DOI: 10.12382/bgxb.2023.0740

融合注意力机制和多层动态形变卷积的多视图立体视觉重建方法

孙凯[1], 张成[1*], 詹天[1], 苏迪[2]
(1. 北京理工大学 宇航学院 飞行器动力学与控制教育部重点实验室, 北京 100081;
2. 杭州极弱磁场国家重大科技基础设施研究院, 浙江 杭州 310051)

摘要: 针对现有多视图立体视觉(Multi-View Stereo, MVS)技术提取弱纹理区域和非郎伯体曲面特征信息不充分及重建效果不理想问题,提出一种融合注意力机制和多层动态形变卷积的 AMDC-PatchmatchNet 方法。构建一种融合坐标注意力的特征提取网络,能更准确地捕捉重建对象的边缘形状和纹理特征,同时融合一种基于动态形变卷积的自适应感受野模块,根据不同尺度的特征自适应调整感受野的大小和形状,获得兼具全局和细节的特征表示。在 DTU 数据集上的测试结果表明,所提方法相较于主流 MVS 方法,点云重建整体性指标提高 2.8%,并且在航空影像数据集上验证了模型的泛化能力。

关键词: 多视图立体视觉; 注意力机制; 动态形变卷积; 深度学习
中图分类号: TP183; TP391 **文献标志码**: A **文章编号**: 1000-1093(2024)10-3631-11

Multi-view Stereo Vision Reconstruction Network with Fusion Attention Mechanism and Multi-layer Dynamic Deformable Convolution

SUN Kai[1], ZHANG Cheng[1*], ZHAN Tian[1], SU Di[2]
(1. Key Laboratory of Dynamics and Control of Flight Vehicle of Ministry of Education, School of Aerospace Engineering, Beijing Institute of Technology, Beijing 100081, China;
2. National Institute of Extremely-Weak Magnetic Field Infrastructure, Hangzhou 310051, Zhejiang, China)

Abstract: The existing multi-view stereo vision technology is not enough to extract the feature information of weak texture region and non-Lambert surface, and its reconstruction effect is not ideal. An AMDC-PatchmatchNet method with fusion attention mechanism and multi-layer dynamic deformable convolution is proposed for the problems above. In this method, a feature extraction network integrating the coordinate attention is constructed, which can capture the edge shape and texture features of reconstructed objects more accurately. At the same time, an adaptive receptive field module based on dynamic deformable convolution is integrated in the feature extraction network, and the size and shape of receptive field can be adjusted adaptively according to different scales of features to obtain both global and detailed feature representation. The generalization ability of the AMDC-PatchmatchNet method is verified on the aerial image data sets. The test results on DTU data sets show that the overall index of point cloud reconstruction of the proposed method is improved by 2.8% compared with those of mainstream MVS methods.

收稿日期: 2023-08-10
* **通信作者邮箱**: zhangcheng@bit.edu.cn

Keywords: multi-view stereo vision; attention mechanism; dynamic deformable convolution; deep learning

0 引言

多视图立体视觉(Multi-View Stereo, MVS)技术是基于一组已知相机位姿信息和固有参数的图像序列,推理出图像中每个像素点的深度信息,从而重建观测场景三维模型的技术,是实现二维图像到三维模型转变的重要途径,在战场态势感知、三维地图构建、智慧城市、视觉导航领域有着广泛的应用[1-5]。然而,在实践应用中,MVS 技术在处理弱纹理区域(如地板、墙面等)和非郎伯体曲面区域(如玻璃、金属表面等)时仍存在挑战。因此,进一步研究和改进 MVS 技术在复杂场景中的应用具有重要意义。

传统 MVS 重建方法[6-7]依赖于手工设计的相似性度量标准和代价空间正则化方法来建立特征匹配关系和恢复三维点空间信息,如绝对误差和(Sum of Absolute Differences, SAD)方法、误差平方和(Sum of Squared Difference, SSD)方法、归一化互相关(Normalized Cross-Correlation, NCC)匹配方法等,这些方法在处理弱纹理、非朗伯体曲面等复杂场景时的鲁棒性较差。传统 MVS 方法通常假设光线是均匀分布的,反射光的强度仅与单位面积上入射光的强度有关,而与光源的方向无关,这对于朗伯体曲面是合理的假设。然而,非郎伯体曲面的镜面反射特性使光线在反射过程中的方向性较强,导致不同视角下的反射光分布不均匀。同时,当光线在表面反射时会集中在某些方向上,形成高光区域[8]。在高光区域缺乏纹理信息的情况下,传统 MVS 方法无法准确地匹配特征点,从而导致深度估计的不准确性。近年来,随着深度学习方法的广泛研究应用,Voronin 等[9]提出通过卷积神经网络可以合并全局语义信息,如镜面反射和反射先验,有助于校正由不同视图的反射引起的三角测量重建误差,实现更准确地重建具有非郎伯体曲面反射特性的三维场景。陈龙等[10]提出通过将全局注意力与不同感受野的深度卷积进行融合,可缓解网络在高维向量中的性能衰减问题,以提升网络对图像特征的感知能力。杜小强等[11]提出引用可变形卷积对区域掩码卷积神经网络模型(Mask Region-based Convolutional Neural Networks, Mask R-CNN)进行优化,通过提升对不同尺度特征提取的适应能力,解决对非结构化目标的检测精度低、鲁棒性较差的问题。在基于深度学习的 MVS 方法中,Yao 等[12]提出了 MVS 网络(Multi-View Stereo Network, MVSNet)方法,通过将立体匹配问题转化为水平方向上逐像素的视差估计问题,采用光度一致性和几何一致性准则来约束重建过程,减少由于非朗伯体曲面和复杂光照条件引起的误差,以提高深度估计的鲁棒性。Chen 等[13]提出 Point-MVSNet,通过将 3D 几何先验信息和 2D 纹理信息融合到具有特征增强的点云中,使用点云数据来优化场景中匹配点的深度预测,有助于提高弱纹理区域多视图立体匹配的准确性。Luo 等[14]提出 P-MVSNet,通过借鉴传统 MVS 方法中 Patch-Wise 思想,将深度图和重建点云的生成过程划分为多个小的局部区域,并对每个局部区域进行独立的深度估计和点云重建,使得深度图和重建点云的准确性和完整性都得到提高。Yang 等[15]提出 CVP-MVSNet,采用由粗到细的策略构建金字塔代价体,通过获取不同尺度下的特征信息,减少弱纹理区域的深度估计误差。Wang 等[16]提出 PatchmatchNet,通过利用图像块之间的相似性,采用传播机制来逐步调整每个像素点的深度估计值,提升弱纹理区域深度预测的准确性。这些基于深度学习改进的方法通过引入不同的网络结构、代价计算策略和优化方法,提高了深度估计的准确性和计算效率,但在处理弱纹理区域和非郎伯体曲面时仍存在一定的局限性。由于弱纹理区域和非郎伯体曲面的有效特征信息有限,随着图像下采样网络层数的增加,关键特征中用于确定特征点位置的信息会进一步损失,因而会出现特征点位置匹配准确性下降的问题。此外,现有基于深度学习的 MVS 方法大多采用金字塔特征网络(Feature Pyramid Network, FPN)结构进行特征提取,FPN 在不同的预测特征层上针对不同尺度的目标进行预测。由于 FPN 主要关注局部邻域内的上下文信息,导致对于全局信息的关注不足,造成丢失场景的整体几何结构和语义信息的问题。

为提高基于深度学习的 MVS 方法在处理弱纹理区域和非郎伯体曲面的重建能力,本文提出一种融合注意力机制和多层动态形变卷积网络(Deformable Convolution Network, DCN)的多视图立体视觉重建深度学习网络(PatchmatchNet with Fusion Attention Mechanism and Multi-Layer Dynamic Deformation Convolution, AMDC-PatchmatchNet):

1)提出融合坐标注意力(Coordinate Attention,CA)机制的 FPN 特征提取网络,使用坐标信息来动态调整通道特征的权重,提升网络对边界信息、纹理信息等细节特征的感知能力;

2)构建基于 DCN 的自适应感受野模块(Adaptive Receptive Field Block, ARFB),该模块可以根据上采样特征融合后的尺度,自适应地调整卷积核的采样位置和权重,使网络获取具有全局和细节语义的多尺度信息。

1　AMDC-PatchmatchNet 网络结构

1.1　总体框架

AMDC-PatchmatchNet 是一种基于 PatchmatchNet 框架进行设计的 MVS 方法,结合了改进的特征提取网络和自适应感受野模块。其整体结构如图 1 所示,包括融合 CA 模块(Coordinate Attention Block,CAB)的特征提取网络、ARFB、基于学习的区域匹配和深度图优化。

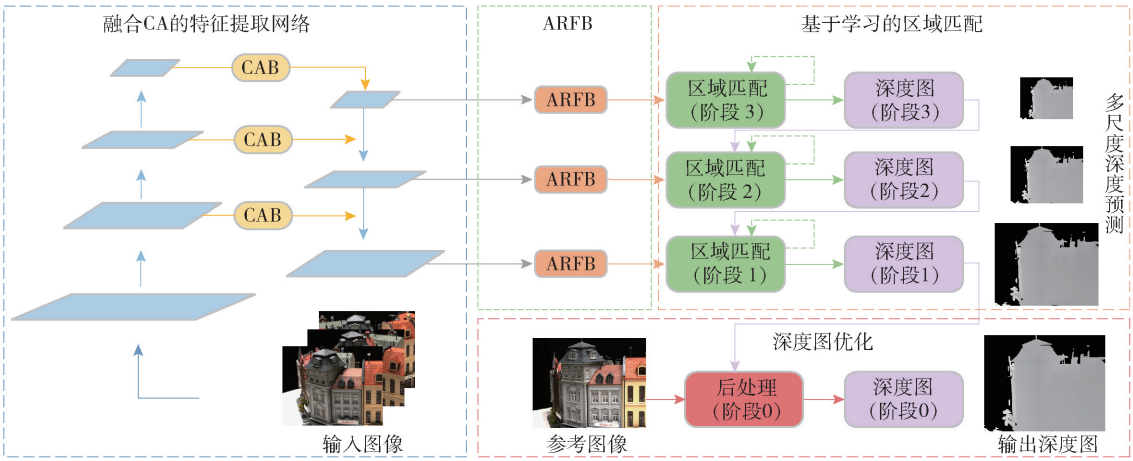


图 1　AMDC-PatchmatchNet 的总体框架

Fig. 1　Overall framework of AMDC-PatchmatchNet

首先,在 FPN 的下采样过程中,AMDC-PatchmatchNet 通过在每个尺度的特征输出中融合 CAB。CAB 包括两个主要部分:坐标信息融合和坐标注意生成。在坐标信息融合中,模型将图像中特征的位置信息编码至通道注意力中;坐标注意生成则用于生成最终的 CA 图,从而帮助模型更准确地捕捉物体的形状和纹理特征。其次,为进一步获得全局和细节特征信息,AMDC-PatchmatchNet 引入一种基于 DCN 的 ARFB。该模块通过引入额外的偏移量来增加卷积的空间采样位置,根据不同尺度的特征自适应地调整感受野的大小和形状,获得兼具全局和细节的特征表示。基于学习的区域匹配用于对深度图像进行估计,并根据匹配误差进行不断优化,从而提高深度估计的准确性。深度图优化则是对估计的深度图进行进一步的调整和改进,以得到更准确的深度图。

1.2　融合 CAB 的特征提取网络

目前基于 FPN 的特征提取网络容易受到图像背景、噪声信息以及纹理特征分布不均等因素的影响,导致网络在弱纹理区域和非郎伯体曲面区域无法准确地捕捉物体的结构和纹理特征。因此,如何增强特征的位置信息来提高网络对物体结构信息的关注也是问题之一。注意力机制参考人类视觉系统主动选择关注对象并集中关注处理的视觉特性,将计算资源分配给具有重要信息的部分,优化卷积神经网络[17]。在现有的研究中,大量研究者将注意机制嵌入到深度神经网络中,并取得了很好的实验结果[18]。将注意力机制嵌入深度神经网络的常见例子包括对象分类[19]、图像分割[20]、目标检测[21]等。CA 是 Hou 等[22]在 2021 年提出的一种方法,主要由两阶段构成,其结构如图 2 所示。

在坐标信息融合阶段,对于来自上层的输入特征图 $\boldsymbol{X}$,首先将其分别在水平和垂直方向进行分解,得到两个一维特征编码。实现过程为对于输入特征图的高度 H 和宽度 W,对每个特征向量通道 C,使用两个尺寸为(H,1)、(1,W)的池化核,沿水平和垂直方向对特征进行编码:

$$z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i<W}x_c(h,i) \tag{1}$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j<H}x_c(j,w) \tag{2}$$

式中:h 和 w 对应为当前输入特征图的高度和宽度,

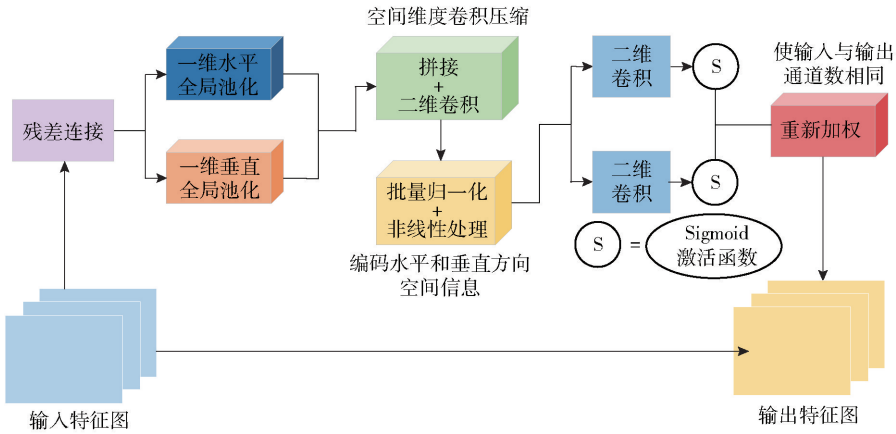


图 2 CAB 结构

Fig. 2 Coordinate attention block structure

该值随着下采样过程而变化；Z_c^h 表示在第 c 通道处高度 h 下的输出特征；Z_c^w 表示在第 c 通道处宽度 w 下的输出特征；i 表示[0,W]范围内某一列的索引；j 表示[0,H]范围内某一行的索引。通过式(1)和式(2)这两个一维特征编码过程，可以提取出特征在水平和垂直方向上的局部结构信息，使坐标信息得到有效融合：

$$\boldsymbol{f}=\delta(F_1([\boldsymbol{z}^h,\boldsymbol{z}^w])) \tag{3}$$

$$\boldsymbol{g}^h=\sigma(F_h(\boldsymbol{f}^h)) \tag{4}$$

$$\boldsymbol{g}^w=\sigma(F_w(\boldsymbol{f}^w)) \tag{5}$$

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \tag{6}$$

式中：$\boldsymbol{f}$ 为具有水平和垂直空间信息的中间特征图；$\delta(\cdot)$ 表示非线性激活函数；$F_1(\cdot)$ 表示 1×1 卷积变换函数；$\boldsymbol{z}^h$、$\boldsymbol{z}^w$ 分别为垂直和水平方向上的输出特征；$\boldsymbol{g}^h$ 和 $\boldsymbol{g}^w$ 分别为垂直和水平方向上生成的注意力权重；σ 为 sigmoid 激活函数；F_h 和 F_w 分别代表垂直方向和水平方向上的 1×1 卷积操作；$\boldsymbol{f}^h$、$\boldsymbol{f}^w$ 分别为垂直和水平方向上的中间特征图；$y_c(i,j)$、$x_c(i,j)$分别为第 c 通道处的输出特征图与输入特征图；$g_c^h(i)$、$g_c^w(j)$分别为第 c 通道处垂直和水平方向上生成的注意力权重。

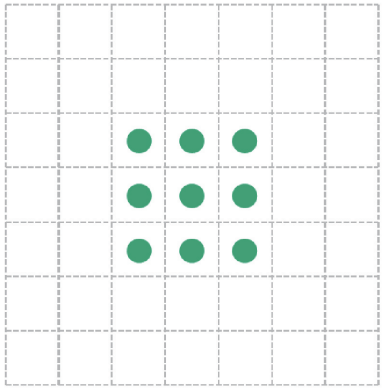
在注意力生成阶段，将获取到的一对具有空间方向性的感知特征图[$\boldsymbol{z}^h$,$\boldsymbol{z}^w$]进行拼接，采用 1×1 卷积操作将特征通道数压缩，使用 ReLU 函数进行非线性激活，将获取到的结果分解为水平注意张量和垂直注意张量，使用两组 1×1 卷积还原特征通道数，再使用 Sigmoid 函数进行非线性激活。此时，将目标位置信息保存到生成的注意力图中，然后通过乘法运算将目标位置信息添加到输入特征图中，得到融合位置权重的输出特征图 $\boldsymbol{Y}$。

融合 CA 注意力的 FPN 特征提取网络由卷积层和注意力层组成，其结构如图 1 所示，通过在水平和垂直方向对输入特征分配权重计算，两个注意力映射中的每个元素都反映对应行和列所在区域是否有关注的特征区域。CAB 模块的输出融合了通道间的信息、横向以及纵向的空间信息，能够实现对特征区域的准确定位，提高对物体边缘特征的检测精度。将临近像素之间的语义信息编码至不同尺度的特征图，可以有效避免输入的图像特征降采样后丢失语义信息。

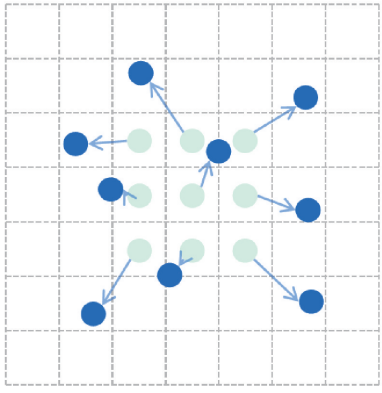
1.3 基于多层 DCN 的 ARFB

在 1.2 节中，特征提取网络采用了多尺度特征融合的策略，旨在充分利用底层特征的高分辨率信息和高层特征的丰富语义信息。网络通过上采样模块将不同尺度的特征图进行融合，从而在输出的特征图中的不同位置包含不同尺度和不同形变的特征信息。然而，原网络采用的卷积操作将特征图分成与所设定卷积核大小相同的网格进行采样。这种固定权重的感受野卷积核导致同一卷积层在处理每张图像的不同位置区域时的感受野尺寸都相同。因此在检测弱纹理区域和非朗伯体曲面区域特征时，由于目标的纹理较弱或非朗伯体曲面的光照变化较大，图像特征将变得更加复杂和不规则，采用固定感受野卷积核的卷积处理对此图像特征的适应能力和调节能力较弱，导致特征提取的不足。研究表明，多感受野机制可以使检测网络更好的学习远程空间关系，并建立隐式空间模型[23]。为此，Dai 等[24]提出了一种可变形卷积的方法，在可变形卷积中，对输入特征图施加额外的卷积层来学习每个像素点的偏移量，从而实现对特征图的非线性变形。这种偏移量可以用来调整卷积核在输入特征图上的采样位置，

从而更好地适应不同的图像结构和纹理特征。可变形卷积采样的偏移示例如图 3 所示。



(a) 标准卷积
(a) Standard convolution



(b) 可变形卷积
(b) Deformable convolution

图 3　两种 3×3 卷积核不同的采样形式

Fig. 3　Two different 3×3 sampling forms of convolutional kernel

传统的标准卷积结构定义为式(7),对于输出特征图 $\boldsymbol{y}$ 中的每一个位置 $\boldsymbol{P}_0$,有

$$\boldsymbol{y}(\boldsymbol{P}_0)=\sum_{\boldsymbol{P}_n\in\boldsymbol{R}}\boldsymbol{\omega}(\boldsymbol{P}_n)\cdot\boldsymbol{x}(\boldsymbol{P}_0+\boldsymbol{P}_n) \tag{7}$$

式中:$\boldsymbol{P}_0$ 是将输出特征图 $\boldsymbol{y}$ 每点对应到卷积核中心,然后映射到输入特征图 $\boldsymbol{x}$ 中的坐标值;$\boldsymbol{R}$ 定义了感受野的大小和扩张率,对于一个大小为 3×3 的感受野和扩张率为 1 的卷积核,$\boldsymbol{R}=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$;$\boldsymbol{P}_n$ 是在 $\boldsymbol{R}$ 中对于 $\boldsymbol{P}_0$ 的相对坐标;$\boldsymbol{\omega}(\cdot)$ 为采样点权重;$\boldsymbol{x}(\cdot)$ 为输入特征图 $\boldsymbol{x}$ 中坐标值到特征向量的映射。

可变形卷积是对常规卷积的采样位置 $\boldsymbol{R}$ 增加偏移量$\{\Delta\boldsymbol{P}_n\mid n=1,\cdots,N\}$进行增广,其中 $N=|\boldsymbol{R}|$,于是可变形卷积定义为

$$\boldsymbol{y}(\boldsymbol{P}_0)=\sum_{\boldsymbol{P}_n\in\boldsymbol{R}}\boldsymbol{\omega}(\boldsymbol{P}_n)\cdot\boldsymbol{x}(\boldsymbol{P}_0+\boldsymbol{P}_n+\Delta\boldsymbol{P}_n) \tag{8}$$

可变形卷积提取特征的过程如图 4 所示,对输入特征图施加额外卷积层来学习偏移量 $\Delta\boldsymbol{P}_n$。在网络训练的过程中,可变形卷积的参数包括用于生成输出特征的卷积核和用于生成偏移量的卷积核,并在同一次前向传播中进行更新。由于加入偏移量后的采样位置通常为非整数,并不对应特征图上实际存在的像素点,需要使用插值法来得到偏移量。通过反向传播算法,网络可以根据损失函数的梯度来调整这两个卷积核的权重,从而使得生成的输出特征和偏移量能够更好地适应目标特征的形态变化,提高卷积操作的灵活性和适应性。

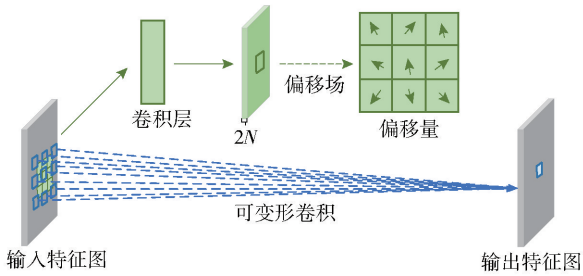


图 4　可变形卷积特征的提取过程

Fig. 4　Extraction process of deformable convolution feature

为使网络能够根据局部上下文信息获取具有不同感受野的特征,加强不同尺度特征层间的信息融合,本文设计一个基于 DCN 的 ARFB 对 1.2 节所获取的多尺度特征进行进一步处理,其结构如图 5 所示,通过常规卷积计算采样位置的二维偏移量,使得可变形卷积的采样位置可以是输入特征图的任意位置,不仅扩大了特征检测网络的感受野,同时增强了网络适应不同尺度特征图的采样能力,提升对弱纹理区域以及非郎伯体曲面特征的提取效果。

1.4　损失函数

本文采用的损失函数同时衡量区域匹配每个阶段输出的预测深度图和真实深度值之间的损失以及经过深度图优化后和真实深度值之间的损失,网络

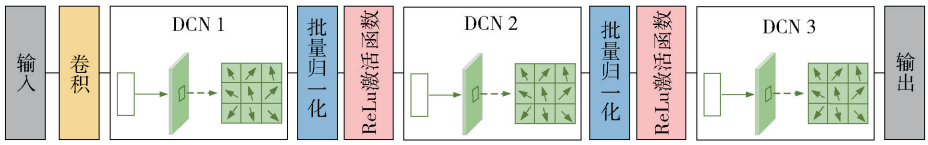


图 5　基于 DCN 的 ARFB 结构

Fig. 5　Structure diagram of DCN-based adaptive receptive field block

采用估计深度图和真实深度值之间的平均绝对误差作为训练损失,具体为

$$Loss = \sum_{k=1}^{3}\sum_{i=1}^{n_k} L_i^k + L_{ref}^0 \tag{9}$$

即在区域匹配的阶段 k 所输出第 i 次迭代得到的深度图与真实深度值之间的损失 L_i^k 和经过深度图优化后得到的深度图与真实深度值之间的损失 L_{ref}^0 之和为最终的损失。

2 实验与结果分析

2.1 数据集

本文采用 DTU 数据集[25]进行实验,该数据集是由丹麦理工大学针对 MVS 任务而拍摄和处理的大型室内数据集。为了获取不同光照条件下具有准确相机位姿参数的视图,研究人员在工业机械臂上搭载了亮度可调的光照设备,从多个视角拍摄获取图像。数据集由 124 个不同场景组成,每个场景包括 49 个视角,每个视角包含 7 种光照条件,形成 343 张图片,每张图片的分辨率为 1 600 × 1 200。数据集划分为训练集(包含 78 个场景)、验证集(包含 18 个场景)、测试集(包含 22 个场景),确保在训练、验证和测试过程中都能涵盖不同的类型场景和视角,从而更全面地评估 MVS 方法的性能。通过使用 DTU 数据集进行训练,可以使模型学习到不同光照条件下的视图之间的关系,并提高模型在实际应用中的泛化能力。同时,由于 DTU 数据集具有较高的质量和丰富的场景多样性,可以为 MVS 方法的研究和改进提供重要的参考和基准。

同时,本文还使用公开数据集 SenseFly[26]来验证改进网络的泛化能力。与 DTU 数据集不同,SenseFly 数据集由 eBee Classic 无人机采集的室外场景图像构成。通过使用 SenseFly 数据集,可以验证改进网络在处理视角变化范围大、存在弱纹理与非朗伯体曲面区域的大规模场景中是否具备较为精确的重建能力,从而更全面地评估改进网络的性能,并进一步验证其在实际应用中的可靠性和有效性。

2.2 实现细节

本文在 Pytorch 环境下使用 2.1 节提出的数据集训练网络,所有实验均在 Intel Xeon(R) Gold 5128@2.30 GHz × 64 CPU 和 2 块 NVIDIA RTX 3090 GPU 条件下进行。

在训练阶段使用 640 × 512 的图像分辨率,测试阶段采用 1 600 × 1 200 的图像分辨率。预设的深度假设范围是[425 mm,935 mm]。阶段 3、阶段 2、阶段 1 的区域匹配迭代次数分别设置为 2、2、1,在深度传播过程中所施加的随机扰动的参数在阶段 3、阶段 2、阶段 1 设为 16、8、1,自适应传播的邻域数设置为 16、8、0。自适应空间代价聚合的所有阶段自适应匹配成本聚合的邻域数设置为 9。网络学习率初始值设为 0.001,待训练到 10 个、12 个、14 个轮次时,对学习率分别进行减半操作,共训练 16 个轮次(迭代次数为 108 380 次),批大小设为 4,使用 Adam 优化器($\beta_1 = 0.9$,$\beta_2 = 0.999$)。

如图 6 所示,将本文方法和 PatchmatchNet 方法在相同的实验环境下进行训练,通过比较网络估计深度值与实际深度值之间的绝对误差,并统计绝对误差值大于 1 mm 的误差像素的比列,得到网络训练过程的深度估计误差率曲线,与 PatchmatchNet 方法相比,本文方法在深度估计误差率曲线的前期下降速度方面表现出相似的趋势,深度估计误差率随着迭代次数的增加逐渐收敛,但本文方法在训练过程中可达到更低的深度估计误差率,结果表明通过在网络中引入本文设计的 CAB 和 ARFB,可以提高深度估计的准确性。

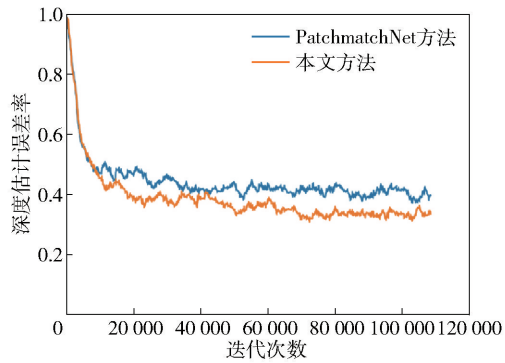


图 6 绝对误差值大于 1 mm 的深度估计误差率曲线图

Fig. 6 Depth estimation error rate with absolute error value greater than 1 mm

2.3 对比实验

为证明本文方法的有效性,与传统方法 Camp[27]、Furu[28]、Tola[29]、Gipuma[6]、Colmap[7],以及基于深度学习的三维重建方法 MVSNet[12]、R-MVS-Net[30]、P-MVSNet[14]、CVP-MVSNet[15]、Fast-MVSNet[31]、PatchmatchNet[16]进行定量实验测试结果对比。

使用 DTU 数据集官方提供的评估准则,对点云重建准确性误差和点云重建完整性误差和点云重建整体性误差进行评估。其中,准确性误差是由三维重建出的预测点云与结构光扫描仪获取的真实点云

之间的匹配误差计算得到,完整性误差计算的是从真实点云到预测点云之间的匹配误差。具体而言,匹配误差是通过计算重建点云中每一个点到真实点云中最近点的距离来得到的。在计算匹配误差时,首先需要将重建点云中的每一个点与真实点云中的最近点进行匹配。然后,计算并记录每一个点到最近点的距离。为了避免部分外点对结果产生过大的影响,将距离大于 20 mm 的观测值进行移除。排除一些异常值或噪声点对结果的影响,使得匹配误差更加准确和可靠。最后,计算所有距离的均值作为该点云到真实点云的距离。

在点云质量评估过程中,考虑高准确性与高完整性存在相互制约的关系,为得到更均衡的评价分析,通过计算准确性和完整性的平均值,使用整体性误差评价整体重建质量。因此,整体性误差是点云重建 3 个指标中最重要的指标,通过避免单一指标的局限性,可以综合考虑点云重建的准确性和完整性,提供更全面和准确的评价分析,为重建结果的质量评估提供更可靠的依据。

对比结果如表 1 所示,评估指标值越低表示重建质量越好。

由表 1 可见:Gipuma 方法在准确性上表现最好,PatchmatchNet 方法在完整性上表现最好,而本文方法在点云重建整体性指标优于所有方法,相较于原网络,在点云重建准确性方面提升了 4.9%,点云重建整体性方面提升了 2.8%。在特征提取模块中,引入 CA 机制以增强对纹理信息的感知能力,使网络聚焦于关键点的位置信息,从而提高特征提取的精确度。而基于可变形卷积的 ARFB 使网络根据局部上下文信息获取具有不同感受野的特征,以增强不同尺度特征之间的信息融合能力,实现重建点云精度的提升。根据表 2 中对比框选区域的重建点云效果,可以观察不同方法在处理弱纹理和非郎伯体曲面时的表现差异。传统方法(Colmap 方法)在弱纹理和非郎伯体曲面区域的重建效果较差,而基于深度学习的方法(MVSNet、PatchmatchNet 方法)在此区域的重建效果有较大改善,但仍存在不同程度的点云空洞和细节缺失。相比之下,本文方法产生的点云在弱纹理和非朗伯体曲面区域的重建结果最为完整,并且对表面细节的恢复也更加精细。

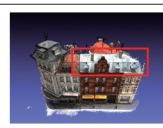
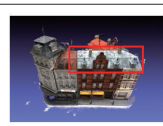
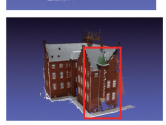
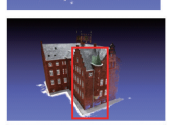
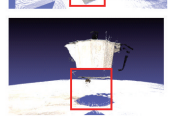
表 1　不同方法在 DTU 数据集上的定量测试结果

Table 1　Quantitative test results of different methods on DTU datasets

方法	准确性误差/mm	完整性误差/mm	整体性误差/mm
Camp	0.835	0.554	0.695
Furu	0.613	0.941	0.777
Tola	0.342	1.190	0.766
Gipuma	**0.283**	0.873	0.578
Colmap	0.532	0.400	0.664
MVSNet	0.396	0.527	0.462
R-MVSNet	0.383	0.452	0.417
P-MVSNet	0.406	0.434	0.420
Fast-MVSNet	0.336	0.403	0.370
CVP-MVSNet	0.296	0.406	0.351
PatchmatchNet	0.427	**0.277**	0.352
本文方法	0.406	0.279	**0.342**

表 2　DTU 数据集部分模型稠密重建点云对比

Table 2　Point clouds reconstructed by different methods on DTU dataset

编号	输入图像	Colmap方法	MVSNet方法	PatchmatchNet方法	本文方法	真实扫描点云
Scan 9						
Scan 24						
Scan 77						

2.4　消融实验

为了进一步验证本文方法的有效性,对网络框架进行了消融实验和定量定性实验分析,对框架中关键组件的优势进行评估。针对本文提出的 CAB

和 ARFB,在 DTU 数据集上进行了 4 组对比实验,输入图片分辨率为 1 600 × 1 200,并使用准确性误差、完整性误差、整体性误差、GPU 内存消耗、特征图对比、深度图对比结果衡量重建质量。实验结果如表 3 所示(得分越低越好),定性结果对比如表 4、图 7 所示。

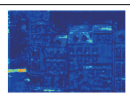
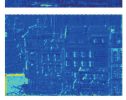
表 3 消融实验定量结果对比

Table 3 Comparison of quantitative results of ablation experiments

方法	准确性误差/mm	完整性误差/mm	整体性误差/mm	GPU 内存消耗/MB
无 CAB、无 ARFB	0.429	0.335	0.382	**10 877**
加入 CAB、无 ARFB	0.420	0.304	0.362	10 885
无 CAB、加入 ARFB	**0.397**	0.321	0.359	10 887
本文方法	0.406	**0.279**	**0.342**	10 885

表 4 融合 CAB 前后不同阶段特征图可视化图像对比

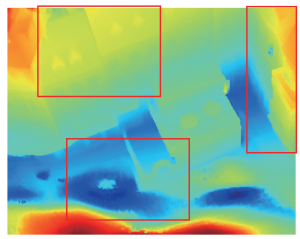
Table 4 Visualization comparison of feature maps at different stages before and after CAB

方法	$\frac{w}{2}\times\frac{h}{2}$	$\frac{w}{4}\times\frac{h}{4}$	$\frac{w}{8}\times\frac{h}{8}$
PatchmatchNet方法			
本文方法			

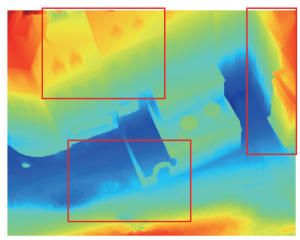
如表 3 所示实验结果进行定量分析,单独使用 CAB 模块后的网络主要在完整性方面优于原网络,由于 CAB 模块能够更好地保留输入图像特征的细节和结构,从而提高点云重建的完整性。单独使用 ARFB 模块后的网络主要在准确性方面有较大提升,ARFB 模块通过反馈机制不断调整卷积核采样的位置,使得网络能够更准确地进行深度估计。本文方法同时使用 CAB 模块和 ARFB 模块的网络点云重建效果最好,且在 GPU 内存消耗上没有较大影响,因此可以在不消耗额外的计算资源的情况下,同时使用这两个模块来提升网络的性能。

结合可视化结果对比对模块的作用进行定性分析。使用 Grad-CAM[32] 对不同尺度阶段生成的特征图进行可视化处理,选用 DTU 数据集中验证集的 scan23 作为测试输入对象,将融合 CAB 前后网络生成的特征图进行对比,如表 4 所示。从表 4 中可以看出,融合 CA 机制模块后可以更精确的定位特征分布区域,通过将通道注意分解为垂直方向和水平方向,聚合成两个独立的方向感知特征映射,减轻二维全局池化造成的位置信息丢失,有效地将空间坐标信息整合到生成的注意图中。这种方法既能充分保持特征图位置信息的完整性,又能对特征分布的空间位置动态赋权,有效提高了空间特征信息的利用率和检测网络对物体边缘信息的关注程度,使得由于深层特征分辨率的降低而引起物体边缘模糊或纹理信息丢失的情况得到改善。

(a) 参考图像
(a) Reference image



(b) PatchmatchNet 方法
(b) PatchmatchNet method



(c) 本文方法
(c) The proposed method

图 7 输出深度图对比

Fig. 7 Comparison of output depth maps

通过对网络预测输出的深度图进行可视化处理,对比加入 ARFB 模块前后网络预测的深度图,如图 7 所示。加入基于 DCN 的 ARFB 模块后,网络首先根据输入特征图学习像素采样点偏移量,使得偏移后卷积运算的感受野能够更好地匹配物体的实际形状。对比图片上端在楼顶与弱纹理区域(墙壁)之间的区分情况以及右侧楼体的轮廓细节,本文方法预测的深度图在场景轮廓分割方面表现更为清晰。同时,对比图片下端在建筑楼体与非郎伯体曲面(反光地面)之间的区分,本文方法预测的深度图在不同深度表面之间的过渡更加平滑且深度估计的结果更为准确。通过分析可知,采用基于多层 DCN

的 ARFB 可以使得网络通过获取具有不同感受野的特征信息,并加强特征层间的信息融合,通过对全局信息和局部信息的综合利用,提高弱纹理和非朗伯体曲面区域的重建质量。

综合来看,实验结果验证了 CAB 模块和 ARFB 模块的有效性,并且同时使用这两个模块可以进一步提升网络的性能。

2.5　网络泛化能力实验

为了验证本文方法的泛化能力,使用在 DTU 数据集上训练的模型,在不进行任何微调的情况下,选用航空影像公开数据集 SenseFly-Merlishachen 的示例数据集进行测试。图像由 eBee Classic 无人机拍摄的倾斜航空影像,共拍摄 297 张图像,图像分辨率大小为 4 608 × 3 456,数据采集覆盖区域面积为 0.57 km^2,飞行高度为 162 m,使用 Colmap 软件进行数据前处理,获取相机位姿。图 8 为实验的航空影像示例,数据集图像包含密集建筑区域、弱纹理区域(草地)、非郎伯体曲面区域(水面)。

(a) 密集建筑区域
(a) Built-up area

(b) 弱纹理区域
(b) Weak texture area

(c) 非郎伯体曲面区域
(c) Non-Lambertian surface region

图 8　航空影像数据集示例

Fig. 8　Example of an aerial imagery dataset

将本文方法生成的重建点云分别与数据集提供的真值点云通过 CloudCompare 软件采用最近点迭代(Iterative Closest Point,ICP)算法进行配准,计算两个点云之间的最近邻距离,搜寻点云中每个点的最近点,并计算两点之间的欧式距离,最终得到所有点的平均距离和标准差,较低的平均距离误差表示生成的点云与真值点云之间存在较小的偏差,较低的标准差表示点云中距离误差的分散程度较小。实验结果如表 5 所示,由于缺少尺度信息,计算结果单位为无量纲距离。与原网络相比,本文方法的平均距离误差减少了 19.12%,标准差减少了 8.7%。本文方法生成的重建点云与真值点云之间的平均距离误差更小,标准差更稳定,具有更好的准确性和一致性。

表 5　泛化能力实验定量结果对比

Table 5　Comparison of quantitative results of generalization ability experiments

方法	平均距离	标准差
PatchmatchNet 方法	0.127 894	0.089 143 6
本文方法	**0.103 431**	**0.081 384 5**

根据表 6 中生成的点云结果观察,原网络生成的重建点云具有 13 295 132 个点,点云整体在弱纹理区域和非郎伯体曲面区域存在孔洞。相比之下,采用本文方法生成的重建点云具有 19 290 474 个点,在重建点云中点的数量上增加了 45.1%。同时,通过与真实点云做对照,本文方法在弱纹理区域和非郎伯体曲面区域的重建效果也有所改善,更好地保留了场景的细节,并提供更丰富的点云信息,验证了本文方法在大范围室外场景重建任务中的有效性。

表 6　航空影像数据集测试结果对比

Table 6　Comparison of aerial imagery dataset test results

方法	整体视角	局部视角1	局部视角2
PatchmatchNet 方法			
本文方法			
真实点云			

3　结论

本文提出了一种 AMDC-PatchmatchNet 方法。得出主要结论如下:

1)通过在通道注意力机制中引入位置编码,可

以改善图像下采样过程中特征点位置信息逐渐丢失的问题。位置编码可以将图像中特征点的位置信息嵌入到通道注意力中,从而帮助网络更准确地捕捉物体的形状和纹理特征。通过引入位置编码,网络可以更好地理解图像中不同位置的特征,并在特征提取过程中保留更多的空间信息,使网络在弱纹理区域中更好地区分物体边界和纹理细节,从而提高深度估计的完整性。

2)本文提出的基于 DCN 的 ARFB,用于根据上采样特征融合后的尺度自适应地调整感受野。传统的固定感受野对于全局信息的获取存在一定的局限性,而 ARFB 可以根据不同尺度的特征自适应地调整感受野的大小,从而获取具有全局和细节语义的多尺度信息,使网络对非朗伯体曲面的重建更加鲁棒,提高深度估计的准确性。

3)在 DTU 数据集上的测试结果表明,与主流的 MVS 方法相比,本文方法在点云重建的整体性方面提升了 2.8%。同时,本文方法在不消耗额外计算资源的情况下,能够提高弱纹理区域和非郎伯体曲面的点云重建质量,在多视图立体视觉重建任务中具有较好的性能和泛化能力。

4)当前 MVS 方法重建场景(尤其是大规模场景)需要耗费大量时间,距离实时重建有一定差距。在未来,研究者可以通过设计高效率的深度学习模型,以及在硬件上的优化,以实现大规模场景的实时三维重建。同时,可将 MVS 与其他传感器信息(如激光雷达、RGB-D 相机等)结合,以获取更全面的场景信息,以提高 MVS 技术对于复杂场景的理解和建模能力。

参考文献(References)

[1] 蒋超, 崔玉伟, 王辉. 基于图像的无人机战场态势感知技术综述[J]. 测控技术, 2021,40(12):14 - 19.
JIANG C, CUI Y W, WANG H. Review of image-based UAV battlefield situation awareness technology[J]. Measurement and Control Technology, 2021, 40(12):14 - 19. (in Chinese)

[2] 纪广, 郝建国, 张振伟. 面向无人机作战的虚拟孪生系统设计方案[J]. 兵工学报, 2022,43(8):1902 - 1912.
JI G, HAO J G, ZHANG Z W. Design scheme of virtual twin system for UAV combat[J]. Acta Armamentarii, 2022, 43(8): 1902 - 1912. (in Chinese)

[3] 龙霄潇, 程新景, 朱昊,等. 三维视觉前沿进展[J]. 中国图象图形学报, 2021,26(6):1389 - 1428.
LONG X X, CHENG X J, ZHU H, et al. Advances in 3D vision [J]. Journal of Image and Graphics, 2021,26(6):1389 - 1428. (in Chinese)

[4] 张宗华, 刘巍, 刘国栋, 等. 三维视觉测量技术及应用进展[J]. 中国图象图形学报, 2021,26(6):1483 - 1502.
ZHANG Z H, LIU W, LIU G D, et al. Progress of 3D visual measurement technology and its application[J]. Journal of Image and Graphics, 2019,26(6):1483 - 1502. (in Chinese)

[5] 赵双赫. 基于双目立体视觉的实时三维重建系统研究[D]. 西安:西安电子科技大学, 2022.
ZHAO S H. Research on real-time 3D reconstruction system based on binocular stereo vision[D]. Xi'an:Xidian University, 2022. (in Chinese)

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

[7] SCHÖNBERGER J L, FRAHM J M. Structure-from-motion revisited[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, US: IEEE, 2016: 4104 - 4113.

[8] 朱红军,高潮,郭永彩. 基于计算机视觉的非朗伯表面三维重构[J]. 强激光与粒子束,2014,26(1):295 - 305.
ZHU H J, GAO C, GUO Y C. 3D reconstruction of non-Lambertian surfaces based on computer vision[J]. High Power Laser and Particle Beams,2014,26(1):295 - 305. (in Chinese)

[9] VORONIN V, FRANTC V, SEMENISHCHEV E, et al. 3D shape object reconstruction with non-Lambertian surface from multiple views based on deep learning[C]// Proceedings of 2022 SPIE The International Society for Optical Engineering. Orlando, FL, US: SPIE, 2022:296 - 303.

[10] 陈龙,张建林,彭昊,等. 多尺度注意力与领域自适应的小样本图像识别[J]. 光电工程,2023,50(4):67 - 80.
CHEN L, ZHANG J L, PENG H, et al. Multi-scale attention and domain adaptive small sample image recognition [J]. Opto-Electronic Engineering, 2023, 50(4):67 - 80. (in Chinese)

[11] 杜小强,李卓林,马锃宏,等. 基于空间注意力和可变形卷积的无人机田间障碍物检测[J]. 农业机械学报,2023,54(2):275 - 283.
DU X Q, LI Z L, MA Z H,et al. Unmanned aerial vehicle field obstacle detection based on spatial attention and deformable Convolution [J]. Transactions of the Chinese Society for Agricultural Machinery,2023,54(2):275 - 283. (in Chinese)

[12] YAO Y, LUO Z X, LI S W, et al. MVSNet: depth inference for unstructured multi-view stereo [C] // Proceedings of 2018 Springer Verlag European Conference on Computer Vision. Munich, Germany:Springer Verlag,2018:767 - 783.

[13] CHEN R, HAN S F, XU J, et al. Point-based multi-view stereo network [C] // Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019, :1538 - 1547.

[14] LUO K Y, GUAN T, JU L 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. Seoul, Korea (South): IEEE, 2019: 10451 -

10460.

[15] YANG J Y, MAO W, LIU M M, et al. Cost volume pyramid based depth inference for multiview stereo[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, US:IEEE, 2020:4876 - 4885.

[16] WANG F J H, 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. Nashville, TN, US:IEEE, 2021:14189 - 14198.

[17] TSOTSOS J K. A computational perspective on visual attention [M]. Cambridge, MA, US:MIT Press, 2021.

[18] LIU J J, HOU Q, CHENG M M, et al. Improving convolutional networks with self-calibrated convolutions [C] // Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, US:IEEE, 2020:10093 - 10102.

[19] BELLO I, ZOPH B, LE Q, et al. Attention augmented convolutional networks [C] // Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South):IEEE, 2019:3285 - 3294.

[20] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, US:IEEE, 2019:3141 - 3149.

[21] SHEN Z, NGUYEN C. Temporal 3D RetinaNet for fish detection [C] // Proceedings of 2020 IEEE Digital Image Computing: Techniques and Applications. Melbourne, Australia: IEEE, 2020:1 - 5.

[22] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, US: IEEE, 2021:13708 - 13717.

[23] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// Proceedings of IEEE Conference Computer Vision and Pattern Recognition. Las Vegas, NV, US: IEEE, 2016:4724 - 4732.

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

[25] JENSEN R, DAHL A, VOGIATZIS G, et al. Large scale multi-view stereopsis evaluation [C] // Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, US:IEEE, 2014:406 - 413.

[26] Discover a wide range of drone datasets senseFly[DS]. (2013-12-26)[2022-07-08]. https://www.sensefly.com/education/datasets/.

[27] CAMPBELL N, VOGIATZIS G, HERNÁNDEZ C, et al. Using multiple hypotheses to improve depth maps for multi-view stereo [M]. Berlin, Germany: Springer-Verlag, 2008: 766 - 779.

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

[29] ENGIN T, CHRISTOPH S, PASCAL F. Efficient large-scale multi-view stereo for ultra high-resolution image sets[J]. Machine Vision and Applications, 2011, 23(5):903 - 920.

[30] YAO Y, LUO Z X, LI S W, 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. Long Beach, CA, US:IEEE, 2019:5520 - 5529.

[31] YU Z H, GAO S H. Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and Gauss-Newton refinement[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, US:IEEE, 2020:1946 - 1955.

[32] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [C] // Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 618 - 626.