[文章编号] 1003－4684(2026)04-0038-06

融合多尺度特征增强和代价体融合的三维重建方法

王云艳[1,2]，朱镇中[1]，熊 超[1]

(1 湖北工业大学 电气与电子工程学院，湖北 武汉 430068；2 襄阳湖北工业大学产业研究院，湖北 襄阳 441100)

[摘 要] 为了提高多视图三维重建的精度，提出了一种基于多尺度交叉感知特征增强模块和多尺度代价体融合模块的重建方法。以 CasMVSNet 为基准网络，首先设计了一个交叉信息自注意力模块，利用纹理丰富的底层特征信息选择性地增强高层信息，将交叉信息自注意力模块嵌入 U-Net 网络中，采用高低层的语义特征融合方法，建立深层特征和浅层特征之间的依赖关系，提高特征提取的完整性，随后输出三种尺度不同大小的特征层；然后考虑到三种阶段的生成代价体相关性不足，设计一款代价体分层融合模块，将小尺度代价体中的信息分离并传递给下一层的代价体进行融合，由粗到精地进行深度估计，使重建精度和完整性都有大幅提高；最后考虑到粗阶段所生成的深度图中深度可能存在较大误差，会导致后续操作持续产生误差，因此在每一层后都添加了一个深度细化模块，使用参考图像作为指导来细化深度图，从而减少累积误差。由实验结果可知，该模型在 DTU 数据集上和基准网络 CasMVSNet 相比，在准确性误差和完整性误差上分别降低了 6.89％和 1.04％，相较于其他模型均有不同程度的提升，验证了该模型的有效性。

[关键词] 深度学习；交叉注意力；分层融合；深度细化；多视图立体匹配

[中图分类号] TN391.41 [文献标识码] A

多视图立体匹配是计算机视觉领域中一个重要的研究方向，其是将一个物体不同视角的多张已知相机位姿信息的图像转换成三维场景的过程。多视图立体匹配在自动驾驶[1]，医学图像处理[2]，虚拟现实[3]和机器人技术[4]等领域具有广泛的应用场景。因此，研究多视图立体匹配具有一定的现实意义。

传统的多视图立体匹配方法采用局部特征描述符，如 SIFT[5]、SURF 或 ORB[6]等来进行特征提取，利用人为定义的标准来确定两个视角像素的对应关系。虽然在某些场景下表现良好，但对纹理缺乏鲁棒性、遮挡等表面重建效果会非常差。

近年来出现了基于深度学习的多视图立体匹配方法，旨在解决传统方法中存在的问题，并提高三维重建的准确性和完整性。Ji 等人提出了基于体素的 SurfaceNet[7]，该方法在代价体中将体素的值用来表示该位置处的匹配代价，通过使用正则化来聚合代价体中的信息。MVSNet[8]首次提出了一种端到端的训练策略来预测深度图，但由于采用 3D 卷积来进行推导工作，对计算机的硬件要求很高，因此无法进行大型场景的重建任务。为了解决这一问题，R-MVSNet[9]采用了 GRU 以连续的方式规范成本量，减少了显存占用但增加了运行时间。CVP-MVSNet[10]和 Cas-MVSNet[11]等一些模型采用了一种由粗到精地策略，自适应地调整深度范围，获得了更多的图像细节，且对 GPU 较为友好，实现高效率的重建。Point-MVSNet[12]中采用由粗到精地两阶段方法生成高分辨率的深度图，但也存在特征通道之间相关性不足等问题，不利于网络完整的学习到更加丰富的信息。Fast-MVSNet[13]考虑到效率也是重建的一个重要的因素，构建一个稀密粗精框架，保证了方法的快速性。HR-MVSNet[14]采用了一种金字塔编解码结构的网络，由粗到精，降低内存需求。

虽然基于深度学习的重建工作已经取得了令人印象深刻的结果，但在特征提取的完整性和深度图估计的准确性方面仍然存在一定的提升空间。在特征提取阶段，由于经过多次卷积和池化，一些深层次特征可能会丢失一些细节信息，特别是在一些无纹理或者重复纹理区域容易出现空洞现象。本文提出了一种改进 U-Net 模型做为特征提取网络，嵌入交叉自注意力模块，使其较为完整的提取图像特征，输出 3 种不同尺度的特征图，采用级联结构由粗到精

[收稿日期] 2023－09－05

[第一作者] 王云艳(1981－)，女，湖北武汉人，湖北工业大学副教授，研究方向为遥感图片目标检测。

[通信作者] 朱镇中(1998－)，男，湖北黄冈人，湖北工业大学硕士研究生，研究方向为多视图三维重建。

的输入后续网络。代价体正则化阶段采用了一个代价体分层融合模块,将低尺度的代价体信息传递给高尺度的代价体,进一步的提高预测深度图的准确性。此外,在粗阶段预测的深度图中也可能存在误差,因为代价体不能保留不同视图之间相似特征的区别,这势必会影响到后续预测高分辨率深度图的精度。因此本文提出了一种深度图细化策略,用于提高粗阶段的深度估计的精度,解决累积误差问题。

1　MC-MVSNet 方法介绍

针对已知的多视图三维重建存在的问题,本文提出了一种基于多尺度交叉信息感知的多视图立体匹配模型 MC-MVSNet。在本节中,将详细介绍基于多尺度交叉信息感知的多视图立体匹配模型及其相应模块,首先设计了一个交叉自注意模块,利用纹理丰富的底层特征选择性地增强高层信息,将该模块嵌入到 U-Net 网络中,由 U-Net 产生三个不同尺度的特征进行输出,构成由粗到精的级联结构,然后与 MVSNet 工作类似,经过单应性变换形成特征体,再对特征体同一通道上各点求方差作为该点取值最终形成代价体。在代价体正则化模块中,利用代价体分层融合模块,将上层代价体中包含的信息向下层进行传输,用以提高最后重建的精度。粗阶段的深度图形成后,对其进行细化再传输到下一级进行重建工作,最终生成高精度的深度图。整体框架如图 1 所示。

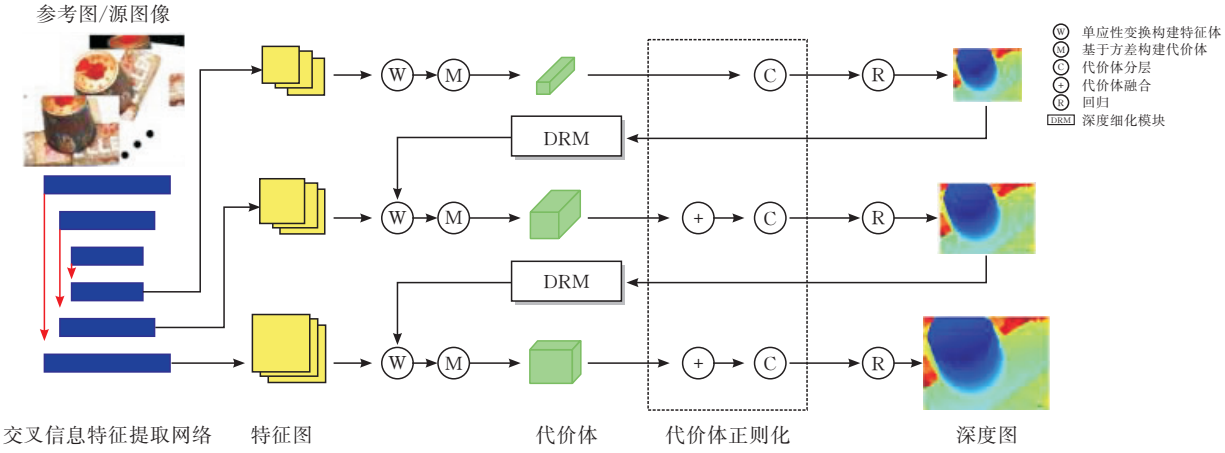

图 1　多尺度交叉信息感知的重建网络 MC-MVSNet

1.1　交叉信息融合特征网络

一般情况下,卷积神经网络中的合并运算和卷积运算会丢失大量底层视觉特征,使得底层结构的细节难以恢复,降低了重建的准确性和完整性,因此本文提出了一种交叉信息融合特征网络,如图 2 所示。现有的大多数方法都是直接使用特征金字塔网络进行特征提取,集成了来自高级和低级特征映射的信息,但是在不同层次之间存在语义鸿沟,直接融合会降低多尺度表达能力。选用 U-Net 为基准网络,U-Net 最早用于图像分割领域,主要特点是将卷积神经网络的编码器部分和解码器部分连接起来,使得网络可以同时进行高层次特征提取和像素级别的分类。在 U-Net 网络中嵌入一个 CSAM 模块,对高级特征的空间信息进行选择性增强,然后再通过 1×1 卷积降维后将它们连接起来。最后,采用逐层融合操作将浅层特征与深层特征连接起来。

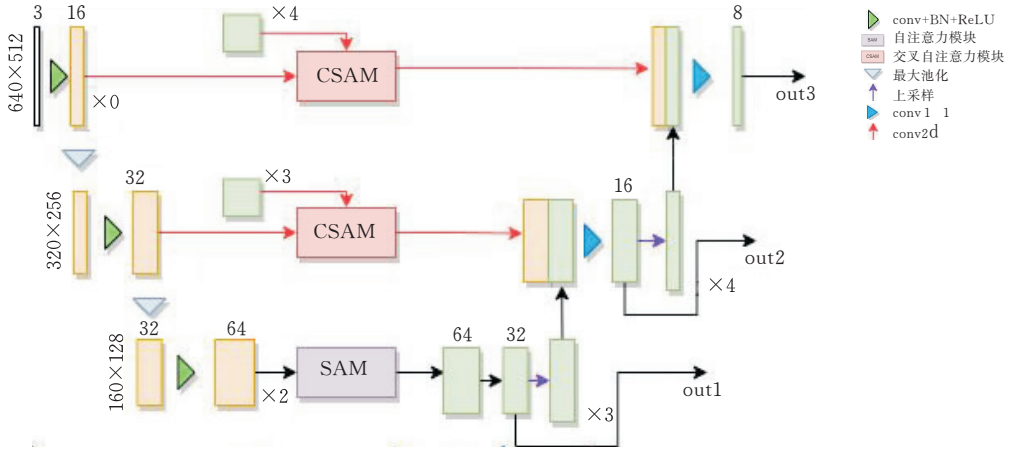

图 2　交叉信息融合特征提取网络

与之前的工作类似，输入 N 张图像，选其一张作为参考图像，N-1 张为源图像，所有输入图像都为 640x512 的 RGB 图像，首先经过卷积操作，将图像的通道变成 8 通道而后再变成 16 通道，再经过两次池化操作和相应卷积，分别生成三个尺寸为原图像 1，1/4，1/16 倍的特征图，底层通过一个 SAM 模块强化特征信息，经过 1×1 卷积进行降维生成一个 out1，随即进行上采样操作，与第二层经过 CSAM 模块的特征图进行拼接，然后生成 out2，重复上述操作，最终得到 3 组不同尺度的特征图，通道数分别为 8，16，32。交叉信息融合特征模块可以保留丰富的细节特征，使后续的深度估计可以更加准确。值得注意的是，在网络的每一层中，它首先通过 CSAM 增强空间信息，然后与低层特征拼接。

注意力机制是一个非常重要且有用的技术，可以帮助模型关注重要的信息，提高任务性能，Fu J 等[16]提出了 SA 模块，通过计算不同位置之间的相互关系来获取上下文信息，从而对注意力进行加权。在此基础上 Wang G[17]在三维点云上提出了类似的方法，受上述论文的启发，本文提出了交叉注意力模块，如图 3 所示。

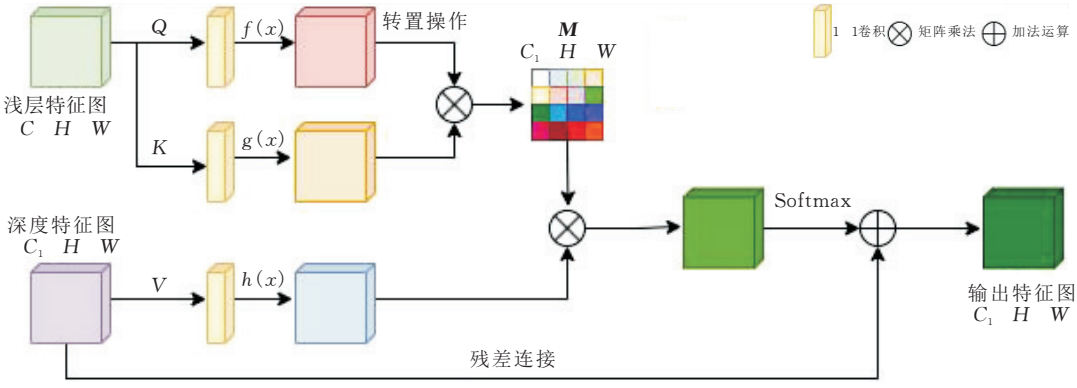

图 3 交叉自注意力 CSA 模块示意图

将浅层特征图 x 输入到 Q 和 K 分支中，并通过 1×1 卷积调整通道数，得到 $f(x)$ 和 $g(x)$，其中 $f(x)=W_f * x$，$g(x)=W_g * x$。这两个特征空间的值将相乘来获得 β，$\beta=(f(x_i)^T * g(x_i))$。将深层特征图映射到 $h(X)$ 与 β 相乘得到交叉注意力矩阵 $\boldsymbol{M}$，再通过 Softmax 函数计算每个像素的加权分数。最后深层特征进行残差连接。整个过程可表示为：

$$y = \mathrm{Softmax}(\beta * h(X)) + s(X) \tag{1}$$

其中，y 是增强空间信息的高层特征图；$s(X)$ 是残差分支的输出。

本文提出了 CAS 模块来提升特征提取部分，由于卷积神经网络会丢失大量底层视觉特征，该模块将底层特征和深层特征信息进行融合，利用纹理丰富的底层特征选择性地增强高层信息，嵌入到 U-Net 模块，将 CAS 模块增强过的特征与上采样的特征再进行逐级融合，尽可能的保留较多的信息。U-Net 中的侧边连接也能够更好的修复目标的细节，使后续的深度估计结果更加准确。

1.2 代价体的构建

在构建 3D 成本代价体时，通常会使用单应性变换将之前提取的特征图映射到每个参考视角上。这个过程类似于经典的平面扫描立体视觉，它会将所有的特征图扭曲到参考相机视锥体中的不同正交平面上。单应性矩阵定义为：

$$H_i(d) = K_i \cdot R_i \cdot \left(I - \frac{(t_0 - t_i) \cdot n_0^{\mathrm{T}}}{d}\right) \cdot R_0^{\mathrm{T}} \cdot K_0^{-1} \tag{2}$$

其中，$H_i(d)$ 是第 i 个特征图与深度 d 处参考特征图之间的单应性变换。此外，K_i、R_i、t_i 指的是相机的内参、外参和平移，n_0 表示参考相机的主轴。然后再将变换后的特征体上同一通道上各点求方差作为该点取值最终形成代价体，公式如下：

$$C = \frac{\sum_{i=1}^{N} (F_i - \bar{F}_i)^2}{N} \tag{3}$$

其中，$\bar{F}_i$ 是所有特征体素的平均体素，对所有的特征层执行此操作，以获得 3 个成本代价体，每个代价体对应不同的尺度。

1.3 代价体正则化模块

在多视图立体匹配任务中，代价体是一个非常关键的中间过程，但在初始代价体的构建中存在大量噪声和误匹配等问题，因此需要对代价体进行正则化。代价体正则化的目的是提高重建结果的精度和鲁棒性。同时，代价体正则化也有助于减少后续体积重建和细化阶段中的计算复杂度和存储空间需求，提高算法的效率和扩展性。MVSNet 中将生成的代价体通过一个类似 U-Net 结构的正则化网络，Yang 等[18]提出了一个分层特征体积解码器，将深

度特征分离与金字塔下一阶段的特征体进行融合，有效的融合了不同层次的细节的特征，基于上述思想，本文也将采取此方法进行代价体正则化操作。如图 4 所示。

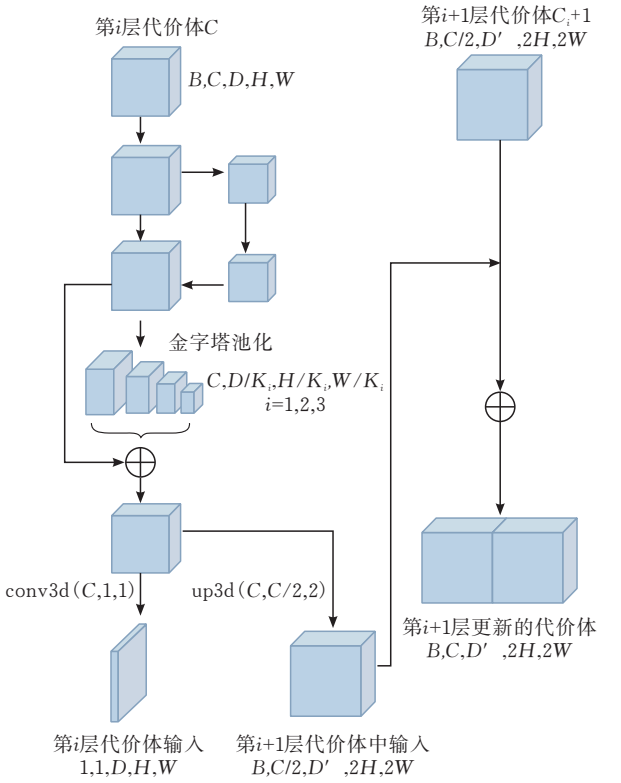

图 4　多尺度代价体融合正则化模块

由于级联网络中会生成三种不同尺度的代价体，除了特征提取模块中的上采样操作，几乎就毫无关联，为了在保持粗尺度信息的同时有效提取不同层次细节的信息，采用了一个多尺度特征融合模块，将上一层代价体进行分层操作传递给下一个层的代价体，使粗精度的代价体信息可以融合到下一层的代价体中。首先将输入代价体经过一个类似两层的金字塔结构，然后进行金字塔池化，捕获足够的上下文信息，采用残差连接，随后进行分层，一部分作为粗阶段的输出，另一部分进行上采样操作输入到下一阶段代价体进行拼接操作，重复上述步骤，得到最终的高精度深度图。由此，将小尺度的代价体信息融合进下一层的代价体中，这样能够增加多尺度特征体之间的信息联系，提高深度图的估计质量。

1.4　深度细化模块

在粗阶段中，由于低分辨率图像不能很好捕捉相邻视图的一些细节特征，因此构成的代价体不一定可靠，会导致预测的深度图不准确，而且这个误差会一直累积到最后，因此提升粗阶段深度图的精度具有很重要的意义。而且正则化过程中涉及的接受场很大，重构边界可能会出现过度平滑的问题。在 MVSNet 网络中，在深度图中应用了深度残差网络来细化深度图，本文借鉴其思想。和 CasMVSNet 不同的是，本文在三个阶段中都应用深度细化模块，使用金字塔池化来提取尽可能全面的信息，将同尺寸参考图作为引导来细化当前阶段的深度图，从而获得精细化的深度图。深度细化模块如图 5 所示。

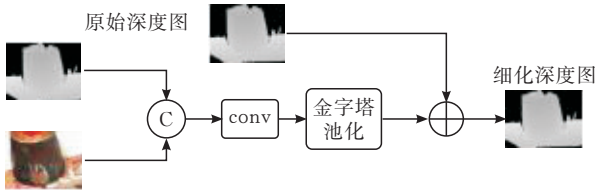

图 5　深度细化模块

1.5　损失函数

本文构建了三层级联结构，关注每一级的损失，总损失定义如下：

$$\text{Loss}=\sum_{k}^{N}\lambda^{k}\times L^{k};(k=1,2,3) \tag{4}$$

在上式中 L^k 是指第 k 阶段的损失；λ^k 是指其对应的损失权重，与 CasMVSNet 相同，损失权重分别设置为 0.5，1 和 2。

2　实验及结果分析

2.1　数据集

本文在 DTU 数据集上进行了实验。DTU 数据集包含 124 个场景，每个场景都包含 49 张图像和相应的深度图，图像分辨率为 1600×1200，深度图分辨率为 800×600。这些图像是在不同的角度和距离下拍摄的，涵盖了各种不同的场景，例如室内、室外、自然和人工物体等。DTU 数据集中的深度图是由激光扫描仪扫描获得的，因此具有高精度和高质量。每个深度图都包含了每个像素点到相机的距离信息，可以用于三维重建和立体匹配等应用。

2.2　实验设置

本文使用 pytorch 框架编写模型，训练数据是由 DTU 划分的训练集，输入图像个数为 3 张，图像分辨率为 640×512，网络采用由粗到精三个阶段，每个阶段的深度假设层为 48，32，8。三个阶段的特征层分辨率分别为原始输入大小的 1/16，1/4，1。遵循与 MVSNet 相同的输入视图选择策略。使用 Adam 优化器。其中 β_1 为 0.9，β_2 为 0.999。初始学习率设置为 0.001，在 10、12 和 14 个周期之后，学习率减半。共训练 16 个周期。

2.3　实验结果

为了验证模型的准确性，通过 DTU 数据集提供的 Matlab 代码来计算重建模型的准确性误差、完整性误差和整体性误差，其中后者是前面二者的总和求平均，这三个指标越低表示重建效果越好。将

本文方法与传统方法和其他基于学习的方法进行比较，点云重建结果如表 1 所示。

表 1 DTU 数据集上的定量评估结果 mm

Methods	Acc	Comp	Overall
Camp	0.8352	0.5544	0.6948
Furu	0.6135	0.9415	0.7775
Tola	0.3425	1.1907	0.7666
Gipuma	0.2837	0.8737	0.5787
Colmap	0.4005	0.6645	0.5325
MVSNet	0.3968	0.5278	0.4623
R-MVSNet	0.3894	0.4358	0.4126
FastMVSNet	0.3365	0.4132	0.3748
CVPMVSNet	0.3105	0.4065	0.3585
EPPMVSNet	0.4135	0.3058	0.3846
CasMVSNet	0.3804	0.3659	0.3731
Vis-MVSNet	0.3596	0.3847	0.3721
MC-MVSNet	0.3115	0.3555	0.3335

表中显示虽然 Gipuma 在准确性方面取得了最佳，但是在完整性方面很差，直观效果并不好，在整体误差上，本文模型取得了最佳，准确性误差和完整性误差均取得了不错的数据，和基准网络 CasMVSNet 相比，在准确性误差和完整性误差上，降低了 6.89%和 1.04%，由此可以证明改进的模型比原模型性能更佳，本文改进的模型是具有有效性的。选取了 scan1 和 scan75 两个场景重建后的效果图与其他网络进行对比，效果图如图 6 所示。

本文相较于 Casmvsnet 网络，消除了很多噪声，与其他网络相比，细节表现的更好，重建的完整也得到了保证。

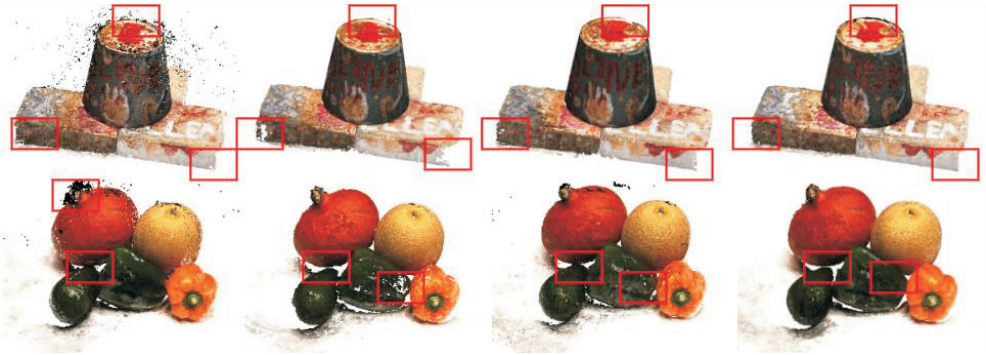

图 6 scan1 和 scan75 可视化对比

2.4 消融实验

针对本文提出的三个改进方法，本文采用消融实验来验证每个模块的有效性，测试结果如表 2 所示，在 DTU 数据集上进行了 5 组对比实验，因本文以 CasMVSNet 模型为基准网络，所以对比实验里包括 CasMVSNet，交叉信息融合提取（CIF）+CasMVSNet，多尺度代价体融合模块（MCB）+CasMVSNet，深度细化模块（DR）+CasMVSNet 以及最终版模型，并使用准确性和完整性误差衡量重建质量。

表 2 消融评价指标的比较 mm

Methods	ACC	Comp	Overall
CasMVSNet	0.3804	0.3659	0.3731
CasMVSNet+CIF	0.3415	0.3589	0.3502
CasMVSNet+MCB	0.3545	0.3589	0.3567
CasMVSNet+DR	0.3785	0.3611	0.3698
McMVSNet	0.3115	0.3555	0.3335

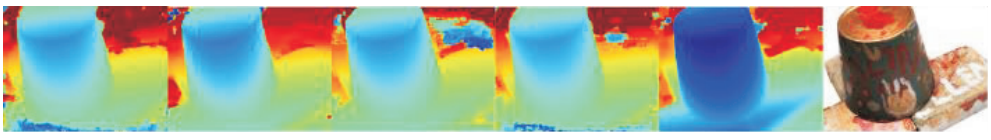

图 7 消融深度估计图的比较

实验表明该模块是有效的，由表 2 可以看出，改进的三处相对于原始网络均有提升，证明了三处改进的有效性。为了直观的观察效果，消融深度估计图比较如图 7 所示。基准网络噪声比较大，直观效果并不好，本文方法在消除噪声上起到了效果，在物体表明轮廓上更突出，如图中杯身表现的更加平滑，直观的证明本文设计模块的有效性。

3 结束及展望

本文提出了一种基于多尺度交叉融合特征网络

和多尺度代价体分层融合模块,通过改进 CasMVS-Net 的特征提取模块和代价体正则化模块。采用了 CSA 模块来加强 U-Net 的提取性能,利用纹理丰富的底层特征选择性地增强高层信息。同时考虑到各层代价体之间的相关性不足,设计了分层融合模块来提高深度图的估计质量,考虑粗阶段产生的深度图不准确,会带来累积误差,在每一层都添加了一个深度细化模块,进一步提高最终深度图的精度。本文方法与其他基于学习的方法在 DTU 上取得比较好的效果,验证本文提出模型的有效性,但是本文专注于提高模型重建精度,忽略了内存消耗,如何在保持相同精度的同时有效减少内存消耗和运营时间成为后续工作的主要目标。

[参 考 文 献]

[1] Lee H, Song S, Jo S. 3D reconstruction using a sparse laser scanner and a single camera for outdoor autonomous vehicle[C]// 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2016: 629-634.

[2] Joseph S S, Aju D. A Comparative Survey on Three-Dimensional Reconstruction of Medical Modalities Based on Various Approaches[C]// Information Systems Design and Intelligent Applications: Proceedings of Fifth International Conference INDIA 2018 Volume 1. Springer Singapore, 2019: 223-233.

[3] Sing K H, Xie W. Garden: A mixed reality experience combining virtual reality and 3D reconstruction[C]// Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2016: 180-183.

[4] Nuchter A, Surmann H, Hertzberg J. Automatic model refinement for 3D reconstruction with mobile robots[C]// Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings. IEEE, 2003: 394-401.

[5] Peng K, Chen X, Zhou D, et al. 3D reconstruction based on SIFT and Harris feature points[C]// 2009 IEEE international conference on robotics and biomimetics (ROBIO). IEEE, 2009: 960-964.

[6] Tafti A P, Baghaie A, Kirkpatrick A B, et al. A comparative study on the application of SIFT, SURF, BRIEF and ORB for 3D surface reconstruction of electron microscopy images[J]. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2018, 6(1): 17-30.

[7] 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.

[8] 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.

[9] 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. 2019: 5525-5534.

[10] 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. 2020: 4877-4886.

[11] 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.

[12] 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.

[13] 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.

[14] Weilharter R, Fraundorfer F. Highres-mvsnet: A fast multi-view stereo network for dense 3d reconstruction from high-resolution images[J]. IEEE Access, 2021, 9: 11306-11315.

[15] Zhang J, Li S, Luo Z, et al. Vis-mvsnet: Visibility-aware multi-view stereo network[J]. International Journal of Computer Vision, 2023, 131(1): 199-214.

[16] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.

[17] Wang G, Zhai Q, Liu H. Cross self-attention network for 3D point cloud[J]. Knowledge-Based Systems, 2022, 247: 108769.

[18] Yang G, Manela J, Happold M, et al. Hierarchical deep stereo matching on high-resolution images[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5515-5524.

(下转第 69 页)

Design and Optimization of Small Solid State Shaped Material Picking Device

SUN Chao, TAO Qi, ZHONG Fei, CHENG Kai

(*School of Mechanical Engineering, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: Aiming at the problems of difficult grabbing and low feeding speed of small-sized special-shaped solid ingredients in the molding process of cigarette composite filter rods, a material picking device adapted to small-sized special-shaped solid ingredients is designed based on the Beckhoff XTS flexible conveying system. Transient dynamic simulation analysis of the device is carried out by using ANSYS Workbench, and the time-varying cloud diagrams and curves of equivalent strain and equivalent stress, as well as the velocity and acceleration curves of the top block and clamping jaw are obtained. The simulation results show that the maximum equivalent stress of the picking device is 39.755 MPa, which is lower than the allowable stress of the material and meets the design strength requirements. The maximum velocities of the top block and clamping jaw are 79.074 mm/s and 155.110 mm/s respectively, indicating smooth and stable motion performance. To further verify the practical application performance, ingredient interval deviation experiments are conducted. After parameter calibration, the ingredient interval deviation rate is less than 5% when the ingredient release height is set to 8.9 mm, which satisfies the technical requirements for ingredient feeding in composite filter rod molding.

Keywords: composite filter rod molding; shaped dosing; flexible conveying system; pickup device; transient dynamics

[责任编校: 张　众]

(上接第 43 页)

Multi-view Stereo Network Reconstruction with Multi-scale Cross-Information Perception

WANG Yunyan[1,2], ZHU Zhenzhong[1], XIONG Chao[1]

(1.*College of Electrical and Electronic Engin., Hubei Univ. of Tech., Wuhan* 430068, *China*;
2. *Xiangyang Industrial Institute of Hubei Univ. of Tech., Xiangyang* 441100, *China*)

Abstract: To improve the accuracy of multi-view 3D reconstruction, a reconstruction method based on multi-scale cross perception feature enhancement module and multi-scale cost volume fusion module is proposed. The method takes CasMVSNet as the baseline network. Firstly, a cross information self attention module is designed to selectively enhance high-level information by incorporating it into the U et network and leveraging rich texture features from lower layers to improve the integrity of feature extraction. Three feature layers of different scales are then outputted to better capture multi-scale information. Secondly, considering the insufficient correlation of generated cost volumes in three stages, a cost volume hierarchical fusion module is designed. It separates and transfers information from small-scale cost volumes to the next layer for fusion, enabling depth estimation from coarse to fine, significantly improving reconstruction accuracy and completeness. Lastly, to address the potential large errors in depth maps generated by the coarse stage that could accumulate errors in subsequent operations, a depth refinement module is added after each layer. It refines the depth maps using reference images as guidance to reduce cumulative errors. Experimental results show that compared to the baseline network CasMVSNet and other models, the proposed method reduces accuracy error and completeness error by 6.89% and 1.04% respectively on the DTU dataset. This demonstrates the effectiveness of the approach, providing a strong reference for further research in the field of multi-view 3D reconstruction.

Keywords: deep learning; cross-attention; hierarchical fusion; depth refinement; multi-view stereo matching

[责任编校: 赵　瞳]