融合粗粒度代价体及双边网格的轻量级多视图三维重建

张啸，董红斌

引用本文

张啸，董红斌. 融合粗粒度代价体及双边网格的轻量级多视图三维重建[J]. 计算机科学, 2023, 50(8): 125-132.

ZHANG Xiao, DONG Hongbin. Lightweight Multi-view Stereo Integrating Coarse Cost Volume and Bilateral Grid [J]. Computer Science, 2023, 50(8): 125-132.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600046

融合粗粒度代价体及双边网格的轻量级多视图三维重建

张 啸 董红斌
哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001
(zhangxiao980516@163.com)

摘 要 针对基于深度学习的多视图立体(Multi-view Stereo,MVS)重建算法内存消耗过大、推理速度慢,以及对病态区域重建效果不佳的问题,提出了一种基于双边网格和融合代价体的轻量级级联的MVS重建网络。首先利用基于双边网格的代价体上采样模块将较低分辨率代价体高效地恢复成高分辨率代价体。随着采用轻量级的动态区域卷积和粗粒度代价体融合模块,提升网络对病态区域特征的表示能力以及对场景整体信息和结构信息的感知能力。实验结果表明,该网络在DTU数据集以及Tanks and Temples数据集上均取得了具有竞争性的结果,并且在内存消耗以及推理速度上都显著优于其他方法。

关键词:三维重建;多视图立体;深度学习;双边网格;轻量级
中图法分类号 TP391

Lightweight Multi-view Stereo Integrating Coarse Cost Volume and Bilateral Grid

ZHANG Xiao and DONG Hongbin
College of Computer Science and Technology,Harbin Engineering University,Harbin 150001,China

Abstract In order to tackle the problems of large memory consumption,poor real-time performance and poor reconstruction quality for low-textured areas of multi-view stereo reconstruction algorithm basedon deep learning,this paper proposes a lightweight cascade MVS reconstruction network based on bilateral grid and fused cost volume. Firstly,it builds the cost volume upsampling module based on learned bilateral grid,which can efficiently restore the low-resolution cost volume to the high-resolution cost volume. Then the dynamic region convolution and coarse cost volume fusion module are used to improve the network's ability to extract the feature of the challenging area and to perceive the global and structural information of the scene. Experimental results show that our method achieves competitive results on DTU dataset and tanks and temples benchmark,and is significantly better than other methods in memory consumption and inference speed.

Keywords 3D reconstruction,Multi-view stereo,Deep learning,Bilateral grid,Lightweight

1 引言

三维重建是计算机视觉中的传统研究热点,其已被研究了数十年之久。三维重建主要包括基于图像的三维重建[1]以及基于扫描点云的三维重建[2]。多视图立体[3]作为基于多视图像三维重建的重要环节也受到了广泛的关注与研究。MVS方法试图根据多张标定过的图像恢复出其对应的深度图,并最终根据深度图恢复出场景的三维结构。深度学习之前,很多学者使用基于几何知识的传统方法[4-5]对其进行了广泛且深入的研究,并且在丰富纹理区域取得了很好的效果,但是传统算法存在着许多缺点,在一些病态区域中,比如反射平面、弱纹理区域,以及遮挡区域中,其最终的重建结果都不尽如人意。

近年来,随着深度学习的快速发展,基于学习的多视图立体重建方法取得了巨大的成功,能够通过训练获得场景的先验信息,使得在病态区域重建的结果得到了改善,并且刷新了很多MVS基准。但是基于学习的方法都存在一些缺点,比如鲁棒性较差,有些方法在室内数据集上效果较好,但在进行室外的大场景重建时,对其中的病态区域重建的效果依然较差。同时,很多方法由于大量使用三维卷积,在进行推理时占用了大量的内存,并且推理速度较慢,造成大量的资源消耗,使其很难部署到移动设备上,这些缺点阻碍了其在大规模三维场景上的实时应用。

为了提高模型的鲁棒性以及病态区域的重建效果,我们需要增强网络的整体感知能力以及结构提取能力。许多方法采取了粗到细的级联结构[6],然而这些方法为了获得更加准确的初始深度值,在起始阶段的代价体往往分辨率仍然较高,没有具备很好的全局以及结构信息感知能力,而更偏向于局部信息,因此训练出的模型对数据集比较敏感。而不同尺度的低分辨率代价体可以覆盖多个尺度的感受野,其中包含的

到稿日期:2022-06-06 返修日期:2022-11-06
基金项目:黑龙江省自然科学基金(LH2020F023)
This work was supported by the Natural Science Foundation of Heilongjiang Province,China(LH2020F023).
通信作者:董红斌(donghongbin@hrbeu.edu.cn)

多尺度信息是可以相互补充的。受此启发，本文在粗到细级联结构的基础之上，在初始阶段使用了代价体融合模块，通过提取更粗粒度的特征构建多个低分辨率的代价体，并将它们融合成初始阶段的代价体，使其能够更好地包含场景的结构以及全局信息，同时，该模块的引入只会带来很小的内存消耗。此外，很多方法为了使提取到的特征能够更好地具备场景的感知信息以及语义信息，在特征提取时采用了类注意力机制以及可变形卷积[6-8]，但是这些方法通常内存消耗较大。为了提升所提取特征的语义表征能力，同时也避免上述方式带来的大量内存消耗，我们在特征提取时引入了动态区域感知卷积[9]。这种卷积方式首先将图像划分成不同的区域，然后在同一区域内使用相同的卷积核，能够在提升特征提取能力与降低计算负担之间达到很好的平衡。

减少 MVS 重建的内存消耗也是近期的研究热点，一些方法[10]在代价体正则化时采用循环网络来达到以时间换空间的效果，还有一些方法采用粗到细的级联结构，通过逐步精炼代价体来减少冗余的计算。这些方法都在提升重建效果的同时降低了部分内存占用，但仍然有很大的提升空间。分析得知，这主要是高分辨率代价体占用内存较大并对其进行三维卷积造成的。为了解决这个问题，本文在受到双边网格上采样[11]的启发的基础上，利用基于双边网格的代价体上采样模块对正则化后的低分辨率代价体进行上采样得到高分辨率代价体，再对它回归出最后的高分辨率深度图。该方法能够在使得到的深度图尽可能准确的同时避免了对高分辨率代价体使用三维卷积所带来的巨大的计算资源消耗。

本文主要有以下几个创新点：

(1)在特征提取模块中加入了动态区域卷积，在尽可能减少内存消耗的同时提升对病态区域特征的表征能力。

(2)在初始阶段构建代价体时采用粗粒度代价体融合模块，提升网络对场景结构以及全局信息的学习能力。

(3)在最后阶段采用了基于双边网络的代价体上采样模块。从低分辨率的代价体中高效地恢复出高分辨率的代价体，在尽可能维持准确性的同时减少了模型的内存消耗并提升了推理速度。

2 相关工作

多视图三维重建方法在过去的几十年内一直是计算机视觉领域的热点问题，根据方法输出的场景表示类别，可以将三维重建算法分为 3 个类别，分别为基于体素的方法[12]、基于点云的方法[13-14]，以及基于深度图的方法。与基于体素以及基于点云的方法相比，基于深度图的方法具有更加灵活、计算负担小的优点，其可以把整个需要重建的场景分解为多个视角的深度图预测，同时，根据获得的深度图，可以将其转换为对应的点云或者体素表示。因此，当前表现最好的几种 MVS 方法，如 COLMAP[15]，ACMH[4]，以及 Gipuma[16] 都采用了基于深度图的表示，并且在重建的精度以及鲁棒性上都有很好的表现。COLMAP 使用了手工设计的特征度量，并利用光照、几何一致性以及像素级视图选择权重来提升每个视图的深度估计的质量。ACMH 使用了多尺度的几何一致性来优化在弱纹理区域的深度估计，并且使用自适应棋盘采样以及多假设联合视角选择方法来提升深度估计的质量。Gipuma 采用了红黑棋盘格的传播方式，充分利用了 GPU 并行计算的优点，极大地加快了深度估计的速度。但是传统方法通常使用手工设置的特征度量，容易受到弱纹理区域、光照变化、反射区域以及遮挡区域的影响。

MVSNet[17] 首先在深度学习中引入了传统算法的平面扫描思想[18]，首创性地使用可微单应性变换将二维特征构建成三维代价体，并通过三维卷积以及 soft-argmax 操作恢复出最终的深度图。该模型能够通过学习获得场景的先验知识，在病态区域的重建上取得较好的效果，但其仍然存在内存消耗大、推理速度慢等问题，之后的工作大部分是对其的改进。Point-MVSNet[19] 网络把 3D 几何先验知识和 2D 纹理信息融合到特征增强点云中，并不断迭代优化形成最终的点云。R-MVSNet[20] 将代价体三维卷积换成了循环神经网络，以此避免代价体正则化 3 倍指数增长所带来的内存消耗，从而使得大场景重建成为可能。由于丧失了一部分计算的并行性，其推理速度也变得缓慢。在此之后出现的网络不仅提高了重建质量，而且在内存占用上的需求也有所缓解。比如 Fast-MVSNet[21] 首先构造了一个稀疏的代价体来获得一个稀疏但高分辨率的深度图，然后通过高斯-牛顿层对其进行优化获得最终的深度图。CasMVSNet[22] 采用了粗到细的级联结构，后面的阶段都会根据上一阶段的深度值缩小深度搜索范围，从而获得更精确的深度以及达到节约显存的目的。之后出现的网络往往都采用级联结构。Vis-MVSNet[23] 将像素级的可视性信息融入单视图代价体聚合过程中，能够提升遮挡场景中的重建效果。UCS-Net[24] 在粗到细的基础之上，使用基于方差的不确定性估计构建自适应薄代价体，使得代价体中每个像素的深度采样范围根据上一阶段的不确定性进行调整。AA-RMVSNet[25] 在 R-MVSNet 的基础上，在特征提取阶段引入了可变形卷积来聚合多尺度的特征，并在构建代价体时采用了视图间聚合模块来降低视角转换过大的干扰。EPP-MVSNet[26] 也在基于粗到细结构的基础上使用了极线装配模块来聚合高分辨率的代价体，并采用了伪三维卷积来将三维卷积解耦成两个方向的二维卷积。

双边网格由双边滤波逐步改进而来，双边滤波通过空间域核函数和范围域核函数共同作用来进行滤波，从而起到保护边缘以及降噪平滑的作用，但是由于双边滤波的非线性，其运行速度较慢。因此，文献[27]提出了双边网格，用来加速双边滤波。主要通过 3 个步骤来使用双边网格。1)创建网格：将图像平面的二维坐标对应于网格的前两个维度，并将图像的强度对应于第三个维度。2)函数处理：使用三维高斯核对网格进行卷积，并且分别指定空间和范围维度的方差。3)切片：将经过函数处理过的双边网格进行三线性差值计算，最终提取出二维图像。有很多利用双边网格加速图像处理的工作，例如，文献[28]使用了双边网格的思想，将图像变换到双边空间进行高斯滤波后，再将其映射回像素空间，由此实现了双目图像的虚化效果。很多基于深度学习的方法也通过结合双边网格进行加速，例如，文献[29]引入了双边网格来进行图像增强，提取图像的全局以及局部特征，将其结合获得双边网格，并通过将原图进行仿射变换获得的引导图像来引导双边网格对空间及颜色的插值，最后将其进行仿射变换得到输出图像。本文的部分灵感来自文献[30]，其使用双边网格来

加速立体匹配网络，通过构建保持边缘特性的代价空间上采样模块，使得网络在保持精度的同时推理速度得到大幅提升。

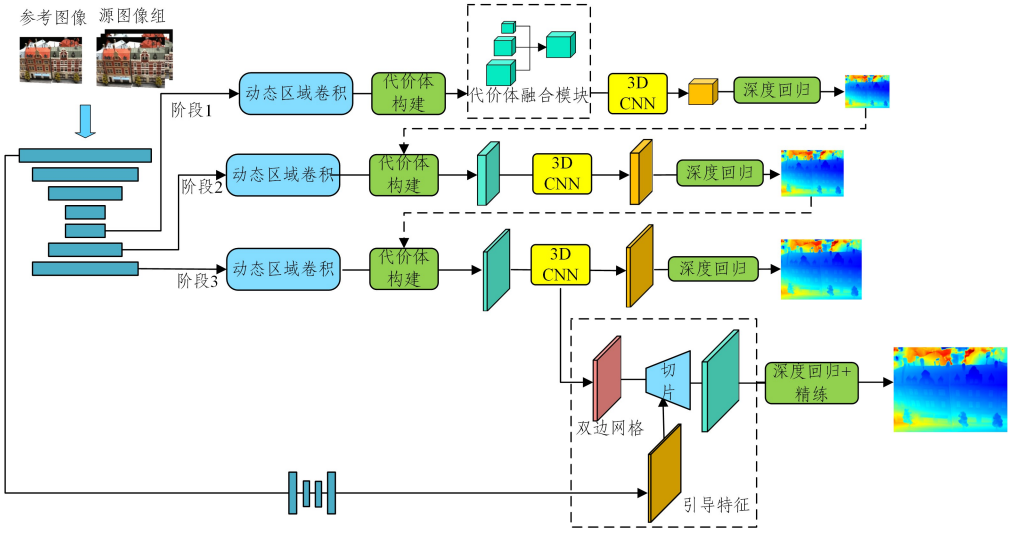

图 1　网络总体结构

Fig. 1　Overall architecture of the proposed network

3　融合粗粒度代价体及双边网格的多视图三维重建模型

本节将重点介绍所提网络的总体结构和主要部分，并且进一步解释动态区域卷积、粗粒度代价体融合模块，以及基于双边网格的代价体上采样模块。

3.1　总体结构

给定输入图像$\{\boldsymbol{I}_0,\boldsymbol{I}_1,\cdots,\boldsymbol{I}_N\}$，$\boldsymbol{I}\in\mathbb{R}^{H\times W}$，令$\boldsymbol{I}_0$为参考图像，$\{\boldsymbol{I}_i\}_{i=1}^N$为源图像集合，以及相机的内外参数$\{\boldsymbol{K}_i,\boldsymbol{R}_i,\boldsymbol{t}_i\}_{i=0}^N$，该网络的目的是根据标定后的参考图像和多张源图像来获得参考图像的深度图D_0。网络的总体结构图如图 1 所示。本模型采用三阶段级联的预测方式，首先采用特征金字塔获得多个尺度的特征，再将其进行动态区域卷积，获得最终的特征$\{\boldsymbol{F}_k\}_{k=1}^3$，其中$\boldsymbol{F}_k$表示在第 k 阶段的特征。为了能够更好地提取场景的全局信息以及结构信息，在起始阶段利用多层粗粒度的特征图去构造多个粗粒度代价体，并将它们融合获得初始阶段的代价体。其他阶段 k 都会利用该阶段的特征$\boldsymbol{F}_k$和深度采样得到的深度序列构建代价体$\boldsymbol{CV}^k$，再对代价体进行三维卷积正则化后恢复出概率体$\boldsymbol{P}_k$，最终回归出该阶段的深度图$\boldsymbol{D}_k$。在第三阶段构建并正则化代价体之后，使用基于双边网格的代价体上采样模块对其进行上采样，由此获得高分辨率的代价体，进而恢复出高分辨率的深度图。

3.2　多尺度区域感知特征提取器

本文特征提取器的主要思想来自 UCS-Net，主要结构包含编码器、解码器以及跳跃连接的轻量级 2DUNet。编码器包含了一系列的二维卷积，并且在每个卷积后都会进行批量正则化以及 ReLu 激活操作，为了加入基于双边网格的代价体上采样模块，本文使用了 3 个步长为 2 的卷积对特征图进行 3 次下采样。解码器与编码器结构类似，相当于编码器的逆过程，包含了两次上采样。给定图像大小为 $W\times H$ 的输入图像$\boldsymbol{I}_i$，可以获得 4 个尺度的特征图$\boldsymbol{F}_{i,0},\boldsymbol{F}_{i,1},\boldsymbol{F}_{i,2},\boldsymbol{F}_{i,3}$，尺度分别为$\frac{W}{8}\times\frac{H}{8},\frac{W}{4}\times\frac{H}{4},\frac{W}{2}\times\frac{H}{2},W\times H$，通道数为 48，32，16，8。在获得多个尺度特征图之后，再利用多个动态区域卷积来增强特征的语义表示能力，该卷积方式能够根据区域中的语义信息对卷积核进行调整。在先前的工作中，AA-RMVSNet 也采用了可变形卷积来提取自适应的语义特征，该方法在每个像素中使用不同形状的卷积来代替标准卷积，可以提升对重要区域的特征提取能力，在第 o 通道的输出特征图可以表示为：

$$\boldsymbol{Y}_{h,w,o}=\sum_{c=1}^{C}\boldsymbol{X}_{h,w,c}*\boldsymbol{W}_{h,w,c}^{(o)} \tag{1}$$

其中，输入$\boldsymbol{X}\in\mathbb{R}^{H\times W\times C}$，$H,W,C$ 分别表示高度、宽度、通道数；$\boldsymbol{Y}\in\mathbb{R}^{H\times W\times O}$表示输出；$\boldsymbol{W}_{h,w,c}$表示在位置$(h,w)$处特有形状的卷积核。但是这种自适应卷积方式会带来大量的参数，造成一定程度的内存以及计算负担。为了进一步节约内存，并且增强对图像语义多样性的建模能力，本文使用了动态区域卷积，其具体结构如图 2 所示。

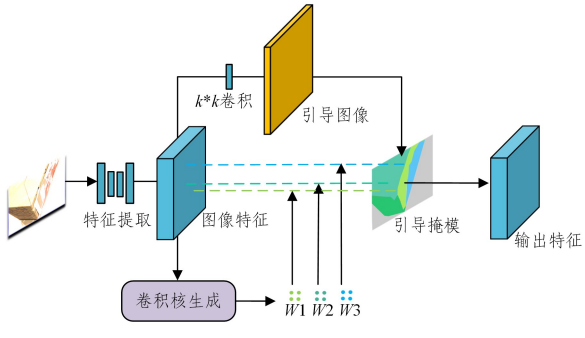

图 2　动态区域卷积示意图

Fig. 2　Dynamic region convolution diagram

动态区域卷积是通过一个可学习的引导掩模将输入特征划分为多个区域，并且在每个区域中使用相同的卷积核，在第 o 通道的输出特征图表示为：

$$\boldsymbol{Y}_{w,h,o}=\sum_{c=1}^{C}\boldsymbol{X}_{w,h,c}*\boldsymbol{W}_{t,c}^{(o)} \tag{2}$$

在每个区域的卷积核可以表示为$\boldsymbol{W}=[\boldsymbol{W}_0,\cdots,\boldsymbol{W}_{m-1}]$，其中第 t 个卷积核$\boldsymbol{W}_t\in R^c$。相对于在每个像素采用不同的卷积核，这种方法可以大大减少参数量。具体来说，对于有 n 个

共享区域的 $k\times k$ 的动态区域卷积，首先使用一个 $k\times k$ 的标准卷积将输入特征转换为引导特征 $\boldsymbol{F}$，$\boldsymbol{F}\in\mathbb{R}^{W\times H\times n}$，然后对其在最后维度进行 argmax 操作，从而获得引导掩模 $\boldsymbol{M}$，$\boldsymbol{M}\in\mathbb{R}^{W\times H}$，操作如式(3)所示：

$$\boldsymbol{M}_{w,h}=\text{argmax}(\boldsymbol{F}^0_{w,h},\boldsymbol{F}^1_{w,h},\cdots,\boldsymbol{F}^{n-1}_{w,h}) \tag{3}$$

使用 argmax 操作可以输出在最后维度上最大值的索引，因此 $0\leqslant\boldsymbol{M}_{w,h}\leqslant n-1$，该值对应着像素 (w,h) 应该属于区域的编号。在该区域之内的全部像素都共享相同的卷积核，这些卷积核通过可学习的子网络生成，卷积核生成以及反向传播过程可参考文献[9]。由于动态区域感知卷积的加入，不仅能够将通道维度的卷积核转换到空间维度上，还能在相同语义区域中采取相同的卷积核，保证了卷积的平移不变性。因此该方法能在提升重要区域卷积表征能力的同时减轻计算的负担。

3.3 代价体构建

本文构建代价体的方式与先前的方法类似，对于参考图像 $\boldsymbol{I}_0$ 和 $N-1$ 张源图像 $\{\boldsymbol{I}_i\}_{i=1}^N$，我们会在每个阶段利用可微的单应性变换构造级联的三维代价体 $\{\boldsymbol{CV}^k\}_{k=1}^3$，在 k 阶段的代价体尺寸为 $\left(\frac{W}{2^{4-k}}\times\frac{H}{2^{4-k}}\times D_k\times C_k\right)$，$D_k$，$C_k$ 为该阶段的深度采样数量以及特征通道数。参考图像特征在 D_k 个深度假设内匹配 $N-1$ 个扭曲的源图像特征，以此来计算每个视图的匹配代价体，深度假设范围获取方式与 UCS-Net 类似。对于源图像 $\boldsymbol{I}_i$，给定其对应相机内参数和外参数矩阵 $\{\boldsymbol{K}_i,\boldsymbol{T}_i\}$，其特征的每个像素在深度假设为 d 时与参考图像特征的匹配关系可由可微的单应性矩阵 $\boldsymbol{H}_i^{(d)}$ 计算得到：

$$\boldsymbol{H}_i^{(d)}=d\boldsymbol{K}_i\boldsymbol{T}_i\boldsymbol{T}_0^{-1}\boldsymbol{K}_0^{-1} \tag{4}$$

然后，利用扭曲后的源图像特征体与参考图像特征体的方差来获得单视图代价体。

$$\boldsymbol{CV}_i^{(d)}=(\boldsymbol{FV}_i^{(d)}-\boldsymbol{FV}_0)^2 \tag{5}$$

其中，$\boldsymbol{CV}_i^{(d)}$ 为在深度假设为 d、视图为 i 时单视图代价体的切片，$\boldsymbol{FV}_i^{(d)}$ 为在深度假设为 d 时扭曲的源图像特征体，$\boldsymbol{FV}_0$ 为参考图像特征体。最后将多个单视图代价体整合获得最终的代价体 $\boldsymbol{CV}$。

3.4 粗粒度代价体融合模块

最近提出的方法大多采用从粗到细的多阶段结构，这些方法通常使用高分辨率的代价体去恢复深度图。由于低分辨率代价体通常不具备足够的信息去恢复精确的初始阶段深度图，因此对低分辨率代价体的利用十分有限，但低分辨率代价体具有更大的感受野，并且包含场景的全局信息以及结构信息，这些信息往往在不同的数据集上是统一的，因此正确使用低分辨率代价体有利于增强模型的鲁棒性以及提升对弱纹理区域的最终重建效果。本文的粗粒度代价体融合模块如图 3 所示。首先，利用经过动态区域卷积得到的第一阶段特征再经过两个步长为 2 的卷积层获得额外两个尺度的特征，再利用这 3 个尺度的特征来构造 3 个粗粒度的稠密代价体，其维度为 $\left(\frac{H_1}{s}\times\frac{W_1}{s}\times\frac{D_1}{s}\times C_1\right)$，$s\in1,2,4$。$H_1$，$W_1$，$D_1$，$C_1$ 为第一阶段特征图的高、宽、深度采样数量和通道数。在获得粗粒度代价体之后，将其输送到代价体融合模块之中。该模块采用了常见的带有跳跃的编码-解码器结构，在每个尺度的代价体都会通过 4 个步长为 1 的 3D 卷积层来获得高阶语义特征。此后将第一层与第二层的代价体通过在特征维度的级联操作融合进第三层中，然后将结果通过一个三维卷积去调整特征维度。最后使用带残差的反卷积进行上采样来获得融合后的初始阶段代价体，维度为 $(H_1\times W_1\times D_1\times C_1)$。其包含了第一阶段全部的深度采样范围，并且融合了场景的全局信息以及结构信息。通过它恢复出的深度图更具有鲁棒性，从而为下阶段预测的深度采样提供了较好的指导作用。

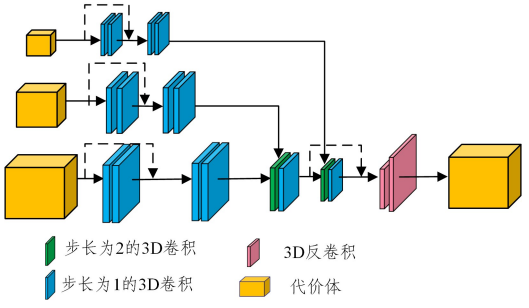

图 3 粗粒度代价体融合模块

Fig. 3 Coarse cost volume fusion module

3.5 代价体正则化及深度预测

在每个阶段，k 获得代价体 $\boldsymbol{CV}^k$ 之后，本文使用 3D UNet 网络对代价体进行正则化，从而减小噪声对其的影响，最终恢复出单通道的概率体 $\boldsymbol{P}^k$，其余维度与该阶段代价体的维度相等。其每个点的值 $P(x,y,d)$ 表示点 (x,y) 在深度为 d 时的概率。然后通过 soft-argmax 操作恢复出第 k 个阶段的深度图 $\boldsymbol{D}^k$。soft-argmax 如式(6)所示：

$$\boldsymbol{D}^k=\sum_{j=1}^{D}d_j\boldsymbol{P}(d_j) \tag{6}$$

其中，j 为深度采样的次序。

3.6 基于双边网格的代价体上采样模块

为了减少网络的内存消耗，提升网络的推理速度，本文采用了基于双边网格的上采样模块，双边网格能够利用引导特征的信息进行上采样，恢复出准确的高分辨率代价体，进而恢复出原始分辨率的深度图。因此，在代价体正则化阶段，可以避免将高分辨率代价体送入内存消耗过大的三维卷积之中，从而减少内存消耗。

本模型仅在最后阶段使用双边网格上采样模块，其结构如图 4 所示，由于切片操作的加入，使用宽高皆为原始图像一半的代价体就可恢复出原分辨率的深度图，该模块主要分为 3 步：获得引导图像、获得双边网格、切片。

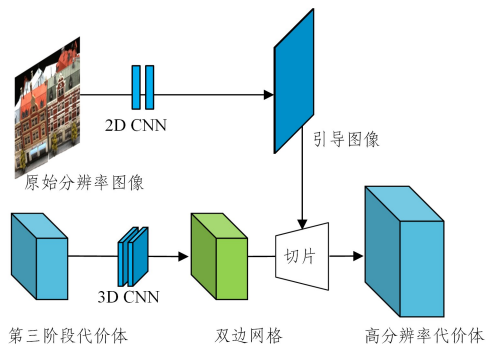

图 4 基于双边网格的代价体上采样模块

Fig. 4 Cost volume upsampling module based on bilateral grid

3.6.1　获得引导图像

首先，对于最细粒度特征图（$H\times W\times C$），使用多个2DCNN去获得引导图像（$H\times W\times 1$）。由于引导信息通过卷积网络获得，与直接利用高分辨率光学图像相比，利用特征图作为引导图像可以包含复杂的语义特征，使得最终的切片操作所产生的代价体能够具有更好的真实性，同时也能更好地保护边缘特征。

3.6.2　利用低分辨率代价体构建双边网格

第三阶段经过正则化后的代价体，其大小为$\left(\frac{H}{2}\times\frac{W}{2}\times D_3\times C_3\right)$，$D_3$和$C_3$分别为第三阶段深度采样数量以及特征通道数量，使用多个三维卷积将其转换至双边网格，双边网格的大小为$(\frac{H}{2}\times\frac{W}{2}\times D_3\times G)$，其中$G$为对应引导图像强度的维数，双边网格的值可以表示为$B(h,w,d,g)$。

3.6.3　利用切片进行代价体上采样

目前，我们利用第三阶段的代价体构建了可学习的双边网格，紧接着需要将其信息转移到高分辨率空间中去形成原始图片大小的高分辨率代价体，因此本节采用了双边网格切片操作的思想。该部分的输入是上文提到的单通道原始分辨率大小的特征图$\boldsymbol{G}$以及利用低分辨率代价体构造的双边网格$\boldsymbol{B}$，切片操作会在最后的双边网格中执行数据依赖的查找，并且最终产生与引导图像$\boldsymbol{G}$分辨率相同的代价体。具体来说，切片操作是在高分辨率引导图$\boldsymbol{G}$的引导下，对双边网格进行线性插值，如式(7)所示：

$$\boldsymbol{C}_H(h,w,d)=B(sh,sw,sd,s_G\boldsymbol{G}(h,w)) \tag{7}$$

其中，s为双边网格宽或高与引导特征图的比值，并且$s\in(0,1)$，s_G为网格灰度与引导图灰度的比值且$s_G(x,y)\in(0,1)$。已知切片操作是没有参数的，因此该操作对内存是友好的。并且由于在双边网格中进行操作会使得本文模型参考引导特征图G中的边缘，使得最终产生的代价体能够有效地保护边缘信息，从而在减少内存消耗、提升推理效率的同时，尽可能产生更准确的深度图。

4　实验

本节首先介绍用于评估的数据集——DTU以及Tanks and Temples，然后介绍评估指标以及实验细节，并将其与先前方法进行比较与分析。最后通过消融实验来证明所提网络结构的有效性。

4.1　数据集与评估指标

DTU数据集[31]是一个室内场景数据集，通过带有结构光传感器的工业级机械臂采集而成。其包含124个不同的场景，其中每个场景都拥有49个或者64个视角，并且每个视角下都有7种不同强度的光照。在基于学习的MVS领域中，该数据集的应用最为广泛，同时也是十分重要的基准。本模型利用该数据集进行训练，测试集和验证集的划分方式与MVSNet相同。

Tanks and Temples基准[32]数据集包含室内以及室外场景，使用工业级激光扫描仪获得了真值数据.但是由于其场景规模较大并且光照和场景反射区域会自然变化，该数据集往往比DTU更具有挑战性。按照场景的复杂程度，该数据集分为两组场景集合，分别为中等集以及高等集，本文使用中等集进行评估。

同之前的方法一致，本文的评估指标分为精确度和完整性以及F-值，其中精确度用来衡量重建点云到真值点云的距离，而完整性用来衡量真值点云到重建点云的距离。对于DTU数据集，总体性能为精确性和完整性的算术平均值，并且得分越低代表模型重建效果越好。对于Tanks and Temples数据集，使用F-值来衡量指标的百分比，F-值为精确性和完整性的调和平均值，更容易受到极端值的影响，能够反映出重建结果的不平衡性，该值越高代表模型重建效果越好。

4.2　实验细节

本文采用pytorch框架实现本模型，并且使用DTU训练集对模型进行训练，完成之后利用DTU验证集对其进行评估，DTU训练集与测试集的划分方式与MVSNet相同，此后使用Tanks and Temples数据集进行测试，从而验证模型的泛化能力以及大场景重建能力。在数据集预处理阶段，本模型视图选择策略以及深度图真值生成策略与MVSNet相同。在训练阶段，将DTU数据集的图片缩放为640×512，输入视图的数量设置为3，3个阶段的深度间隔数量分别为$D_1=48$，$D_2=32$，$D_3=8$，初始阶段的深度采样范围为[425 mm，935 mm]。在网络训练时使用SGD优化器，初始学习率设置为0.01，每经过3轮学习率衰减为原来的1/2。整个网络在RTX3090上共训练30轮，batchsize设置为10。在测试阶段，将输入的图片大小调整为$1\,600\times1\,156$，将输入视图数量设置为5。在获得测试集的预测深度图后，为便于进行评估，需对生成的深度图进行融合操作，从而生成稠密点云，具体操作与文献[17]的后处理操作相同，主要包含光度一致性滤波、几何一致性滤波以及深度图融合。然后使用DTU数据集提供的官方MATLAB评估代码来获得重建的密集点云的精确度、完整度，以及它们的平均值。对于Tanks and Temples基准数据集，本文将重建的稠密点云上传至在线官方评估网站来获得最终的F-值。

4.3　DTU数据集的评估结果

为了证明本文提出的网络的有效性，我们将最终的结果与传统方法和最近提出的基于学习的方法进行比较，具体结果如表1所列。

表1　在DTU评估数据集上的定量分析结果

Table 1　Quantitative results on DTU evaluation dataset

Methods	Acc.	Comp.	Overall
CVP-MVSNet[6]	0.296	0.406	0.351
Furu[14]	0.613	0.941	0.777
Gipuma[15]	**0.283**	0.873	0.578
Camp[16]	0.835	0.554	0.695
Point-MVSNet[19]	0.342	0.411	0.376
R-MVSNet[20]	0.383	0.452	0.417
Vis-MVSNet[23]	0.369	0.361	0.365
UCSNet[24]	0.338	0.349	**0.344**
Fast-MVSNet[33]	0.336	0.403	0.370
SurfaceNet[34]	0.450	1.040	0.745
CIDER[35]	0.417	0.437	0.427
Ours	0.398	**0.332**	0.365

可以发现传统方法Gipuma在准确性上处于领先，这是由于其手工设置的特征度量相较于基于学习的方法在纹理

丰富区域具有优势，但就完整性而言，由于其缺乏能够学习到的先验知识，因此其在病态区域的最终重建效果远不及基于学习的方法。与基于学习的方法相比可以发现，本模型在效率较高的基础之上取得了具有竞争性的结果，并且由于动态区域卷积以及粗粒度代价体融合模块的加入，本模型取得了最好的完整性以及具体竞争性的总体性能。但是，本文模型在准确性指标上仍然有较大的提升空间，原因可能是本文模型对代价体进行融合以及上采样操作给所包含的深度信息带来了一定程度的损耗。对于 DTU 测试集的部分场景的具体重建结果如图 5 所示。

图 5 DTU 数据集的点云重建结果

Fig. 5 Point cloud reconstruction results on DTU dataset

4.4 运行时间和 GPU 显存消耗

表 2 列出了本文方法和其他方法在运行时间和 GPU 显存消耗的对比结果，为了公平起见，尽可能地使输入图像分辨率与其他方法保持一致，因此将输入图像分辨率设置为 1 600×1 152。本模型与 Point-MVSNet，CasMVSNet 以及 UCSNet 相比，内存消耗分别减少了约 58%，42%，27%，运行时间分别减少了约 74%，49%，48%。可以发现，本文方法的运行时间更短并且显存占用更低，这是由于其他方法都将高分辨率的代价体送入计算资源消耗较大的三维卷积之中进行正则化，而本文提出的基于双边网格的代价体上采样模块能够使用较低分辨率代价体恢复出高分辨率的深度图，且减少了高分辨率代价体在正则化过程中使用三维卷积所进行的大量计算，从而节约了大量的内存并且加快了推理速度。

表 2 在 DTU 验证集上的运行时间和 GPU 内存占用结果比较

Table 2 Runtime and GPU memory usage results on DTU validation set

Methods	Input sizes	Depth map size	Depth number	GPU Memory/GB	Runtime/s	Overall
CVP-MVSNet[6]	1 600×1 152	1 600×1 152	48,8	8.6	1.72	0.351
D^{2}HC-RMVSNet[10]	1 600×1 200	1 600×1 196	256	6.6	29.15	0.386
MVSNet[17]	1 600×1 184	400×288	256	21.9	2.76	0.462
PointMVSNet[19]	1 600×1 184	800×576	96	12.7	1.72	0.376
R-MVSNet[20]	1 600×1 184	400×296	512	6.7	2.35	0.417
CasMVSNet[22]	1 600×1 184	1 600×1 184	48,32,8	9.9	0.89	0.351
UCSNet[24]	1 600×1 184	1 600×1 184	96	7.3	0.87	0.344
Fast-MVSNet[33]	1 280×960	640×480	192	5.3	0.60	0.370
PVA-MVSNet[36]	1 600×1 184	1 600×1 184	192	24.8	1.01	0.357
BH-RMVSNet[37]	1 600×1 200	1 600×1 200	512	10.4	104.10	**0.343**
Ours	1 600×1 152	1 600×1 152	48,32,8	**5.3**	**0.45**	0.365

4.5 消融实验

为了进一步验证所提模块的有效性及其所带来的内存消耗和运行时间，本文进行了相关消融实验以及定量分析。该实验运行在 DTU 数据集上，实验设置与之前相同，针对本文提出的基于双边网格的上采样模块、代价体融合模块以及动态区域卷积模块设计了 3 组实验，将这 3 个模块逐个相加，实验结果如表 3 所列。

表 3 在 DTU 验证集上的消融实验结果

Table 3 Ablation study results on DTU validation set

Methods	Acc.	Comp.	overall	GPU Memory/GB	Runtime/s
Baseline	**0.338**	0.349	**0.344**	7.3	0.87
Baseline+CU	0.398	0.346	0.372	**5.0**	**0.42**
Baseline+CU+FC	0.401	0.335	0.368	5.2	0.43
Baseline+CU+FC+DC	0.398	**0.332**	**0.365**	5.3	0.45

其中，Baseline 为本文所选取的基线模型，CU 为基于双边网格的代价体上采样模块，FC 为粗粒度代价体融合模块，DC 为动态区域卷积。当加入基于双边网格代价体上采样模块时，所提模块能够在损失 8%的总体性指标的基础上节约 46%的内存消耗和 32%的运行时间，这足以证明该模块能够在减少内存消耗以及运行时间的同时，尽可能地保持重建效果。当加入代价体融合模块以及动态区域卷积时，能够提升重建结果的完整性以及总体性指标，并且二者只会带来很少的内存消耗以及运行时间损耗，由此证明这两个模块是轻量级的、有效的。

4.6 Tanks and Temples 基准的评估结果

为了验证本网络的泛化能力，我们使用经过 DTU 数据集训练后的模型预测 Tanks and Temples 的中等集。该数据集全部为大型的室外场景。将输入图像的大小设置为 1 920×1 056，将输入的视图数设置为 7，同时利用 OpenMVG[39] 获得相机参数以及稀疏点云。实验结果如表 4 所列，可以发现本文模型超过了最好的传统方法，如 ACMH 以及 COLMAP。

在与其他基于学习的 SOTA 模型的定量比较中发现，本文模型在某些场景中也取得了优于对比模型的结果，并且在总体性能上也取得了具有竞争力的结果。这证明了所提网络具有较好的泛化性，以及在室外大型场景中重建效果的有效性。对于 Tanks and Temples 基准的部分场景的具体重建结果如图 6 所示。

表 4　在 Tanks and Temples 中等集上的定量分析结果

Table 4　Quantitative results on Tank and Temples intermediate dataset

Methods	Mean	Francis	Family	Horse	Lighthouse	M60	Panther	Playground	Train
ACMH[4]	54.82	49.45	69.99	45.12	59.04	52.64	52.37	58.34	51.61
CVPMVSNet[6]	54.03	47.74	**76.50**	36.34	55.12	57.28	54.28	57.43	47.54
COLMAP[16]	42.14	22.25	50.41	25.63	56.43	44.83	46.97	48.53	42.04
MVSNet[17]	43.48	28.55	55.99	25.07	50.79	53.96	50.86	47.90	34.69
R-MVSNet[20]	48.40	46.65	69.96	32.59	42.95	51.88	48.80	52.00	42.38
CasMVSNet[22]	**56.84**	**76.37**	58.45	**46.26**	55.81	56.11	54.06	58.18	49.51
UCSNet[24]	54.83	53.16	76.09	43.03	54.00	55.60	51.49	57.38	47.89
CIDER[35]	46.76	32.39	56.79	29.89	54.67	53.46	53.51	50.48	42.85
P-MVSNet[38]	55.62	44.64	70.04	40.22	**65.20**	55.08	55.17	**60.37**	**54.29**
Ours	55.75	55.68	74.51	45.38	55.39	**58.77**	**55.69**	54.76	45.77

图 6　Tanks and Temples 中间集点云重建结果

Fig. 6　Point cloudreconstruction resultson Tanks and Temples intermediate dataset

结束语　本文提出了一种轻量级的多阶段 MVS 重建深度学习模型，首先利用动态区域卷积和粗粒度代价体融合模块增强网络对语义信息的建模能力以及对场景全局和结构信息的提取能力，然后使用基于双边网格的上采样模块使得网络能够在尽可能保证重建效果的同时，减少内存消耗并加快推理速度。实验结果表明，本网络在获得更高效率的同时取得了具有竞争性的结果。未来将探索如何在保持高效率的基础上进一步提升模型最终的重建效果。

参 考 文 献

[1] ZHENG T X, HUANG S, LI Y F, et al. Key Techniques for Vision Based 3D Reconstruction: a Review[J]. Journal of Automation, 2020, 46(4): 631-652.

[2] HE Y, YANG J, HOU X, et al. ICP registration with DCA descriptor for 3D point clouds[J]. Optics Express, 2021, 29(13): 20423-20439.

[3] WANG X, WANG C, LIU B, et al. Multi-view stereo in the Deep Learning Era: A comprehensive review[J]. Displays, 2021, 70: 102102.

[4] XU Q, TAO W. Multi-scale geometric consistency guided multi-view stereo[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Press, 2019: 5483-5492.

[5] CHEN K, LIU X G. Global Optimized Multi-view 3D Reconstruction Method Based on Rays[J]. Computer Engineering, 2013, 39(11): 235-239.

[6] YANG J, MA W, ALVAREZ J M, et al. Cost volume pyramid based depth inference for multi-view stereo[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 4877-4886.

[7] LUO K, GUAN T, JU L, et al. Attention-aware multi-view stereo[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 1590-1599.

[8] LIU H J, BAI Z Y, CHENG W, et al. Fusion attention mechanism and multilayer U-Net for multiview stereo[J]. Chinese Journal of Image and Graphics, 2022, 27(2): 475-485.

[9] CHEN J, WANG X, GUO Z, et al. Dynamic region-aware convolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Press, 2021: 8064-8073.

[10] YAN J, WEI Z, YI H, et al. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking[C]// European Conference on Computer Vision. Cham: Springer, 2020: 674-689.

[11] XU B, XU Y, YANG X, et al. Bilateral grid learning for stereo matching networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Press, 2021: 12497-12506.

[12] KUTULAKOS K N, SEITZ S M. A theory of shape by space carving[J]. International Journal of Computer Vision, 2000, 38(3): 199-218.

[13] LHUILLIER M, QUAN L. A quasi-dense approach to surface reconstruction from uncalibrated images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 418-433.

[14] FURUKAWA Y, PONCE J. Accurate, dense, and robust multi-view stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(8): 1362-1376.

[15] GALLIANI S, LASINGER K, SCHINDLER K. Massively parallel multiview stereopsis by surface normal diffusion[C]// Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE Press, 2015: 873-881.

[16] SCHONBERGER J L, ZHENG E, FRAHM J M, et al. Pixelwise view selection for unstructured multi-view stereo[C]// European

Conference on Computer Vision. Cham: Springer, 2016: 501-518.
[17] YAO Y, LUO Z, LI S, et al. Mvsnet: Depth inference for unstructured multi-view stereo[C]// Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 767-783.
[18] GALLUP D, FRAHM J M, MORDOHAI P, et al. Real-time plane-sweeping stereo with multiple sweeping directions[C]// 2007 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2007: 1-8.
[19] CHEN R, HAN S, XU J, et al. Point-based multi-view stereo network[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Press, 2019: 1538-1547.
[20] YAO Y, LUO Z, LI S, et al. Recurrent mvsnet for high-resolution multi-view stereo depth inference[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Press, 2019: 5525-5534.
[21] YU Z, GAO S. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 1949-1958.
[22] GU X, FAN Z, ZHU S, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 2495-2504.
[23] CHEN R, HAN S, 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.
[24] CHENG S, XU Z, ZHU S, et al. Deep stereo using ADAPTIVE thin volume representation with uncertainty awareness[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 2524-2534.
[25] WEI Z, ZHU Q, MIN C, et al. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 6187-6196.
[26] MA X, GONG Y, WANG Q, et al. EPP-MVSNet: Epipolar-Assembling Based Depth Prediction for Multi-View Stereo[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 5732-5740.
[27] CHEN J, PARIS S, DURAND F. Real-time edge-aware image processing with the bilateral grid[J]. ACM Transactions on Graphics(TOG), 2007, 26(3): 103-112.
[28] BARRON J T, ADAMS A, SHIH Y C, et al. Fast bilateral-space stereo for synthetic defocus[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2015: 4466-4474.
[29] GHARBI M, CHEN J, BARRON J T, et al. Deep bilateral learning for real-time image enhancement[J]. ACM Transactions on Graphics(TOG), 2017, 36(4): 1-12.
[30] XU B, XU Y, YANG X, et al. Bilateral grid learning for stereo matching networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 12497-12506.
[31] AANES H, JENSEN R, VOGIATZIS G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.
[32] 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.
[33] YU Z, GAO S. Fast-mvsnet: Sparse-to-dense MULTI-view stereo with learned propagation and gauss-newton refinement[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 1949-1958.
[34] JI M, GALL J, ZHENG H, et al. Surfacenet: An end-to-end 3d neural network for multiview stereopsis[C]// Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE Press, 2017: 2307-2315.
[35] XU Q, TAO W. Learning inverse depth regression for multi-view stereo with correlation cost volume[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34(7): 12508-12515.
[36] YI H, WEI Z, DING M, et al. Pyramid multi-view stereo net with self-adaptive view aggregation[C]// European Conference on Computer Vision. Cham: Springer, 2020: 766-782.
[37] WEI Z, ZHU Q, MIN C, et al. Bidirectional Hybrid LSTM Based Recurrent Neural Network for Multi-view Stereo[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(1): 1-12.
[38] LUO K, GUAN T, JU L, et al. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 10452-10461.
[39] MOULON P, MONASSE P, PERROT R, et al. Openmvg: Open multiple view geometry[C]// International Workshop on Reproducible Research in Pattern Recognition. Cham: Springer, 2016: 60-74.

ZHANG Xiao, born in 1998, postgraduate. His main research interests include deep learning and 3D reconstruction.

DONG Hongbin, born in 1963, professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include evolutionary computation, machine learning and multi-agent system.

(责任编辑:何杨)