计算机科学
COMPUTER SCIENCE

基于多粒度特征聚合与二分搜索的高效多视图立体重建

许立君, 赵宇杰, 赵敏, 马为鹃, 陈侃松

引用本文

许立君, 赵宇杰, 赵敏, 马为鹃, 陈侃松. 基于多粒度特征聚合与二分搜索的高效多视图立体重建[J]. 计算机科学, 2026, 53(3): 257-265.

XU Lijun, ZHAO Yujie, ZHAO Min, MA Weixuan, CHEN Kansong. Efficient Multi-view Stereo Reconstruction Based on Multi-granularity Feature Aggregation and Binary Search [J]. Computer Science, 2026, 53(3): 257-265.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250200094

基于多粒度特征聚合与二分搜索的高效多视图立体重建

许立君　赵宇杰　赵　敏　马为駧　陈侃松
湖北大学计算机学院　武汉 430062
(xulijun@hubu.edu.cn)

摘　要　在基于深度学习的多视图立体重建方法中,代价体构建面临高计算复杂度和内存消耗的挑战。现有研究多采用级联架构或迭代优化方法降低内存消耗,但级联架构的粗到细采样策略可能导致细节信息丢失,削弱关键特征感知能力。为此,提出了一种基于级联结构的二分搜索与多粒度特征聚合的多视图立体网络框架。该框架通过级联架构减少内存占用,利用二分搜索策略将深度范围划分为多个预选区域,并通过离散分类方法压缩深度值搜索空间,提高深度检索效率并降低内存需求。此外,提出了多粒度特征信息聚合策略,将粗粒度全局语义信息嵌入细粒度代价体构建中,同时关注细粒度局部纹理信息。通过融合不同层次的特征表示,并在聚合模块中引入视图内自适应聚合和逐视图自适应加权策略,增强了模型对全局结构和局部细节特征的感知能力。实验结果表明,在 DTU 和 Tanks & Temples 公共数据集上,此方法在保持低内存消耗的同时,实现了优异的点云重建效果。

关键词:多视图立体;二分搜索策略;多粒度特征信息聚合策略
中图分类号　TP391.41

Efficient Multi-view Stereo Reconstruction Based on Multi-granularity Feature Aggregation and Binary Search

XU Lijun,ZHAO Yujie,ZHAO Min,MA Weixuan and CHEN Kansong
School of Computer Science,Hubei University,Wuhan 430062,China

Abstract　In deep learning-based multi-view stereo(MVS) reconstruction,cost volume construction faces challenges of high computational complexity and memory consumption. Existing studies often employ cascade architectures or iterative optimization methods to reduce memory usage. However,the coarse-to-fine sampling strategy in cascade structures may lead to the loss of fine-grained details,weakening the perception of critical features. To address this,this paper proposes a novel multi-view stereo network framework based on a cascade structure with binary search and multi-granularity feature aggregation. The proposed framework reduces memory overhead through a cascade architecture while employing a binary search strategy to partition the depth range into multiple candidate regions. A discrete classification method is introduced to compress the depth search space,improving depth retrieval efficiency and lowering memory requirements. Furthermore,this paper proposes a multi-granularity feature aggregation strategy that embeds coarse-grained global semantic information into fine-grained cost volume construction while preserving attention to fine-grained local texture details. By fusing multi-level feature representations and incorporating intra-view adaptive aggregation and view-wise adaptive weighting strategies in the aggregation module,the proposed model enhances the perception of both global structures and local detailed features. Experimental results on the DTU and Tanks & Temples benchmark datasets demonstrate that the proposed method achieves superior point cloud reconstruction quality while maintaining low memory consumption.

Keywords　Multi-view stereo,Binary search strategy,Multi-granularity feature aggregation

1　引言

多视图立体重建(Multi-View Stereo,MVS)旨在利用多幅经过严格校准的二维图像精确重构场景的三维几何形态。从数学本质上看,这一过程可被视为一个高维度的搜索问题。近年来,基于学习的 MVS 方法通过卷积神经网络进行密集深度预测。其中,3D 代价体为该类方法的核心。以 MVS-Net[1] 和 R-MVSNet[2] 为例,在预定义深度范围内构建离散化

到稿日期:2025-02-24　返修日期:2025-05-23
基金项目:武汉市知识创新专项——曙光计划项目(2022010801020327);湖北省重点研发计划项目(2022BAA045)
This work was supported by the Knowledge Innovation Program of Wuhan—Shuguang Project(2022010801020327) and Key Research and Development Program of Hubei Province(2022BAA045).
通信作者:陈侃松(kschen1999@aliyun.com)

的三维代价体(3D Cost Volume),该数据结构由空间维度($H\times W$)与深度假设维度 D 共同构成三维网格,每个网格单元通过可微分单应性变换聚合多视图特征差异,量化表征对应空间点在特定深度假设下的匹配置信度。该代价体通过多视图特征投影匹配编码场景几何分布,经 3D 卷积网络正则化优化概率分布并抑制噪声,最终沿深度维度回归,获得像素级深度估计。因此,3D 代价体构造的好坏,将显著影响多视图立体重建网络对复杂几何的建模能力与计算效率。

然而,在代价体构建过程中,内存资源的高需求成为一个关键挑战。因此,在保证模型性能的前提下,有效降低内存消耗,已成为当前研究的重点。构造代价体时,其内存占用与深度假设的数量直接相关,因此合理设定深度评估等级是控制内存占用的关键因素。MVSNet(Multi-view Stereo Network)系列方法以并行的方式密集搜索所有深度评估等级,造成了极大的内存压力。为了减少一次性的内存占用问题,CasMVSNet[3]采用了级联式的粗到细策略:先通过粗粒度的深度预测限定搜索范围,减小无效深度评估区域,从而减少了内存占用;再在限定搜索范围内进行细粒度的深度预测,以增加细节信息,保证深度预测的精确性。然而,这种方法还是要搜索全部粗粒度深度评估等级周围的区域,会消耗较多的内存,且没有考虑到图像形态信息;此外,当预测空间变大时,由于粗-细粒度的变化过大,密集搜索的准确性将降低,且细粒度搜索范围增大会增加内存消耗。GBi-Net[4]通过自适应动态调整深度搜索范围,减少了对全部深度评估等级进行搜索的内存开销,同时结合图像形态信息,使得搜索范围更精准,显著提升了深度预测的精度。然而,GBi-Net 通过多阶段迭代进行优化,虽然能提高预测精度,但也带来了额外的内存开销和计算负担。动态调整机制的引入进一步增加了模型的计算复杂性,导致推理速度下降;此外,GBi-Net 对超参数的设置较为敏感,尤其是在多阶段迭代过程中,不合理的参数可能使内存管理失控,或者影响动态调整的有效性;如果初始粗粒度预测范围不够准确,还可能限制后续阶段的优化效果。因此,设计高效的自适应动态调整机制,在减少无效深度搜索区域的同时兼顾全局优化和跨阶段信息,是减少内存开销并提高深度预测精度的关键。

针对以上问题,本文采用一种基于级联结构的二分搜索策略,将深度预测范围划分为多个预选区域,然后通过二进制分类来压缩深度值的搜索区域并构建 3D 代价体。此方法通过降维和减少深度估计数量,将代价体大小降低到更小的边界。在每次二分搜索的过程中,逐步排除一半的无效深度范围,经过粗粒度、中等粒度和细粒度 3 个逐级优化阶段,快速收敛到高精度的深度估计结果。

为了获取多视角图像之间的关联和长距离上下文信息,本文提出了多粒度特征信息聚合策略,将粗粒度的全局特征引入细粒度的代价体构建中,以引导网络构建包含全局语义和局部细节的代价体,从而保留粗阶段的语义特征和细阶段的纹理特征,保证每个阶段都能从全局和局部混合信息中收益,以提高代价体的稳定性。

本文的主要贡献如下:

(1)为了压缩 3D 代价体的大小,并保证模型的有效性,采用了一种二分搜索策略,该方法通过联级结构减小代价体空间,并在每个搜索阶段消除一半的搜索空间,进一步压缩代价体的大小,减少内存占用;

(2)设计了一种多粒度特征信息聚合方法,融合局部和全局信息,增强了代价体的稳定性,提升了深度预测的精确度。

在 DTU 和 Tanks & Temples 公共数据集上的实验结果表明,本文方法在保持较小内存消耗的情况下,实现了更加优秀的点云重建效果和良好的泛化性。

2 相关工作

2.1 基于学习的 MVS 方法

传统的 MVS 方法主要采用光度一致性来推断物体的几何结构,但是在无纹理和反光区域,由于特征难以提取,因此点云重建结果不佳。基于学习的 MVS 方法通过特征金字塔等网络结构,能够在纹理、光照影响的情况下提取多视图特征信息。Ji 等[5]提出了基于体素的点云重建网络模型 SurfaceNet,为使用深度神经网络重建多视角图像提供了基础思路。该方法将整个空间分解成立方体,并以立方体为单位对曲面进行回归,利用卷积神经网络结合全局语义信息来实现三维重建。Yao 等[1]将 MVS 中的图形学基础和深度神经网络相结合,开创性地提出了一种基于深度神经网络的多立体视图视觉网络 MVSNet,用于生成三维稠密点云。该方法采用了特征金字塔网络(Feature Pyramid Network,FPN)代替传统尺度不变特征变换(Scale-Invariant Feature Transform,SIFT)方法提取多视角图片中的特征点,缓解了纹理单一和光照影响产生的特征识别难题;使用单应性变换将多视角图像对齐到参考图像空间构建 3D 代价体,并使用三维卷积神经网络对代价体进行正则化形成概率体,最后通过回归网络预测特征点的深度值。MVSNet 利用深度神经网络强大的隐式特征提取能力,显著提升了稠密点云重建的精度。然而,该算法构建 3D 代价体时每个视图和每个深度评估等级都需要被存储,内存消耗巨大,难以完成更多视图和更细腻的深度预测任务,这也是 MVSNet 系列算法亟待解决的关键难点问题。

R-MVSNet 采用了循环神经网络(RNN)结构处理多视角图像之间的关联性,从而在同样的内存消耗情况下更有效地利用多视角数据中的信息。但此方法并没有缩减 MVSNet 的内存占用。Fast-MVSNet[6]通过构建一个稀疏代价体来学习稀疏的高分辨率深度图,通过利用小尺度卷积对局部区域像素的深度依赖进行编码来处理稀疏的深度图,减少了内存占用;但是重建的精度并不太理想。TransMVSNet[7]利用 Transformer[8]来捕获长程全局上下文信息,以提高深度估计和立体匹配的性能,显著提升了 MVSNet 的深度预测精度。GBi-Net 通过自适应地动态调整深度搜索范围的方式,在多阶段级联结构中逐步缩小无效搜索空间,结合图像形态信息动态优化深度评估范围,显著减少了内存占用,同时提升了深度预测的精度。然而,GBi-Net 的多阶段迭代方法在提升精度的同时,也增加了计算复杂性和一定的内存开销。

2.2 级联的 MVSNet

使用级联结构[3]的 MVSNet 能够提高重建分辨率,兼顾内存效率和计算成本。这类方法借鉴了特征金字塔的策略,

从粗到细在多个阶段逐步优化和细化深度估计。在粗粒度阶段，通过加大深度间隔距离来预测低分辨率深度图，减小深度搜索范围；在细粒度阶段，通过上采样细化深度图，提升深度预测精度。CVP-MVSNet[9]使用基于代价体的特征金字塔结构，在有限内存下回归高分辨率深度图。UCS-Net[10]通过自适应地构建薄体积表示来改进 MVSNet 中使用立方体来表示场景的方式，该自适应薄体积表示方法可以根据场景的特点和上下文信息动态调整体积的大小和分辨率，从而更有效地利用图像信息，提高深度估计的准确性和鲁棒性。尽管由粗到细的架构能够减小深度搜索范围，从而减少代价体的内存消耗，但在深度范围大的场景中，由于粗到细阶段差异大，深度搜索范围依然难以压缩。若为了压缩深度搜索范围而固定深度评估等级，则会导致细节信息丢失，进而降低深度预测的精度。Patchmatch[11]是一种深度假设传播和评估机制，基于局部像素间的相似性来估计深度，降低了计算复杂度，减少了内存占用，通过舍弃不必要的 3D 代价体积繁重的正则化过程，实现了模型效率的大幅提升。

3 本文方法

3.1 模型结构

基于多粒度特征聚合与二分搜索策略的多视图立体视觉网络采用联级结构，总体结构如图 1 所示，主要由特征提取、代价体构建和深度估计构成。首先，使用 3 层特征金字塔提取参考图像和多视角源图像的多尺度特征。然后，通过单应性变化将源图像特征映射到参考图像空间构建 3D 代价体，并对其进行融合。本文算法使用特征金字塔网络(FPN)[12]作为提取图像的多尺度特征，给联级结构提供不同粒度的特征图。对于每张图片，FPN 可以获得一个包含 3 种不同尺度的特征图。此外，在特征金字塔的每个输出层采用可变形的卷积网络(Deformable Convolutional Network，DCN)[13]作为输出层，以便适应复杂几何形状和非刚性形变。

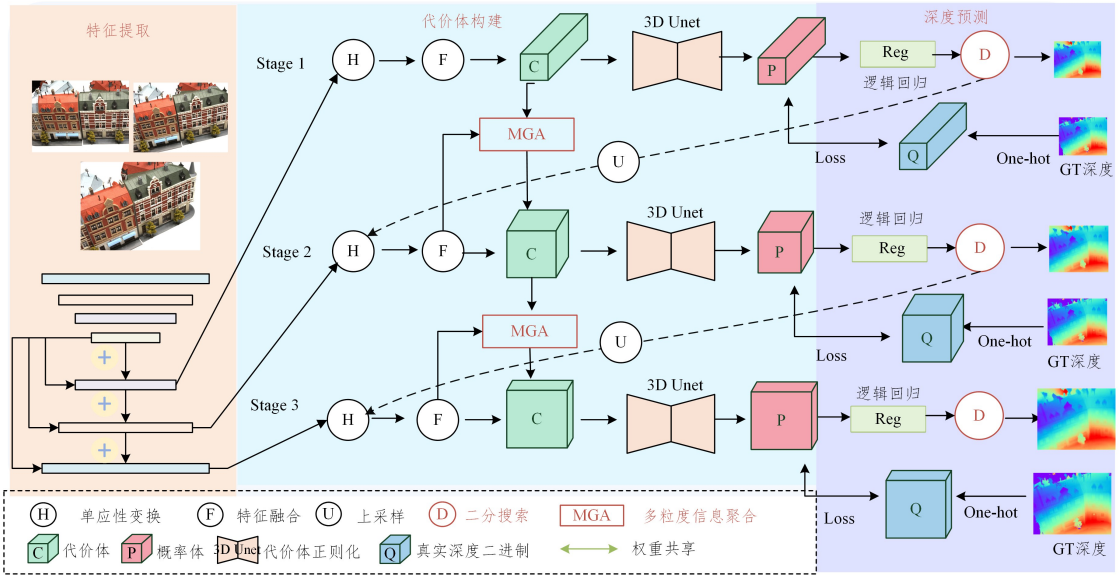


图 1 模型架构

Fig. 1 Architecture of the proposed model

3.1.1 代价体构建

在 MVS 任务中，通过存储不同深度假设下的匹配，为多视图立体匹配提供了一个结构化的三维表示，以便网络进行深度图的估计和优化。

首先，对于 FPN 提取得到的特征图，通过单应性变换依次将源视图特征图映射到参考图像锥形视角下的深度假设平面，形成 $N-1$ 个特征体$\{\boldsymbol{V}_i\}_{i=1}^{N-1}$。源图像中通过深度假设操作 d 的变换像素 $\bar{\boldsymbol{q}}$ 如式(1)所示：

$$\bar{\boldsymbol{q}}=\boldsymbol{K}_i\cdot(\boldsymbol{R}_{F_0\to F_i}\cdot\boldsymbol{K}_0^{-1}\cdot q\cdot d+\boldsymbol{T}_{F_0\to F_i}) \tag{1}$$

通过多粒度特征信息聚合模块将粗粒度代价体特征传递到细粒度阶段。最后，使用 3D-Unet 将代价体正则化为概率体，并使用逻辑回归预测深度评估等级，利用二分搜索策略压缩下一阶段的深度搜索范围。

3.1.2 特征提取

多视角图片输入$\{\boldsymbol{X}_i\}_{i=0}^{N-1}$由 N 个图像组成，其中 $\boldsymbol{X}_0$ 为参考图像，$\{\boldsymbol{X}_i\}_{i=1}^{N-1}$为包含 $N-1$ 个源图像的集合。其中，$\boldsymbol{F}_0$ 表示参考特征图，F_i 表示源特征图，q 为参考图像的像素，$\boldsymbol{R}_{F_0\to F_i}$ 和 $\boldsymbol{T}_{F_0\to F_i}$ 分别表示 $\boldsymbol{F}_0$ 到 $\boldsymbol{F}_i$ 相对于相机的旋转操作和平移操作。

然后，假定参考特征图 $\boldsymbol{F}_0$ 和源特征图 $\boldsymbol{F}_i$ 具有相同的通道维度 N_C，将特征图的通道维度划分成 N_G 组，每个特征组的通道数为 N_C/N_G。计算第 k 搜索阶段源特征的第 i 个代价体，如式(2)所示：

$$\boldsymbol{V}_i(k,\boldsymbol{q},g)=\frac{N_G}{N_C}\langle\boldsymbol{F}_0^g(\boldsymbol{q}),\boldsymbol{F}_i^g(\bar{\boldsymbol{q}})\rangle \tag{2}$$

其中，$\boldsymbol{F}_0^g$ 表示参考特征的第 g 个特征组，$\boldsymbol{F}_i^g$ 表示第 i 张源图像特征的第 g 个特征组，q 为参考图像的像素，$\langle\cdot,\cdot\rangle$表示内积操作。通过内积操作，模型可以更有效地处理大规模的特征图，从而更有效地构建代价体。

最后，采用三维卷积神经网络来预测像素权重矩阵$\{\boldsymbol{W}_i\}_{i=1}^{N-1}$，并将权重矩阵与代价体集合融合，沿着深度方向执行 Softmax(·)函数，生成用于深度预测的概率体 $\boldsymbol{P}$。

$$\boldsymbol{P}(k,\boldsymbol{q},g)=\mathrm{Softmax}\left(\frac{\sum_{i=1}^{N-1}\boldsymbol{W}_i(\boldsymbol{q})\times\boldsymbol{V}_i(j,\boldsymbol{q},g)}{\sum_{i=1}^{N-1}\boldsymbol{W}_i(\boldsymbol{q})}\right) \tag{3}$$

在代价体构建过程中，创新性地引入了联合自适应和逐项加权的多粒度特征提取模块（详见 3.2 节）。该模块能够同时获取局部和全局信息，从而提升代价体的稳定性和模型的鲁棒性。

3.1.3 深度估计

在获取概率体后，采用逻辑回归模块预测每个体素最可能的深度评估等级，并生成深度图。在本文模型框架中，规定深度评估等级为 128，深度范围归一化为[0,1]。在每一阶段的深度估计过程中，深度评估等级不变，通过二分搜索动态压缩深度搜索范围，达到减少代价体存储量并提升深度搜索效率的目的（详见 3.3 节）。

3.2 多粒度特征信息聚合模块

本文提出的多粒度特征信息聚合策略如图 2 所示。在细采样阶段，模型进行自适应采样时，由于数据分布不均匀，在特征融合为代价体时会丢失部分特征信息，从而影响深度估计结果。为了解决这个问题，本文提出了多粒度特征信息聚合策略来丰富代价体的特征信息。首先，通过一个 3×3 的二维卷积神经网络，将前一阶段代价体中的粗粒度全局语义特征与当前阶段生成代价体时提取的局部纹理特征进行融合，并通过一个 1×1 的二维卷积神经网络对融合后的代价体进行正则化。最终获得的代价体 $\hat{\boldsymbol{V}}^l$ 如式(4)所示：

$$\hat{\boldsymbol{V}}^l=\boldsymbol{CNN}_{1\times1}[\boldsymbol{CNN}_{3\times3}(\boldsymbol{V}^{l-1}),\boldsymbol{CNN}_{3\times3}(\boldsymbol{V}^l),\boldsymbol{V}^l] \tag{4}$$

其中，$\boldsymbol{V}^{l-1}$ 表示从前一阶段采样得到的代体，$\boldsymbol{V}^l$ 表示当前阶段的代价体，$[\cdot,\cdot]$表示链接操作，$\boldsymbol{CNN}_{1\times1}$ 表示 1×1 的二维卷积神经网络，$\boldsymbol{CNN}_{3\times3}$ 表示 3×3 的二维卷积神经网络。在这里，1×1 的二维卷积神经网络增强了图像的特征表示，能促进跨通道特征的信息交互，提升模型对图像的抽象能力，并有效减少参数量和降低计算复杂性，使得模型更轻量和高效；而3×3的二维卷积神经网络则能够有效捕捉图像中的空间交互信息、局部结构和上下文信息的差异性，有益于模型详细理解场景。

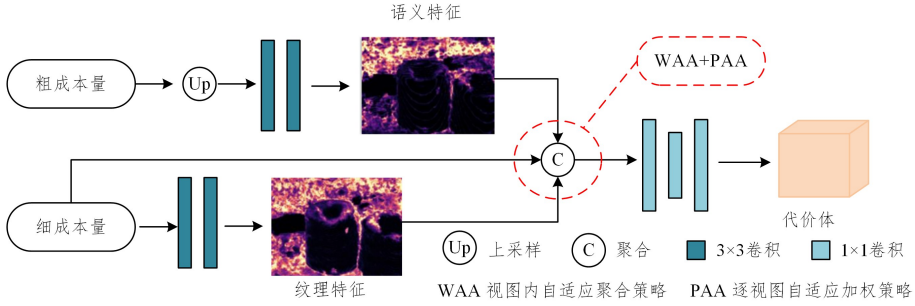


图 2 多粒度特征信息聚合模块

Fig. 2 Multi-granularity feature aggregation module

此外，在聚合部分采用了视图内自适应聚合策略（WAA）和逐视图自适应加权策略（PAA）来获取局部和全局信息。

（1）视图内自适应聚合策略

在代价体构建过程中，由于 2D 卷积神经网络在固定的 2D 网格上运行，无法实现图 3 中视图内自适应聚合策略（WAA）所需的感受野自适应变化能力，因此很难处理反射表面和低纹理或者无纹理区域。基于此问题，引入了一种创新的视图内自适应聚合策略，如图 3 所示。该策略旨在优化特征提取过程，特别是针对不同纹理丰富度区域的处理。

视图内自适应聚合策略基于不同纹理区域对感受野需求差异的直观观察，即纹理缺乏的区域需要较大的感受野以捕获更多全局特征，而纹理丰富的区域则适合较小的感受野进行精细化处理。该模块利用可变形卷积（DCN）动态调整卷积核的采样位置，以适应不同区域的特征提取需求。可变形卷积的定义如式(5)所示。具体实现中，输入特征图首先进行多尺度处理，并分别传入 3 个具有独立参数的可变形卷积网络，这些网络根据局部动态结构自适应地调整采样位置，从而更高效地提取多尺度特征。接着，通过双线性插值将较小的特征图调整到统一大小，并融合生成一个通道数为 32、大小为 $H\times W$的特征图。该融合过程充分整合了不同尺度的信息，增强了特征的表达能力，为代价体构建提供了稳定且全面的特征基础。

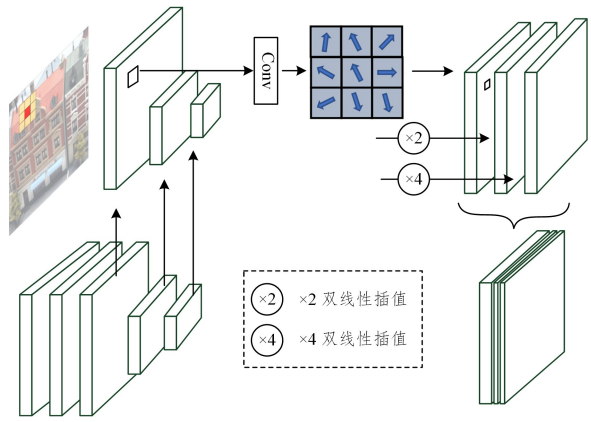


图 3 视图内自适应聚合策略（WAA）

Fig. 3 Intra-view adaptive aggregation strategy(WAA)

$$f'(p)=\sum_k \boldsymbol{w}_k\times f(p+p_k+\Delta p_k)\times\Delta m_k \tag{5}$$

其中，$f(p)$表示特征像素 p 的值；$\boldsymbol{w}_k$ 和 p_k 分别表示传统卷积操作中定义的核参数和固定偏移量；Δp_k 和 Δm_k 是可变形卷积网络自适应产生的偏移量和调节权重。将较小的特征映射插值到 $H\times W$ 特征图上，分别得到 3 个具有 16，8 和 8 个通道的特征映射，并将这些特征连接起来，构成更高维度（$H\times W\times32$）的特征图。特征图的融合，增加了特征的丰富性和多样性，为后续的代价体构建提供了更具鲁棒性和全面性的特征基础，从而能够更好地捕捉复杂的几何信息和纹理细节。

(2)逐视图自适应加权策略

在获得逐视图代价体后，下一步是将所有代价体聚合成一个用于正则化的代价体。通常的做法是对各视图的代价体取平均，以确保每个视图在最终深度估计中的贡献相同，不会出现某个视图被过度放大或被忽略的情况。但是这种简单的平均并不足够合理，不同的拍摄角度可能会导致遮挡问题，使得某些区域在某些视图中缺乏可见性，从而影响深度估计的准确性。为此，设计了一个逐视图自适应加权策略来全面融合不同视图的信息，如图 4 所示。

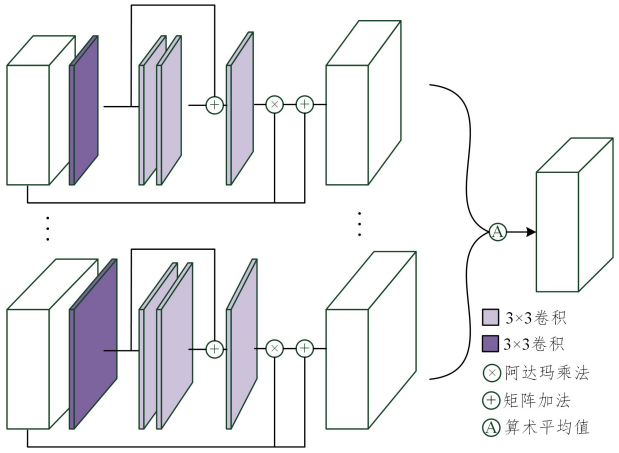


图 4 逐视图自适应加权策略(PAA)

Fig. 4 Per-view adaptive weighting strategy(PAA)

逐视图自适应加权策略通过通道数为 4，4，4 和 1 的结构逐步捕获和压缩多层次特征信息，增强对多源视图信息的建模能力。随后，利用大小为 $H\times W$ 的注意力图灵活调整像素权重，优化深度估计的关键上下文信息。最后，对代价体进行加权求和并归一化，完成特征整合(见式(6))。

$$\boldsymbol{C}^{(d)}=\frac{1}{N-1}\sum_{i=1}^{N-1}[1+\boldsymbol{w}(\boldsymbol{c}_i^{(d)})]\boldsymbol{e}\boldsymbol{c}_i^{(d)} \tag{6}$$

逐视图自适应加权模块利用 Hadamard 乘法，将代价体中的元素与注意力图中的对应元素相乘，生成像素级别的加权结果。注意力图根据视图代价体自适应生成，动态调整像素贡献：抑制易引起匹配混淆的像素权重，突出关键上下文信息，从而更好地适应不同像素特征。同时，通过引用全局信息避免过度平滑，保留深度估计中的细节。

3.3 基于级联结构的二分搜索策略

在构建代价体的过程中，MVSNet 通常需要为每个像素密集采样深度假设，导致内存消耗巨大。针对这个问题，基于级联的 MVSNet 方法[3]被提出。该方法通过逐层粗构建代价体的方式，在一定程度上减少了内存使用；然而，在每个迭代阶段，采样仍然相当密集，因此模型效率仍不够理想。为了进一步减少模型对内存的占用，本文采用了一种二分搜索策略，其过程如图 5 所示。

该策略在采样深度值的过程中，将给定深度范围动态划分成多个区域，通过二分法搜索真实的深度。下面介绍该策略的具体过程。

在第 K 个搜索阶段，将深度范围划分成两个相同的区域 $\{\boldsymbol{D}_{k,j}\mid j=1,2\}$，第一阶段中，每个区域的宽度为 $R/2$。在深度假设的过程中需要在深度范围内进行连续插值和变换操作，直接使用离散的区域会导致图像信息的损失和不准确的深度假设。对此，本文通过采样这两个区域的中心点来表示深度假设。对于这两个区域，它们共有 3 条边 $\{e_{k,m}\mid m=1,2,3\}$，其中区域 $\boldsymbol{D}_{k,j}$ 的两条边为 $e_{k,j}$ 和 $e_{k,j+1}$。两个区域的深度假设 $d_{k,j}$ 的计算如式(7)所示：

$$d_{k,j}=\frac{e_{k,j}+e_{k,j+1}}{2},j=1,2 \tag{7}$$

接着，根据式(2)构建代价体，并对区 $\boldsymbol{D}_{k,1}$ 和区域 $\boldsymbol{D}_{k,2}$ 进行标签预测。深度假设的预测标签表示真实深度值是否在对应的区域内。在第 K 个搜索阶段，网络输出概率体 P 以后，对其进行 argmax(·)运算，得到标签 j，该标签表示真实深度值位于区域 $\boldsymbol{D}_{k,j}$ 内。在 $K+1$ 搜索阶段，将区域 $\boldsymbol{D}_{k,j}$ 进一步划分成两个相等的新区域 $\boldsymbol{D}_{k+1,1}$ 和 $\boldsymbol{D}_{k+1,2}$。在这个阶段，新区域的 3 条边可以定义为：

$$e_{k+1,1}=e_{k,j},e_{k+1,2}=\frac{e_{k,j}+e_{k,j+1}}{2},e_{k+1,3}=e_{k,j+1} \tag{8}$$

对 $\boldsymbol{D}_{k+1,1}$ 和 $\boldsymbol{D}_{k+1,2}$ 的中心点进行新的采样来表示 $K+1$ 搜索阶段的深度假设。在二分搜索策略的过程中，初始搜索阶段的区域宽度则为 $R/2$；而在第 K 个搜索阶段，区域宽度为 $R/2^K$。

通过计算像素真实深度值确定其所属区域，并利用独热编码生成代价体计算损失。仅保留当前区域内(满足式(9))的有效像素用于损失计算，利用有效像素梯度更新网络参数，避免无用梯度累积，提高优化效率。如果像素的真实深度值在当前区域内，则认为该像素是真实有效的。具体流程如下：假设像素的真实深度值为 d_{gt}，而当前的区域边缘为 $\{e_{k+1,m}\mid m=1,\cdots,D+1\}$，只有当像素满足式(7)时，才是有效的。掩码机制保证了只有在当前搜索阶段具有有效标签的像素才会对损失进行贡献，通过有效像素的损失梯度更新网络中的参数，能够有效避免无用的像素梯度累积。

$$e_m\leqslant d_{gt}\leqslant e_{m+1} \tag{9}$$

通过二分搜索策略，3D 代价体的深度维度成功减少到 2，使得整个代价体的大小大幅度降低，极大地减少了内存的消耗。该策略不仅使得整个 MVS 网络的内存开销由 2D 图像特征提取器所主导，而且不再受 3D 代价体的限制。

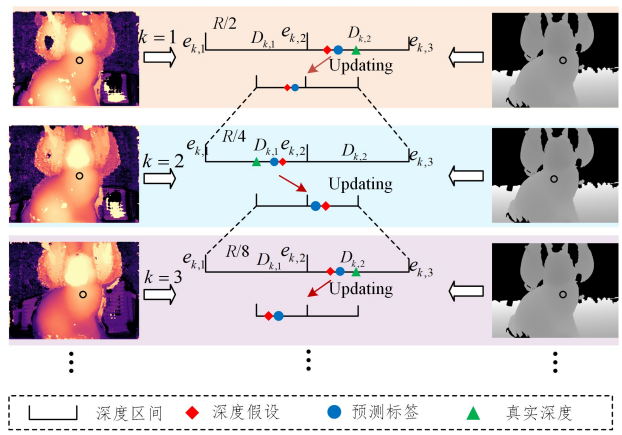


图 5 二分搜索策略流程图

Fig. 5 Flowchart of binary search strategy

3.4 损失函数

为将深度预测任务适配于深度学习框架的优化特性，本

文算法采用离散化深度假设空间策略，将传统回归问题转换为多分类任务。因此，本文通过交叉熵损失函数来计算训练损失，如式(10)所示：

$$Loss=\sum_{\tilde{q}\in j}\sum_{j=1}^{D}-\boldsymbol{G}(j,\tilde{q})\log\boldsymbol{P}(j,\tilde{q}) \tag{10}$$

其中，j 表示通过掩码图获取的有效像素集合，$\tilde{q}$ 表示一个像素值，D 表示最终的区域数目，$\boldsymbol{G}(j,\tilde{q})$ 表示深度假设的真实有效概率值。

4 实验与结果分析

4.1 数据集和评价指标

为了验证所提模型的有效性，在数据集 DTU[14] 和 Tanks and Temples[15] 上进行一系列对比实验，主要包括：基于 DTU 数据集的对比实验，基于 Tanks and Temples 数据集的对比实验，内存参数比较以及消融实验。

DTU 是一个包含多视图图像和相机姿态的大规模立体数据集，包含 124 个不同的场景，每个场景拥有 49 个视角和 7 种光照条件，其中每张图像的分辨率为 1600×1200。DTU 数据集涵盖了室内、室外和复杂的几何结构等多种场景，因此非常适合在现实条件下训练和测试深度学习 MVS 方法。

Tanks and Temples 是一个包含室外各种场景的大规模数据集，它包含一个中级子集和高级子集。中级子集包含 Family，Francis 和 Horse 等不同场景。不同场景拥有不同尺度及表面反射和曝光条件。通过将重建点云提交到官方网站对该数据集进行基准在线评估，以验证本文模型的泛化能力。

对 DTU 数据集的评估，本文采用准确度 Acc.(Accuracy)、完整度 Comp.(Complrtion)、综合度 Overall 以及内存参数 Mem.(Memory)作为指标。其中，Acc. 表示重建三维点云到真实点云的最近距离的平均值，如式(11)所示：

$$Acc.=\frac{1}{|\boldsymbol{S}_1|}\sum_{x\in S_1,y\in S_2}\min\|x-y\|^2 \tag{11}$$

其中，$\boldsymbol{S}_1$ 表示重建三维点云集合，$\boldsymbol{S}_2$ 表示真实点云集合，x 表示重建三维点云中的一个空间点，y 表示真实点云中的一个空间点。

完整度表示重建点云与真实点云的匹配程度，如式(12)所示：

$$Comp.=\frac{1}{|\boldsymbol{S}_1|}\sum_{x\in S_1,y\in S_2}\min\|y-x\|^2 \tag{12}$$

为了对重建点云的综合性能进行全面评估，将准确度和完整度相加，然后取和的平均值得到综合评价分数，如式(13)所示：

$$Overall=\frac{Acc.+Comp.}{2} \tag{13}$$

对数据集 Tanks and Temple 的评估，本文利用 F-score 值作为重建点云的评估指标。

4.2 实验设置

4.2.1 训练参数设置

本文模型在 PyTorch 1.90，Python 3.8，RTX 3090 24 GB GPU 环境下进行了一系列实验。在模型的训练阶段，本文模型的训练参数设置如表 1 所列。

表 1 训练参数设置

Table 1 Training parameters settings

参数	设置
图输入图像大小	640×512
迭代输入图像数量	5
深度假设数量	48,32,8
深度间隔设置	4,2,1
深度采样范围	[425,935]mm
初始深度平面数	192
训练周期	16
Learning rate	0.001
Batch size	4

4.2.2 复杂度测量方法

为了评估本文方法及其他对比方法的计算复杂度，对推理阶段的最大 GPU 显存占用(Mem.)进行了统计，其单位为 MB。Mem. 主要受 cost volume 构建、3D CNN 处理及深度回归等模块的影响，反映了算法的内存需求。在 DTU 数据集的标准测试视角下，保持相同的输入尺寸和 batch size ($batch=4$)，并采用 nvidia-smi 工具实时监测 NVIDIA RTX 3090 显卡的显存状态，截取从模型加载至深度图输出全过程的显存峰值作为最终 Mem. 值。该测量方式与 MVSNet 论文中的方法保持一致，确保了实验结果的可比性和复现性。

4.3 对比实验

4.3.1 基于 DTU 数据集的实验结果分析

为了验证本文模型的有效性，在两种不同类型的度量标准上对 DTU 数据集进行评估实验。首先，通过 DTU 官方评估脚本对真实点云和生成点云之间的距离进行综合评估，对比结果如表 2 所列，其中黑体字表示指标的最优值，下划线表示指标的次优值，实验的可视化结果如图 6 所示。

表 2 基于 DTU 数据集的实验对比结果

Table 2 Experimental comparison results on DTU dataset

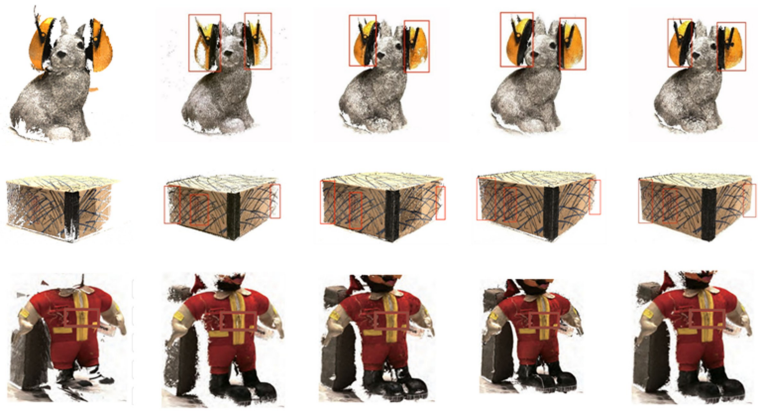
图像方法	评价指标			
	Acc.	*Comp.*	*Overall*	*Mem.*/MB
R-MVSNet[2]	0.385	0.459	0.422	9384
CasMVSNet[3]	<u>0.325</u>	0.385	0.355	4591
AA-RMVSNet[16]	0.376	0.339	0.357	11973
Patchmatchnet[12]	0.427	0.277	0.352	**1629**
UniMVSNet[17]	0.352	0.278	0.315	4057
GBi-Net[4]	0.327	0.268	<u>0.298</u>	2108
CT-MVSNET[18]	0.341	**0.264**	0.302	5623
DMVSNET[19]	0.338	0.272	0.305	3126
本文模型	**0.322**	<u>0.265</u>	**0.294**	<u>2102</u>

从表 2 可以得知，本文模型虽然在完整度上表现次优，但是在整体性能的综合度上取得最优结果。本文模型的完整度得分相比于 UniMVSNet 的 0.278 提升到了 0.265，虽然较 CT-MVSNet 的完整度差了 0.001，但是在综合度上是所有方法中最好的。实验结果表明，本文模型采用的动态采样策略压缩了代价体的大小，降低了计算复杂度。此外，本文模型利用多粒度特征信息聚合策略加强对不同粒度信息的利用，增强特征信息表示，在减少参数的情况下，整体性能仍表现最佳，这充分说明其更具竞争力。

本文方法在使用更少的内存的情况下显示出重建质量的改进，与 R-MVSNet 相比，内存占用减少了 77.6%，与 CasMVSNet 相比减少了 54.2%，与 AA-RMVSNet 相比减少

了 82.4%，与 CT-MVSNet 相比减少了 62.6%，与 DMVSNet 相比减少了 32.8%，与 GBi-Net 相比减少约 0.3%。虽然本文方法的内存比 Patchmatchnet 略大 473 MB，但其在精准度、完整度以及综合性能方面都有很大优势。

本文模型与 CasMVSNet 和 UniMVSNet 在 DTU 数据集场景 33、场景 10 和场景 49 的可视化对比结果如图 6 所示。



(a)ground truth (b)CasMVSNet (c)TransMVSNet (d)UniMVSNet (e)本文模型

图 6 场景 33、场景 10 和场景 49 的可视化对比结果图

Fig. 6 Visual comparison result diagrams of Scene 33, Scene 10, and Scene 49

由图 6 可知，本文模型比其他模型在低纹理和边缘表面上能够产生更加明显且准确的深度估计和全面的点云。实验结果证明了本文方法的有效性，多粒度特征信息聚合策略能够使得网络在构建代价体的过程中考虑全局语义信息，提高对低纹理和边缘细节的感知，同时考虑当前阶段初始代价体中的局部纹理信息，增强了网络对全局和局部特征的捕获能力。

4.3.2 基于 Tanks and Temples 数据集的实验结果分析

数据集 DTU 在室内采集，缺少外界因素的干扰。为了验证本文模型的鲁棒性和泛化能力，本小节在 Tanks and Temples 数据集上对其进行基准测试。本文模型与其他模型在高级子集和中级子集上的实验结果对比如表 3 与表 4 所列（第一名用黑体表示，第二名用下划线表示）。

表 3 Tanks and Temples 数据集中级子集实验结果对比

Table 3 Experimental comparison results on the medium subdataset of the Tanks and Temples dataset

图像方法	Mean	Family	Francis	Horse	Light-huse	M60	Panther	Playground	Train
R-MVSNet[2]	50.14	73.41	54.46	43.63	43.43	46.83	46.69	50.53	45.22
CasMVSNet[3]	56.55	76.01	58.46	46.42	55.88	56.82	54.69	58.87	46.25
Patchmatchnet[11]	53.42	66.36	52.54	43.20	54.53	52.11	49.02	54.17	50.56
AA-RMVSNet[16]	61.15	77.99	59.64	51.24	**64.87**	**64.05**	59.54	60.21	**55.81**
EPP-MVSNet[20]	56.51	72.66	51.53	51.53	58.05	58.63	56.47	57.85	49.50
TransMVSNet[7]	**63.88**	80.21	65.02	**56.78**	62.63	61.37	60.21	60.07	51.89
GBi-Net[4]	61.42	79.77	67.69	51.81	61.25	60.37	55.87	60.67	53.89
MFE-MVSNe[21]	60.02	79.28	62.23	49.47	61.43	61.46	57.34	57.45	51.49
本文模型	63.31	**82.10**	**67.69**	51.58	61.25	63.08	**65.07**	**60.83**	54.89

表 4 Tanks and Temples 数据集高级子集实验结果对比

Table 4 Experimental comparison results on the advanced subdataset of the Tanks and Temples dataset

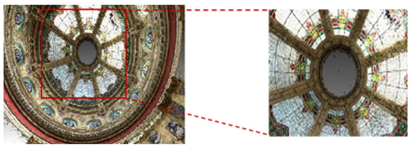
图像方法	Mean	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
R-MVSNet[2]	29.45	20.04	31.05	29.56	42.33	22.85	30.99
CasMVSNet[3]	31.08	19.88	38.52	29.43	43.56	27.33	28.32
Patchmatchnet[11]	32.11	23.07	37.46	30.10	41.87	28.36	32.11
AA-RMVSNet[16]	33.31	20.69	40.73	32.04	46.80	29.31	32.71
EPP-MVSNet[21]	34.53	20.96	42.15	33.05	45.01	29.28	35.71
TransMVSNet[7]	37.39	24.22	**44.39**	34.77	46.08	**34.69**	36.09
GBi-Net[4]	37.32	**29.77**	42.12	36.30	47.69	31.11	36.93
MFE-MVSNet[21]	36.04	24.16	41.13	33.47	**48.64**	32.10	36.72
本文模型	**37.76**	29.75	42.20	**36.31**	47.49	31.69	**39.12**

从表 3 与表 4 可以得知，本文模型在高级子集中取得了最佳的 Mean 值 38.4，并在 Auditorium，Courtroom 以及 Temple 场景下表现出更好的性能，表明本文模型在不同大规模室外场景中具有卓越的性能，通过引入二分搜索策略，加大深度范围，提高了深度假设的准确率，降低了计算复杂度。此外，还添加掩码梯度优化，避免无用的像素梯度累积。而在中级子集中，本文模型虽然未能获得最好的 Mean 结果，但在内存显著减少的情况下，平均得分仅比 TransMVSNet 低 0.7；此外，在 Family，Francis，Panther 以及 Playground 场景中仍获得最佳表现，模型保持了竞争力。

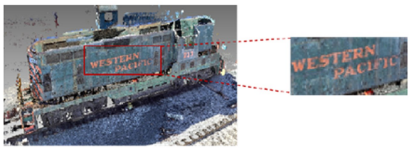
为了更直观地显示本文方法在复杂环境下的泛化能力，本文展示了在中间级 3 个不同场景的三维点云重建结果，如图 7 所示。可以看出，本文模型能够在不同复杂场景下生成完整且平滑的点云，这得益于二分搜索策略能够在深度值范

围内快速找到真实深度值，有效降低了计算复杂度。同时，多粒度特征信息融合策略充分利用了全局和局部特征信息，提高了 3D 代价体的质量，使其更适合深度图的生成，提升了模型在复杂场景中的表现，并确保重建过程中细节的保留和精度的提升。

(d) Family



(b) Palace



(c) Train

图 7 Tanks and Temples 三维点云重建可视化结果

Fig. 7 3D reconstruction visualization results of Tanks and Temples

4.4 消融实验

为了验证本文方法的有效性，本节在数据集 DTU 上设置不同的实验条件来进行消融实验，通过逐步移除特征融合模块等设置，分析各组件对三维重建精度的影响。

4.4.1 搜索策略有效性评估

为了验证本文二分搜索策略的有效性，本小节设计了一组消融实验来比较不同检索策略的效果，其中包括密集线性搜索策略（Dense LS）[1]、密集粗到细搜索策略（Dense C2F）[3]。以本文模型为基线，去除二分搜索策略的模型记为 No-G，去除二分搜索策略和多粒度特征信息聚合策略的模型记为 No-All，实验结果如表 5 所列。

表 5 不同搜索策略的实验结果对比

Table 5 Experimental comparison results of search strategies

图像方法	评价指标			
	Acc.	*Comp.*	*Overall*	*Mem.*/MB
No-All	0.396	0.386	0.462	6886
No-G	0.354	0.287	0.332	6884
Dense LS	0.398	0.530	0.482	9562
Dense C2F	0.346	0.396	0.362	5238
本文模型	**0.322**	**0.265**	**0.294**	**2102**

从表 5 可以观察到，本文二分搜索策略在深度性能和内存占用上优于 Dense LS 和 Dense C2F，这充分验证了所提出二分搜索策略的有效性。在深度性能上，Dense LS 和 Dense C2F 均采用回归方式处理深度值，对噪声和异常值敏感，而二分搜索策略采用离散分类的方式，主要关注样本所属的类别，不受具体的深度值影响，因此分类更为高效；在内存占用方面，Dense LS 和 Dense C2F 需要存储整个密集深度图，而二分搜索策略只需要存储离散的散度深度值信息，从而降低了内存占用。

模型 No-G 重建点云相较基准模型的准确度、完整度和综合度指标分别提升了 0.042，0.099 和 0.13。同时，使用二分搜索策略和多粒度特征信息聚合策略时，重建点云相较基准模型的准确度、完整度和综合度指标分别提升了 0.054，0.121 和 0.159。

二分搜索策略首先基于当前深度区间的中心点进行初始深度估计，通过计算该点的匹配代价评估其可靠性。若匹配代价较低，则保留该中心点作为候选深度，并以指数衰减方式逐步收缩搜索区间，以提高定位精度；若匹配代价较高，则根据误差方向动态调整搜索区间边界，向更可能的深度范围偏移。这种动态区间调整机制能够自适应不同场景的深度分布特性，在纹理丰富区域快速收敛，在弱纹理区域保持稳定搜索，从而在保证精度的同时显著提升计算效率。

4.4.2 多粒度特征信息有效性评估

为了验证本文多粒度特征信息聚合策略的有效性，本小节在构建代价体过程中对全局语义信息（SF）和局部纹理信息（TF）进行消融实验，以评估不同粒度特征信息在深度估计中的贡献。以本文模型为基线，去除多粒度特征信息聚合策略的模型记为 No-M，实验结果如表 6 所列。

表 6 关于构建代价体过程中 SF 和 TF 的消融实验结果对比

Table 6 SF and TF ablation results in cost volume construction

多粒度特征信息聚合策略	*Acc.*	*Comp.*	*Overall*
No-All	0.396	0.386	0.462
No-M	0.367	0.332	0.365
SF	0.322	0.272	0.314
TF	0.338	0.268	0.304
SF+TF	**0.322**	**0.265**	**0.294**

由表 6 可以观察到，当模型通过多粒度特征信息聚合策略融合 SF 和 TF 时，性能表现优于单独使用 SF 和 TF 时的性能。原因在于，多粒度特征信息聚合策略将 SF 和 TF 进行整合，在不同阶段的代价体之间建立了链接，促进了关键信息的恢复和补充，从而提升了深度估计的准确性和鲁棒性。

模型 No-M 重建点云相较基准模型 No-All 的准确度、完整度和综合度指标分别提升 0.029，0.054 和 0.097。

多粒度特征信息聚合策略能够融合全局语义信息和局部纹理信息，并对其进行进一步加工和优化，以提高最终重建点云的质量。综上所述，本文二分搜索策略和多粒度特征信息聚合策略均有助于重建点云变得更加准确、完整，同时使用二者，能进一步提升重建点云的增益。

4.4.3 聚合策略实验分析

多粒度特征信息聚合模块聚合时，采用视图内自适应聚合策略（WAA）动态调整上下文信息权重，并通过逐视图自适应加权策略（PAA），使用不同权重聚合不同视图特征。为了验证这两种策略的可行性，在数据集 DTU 上分别对相关模型进行实验对比。实验结果如表 7 所列，其中 baseline 为不使用多粒度信息聚合模块。

表 7 实验结果对比

Table 7 Experimental comparison results

模型/评价指标	*Acc.*	*Comp.*	*Overall*
baseline	0.342	0.265	3.03
baseline-PAA	0.337	0.271	0.300
baseline-WAA	0.338	0.269	0.299
baseline-PAA-WAA	**0.322**	**0.265**	**0.294**

结束语 本文提出了一种基于级联的二分搜索和多粒度特征信息聚合策略的多视图立体网络，该网络设计一种多粒度特征信息聚合策略，用于构建具有语义感知和几何结构细节的代价体；采用一种二分搜索策略，将深度范围划分为多个预选区域，并通过二进制离散分类来压缩深度值的搜索区域，提高检索真实深度值的效率，减小了代价体的内存占用；采用梯度掩码优化方法，计算有效像素的地图损失，并更新网络参数，避免无用梯度累积。在数据集 DTU 和 Tanks and Temples 上的实验结果，验证了本文模型在 3D 重建方面表现出的卓越性能。

参 考 文 献

[1] 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. Springer, 2018: 767-783.

[2] 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. Piscataway, NJ: IEEE, 2019: 5525-5534.

[3] 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. Piscataway, NJ: IEEE, 2020: 2495-2504.

[4] MI Z, DI C, XU D. Generalized binary search network for highly-efficient multi-view stereo[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12991-13000.

[5] 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. Piscataway, NJ: IEEE, 2017: 2307-2315.

[6] YU Z, GAO S. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation andGauss-Newton refinement[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1949-1958.

[7] DING Y, YUAN W, ZHU Q, et al. TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8575-8584.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010.

[9] YANG J, MAO 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. Piscataway, NJ: IEEE, 2020: 4876-4885.

[10] 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(CVPR). Piscataway, NJ: IEEE, 2020: 2521-2531.

[11] WANG F, GALLIANI S, VOGEL C, et al. Patchmatchnet: Learned multi-view patchmatch stereo[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14194-14203.

[12] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2117-2125.

[13] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 764-773.

[14] 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.

[15] 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.

[16] 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. Piscataway, NJ: IEEE, 2021: 6187-6196.

[17] PENG R, WANG R, WANG Z, et al. Rethinking depth estimation for multi-view stereo: A unified representation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8645-8654.

[18] WANG S, JIANG H, XIANG L, et al. CT-MVSNet: Efficient multi-view stereo with cross-scale transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8645-8654.

[19] YE X, ZHAO W, LIU T, et al. Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 17661-17670.

[20] 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. Piscataway, NJ: IEEE, 2021: 5732-5740.

[21] LAI H W, YE C L, LI Z, et al. MFE-MVSNet: Multi-scale feature enhancement multi-view stereo with bi-directional connections[J]. IET Image Processing, 2024, 18(3): 1234-1245.

XU Lijun, born in 1991, Ph.D, associate professor, is a member of CCF (No. 62672M). Her main research interests include computer vision, artificial intelligence and digital twins.

CHEN Kansong, born in 1972, Ph. D, postdoctoral researcher. His main research interests include artificial intelligence, digital twin, industrial Internet and related fields.

（责任编辑：柯颖）