计算机科学
COMPUTER SCIENCE

# 融合上下文引导代价体和深度细化的多视图立体重建

陈光远，王朝辉，程泽

引用本文

陈光远，王朝辉，程泽. 融合上下文引导代价体和深度细化的多视图立体重建[J]. 计算机科学，2025，52(3)：231-238.

CHEN Guangyuan, WANG Zhaohui, CHENG Ze. Multi-view Stereo Reconstruction with Context-guided Cost Volume and Depth Refinemen [J]. Computer Science, 2025, 52(3): 231-238.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200111

# 融合上下文引导代价体和深度细化的多视图立体重建

陈光远　王朝辉　程　泽
武汉科技大学计算机科学与技术学院　武汉 430081
(Cgy2001@wust.edu.cn)

**摘　要**　针对基于深度学习的多视图立体(Multi-view Stereo,MVS)重建算法仍然存在图像特征提取不全面、代价体匹配模糊以及深度误差不断积累而导致在无纹理和重复纹理区域重建效果差的问题,提出了基于上下文引导的代价体构建和深度细化的级联MVS网络。首先,利用基于无参注意力的特征融合模块过滤无用特征并通过特征融合来解决多尺度特征不一致的问题;然后,利用基于上下文引导的代价体模块融合全局信息来提高代价体匹配的完整性和鲁棒性;最后,利用深度细化模块学习深度残差来提升低分辨下深度图的准确性。实验结果表明,在DTU数据集上,该网络相比MVSNet完整度误差减小了24.4%,准确度误差减小了4.1%,整体误差减小了14.3%,其在Tanks and Temples数据集上性能也优于大多数算法,展现出强大的竞争力。

**关键词**:多视图立体;特征融合;上下文引导;代价体匹配;深度细化
**中图分类号**　TP391

## Multi-view Stereo Reconstruction with Context-guided Cost Volume and Depth Refinemen

CHEN Guangyuan,WANG Zhaohui and CHENG Ze
School of Computer Science and Technology,Wuhan University of Science and Technology,Wuhan 430081,China

**Abstract**　In response to the challenges in deep learning-based multi-view stereo(MVS) reconstruction algorithms,which include incomplete image feature extraction,ambiguous cost volume matching,and the accumulation of depth errors leading to poor reconstruction results in textureless and repetitive texture regions,a cascaded MVS network based on context-guided cost volume construction and depth refinement is proposed. First,the feature fusion module based on non-reference attention is used to filter out irrelevant features and address the inconsistency in multi-scale features through feature fusion. Then,the context-guided cost volume module is used to fuse global information to enhance the accuracy and robustness of cost volume matching. Finally,the depth refinement module is employed to learn and reduce depth errors,to improve the accuracy of the low-resolution depth maps. The experimental results show that compared with MVSNet,the integrity error of the network on the DTU dataset is reduced by 24.4%,the accuracy error is reduced by 4.1 %,and the overall error is reduced by 14.3 %. The performance on the Tanks and Temples dataset is also better than most algorithms,showing strong competitiveness.

**Keywords**　Multi-view stereo,Feature fusion,Context-guide,Cost volume matching,Depth refinement

## 1　引言

多视图立体重建[1](MVS)是计算机视觉领域的一个重要研究方向,其目标是在已知相机参数的多个视角图像中重建三维场景的几何结构。传统的MVS方法[2]通常基于手工设计的特征提取和匹配算法,依赖于领域专家的知识和强先验信息,仅在某些特定场景中表现良好,在纹理结构复杂的大型场景中会面临巨大挑战。

近年来,随着深度学习的不断发展,卷积神经网络[3]在计算机视觉领域的各个方向得到广泛应用并取得了良好的效果,MVSNet[4]引入卷积神经网络并首次提出端到端的架构来预测深度图,基于平面扫描的思想,利用可微单应性变换来构建三维代价体,并进行代价体正则化以及深度图回归。但代价体正则化过程中内存需求随着模型分辨率的增加呈立方增长。为了减少内存消耗,近期的工作提出了由粗到细的级联结构[5-7],以粗到细的方式计算不同分辨率的深度图,逐渐缩小假设平面引导以降低计算复杂性。然而这些方法仍然存在缺陷。现有的MVS方法都是通过特征金字塔网络[8]对图像进行多尺度特征提取,并没有考虑到多尺度特征之间的不一致性会降低金字塔的有效性,从而导致代价体中的匹配信息失真,对最终重建造成影响。此外,由于卷积神经网络的局部感受野限制了图像特征的全局编码能力,使得在处理遮挡

到稿日期:2023-12-18　返修日期:2024-05-29
基金项目:国家自然科学基金(62302351)
This work was supported by the National Natural Science Foundation of China(62302351).
通信作者:王朝辉(zhwang_pdoc@163.com)

或非常局部的变化时无法很好地捕捉物体之间的关系，在处理大面积场景或者纹理不清晰区域时无法聚合足够多的全局信息，导致出现代价体模糊匹配、匹配失误的情况。为了解决以上问题，本文提出了基于无参注意力的特征融合模块和上下文引导的代价体模块。基于无参注意力的特征融合模块通过无参注意力过滤无用特征，自适应学习每个尺度特征图的融合空间权重，对不同尺度特征进行融合，使网络能够直接学习如何在其他级别对特征进行空间滤波，从而仅保留有用的信息以进行组合。上下文引导的代价体模块显式地构建代价体中像素之间的长距离依赖关系，并将其作为平滑约束来指导更新代价体，以此提高代价体中像素匹配的一致性和完整性，减小由像素级特征匹配困难造成的误差。

此外，粗糙阶段预测的深度图的不准确性会导致深度图的误差不断累积，进而影响最终阶段的深度图。针对这一问题，利用深度细化模块汇聚注意参考特征来预测残差深度值，从而提高深度图的准确性。然后设计了一个轻量化的反投影[9]上采样网络来避免深度图在上采样过程中潜在的深度信息缺失和估计误差的问题，通过减小粗糙阶段深度图的误差来提升最终深度图的准确性。

本文的主要贡献包括：

1）提出基于无参注意力的特征融合模块，融合不同尺度特征来增强特征有效性。

2）提出基于上下文引导的代价体模块，构建代价体中像素之间的长距离依赖关系，将其作为平滑约束来指导更新代价体，改善代价体模糊匹配问题。

3）在粗糙阶段提出深度细化模块，聚合注意参考特征来提高粗糙阶段深度估计精度，并设计轻量级的反投影上采样网络来减少上采样过程的深度信息缺失。

## 2 相关工作

近年来，关于多视图立体重建的研究越来越多，传统的MVS算法可以分为4种类别：基于体素的方法[10]、基于表面估计的方法[11]、基于块的方法[12-13]和基于深度图的方法[14-16]。基于体素的方法将整个三维空间离散成具有规则形状的体素，然后通过光度一致性度量来判断每个体素是否属于表面，但巨大的存储消耗导致这种表示方法不能扩展到大规模场景。基于表面的方法通过直接重建表面网格，使结果变得更加平滑但也损失了详细信息。基于块的方法将表面视为一组块，并首先匹配包含更容易区分特征的块，然后再传播到无纹理区域。基于深度图的方法是最灵活的一种，该方法将复杂的三维几何重建问题归结为在二维领域中的深度图估计。此外，作为中间表示，所有单个图像的估计深度图可以合并成一致的点云或者体积重建，甚至可以进一步重建成网格。尽管传统的MVS方法取得了令人印象深刻的结果，但其使用的手工制作特征并不适用于非朗伯表面、纹理较弱和无纹理区域，这些区域的光度一致性并不可靠。

Surface-Net[17]首次提出基于学习的体素表示MVS流程，用于从三维空间回归出表面体素，其缺点是受到了体积表示常见的内存限制。MV-SNet是首个实现基于3D代价体的端到端流程的网络，在此之后，基于学习的MVS方法得到了快速发展。为了进一步挖掘该网络的潜力，研究者们提出了MVSNet的一系列变体。例如，为了减少内存消耗，处理大规模场景，R-MVSNet[18]使用GRU代替3D CNN进行代价体正则化，Fast-MVSNet[19]提出了用于深度估计的稀疏到密集框架来缩短运行时间。CVP-MVSNet[6]采用了一种新颖的由粗到细的策略来减少内存的消耗，CasMVSNet[5]和UC-SNet[7]提出了基于特征金字塔的级联架构，并以粗到细的方式来估计深度图。具体而言，首先建立具有大深度范围的代价体，以估计最粗糙分辨率的深度图，然后根据前一阶段的估计逐渐缩小代价体的深度范围，最终产生高精度的高分辨率深度图。以粗到细的方式预测多视图深度图是有效降低计算成本的很有前景的方法，此后的许多网络也是在其基础上的改进。Patchmatch-Net[20]将传统的Patchmatch立体匹配融入网络之中，学习了深度假设的自适应传播和评估并且通过轻量级网络替换3D CNN来进行代价体正则化，从而实现了高效的模型；UniMVSNet[21]设计了一种新的损失函数来统一分类和回归；GBI-Net[22]设计了二分搜索网络来提高模型整体效率；MVSF-ormer[23]提出了一个预训练的视觉变换器来增强网络；TransMVSNet[24]使用Transformer来收集全局上下文感知信息；EPP-MVSNet[25]使用极线变换器来学习语义特征，利用熵来选取采样区间从而缩小范围，并用伪3D卷积替代普通的3D卷积运算符来减少冗余计算和降低高成本。

在光流估计方向中，处理在参考帧中成像但在匹配帧中不可见的3D遮挡点一直都是该领域的一个重要难题。KPAFlow[26]提出了基于核块注意力的光流估计方法，对特征图的每个局部块进行操作，通过显式地利用局部场景内容和空间的亲和关系来引导挖掘局部信息，减小由像素特征匹配困难所造成的误差。受到该研究的启发，本文设计了一个上下文编码器来提取上下文参考特征信息，并使用基于核块注意力的分支来挖掘其局部周围环境的特征联系，利用挖掘的特征联系来指导并更新代价体，通过引导代价体聚合全局信息来增强代价体匹配的完整性和鲁棒性。

## 3 融合上下文引导代价体和深度细化的联级网络

本文提出了融合上下文引导代价体和深度细化的级联网络。在本章中，首先介绍该网络的总体架构，然后进一步介绍基于无参注意力的特征融合模块、基于全局上下文引导的代价体模块及深度细化模块的相关细节，最后介绍文中所使用的损失函数。

### 3.1 网络结构

本文采用了由粗到细的策略通过三阶段逐级预测深度图，具体网络结构如图1所示。网络整体流程包括特征提取、代价体构建、上下文引导代价体、代价体正则化与深度回归和深度图细化这5个步骤。具体来说，假设有 $N$ 张图像输入，$\boldsymbol{I}_1$ 和 $\{\boldsymbol{I}_i\}_{i=2}^{N}$ 分别表示参考图像和源图像组。首先，在特征提取阶段利用FPN特征金字塔网络提取多尺度特征，并通过基于无参注意力的特征融合模块输出最终经过融合的多尺度特征 $\{\boldsymbol{F}_i\}_{i=1}^{N}$，其中 $\boldsymbol{F}_i$ 代表每个图像的多尺度特征。通过可微的单应性变化将多尺度特征投影到参考图像的视角上形成 $N$

个特征体$\{\boldsymbol{S}_i\}_{i=1}^{N}$，并以基于方差的方式将 $N$ 个特征体聚合成单个代价体$\boldsymbol{V}\in R^{C\times M\times H\times W}$，其中 $C$ 代表通道数，$M$ 代表当前深度假设数量。在第一阶段，使用基于上下文引导的代价体模块对代价体 $\boldsymbol{V}$ 进行上下文引导使其匹配更多丰富的全局信息，再通过 3D CNN 进行代价体正则化得到概率体 $\boldsymbol{P}$，对其沿着深度维度利用 soft-argmin 操作来回归粗糙深度图，最后通过深度细化模块来提高粗糙深度图的准确性以及减少上采样过程中的深度损失。在第二阶段，将细化后的粗糙深度图作为输入来构建代价体，并在代价体正则化后通过深度回归得到更高分辨率的深度图。在第三阶段，将上一阶段的深度图上采样后作为输入进行代价体构建和正则化，最后通过深度回归得到高分辨率的精确深度图。

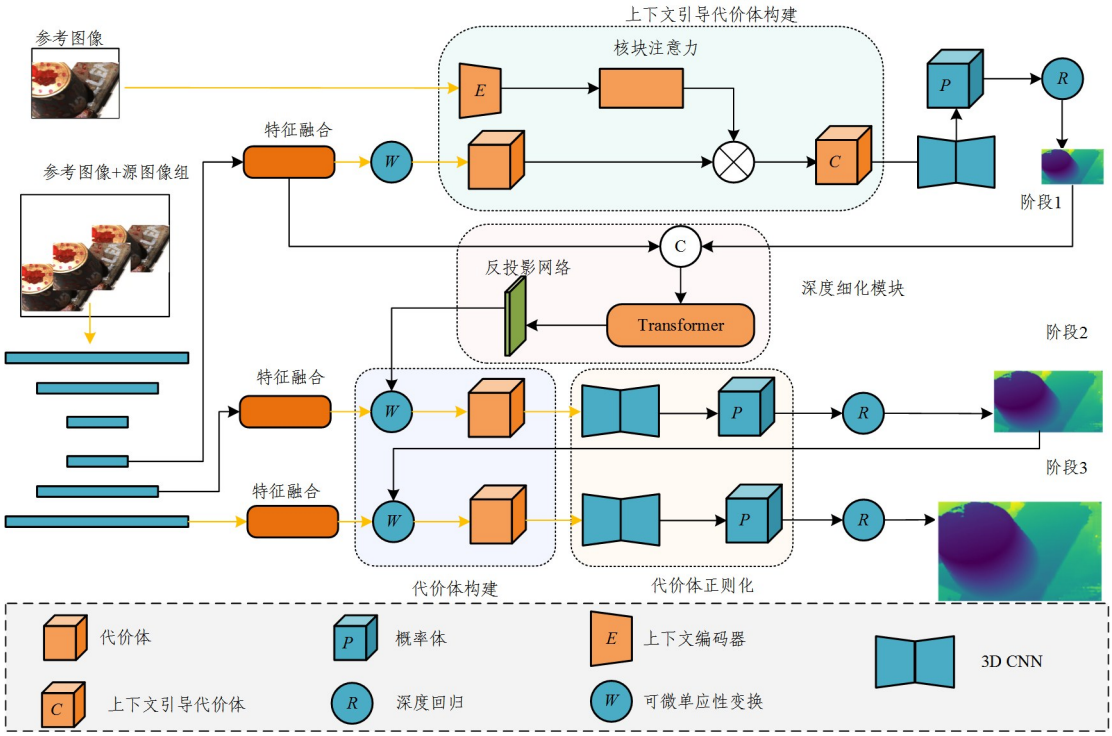


图 1　网络整体结构

Fig. 1　Overall architecture of the proposed network

### 3.2　基于无参注意力的特征融合模块

FPN 特征金字塔网络提取的输入图像组的多尺度特征具有不同尺度特征之间不一致的缺点。具体来说，一旦图像中既含有大目标又含有小目标时，不同层级之间的特征存在冲突(大目标通常与较高特征图相关联，而小目标通常与较低特征图相关联)。这种冲突会对网络训练时的梯度计算产生干扰并降低特征金字塔的有效性，进而可能导致代价体中的匹配信息失真，对最终的重建效果造成影响。针对这一问题，本文提出了基于无参注意力的特征融合模块来指导网络过滤不相关的特征，以增强特征的有效性，并且通过自适应学习每个尺度特征图的融合空间权重对不同尺度特征进行融合。该模块的详细结构如图 2 所示。

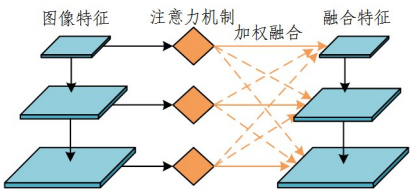


图 2　基于无参注意力的特征融合模块

Fig. 2　Feature fusion modules based on non-reference attention

首先，引入一种简单而有效的 SimAM 注意力机制[27]，在不增加网络参数的情况下为特征图推断 3D 注意力权重。多尺度特征$\{\boldsymbol{F}_i\}_{i=1}^{N}$通过 SimAM 注意力机制过滤一些不相关特征，以增强特征的有效性。对于某一层特征 $\boldsymbol{F}_i^l$，其中 $l\in\{1,2,3\}$，将其他层的特征经过上采样、下采样以及插值等操作调整到该层相对应的分辨率，得到相同分辨率的特征图，然后在每个空间位置上将其自适应地融合在一起。假如某位置携带矛盾信息，这些特征就会被过滤；若某位置的特征带有更多的区分性线索，则这些特征将会被增强。更具体来说，$x_{i,j}^{n\to l}$代表从第 $n$ 层调整到第 $l$ 层的特征图中位置$(i,j)$处的特征向量。特征融合策略如式(1)所示：

$$y_{i,j}^{l}=\boldsymbol{\alpha}_{i,j}^{l}x_{i,j}^{1\to l}+\boldsymbol{\beta}_{i,j}^{l}x_{i,j}^{2\to l}+\boldsymbol{\gamma}_{i,j}^{l}x_{i,j}^{3\to l} \tag{1}$$

其中，$y_{i,j}^{l}$代表第 $l$ 层的输出特征图 $y^l$ 在位置$(i,j)$处的特征向量，它是融合了前 3 个特征图$(i,j)$处的特征向量的加权融合；$\boldsymbol{\alpha}_{i,j}^{l}$，$\boldsymbol{\beta}_{i,j}^{l}$和 $\boldsymbol{\gamma}_{i,j}^{l}$是特征图在 3 个不同层到 $l$ 层中由网络自适应学习得到的空间重要性权重，它们在所有通道间是共享的。限制$\boldsymbol{\alpha}_{i,j}^{l}+\boldsymbol{\beta}_{i,j}^{l}+\boldsymbol{\gamma}_{i,j}^{l}=1$ 且$\boldsymbol{\alpha}_{i,j}^{l},\boldsymbol{\beta}_{i,j}^{l},\boldsymbol{\gamma}_{i,j}^{l}\in[0,1]$，其中$\boldsymbol{\alpha}_{i,j}^{l}$的表达式如式(2)所示：

$$\boldsymbol{\alpha}_{i,j}^{l}=\frac{e^{\lambda_{\alpha_{i,j}}^{l}}}{e^{\lambda_{\alpha_{i,j}}^{l}}+e^{\lambda_{\beta_{i,j}}^{l}}+e^{\lambda_{\gamma_{i,j}}^{l}}} \tag{2}$$

其中，$\boldsymbol{\alpha}_{i,j}^{l}$，$\boldsymbol{\beta}_{i,j}^{l}$，$\boldsymbol{\gamma}_{i,j}^{l}$通过分别将$\lambda_{\alpha_{i,j}}^{l}$，$\lambda_{\beta_{i,j}}^{l}$，$\lambda_{\gamma_{i,j}}^{l}$作为控制参数，再使用 softmax 函数定义；权重标量映射$\lambda_{\alpha_{i,j}}^{l}$，$\lambda_{\beta_{i,j}}^{l}$，$\lambda_{\gamma_{i,j}}^{l}$通过 $x^{1\to l}$，$x^{2\to l}$，$x^{3\to l}$经过 1×1 卷积层后得到。

### 3.3 代价体构建

获得输入图像组的多尺度融合特征后，通过可微单应性变换将所有特征图变换到参考相机的截锥体的法向平面上，构建代价体。具体来说，给定参考图像 $\boldsymbol{I}_1$ 和源图像组 $\{\boldsymbol{I}_i\}_{i=2}^{N}$，以参考图像 $\boldsymbol{I}_1$ 为基准，利用单应性变换将第 $i$ 个视角图像的多尺度特征图 $\boldsymbol{F}_i$ 通过映射投影到与参考图像 $\boldsymbol{I}_1$ 对应的多尺度特征的平行平面上，以形成特征体 $\boldsymbol{S}_i$。单应性变换过程如式(3)所示：

$$\boldsymbol{H}_i(d)=\boldsymbol{K}_i\cdot\boldsymbol{R}_i\cdot\left(\boldsymbol{I}-\frac{(\boldsymbol{t}_1-\boldsymbol{t}_i)\cdot\boldsymbol{n}_1^{\mathrm{T}}}{d}\right)\cdot\boldsymbol{R}_1^{\mathrm{T}}\cdot\boldsymbol{K}_1^{\mathrm{T}} \tag{3}$$

其中，$\boldsymbol{H}_i(d)$ 表示多尺度特征图 $\boldsymbol{F}_i$ 与参考特征图在深度 $d$ 处的单应性；参数 $\{\boldsymbol{K}_i,\boldsymbol{R}_i,\boldsymbol{t}_i\}$ 分别是源图像的相机内参矩阵、相机旋转矩阵和位移向量；$\boldsymbol{I}$ 是单位矩阵；$\boldsymbol{n}_1^{\mathrm{T}}$ 是参考图像平面法向量的转置。

为了适应任意数量的输入视图，采用基于方差的方法来衡量 $N$ 视图之间特征体的相似性，将多个特征体 $\{\boldsymbol{S}_i\}_{i=1}^{N}$ 聚合为一个代价体，即：

$$\boldsymbol{V}=\frac{\sum_{i=1}^{N}(\boldsymbol{S}_i-\bar{\boldsymbol{S}}_i)^2}{N} \tag{4}$$

其中，$\boldsymbol{V}$ 是代价体，$\bar{\boldsymbol{S}}_i$ 是所有特征体的平均值，$N$ 是输入视图的数量。

### 3.4 基于上下文引导的代价体模块

由于卷积神经网络有限的局部感受野限制了图像的全局编码能力，因此在处理大型低纹理或反射表面等困难场景时会导致代价体模糊匹配。针对这一缺陷，设计了基于上下文引导的代价体模块，利用上下文参考特征构建代价体中像素之间的长距离依赖关系，以提高代价体中像素匹配的一致性和完整性。该模块的详细结构如图 3 所示。

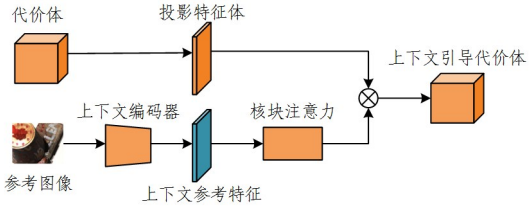


图 3 基于上下文引导的代价体模块

Fig. 3 Context-guided cost volume modules

首先参考图像通过上下文编码器提取上下文参考特征。为了建立上下文参考特征与代价体的联系，将构建的代价体 $\boldsymbol{V}\in R^{C\times M\times H\times W}$ 通过投影和形状变换来融合深度维度与通道维度，得到投影特征体 $\boldsymbol{F}_r\in R^{C\times MH\times W}$。随后将参考特征划分为不重叠的块窗口，通过核块注意力在其各个块窗口上挖掘与局部周围块窗口的特征联系，利用挖掘的特征联系来指导并更新投影特征体对应块窗口的信息。核块注意力的具体实现过程如图 4 所示，具体来说，对上下文参考特征的核窗口中心块区域 $e$ 的特征 $\bar{\boldsymbol{F}}_c$ 和核窗口区域的特征 $\boldsymbol{F}_c$ 使用权重函数 $\boldsymbol{W}_{(t_i)}(\cdot)$ 来生成自适应权重，并通过尺度函数 $\boldsymbol{S}(t_i,\hat{x}_n)$ 获得尺度因子掩码，对 $t_i$ 和 $\hat{x}_n$ 之间的距离进行空间约束。权重函数 $\boldsymbol{W}_{(t_i)}(\cdot)$、尺度函数 $\boldsymbol{S}(t_i,\hat{x}_n)$ 分别如式(5)和式(6)所示：

$$\boldsymbol{W}_{(t_i)}(\bar{\boldsymbol{F}}_c,\boldsymbol{F}_c)=\frac{e^{\theta(\bar{\boldsymbol{F}}_c^{u})^{\mathrm{T}}\cdot\lambda(\boldsymbol{F}_c^{v})}}{\sum_{\forall v}e^{\theta(\bar{\boldsymbol{F}}_c^{u})^{\mathrm{T}}\cdot\lambda(\boldsymbol{F}_c^{v})}} \tag{5}$$

$$\boldsymbol{S}(t_i,\hat{x}_n)=\max\left(0,l+abs\left(b\left(\|\boldsymbol{t}_i-\hat{\boldsymbol{x}}_n\|-\frac{k}{2}\right)\right)\right) \tag{6}$$

其中，$t_i=x_i-x$，$x_i$ 代表核窗口内的位置 $i$；$\theta(\boldsymbol{F})=\boldsymbol{W}_\theta\boldsymbol{F}$ 和 $\lambda(\boldsymbol{F})=\boldsymbol{W}_\lambda\boldsymbol{F}$ 是执行特征嵌入的两个线性投影；$\hat{x}_n$ 代表核窗口中位置 $n$ 的坐标；$l$ 是一个基本标量；$b$ 是一个可以学习的参数，用于表示点间距离的影响。

利用核函数构建核窗口中心块区域的特征 $\bar{\boldsymbol{F}}_c$ 和核窗口区域的特征 $\boldsymbol{F}_c$ 的特征联系。投影特征体 $\boldsymbol{F}_r$ 对应核窗口区域 $\boldsymbol{F}_m$，将该特征联系作为平滑约束，以此来引导学习残差特征 $\boldsymbol{F}'^{e}_m$，并利用得到的残差特征对 $\boldsymbol{F}_r$ 对应核窗口中心区域的原始特征 $\boldsymbol{F}^{e}_m$ 进行更新，如式(7)－式(9)所示：

$$\boldsymbol{K}_{(t_i)}(\bar{\boldsymbol{F}}_c,\boldsymbol{F}_c)=\boldsymbol{S}(t_i,\hat{x}_n)\boldsymbol{W}_{(t_i)}(\bar{\boldsymbol{F}}_c,\boldsymbol{F}_c) \tag{7}$$

$$\boldsymbol{F}'^{e}_m=\sum_{t_i\in E}K_{(t_i)}\,p(\boldsymbol{F}_m) \tag{8}$$

$$\boldsymbol{F}^{\wedge e}_m=\boldsymbol{F}^{e}_m+\alpha\boldsymbol{F}'^{e}_m \tag{9}$$

其中，$E$ 代表核窗口中所有的块；$e$ 代表核窗口的中心区域块；$K_{(t_i)}$ 代表核函数，用于构建参考特征中块窗口之间的特征联系；$p(\cdot)$ 是将输入的特征 $\boldsymbol{F}_m$ 映射到嵌入空间的线性投影；$\alpha$ 表示一个可学习参数，其初始值为 0，并逐渐执行加权求和。为了保证投影特征体的每个块区域能够被引导更新，使用滑动窗口机制在空间域上移动核窗口，不断更新核窗口中心区域块的特征 $\boldsymbol{F}^{\wedge e}_m$ 直至更新完投影特征体 $\boldsymbol{F}_r$ 的整个空间域 $(H\times W)$，从而实现上下文引导代价体的构建。

此外，为了更好地丰富代价体的全面信息以及减轻代价体正则化过程中的噪声影响，本文选择将上下文参考特征和更新前的代价体以及上下文引导的代价体在通道维度上进行拼接，进而得到最终代价体，并将其送入 3D 正则化网络进行代价体正则化。这使得网络在特定像素位置根据需要动态选择或组合局部和全局特征，增强网络的灵活性以便可以更好地适应不同像素位置的特定要求。

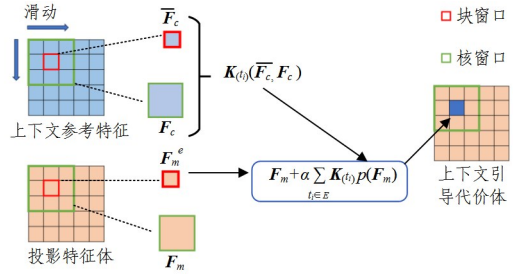


图 4 核块注意力的实现过程

Fig. 4 Implementation process of the kernel patch attention

### 3.5 代价体正则化与深度回归

由于非朗伯面、遮挡等，代价体往往包含噪声。为了减弱噪声对代价体中匹配信息的影响，使用 3D U-Net 网络对代价体进行正则化，对代价体进行编码和解码并压缩通道数得到单通道的概率体 $\boldsymbol{P}$。对其沿着深度维度使用 soft-argmin 操作获得每个像素点 $(x,y)$ 在深度维度中的概率估计，将所有深度假设值与对应的概率估计加权回归得到深度图 $\boldsymbol{D}$。

具体操作如式(10)所示：

$$\boldsymbol{D}=\sum_{j=1}^{M} d_j \cdot \boldsymbol{P}(d) \tag{10}$$

其中，$M$ 代表深度假设数量；$d_j$ 代表当前深度假设值；$\boldsymbol{P}(d)$ 代表每个像素点 $(x,y)$ 在深度 $d$ 下的概率估计值。

### 3.6 深度细化模块

粗层特征不能很好地捕捉相邻视图中的微小平移和变形，使得预测的深度图并不准确，这会导致在后续细化阶段的误差不断累积，进而影响最终深度图的准确性。本文提出深度细化模块，利用参考特征与初始深度图融合并通过轻量级 Transformer 得到注意力参考特征，从而在空间域中聚合一个大的感受野来接收全局信息。最后经过卷积学习残差深度值来更新深度图。此外，设计了一个轻量的参数化反投影网络代替上采样操作，减少了上采样过程的深度信息缺失和估计误差。该模块的网络结构如图 5 所示。

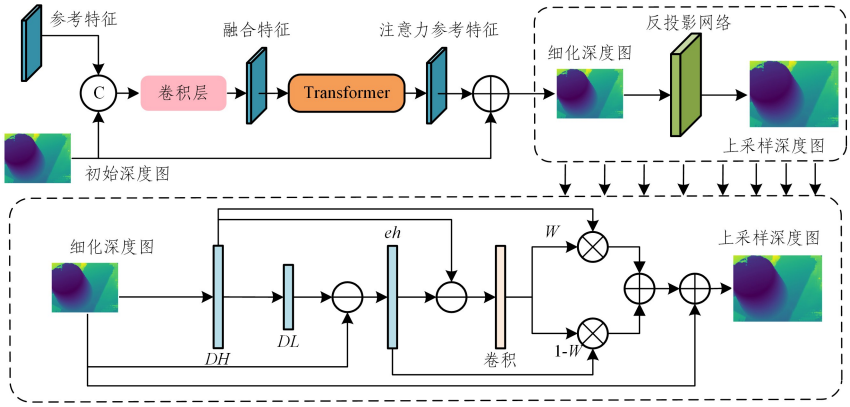


图 5 深度细化模块的网络结构

Fig. 5 Network architecture of depth refinement module

具体来说，首先将参考特征 $\boldsymbol{F}$ 与初始深度图 $\boldsymbol{D}$ 在通道维度进行拼接，通过卷积层减少通道数，得到包含高分辨率和局部像素信息的融合特征。如图 6 所示，融合特征通过卷积层来减少通道数并张量扁平化作为 Transformer 编码器的输入，经过轻量的 Transformer 编码器来输出包含更多有效全局信息的嵌入特征。然后，在通道维度上对融合特征和嵌入特征进行点积运算来计算注意力参考特征。最后，使用卷积操作来学习残差深度值并与初始深度图相加得到细化深度图 $\boldsymbol{D}^r$。

轻量的参数化反投影网络利用上采样、下采样和卷积等操作生成权重掩码，学习深度图中不同区域的权重来减小上采样过程中的深度误差。具体操作如式(11)和式(12)所示：

$$\boldsymbol{e}^{\wedge}\boldsymbol{h}=(1-\boldsymbol{W})\cdot\boldsymbol{eh}+\boldsymbol{W}\cdot\boldsymbol{DH} \tag{11}$$

$$\boldsymbol{D}^{\wedge}=\boldsymbol{D}^r+\boldsymbol{e}^{\wedge}\boldsymbol{h} \tag{12}$$

其中，$\boldsymbol{DH}$ 是经过上采样的高分辨率深度图，其下采样后与细化深度图 $\boldsymbol{D}^r$ 相减得到低分辨率误差，再将其上采样后得到高分辨率误差 $\boldsymbol{eh}$。$\boldsymbol{W}$ 是通过($\boldsymbol{DH}$-$\boldsymbol{eh}$)后经过卷积与 softmax 操作得到的权重掩码。通过捕捉上下采样过程中的深度信息缺失来优化深度图，为下一阶段提供更精确的深度假设。注意，上采样和下采样操作都是简单的卷积操作。

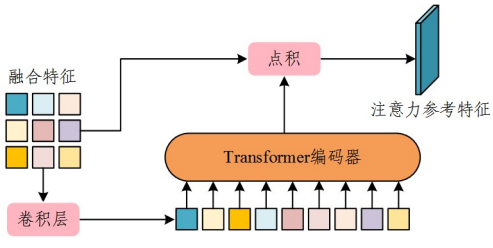


图 6 注意力参考特征的实现过程

Fig. 6 Implementation process of attentiol reference features

### 3.7 损失函数

本文采用监督学习策略，利用损失函数来限制由深度回归预测的深度图与真实深度之间的误差。与现有的 MVSNet 框架类似，使用 L1 范数度量所有阶段的误差，通过不同的权重将所有阶段的误差融合，总损失定义为：

$$Loss=\sum_{l=0}^{L}\sum_{p\in P_{\text{valid}}}\|d(p)-d^p(p)\|_1+\alpha\sum_{p\in P_{\text{valid}}}\|d(p)-d^{\text{refine}}(p)\|_1 \tag{13}$$

其中，$P_{\text{valid}}$ 代表在真实深度图中的有效像素集；$d(p)$ 代表真实深度图像素 $p$ 的深度值；$d^p(p)$ 表示网络预测深度图像素 $p$ 的深度值；$d^{\text{refine}}(p)$ 表示在粗糙阶段经过深度细化后的深度图像素 $p$ 的深度值；$l$ 代表金字塔的第 $l$ 阶段；$\alpha$ 代表更新权重。在训练过程中，将 $L$ 设置为 2。

## 4 实验

在本章中首先介绍数据集、评估指标和实验细节。然后展示本文方法与传统 MVS 方法和基于学习的先进的 MVS 方法的比较结果，以及通过定性的可视化分析来证明所提方法的有效性。最后通过多组消融实验，验证了提出的每个模块的有效性。

### 4.1 数据集

在 DTU 数据集[28]和 Tanks and Temples(T&T)数据集[29]上进行了实验。在 DTU 数据集上将本文方法与基线方法以及最先进的方法进行性能评估和比较，Tanks and Temples 数据集用于验证模型的泛化能力。DTU 数据集是一个大型室内数据集，由 124 个不同场景组成，每个场景从 49 个不同视角以及 7 个不同亮度级别进行拍摄。该数据集提供了由高精度结构光扫描仪获取的参考模型以及高分辨率 RGB 图像，本文使用和 MVSNet 中相同划分方式的验证集和测试集进行训练和评估。Tanks and Temples 数据集包含了更具

挑战性的现实环境，具有大规模变化和光照变化，它包含一个包括 8 个场景的中级子集以及一个包括 6 个场景的高级子集。

### 4.2 评估指标

采用距离度量的准确性(Acc)、完整性(Comp)以及整体性(Overall)来评估 DTU 数据集的定量结果。其中，准确性指重建结果与真实点云的接近程度，即每个重建点到真实点云中最近点的绝对距离的均值或中值；完整性是看真实模型中有多少点被重建结果覆盖，即从真实点云到重建点云的绝对距离的均值或中值；整体性是准确性和完整性的算数平均值。这 3 个指标越低，代表模型重建效果越好。在 Tanks and Temples 数据集上，采用 F-score，Mean 来作为衡量完整性和准确性的百分比度量指标。其中，Mean 是所有场景 F-score 的平均分，该指标越高，代表模型重建效果越好。

### 4.3 实验细节

本文所用的基准模型是 CasMVSNet，在训练和评估中遵循与其相同的输入视图选择和数据预处理策略。在 DTU 的训练集上训练网络，并在 DTU 的测试集上进行评估，将在 DTU 的训练集上训练的网络进行微调后在 Tanks and Temples 数据集上进行评估。在训练过程中，对于 DTU 数据集，输入的图像数量为 $N=5$，分辨率为 $640\times512$。使用 $L=3$ 层金字塔，在 3 个阶段中，深度范围假设的数量分别为 48，32，8，深度采样范围为 425～935 mm，更新权重 $\alpha$ 为 3.0，使用 L1 范式作为损失函数来衡量真实深度和每个阶段预测深度之间的绝对差异。对于 Tanks and Temples 数据集，将输入图像的分辨率设置为 $1\,920\times1\,024$，图像数量 $N$ 设置为 7，其余设置与 DTU 数据集设置相同。本文网络使用 Pytorch 实现，并在 NVIDIA Tesla V100 GPU 上使用 Adam 优化器进行了 16 轮模型训练，批量大小为 2，初始学习率为 0.001，在第 10，12 和 14 轮将学习率逐步减半以避免陷入局部最优解。本文网络在获得预测深度图后，在将预测深度图转换为稠密点云前还需要使用光度一致性和几何一致性来进行鲁棒的深度图过滤。最后，使用深度融合的方式将多视图的预测深度图融合为三维空间的点云。

### 4.4 DTU 数据集的评估结果

设置输入的图像数量 $N=5$，将图像裁剪为 $1\,600\times1\,152$ 并在 DTU 测试集上进行评估。将评估结果与传统方法和近期基于学习的方法进行了比较，定量比较结果如表 1 所列。

由表 1 可知，本文方法在所有方法中完整性排名第二且在整体性上优于其他方法。相比基准模型 CasMVSNet，本文方法的准确度误差增加了 3%，完整度误差和整体误差分别减小了 10.2%和 3.6%。本文方法与近期基于学习的方法的部分场景重建效果如图 7 所示，细节部分由红色矩形框突出显示。从图中可以看出本文方法能够估计出更完整的密集点云来保留原有的细节部分，并且在无纹理区域也能表现良好，这得益于代价体构建过程通过引导上下文聚合了丰富的全局信息。此外，本文方法还可以为具有更多细节的复杂场景保留更完整的点云和更清晰的纹理结构。其他不同场景的密集点云重建效果如图 8 所示。

表 1 DTU 数据集上不同方法的定量分析结果

Table 1 Quantitative analysis results of different methods on DTU dataset

| Method | Acc | Comp | Overall |
|---|---|---|---|
| Camp[30] | 0.835 | 0.554 | 0.695 |
| Gipuma[31] | **0.283** | 0.873 | 0.578 |
| Colmap[32] | 0.411 | 0.657 | 0.534 |
| MVSNet[4] | 0.396 | 0.527 | 0.462 |
| PointMVSNet[33] | 0.342 | 0.411 | 0.376 |
| CasMVSNet[5] | 0.325 | 0.385 | 0.355 |
| CVP-MVSNet[6] | 0.296 | 0.406 | 0.351 |
| EPP-MVSNet[25] | 0.413 | 0.296 | 0.355 |
| PatchmatchNet[20] | 0.427 | **0.277** | 0.352 |
| AA-RMVSNet[34] | 0.376 | 0.339 | 0.357 |
| UCSNet[7] | 0.338 | 0.349 | 0.344 |
| DDR-Net[35] | 0.339 | 0.320 | 0.329 |
| Ours | 0.355 | 0.283 | **0.319** |

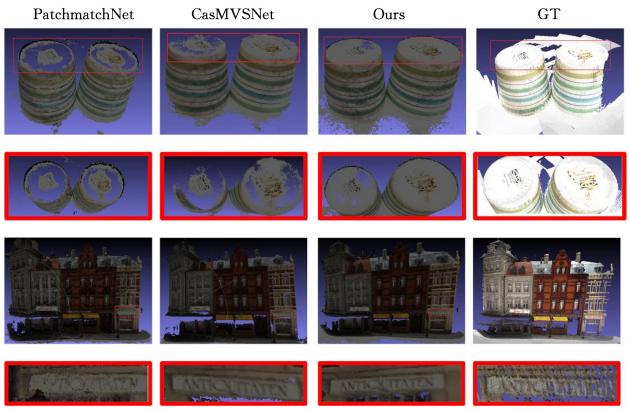


图 7 DTU 数据集上场景 48 和场景 9 的定性比较(电子版为彩图)

Fig. 7 Qualitative comparison of Scene48 and Scene9 on DTU dataset

图 8 DTU 数据集上不同场景的点云重建效果

Fig. 8 Point cloud reconstruction results of different scenes on DTU dataset

### 4.5 消融实验

为了进一步验证本文方法的有效性，在 DTU 数据集上进行了多组消融实验以定量分析并验证本文方法中各个模块的有效性。实验结果如表 2 所列，其中 FFA 代表特征融合模块，CGPA 代表基于上下文引导代价体模块，UPDR 代表深度细化模块，粗体代表最佳结果。

从表 2 可以看出，在基准模型的基础上增加 FFA 模块后，准确度误差减小了 0.2%，完整度误差减小了 1.2%，整体误差减小了 0.7%，证明该模块通过生成更加有效的多尺度特征，对后续的代价体匹配过程以及最终的重建过程产生了积极的影响。在增加 UPDR 模块后与基准模型相比，在增加小部分准确度误差的情况下，完整度误差减小了 9.6%，整体误差减小了 3.3%。这归功于该模块在空间域中聚合了一个大的感受野，通过卷积层学习残差深度值来更新深度图，并利用设计的上采样网络减少了上采样的深度损失。在加入 CGPA 模块后通过引导参考特征的上下文信息来构建具有丰富信息的全局代价体使得完整度误差与整体误差大幅减小。最后，同时加入本文提出的 3 个模块后，以牺牲部分内存占用

达到了最佳性能。

表 2 DTU 测试集上各模块的消融结果对比

Table 2 Comparison of ablation results of each module on DTU test dataset

| Method | Acc | Comp | Overall | Memory/MB | Runtime/(s/view) |
|---|---|---|---|---|---|
| Baseline | 0.325 | 0.385 | 0.355 | 10690 | 0.29 |
| Baseline+FFA | **0.323** | 0.373 | 0.348 | 11820 | 0.32 |
| Baseline+CGPA | 0.361 | 0.285 | 0.323 | 14821 | 0.48 |
| Baseline+UPDR | 0.355 | 0.289 | 0.322 | 11206 | 0.29 |
| Baseline+FFA+UPDR | 0.353 | 0.290 | 0.321 | 12408 | 0.34 |
| Baseline+FFA+CGPA | 0.356 | 0.288 | 0.322 | 15521 | 0.56 |
| Ours | 0.355 | **0.283** | **0.319** | 16223 | 0.57 |

### 4.6 Tanks and Temples 数据集的评估结果

为了验证本文网络的泛化能力，使用微调后的模型对 Tanks and Temples 数据集的中间集进行点云重建来评估泛化效果。评估中将输入图像大小设置为 1920×1024，视图数 $N$ 设置为 7。中间集的定量结果如表 3 所列，其中粗体数字表示最优得分。从表中可以发现，本文方法在大部分场景中获得了最优得分，整体上展现出具有竞争力的结果。其在部分场景中保留了大量重复纹理区域的更多细节以及生成了更多的点来丰富细节，证明了本文方法不仅在 DTU 数据集上展现出良好的性能，还在大规模室外场景数据集上展现出了优秀的泛化能力。

表 3 Tanks and Temples 中间集的定量结果

Table 3 Quantitative results of Tanks and Temples intermediate set

| Method | Mean | Francis | Horse | Family | Lighthouse | M60 | Panther | Playground | Train |
|---|---|---|---|---|---|---|---|---|---|
| COLMAP[32] | 42.14 | 22.25 | 25.63 | 50.41 | 56.43 | 44.83 | 46.97 | 48.53 | 42.04 |
| MVSNet[4] | 43.48 | 28.55 | 25.07 | 55.99 | 50.79 | 53.96 | 50.86 | 47.90 | 34.69 |
| R-MVSNet[18] | 48.40 | 46.65 | 32.59 | 69.96 | 42.95 | 51.88 | 48.80 | 52.00 | 42.38 |
| UCS-Net[7] | 54.83 | 53.16 | 43.03 | 76.09 | 54.00 | 55.60 | 51.49 | 57.38 | 47.89 |
| CasMVSNet[5] | 56.42 | **58.45** | **46.20** | 76.36 | 55.53 | 56.11 | 54.02 | 58.17 | 46.56 |
| PatchmatchNet[20] | 53.15 | 52.64 | 43.24 | 66.99 | 54.87 | 52.87 | 49.54 | 54.21 | **50.81** |
| CVP-MVSNet[6] | 54.03 | 47.74 | 36.34 | 76.5 | 55.12 | 57.28 | 54.28 | 57.43 | 47.54 |
| Ours | **57.78** | 56.72 | 45.26 | **77.48** | **58.05** | **60.16** | **56.25** | **61.12** | 47.22 |

在 Tanks and Temples 数据集重建的点云如图 9 所示。

图 9 Tanks and Temples 数据集上不同场景的点云重建效果

Fig. 9 Point cloud reconstruction results of different scenes on Tanks and Temples dataset

**结束语** 本文提出了一个新颖的融合上下文信息引导代价体构建和深度细化的由粗到细的网络。首先，利用基于无参注意力的特征融合模块来解决多尺度特征不一致的问题，在特征提取方面提供更加有效的特征；然后，使用基于上下文引导代价体模块，利用参考特征的上下文信息来显式指导代价体的构建，通过更多的全局信息来增强代价体匹配的鲁棒性和完整性；最后，使用深度细化模块以提供一个更为准确的深度图给下一阶段。实验结果表明，本文方法在无纹理、重复纹理区域能够生成更为准确的深度图，在公开的 MVS 基准测试中，本文方法也取得了不错的成绩。全面的消融实验证明了提出的模块在 MVS 网络的有效性，但由于真实深度以及点云的缺乏，未来打算在无监督或者自监督的 MVS 领域中应用本文提出的网络并进一步发掘其潜力。

## 参 考 文 献


[1] WANG X, WANG C, LIU B, et al. Multi-view stereo in the deep learning era: A comprehensive review[J]. Displays, 2021, 70: 102102.

[2] FURUKAWA Y, HERNANDEZ C. Multi-view stereo: A tutorial[J]. Foundations and Trends © in Computer Graphics and Vision, 2015, 9(1/2): 1-148.

[3] GU J, WANG Z, KUEN J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[4] 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(ECCV). 2018: 767-783.

[5] 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. 2020: 2495-2504.

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

[7] 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. 2020: 2524-2534.

[8] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature p yramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[9] HARIS M, SHAKHNAROVICH G, UKITA N. Deep back-projection networks for super-resolution[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 16641673.

[10] SINHA S N, MORDOHAI P, POLLEFEYS M. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh [C]// 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007: 1-8.

[11] FURUKAWA Y, PONCE J. Carved visual hulls for image-based modeling[C]// Computer Vision – ECCV 2006: 9th European

Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. Springer Berlin Heidelberg, 2006: 564-577.

[12] SCHONBERGER J L, FRAHM J M. Structure-from-motion revisited[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4104-4113.

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

[14] CAMPBELL N D F, VOGIATZIS G, HERNANDEZ C, et al. Using multiple hypotheses to improve depth- maps for multi-view stereo[C]// Computer Vision ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10. Springer Berlin Heidelberg, 2008: 766-779.

[15] TOLA E, STRECHA C, FUA P. Efficient large-scalemulti-view stereo for ultra high-resolution image sets[J]. Machine Vision and Applications, 2012, 23: 903-920.

[16] KANG S B, SZELISKI R, CHAI J. Handling occlusions in dense multi-view stereo[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. IEEE, 2001.

[17] 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. 2017: 2307-2315.

[18] 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 PaRtern recognition. 2019: 5525-5534.

[19] 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. 2020: 1949-1958.

[20] 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. 2021: 14194-14203.

[21] 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. 2022: 8645-8654.

[22] 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. 2022: 12991-13000.

[23] CAO C, REN X, FU Y. Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo[J]. arXiv: 2208. 02541, 2022.

[24] Ding Y, YUAN W, Zhu Q, et al. Transmvsnet: Globalcontext-aware multi-view stereo network withtransformers[C]// Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition. 2022: 8585-8594.

[25] 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. 2021: 5732-5740.

[26] LUO A, YANG F, LI X, et al. Learning optical flow with kernel patch attention[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 8906-8915.

[27] YANG L, ZHANG R Y, LI L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks[C]// International Conference on Machine Learning. PMLR, 2021: 11863-11874.

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

[29] 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.

[30] CAMPBELL N D F, VOGIATZIS G, HERNANDEZ C, et al. Using multiple hypotheses to improve depth-maps for multi-view stereo[C]// Computer Vision ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 1218, 2008, Proceedings, Part I 10. Springer Berlin Heidelberg, 2008: 766-779.

[31] GALLIANI S, LASINGER K, SCHINDLER K. Gipuma: Massively parallel multi-view stereo reconstruction[J/OL]. https://www. dgpf. de/src/tagung/jt2016/proceedings/papers/34 _ DLT2016_Galliani_et_al. pdf.

[32] SCHONBERGER J L, FRAHM J M. Structure-from-motion revisited[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4104-4113.

[33] 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. 2019: 1538-1547.

[34] 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. 2021: 6187-6196.

[35] YI P, TANG S, YAO J. DDR-Net: Learning multi-stage multi-view stereo with dynamic depth range[J]. arXiv: 2103. 14275, 2021.

**CHEN Guangyuan**, born in 2001, postgraduate. His main research interests include multi-view stereo and 3D reconstruction.

**WANG Zhaohui**, born in 1967, professor, Ph.D supervisor. His main research interests include advanced computer control technology and biomedical information processing.

（责任编辑：何杨）