视图中的残差向量;$\boldsymbol{J}\sim\{J_1(p),J_2(p),\cdots,J_{N\text{-}1}(p)\}$表示由式(9)对每张源视图基于残差向量对$D(p)$求一阶导数构成的雅可比矩阵;$D_{\mathrm{out}}(p)$即为经过第$k$次迭代计算后,像素点$p$校正后的深度值。对参考视图特征图上的所有像素点均进行以上迭代,即可实现对输出深度图的最优化校正,提高重建网络生成三维点云的准确度。

3.5　损失函数

基于级联的深度评估结构生成深度图,并在网络的输出部分引入深度图校正模块,因此本文分为两个部分对模型的损失进行计算。

在级联深度评估结构的输出部分,基于概率体对各像素沿深度方向加权求和作为其估计的深度值,同时考虑到各阶段输出的深度图精度是迭代增加的,因此在该部分计算每个阶段输出深度图与真实深度图之间的Smooth L_1损失函数,用于鼓励生成深度值的平滑性,并且也为每个阶段的损失函数均赋予不同的权重[22]。Smooth L_1损失函数的计算方法为

$$f(x)=\begin{cases}0.5x^2 & |x|<1\\ |x|-0.5 & |x|\geqslant 1\end{cases} \tag{10}$$

因此第1部分的损失计算表示为

$$\mathrm{Loss}_1=\sum_{k=1}^{3}\lambda_k\sum_{p}\|D_k(p)-D_{\mathrm{GT}}(p)\|_1 \tag{11}$$

式中：$\|\cdot\|_1$表示Smooth L_1损失计算；k表示深度评估结构的阶段数；λ_k表示每个阶段损失函数计算的权重值；$D_k(p)$表示估计深度图中各像素的深度估计值；$D_{GT}(p)$为其对应的深度真值，由真值深度图提供。每个阶段使用的深度真值图均与其输出的估计深度图尺度信息相同。

在深度图的校正部分，还需将优化后输出的深度图$D_{out}(p)$与真值深度图再进行一次损失计算，因此该第2部分的损失计算表示为

$$\mathrm{Loss}_2=\sum_p\|D_{out}(p)-D_{GT}(p)\|_1 \tag{12}$$

因此改进网络的整体损失函数计算表示为

$$\mathrm{Loss}=\mathrm{Loss}_1+\mathrm{Loss}_2 \tag{13}$$

4 实验结果及分析

4.1 典型测试数据集

选取DTU数据集[17]对改进网络进行实验结果分析。其中DTU数据集用于对本文DA-MVSNet网络进行训练以及输出结果的验证测试。同时为保证输入实验数据的一致性，在此参考CasMVSNet网络的数据集处理方式，将DTU数据集划分为3个部分：验证集、基准测试集以及训练集。同时利用Tanks & Temples数据集[17]对DA-MVSNet网络进行评估，将DTU数据集训练的网络模型直接对Tanks & Temples数据集中的未知场景进行重建，可以验证其泛化性能[17]。

4.2 实验环境及参数配置

本视图重建网络的实验环境基于AutoDL云服务器搭建，服务器配置如下：CPU为Intel Xeon第3代 Platinum 8350C处理器；GPU为NVIDIA GeForce RTX 3090显卡，显存为24 GB。操作系统基于Ubuntu 20.04使用Pytorch构建算法框架并通过Python编写代码，利用Anaconda对多个实验环境进行管理。使用Matlab R2022a运行DTU官方数据集提供的评估代码以计算三维视图重建评估指标，其中Matlab的运行平台为联想拯救者R7000p 2021。

如前所述，在模型实验参数的配置上，遵循CasMVSNet网络的策略。在模型训练阶段，利用DTU数据集划分的78个场景训练集对模型进行训练，统一将输入图像分辨率设置为640像素×512像素。每次输入视图数量为$N=3$，模型初始学习率为0.001并且训练次数16轮。在训练至第10、12和14轮时，其学习率分别降低至原来的一半。

4.3 DTU数据集重建点云基准测试

4.3.1 基于DTU测试集部分场景的重建效果基准测试

测试实验主要利用DTU数据集中78个场景的训练集对DA-MVSNet进行训练，训练后的网络基于DTU测试集的参考视图与源视图生成深度图，并通过后处理对深度图进行滤波、融合后得到三维视图重建点云。其中部分场景的可视化视图重建点云如图5所示。由图5可以看到，DA-MVSNet的重建点云能基本完整地还原每个场景，各场景点云在还原图像表面无明显缺失，在各部分边缘交界处也较为分明。同时也观察到，重建点云的表面以及边缘处也拥有较少的黑色噪点，进而在DTU数据集的重建点云基准测试中初步证明本文DA-MVSNet网络可实现完整度以及精度较好的三维视图重建。

4.3.2 DTU数据集重建点云定量对比实验

将DA-MVSNet网络与CasMVSNet和CVP-MVSNet网络的重建点云分别在较小的室内场景以及较大规模的室内场景中进行定性对比实验，综合多组对比结果证明了本文改进网络DA-MVSNet在点云完整度上的较好优越性。本文实验将基于DTU官方数据集提供的Matlab代码，对DA-MVSNet

图5　DA-MVSNet基于DTU测试集部分场景的重建效果基准测试

Fig.5　Benchmark testing of reconstruction performance of DA-MVSNet for DTU dataset

生成的三维点云与其他现有网络模型进行定量的点云质量数值评估对比,通过直观的数据来验证本文网络生成点云在准确度、完整度以及整体质量上的表现。评价指标为常用的准确性误差(Acc)、完整性误差(Comp)和整体性误差(Overall)。DA-MVSNet同其他基于传统方法和基于深度图的深度学习方法生成的三维点云在DTU数据集上的评价指标对比如表1所示,其中加粗数值为每列评价指标中的最优值。由表1可知,准确性误差最低的是基于传统方法的Gipuma,而基于深度学习方法准确性误差最低的是CVP-MVSNet。本文提出的DA-MVSNet在准确性误差上高于CVP-MVSNet但低于基准模型CasMVSNet,相较于CasMVSNet降低了约1.2%,证明本文引入的深度最优化机制模块可以对视图重建点云的准确性误差进行有效限制。相较于CasMVSNet与CVP-MVSNet,本文网络完整性误差分别降低了13.8%和18.2%,证明DA-MVSNet模型引入的特征增强网络模块可以较好地提升重建点云的完整度。另外DA-MVSNet在准确性误差以及完整性误差的单项表现上并不突出,但整体性误差达到最低的0.327 mm,证实在视图重建点云整体性指标上,DA-MVSNet比其他视图重建模型优秀。

表1　不同模型在DTU数据集上的重建点云数值评估

Table 1　Numerical evaluation of reconstructed point clouds using different models on DTU dataset　mm

网络模型	Acc	Comp	Overall
Gipuma	**0.283**	0.873	0.578
COLMAP	0.400	0.664	0.532
MVSNet	0.396	0.527	0.462
R-MVSNet	0.383	0.452	0.417
Point-MVSNet	0.342	0.411	0.376
Fast-MVSNet	0.336	0.403	0.370
UCSNet	0.338	0.349	0.344
PatchmatchNet	0.427	**0.277**	0.352
VisMVSNet	0.369	0.361	0.365
AA-RMVSNet	0.369	0.339	0.357
MSCVP-MVSNet	0.379	0.278	0.328
CVP-MVSNet	0.296	0.406	0.351
CasMVSNet	0.325	0.385	0.355
DA-MVSNet	0.321	0.332	**0.327**

4.3.3　网络模型重建效率定量对比实验

本节主要对DA-MVSNet、CasMVSNet以及CVP-MVSNet网络在DTU数据集上对比单张输入视图随着分辨率的增加其重建过程在内存占用和时间消耗上的效率。在实验中采用与上文相同的参数配置,得到各重建网络模型的内存占用以及时间消耗结果分别如图6和图7所示。

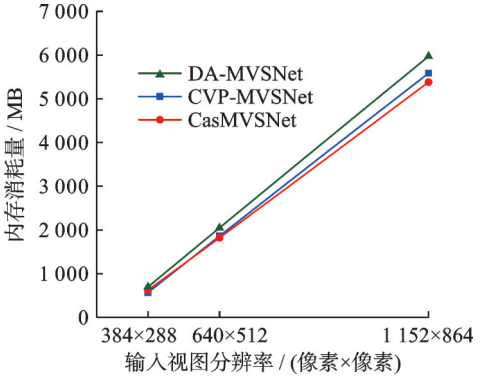

图6 网络重建模型的内存消耗量对比

Fig.6 Comparison of memory consumption for reconstruction network models

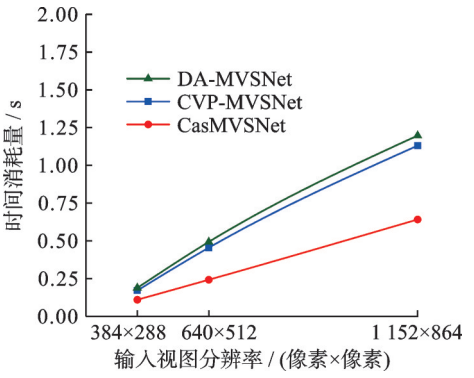

图7 网络重建模型的时间消耗量对比

Fig.7 Comparison of time consumption for reconstruction network models

由图6可知，在内存占用量上，本文的改进网络DA-MVSNet由于采用与CasMVSNet相同的深度评估结构，当输入视图分辨率较小时，DA-MVSNet与CasMVSNet和CVP-MVSNet的内存消耗量曲线趋于拟合。由图7所示，随着输入视图分辨率的增加，引入了特征增强网络以及深度图校正网络的DA-MVSNet对分辨率为1 152像素×864像素的单张视图进行重建时，其内存消耗量稍高于CasMVSNet以及CVP-MVSNet，且对分辨率越高的视图其消耗量增涨幅度越大。

同样在单张输入视图的时间消耗量上，DA-MVSNet网络的消耗时间约高于CVP-MVSNet与CasMVSNet网络。分析得知，由于DA-MVSNet网络中引入了特征增强网络以及深度图最优化校正网络，使得DA-MVSNet网络虽然在消耗内存和时间的效率上稍低于其他两个网络模型，但以此较小的成本量增加换取了更优的重建点云质量，说明该模型能更好地平衡模型的综合性能。因此DA-MVSNet网络具备了良好竞争力，且其视图重建综合性能得到明显提升。

4.4 消融实验与分析

本文对DA-MVSNet中引入的特征增强网络以及深度校正网络进行消融实验，通过控制变量来证明各网络模块的有效性。由于本文网络是在DTU数据集上进行训练的，因此消融实验均在DTU数据集上进行验证。针对改进模块的不同组合进行消融实验，在相同的实验环境及参数条件下，对DA-MVSNet各模块基于DTU数据集的评价指标进行视图重建性能数值对比。如表2所示，其中标记“√”的部分表示引入此网络模块，加粗数值代表最优化值。由表2结果可知，DA-MVSNet在仅引入特征增强网络时其完整性误差降低了8.1%，但准确性误差却增加了5.0%，如前文对特征增强网络原理的介绍，其额外提取的非相关特征信息也会随之增加，因此会导致重建点云准确度下降。在只引入深度校正网络的情况下，其重建点云的准确性误差与完整性误差分别降低了2.5%和4.7%，证明深度校正网络可以有效纠正输出深度图的错误深度信息，使其重建点云的空间位置信息更加准确，并且在后处理的深度图滤波中，被剔除的错误像素点数量也随之减少，增加了参与点云重建的像素点数量，进而提升了完整度。通过以上数据

表2 DA-MVSNet引入不同的网络组合在DTU数据集上的评估数值

Table 2 Evaluation of introducing different network combinations on DTU dataset mm

重建网络名称	特征增强网络	深度校正网络	Acc	Comp	Overall
CasMVSNet			0.325	0.385	0.355
	√		0.341	0.354	0.348
		√	**0.317**	0.367	0.342
DA-MVSNet	√	√	0.321	**0.332**	**0.327**

比对,在同时引入特征增强和深度校正机制网络模块时,DA-MVSNet重建点云的完整性误差以及整体性误差均为最优值。由此说明增加的模块可以进行有效的数据融合,并增强重建点云的整体质量。

5　结束语

本文在CasMVSNet的基础上提出了一种基于扩张注意力与深度最优化校正机制的多视图三维重建网络DA-MVSNet,重点引入了融合深度可分离卷积的并行空洞卷积与注意力模块构成的特征增强网络模块,提升了对输入视图的全局特征提取能力,降低了重建点云的完整性误差。在DA-MVSNet输出模块中,引入了基于非线性最小二乘的最优化校正机制模块,获得了重建点云的完整性误差以及整体性误差最优值。通过在室内场景数据集DTU上进行重建点云定量结果对比和消融实验结果分析,证明了DA-MVSNet改进网络的重建点云在完整度上有着显著提升,并能兼顾准确度指标以及拥有良好的重建点云整体质量和综合性能。后续研究将关注如何进一步提升网络重建点云的准确度,并进一步改善其重建效率和泛化能力。

参考文献:

[1] JI M, GALL J, ZHENG H, et al. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2326-2334.

[2] CHARLES R Q, HAO S, MO K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 77-85.

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

[4] YAO Y, LUO Z, LI S, et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5520-5529.

[5] LUO K, GUAN T, JU L, et al. P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019: 10451-10460.

[6] YI H, WEI Z, DING M, et al. Pyramid multi-view stereo net with self-adaptive view aggregation[C]//Proceedings of Computer Vision-ECCV 2020: 16th European Conference. Glasgow, UK: Springer International Publishing, 2020: 766-782.

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

[8] GU X, FAN Z, ZHU S, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 2492-2501.

[9] ZHANG J, LI S, LUO Z, et al. Vis-MVSNet: Visibility-aware multi-view stereo network[J]. International Journal of Computer Vision, 2023, 131(1): 199-214.

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

[11] WEI Z, ZHU Q, MIN C, et al. AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 6167-6176.

[12] GAO S, LI Z, WANG Z. Cost volume pyramid network with multi-strategies range searching for multi-view stereo[C]//Proceedings of Advances in Computer Graphics. Cham: Springer Nature Switzerland, 2022: 157-169.

[13] CAO C, REN X, FU Y. MVSFormer: Multi-view stereo by learning robust image features and temperature-based depth[EB/OL]. (2022-08-04). https://arxiv.org/abs/2208.02541.

[14] LIU T, YE X, ZHAO W, et al. When epipolar constraint meets non-local operators in multi-view stereo[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 18042-18051.

[15] XIANG T Z, XIA G S, ZHANG L. Mini-UAV-based remote sensing: Techniques, Applications and Prospectives[EB/OL]. (2018-12-19). https://arxiv.org/abs/1812.07770v1.

[16] 庞彦伟, 苏畅, 龙涛. 自适应构造与聚合多尺度代价体的双目立体匹配[J]. 东北大学学报(自然科学版), 2023, 44(4): 457-468.
PANG Yanwei, SU Chang, LONG Tao, et al. Adaptive multi-scale cost volume construction and aggregation for stereo matching[J]. Journal of Northeastern University(Natural Science), 2023, 44(4): 457-468.

[17] 尉婉青, 禹晶, 柏鳗晏, 等. SSD与时空特征融合的视频目标检测[J]. 中国图象图形学报, 2021, 26(3): 542-555.
YU Wanqing, YU Jing, BAI Manyan, et al. Video object detection using fusion of SSD and spatiotemporal features[J]. Journal of Image and Graphics, 2021, 26(3): 542-555.

[18] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 11531-11539.

[19] FURUKAWA Y, PONCE J. Accurate, dense, and robust multi-view stereopsis[C]//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA: IEEE, 2007: 1-8.

[20] 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.

[21] 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.

[22] 黄裕青, 李华锋, 原铭, 等. 基于卷积神经网络梯度和纹理补偿的单幅图像超分辨率重建[J]. 数据采集与处理, 2023, 38(5): 1112-1124.
HUANG Yuqing, LI Huafeng, YUAN Ming, et al. Super-resolution reconstruction of single image based on convolutional neural network gradient and texture compensation[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 1112-1124.

作者简介:

徐蕾(1969-),**通信作者**,女,高级工程师,硕士生导师,研究方向:智能算法、动态视图处理、软件设计管理与开发,E-mail: xulei@nuist.edu.cn。

雷有元(1999-),男,硕士研究生,研究方向:三维视图重建、深度学习与人工智能,E-mail: 1817630714@qq.com。

朱军(1999-),男,硕士研究生,研究方向:图像处理,神经网络与深度学习、动态视图三维重建。

周杰(1964-),男,教授,博士生导师,研究方向:移动通信理论、无线传感网络和无线接入网,E-mail: zhoujie@nuist.edu.cn。

邵根富(1962-),男,教授,博士生导师,研究方向:物联网通信理论、控制理论与方法和数据传输。

张家铭(1999-),男,硕士研究生,研究方向:人工智能、操作系统与系统管理软件。

(编辑:张黄群)