DOI:10.19651/j.cnki.emt.2517835

基于选择性融合上下文信息的立体匹配网络*

宁安琪[1] 於跃成[1] 杨 帆[1] 李 响[2]
(1.江苏科技大学计算机学院 镇江 212000;2.中国移动通信集团有限公司网络事业部 北京 100032)

摘 要:目前基于深度学习的立体匹配网络虽然具备较高的精度,但是网络中复杂的结构导致计算时间的急剧增加。为了平衡网络的匹配速度与精确度,本文提出了基于选择性融合上下文信息的立体匹配网络。首先,通过相关层方法构建成本体,进而在聚合模块中采用单编码器-解码器结构,以降低模型复杂度。其次,在编码器中融合多尺度成本体,以捕捉不同层级的视差信息;同时,在解码器中设计选择性融合上下文信息模块,利用参考图像的上下文特征引导几何信息的高质量解码。最后,设计多分支聚合金字塔池化模块,增强编码-解码模块理解全局语境的能力。实验结果表明,本文算法在 KITTI2015 数据集上全部区域的误匹配率为 1.97%,在 KITTI2012 数据集上的三像素误差为 1.50%。与其他算法相比,在满足算法实时性要求的同时,实现了更精准的立体匹配精度。

关键词:立体匹配;注意力模块;多尺度融合;视差图;成本聚合
中图分类号:TP391;TN911 **文献标识码**:A **国家标准学科分类代码**:520.6040

Stereo matching network based on fusing contextual information selectively

Ning Anqi[1] Yu Yuecheng[1] Yang Fan[1] Li Xiang[2]
(1. School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212000, China;
2. Department of Network China Mobile Communications Corporation, Beijing 100032, China)

Abstract: At present, although the stereo matching network based on deep learning has high accuracy, the complex model structure in the network leads to a sharp increase in computing time. In order to balance the matching speed and accuracy of the network, this paper proposes a stereo matching network based on fusing contextual information selectively. First, the cost volume is constructed through the correlation layer method, and then the single encoder decoder structure is used in the aggregation module to reduce the complexity of the model. Secondly, multi-scale cost bodies are fused in the encoder to capture different levels of parallax information; a selective context information fusion module is designed in the decoder, which uses the context features of the reference image to guide the generation of high-quality geometric information. Thirdly, multi-scale cost volume is fused in the encoder to capture different levels of parallax information; at the same time, fusing contextual information selectively module is designed in the decoder, which uses the context features of the reference image to guide the high-quality decoding of geometric information. Finally, the multi branch aggregation pyramid pooling module is designed to enhance the ability of the encoding-decoding module to understand the global context. The experimental results show that the mismatch rate of all regions on the KITTI2015 dataset is 1.97%, and the three pixel error on the KITTI2012 dataset is 1.50%. Compared with other algorithms, our algorithm achieves more accurate stereo matching accuracy while meeting the real-time requirements.

Keywords: stereo matching; attention module; multi-scale fusion; disparity map; cost aggregation

0 引 言

双目视觉通过计算物体在不同视点下投影位置之间的差异来估计视差,是机器人导航、自动驾驶和三维重建等计算机视觉领域的一项关键任务[1]。立体匹配作为双目视觉的核心技术,其核心是从一对校正后的立体图像中生成视差图,以用于深度信息的估计[2-4]。传统的立体匹配算法依靠手工设计的特征或优化函数来估计视差,但是算法在不

收稿日期:2025-01-07
* 基金项目:国家重点研发计划项目(2023YFC2809700)资助

适应区域的表现效果并不理想。随着深度学习的发展，卷积神经网络被用于替换传统立体匹配算法的部分模块。Žbontar 等[5]设计了快速匹配的 MC-CNN-Fast 和精确匹配的 MC-CNN-Acc 两种网络结构，利用卷积神经网络学习图像块的相似性来计算匹配成本。Park 等[6]利用金字塔池化来扩大网络的感受野，通过引入更广泛的上下文信息来减少局部匹配的错误。这类方法在效果上虽有有一定程度的提升，但是算法仍然依赖于传统框架，并未能充分利用深度学习的优势。目前，这类传统的深度学习算法已被端到端的立体匹配算法所替代。

基于深度学习的端到端立体匹配算法利用左右图像的特征计算匹配成本，对构建出的成本体进行聚合操作，最后回归出视差图。根据成本体构建维度的不同，端到端立体匹配算法主要分为两类。第一类是基于相关层的 3D 成本体方法。Mayer 等[7]首次在立体匹配任务上使用相关层方法，将输入图像转换成特征向量，以余弦相似值刻画左右图之间的相关性，进而通过相关性计算构建 3D 成本体，最后完成端到端的视差预测。Xu 等[8]设计了一种自适应聚合的高效立体匹配网络（adaptive aggregation network for efficient stereo matching，AANet），将可变形卷积引入立体匹配网络。使用自适应的同尺度和跨尺度聚合模块来实现高效的成本聚合，减少计算时间的同时保持了较高的准确率。宋昊等[9]在 AANet 的基础上进行改进，针对在 AANet 中所提取的特征可能存在信息不足和冗余等问题，设计了更适应特征匹配的模块。这类方法的网络模型结构简单，运行速度快，仅需有限的计算资源便可获得较高的匹配精度。然而，由于 3D 成本体中仅保存了左右图的相关性信息，降低了网络捕捉复杂特征关系的能力，进而影响算法精度的进一步提升。Zhang 等[10]通过分组融合来处理不同的特征信息，应用从粗到细的策略来准确、高效地聚合多通道成本体。Song 等[11]提出的 EdgeStereo 将边缘信息整合到视差主干网络中，训练一个子网络来预测边缘图，利用边缘损失感知促进视差网络和边缘检测网络的互相监督，在薄结构和边缘区域表现较好，但模型的推理时间较长。第二类是基于级联或分组相关的 4D 成本体方法。Chang 等[12]提出了金字塔立体匹配网络（pyramid stereo matching network，PSMNet），将提取到的左右特征根据视差变化范围级联在一起，进而构建 4D 成本体，设计了堆叠的编码器-解码器网络对成本体进行正则化。Guo 等[13]在 PSMNet 基础上提出了分组相关立体网络（group-wise correlation stereo network，GWCNet），将左右特征图按照通道划分为若干组，设计了分组相关的方法构建成本体，减少了网络参数的同时提高了预测精度。相比于 3D 成本体方法，4D 成本体方法在成本体构建阶段为后续的成本聚合提供了更多可用信息，但成本聚合阶段的深层次堆叠操作需要耗费大量的内存和计算资源。此外，Xu 等[14]利用几何和上下文信息构建了一个组合的迭代几何编码体立体网络（iterative geometry encoding volume for stereo matching，IGEV-Stereo），并使用 ConvGRU 迭代更新视差图，实现了较高的精度，但是难以满足实时性的要求。

为此，端到端立体匹配网络的轻量化设计成为当前的主流工作之一。显然，构建低分辨率的成本体可以有效减少成本聚合阶段 3D 卷积所带来的参数增长。Khamis 等[15]提出了快速的立体匹配网络，首先在低分辨率的成本体上获取初始视差，然后通过迭代优化的方式逐步提升视差图的精度。类似地，Xu 等[16]提出了双边网格学习的立体匹配网络（bilateral grid learning for stereo matching networks，BGNet），通过在低分辨率的成本体上进行聚合操作，同时，借助于可学习的双边网络上采样模块，以获得高质量的高分辨率成本体，虽然 BGNet 实现了匹配精度的提升，但对复杂场景的效果并不理想。由于卷积操作是立体匹配网络中的主要操作，为此，Shamsafar 等[17]采用可分离卷积替换传统卷积设计了 MobileStereoNet，有效降低了堆叠编码器-解码器中 3D 卷积的计算量。Xue 等[18]提出轻量级的多尺度 2D 和 3D 模块分别用于特征提取和视差计算，最后利用多尺度的 RGB 图像来细化视差，但网络在细小结构上难以进行准确估计。

近年来，同时关注立体匹配网络轻量化和匹配精度的工作不断出现。事实上，立体匹配网络中特征提取部分的信息有助于网络对特征图中关键信息的捕捉。Bangunharcana 等[19]提出了通过成本体激励的实时匹配网络（correlate-and-excite：Real-time stereo matching via guided cost volume excitation，CoEx），在成本聚合阶段仅采用单个 3D 编码器-解码器结构来降低模型复杂度。同时，依据图像特征计算得到的注意力权重，CoEx 激励成本体的特征通道，改善了成本聚合的效果。Xu 等[20]通过构建级联和组相关串联的成本体，并以特征提取部分的通道图生成注意力权重，设计了 Fast-ACVNet 抑制串联成本体中的冗余信息，增强了相关匹配信息的有效表示。此外，Guo 等[21]提出的 LightStereo 专注于优化 3D 成本体的通道维度，提出多尺度卷积注意力模块，利用左侧图像提取的特征来增强成本聚合，提升了立体匹配的性能。

受上述工作的启发，本文同样借助特征提取部分的上下文信息，通过改进 4D 成本体方法中的成本体构建和成本聚合模块，设计了一种轻量高效的立体匹配网络（fusing contextual information selectively stereo matching network，FCS-Stereo）。首先，利用相关层方法构建 3D 成本体，然后利用通道扩张操作来生成更低复杂度的 4D 成本体。其次，在成本聚合阶段的编码器中设计多尺度组相关成本融合模块（multi-scale group-wise correlation volume fusion，MGCVF），以利用不同尺度的成本体补充初始成本体中所缺乏的的高层次信息。同时，在解码器中设计选择性融合上下文信息模块（fusing contextual information selectively，FCS），以利用特征提取阶段中的上

下文特征来指导解码器中几何特征的上采样操作,从而实现更高质量几何信息的还原。最后,在编码-解码模块中设计多分支聚合金字塔池化模块(multi-branch aggregation pyramid pooling module,MAPPM),以增强整体网络的特征表达能力。

1 网络结构设计

FCS-Stereo 网络的总体架构如图 1 所示。总体上,FCS-Stereo 由特征提取、成本体构建、成本聚合和视差回归 4 个模块组成。在特征提取阶段,FCS-Stereo 包含了下采样和上采样操作,其中下采样采用了预训练的 MobileNetV2[22] 网络。在成本体构建阶段,FCS-Stereo 构建了两类成本体。一个是先利用相关性计算,然后对特征层进行通道扩张生成的相关层成本体,另一个是利用分组相关计算得到的不同尺度的组相关成本体。由于 FCS-Stereo 网络在成本聚合阶段仅采用了单编码器-解码器结构,首先,在编码过程中不断融入不同尺度的低分辨率成本体。其次,在解码过程中引入选择性融合上下文信息模块,对网络逐层解码,实现高质量的上采样。最后,在编码-解码网络中引入多分支聚合金字塔池化模块,完成整个聚合网络的 3D 正则化。在视差回归阶段中,本文借鉴 CoEx 中提出的视差计算方法,然后利用每个像素周围超像素的加权平均值进行上采样,实现了全分辨率视差图的回归。

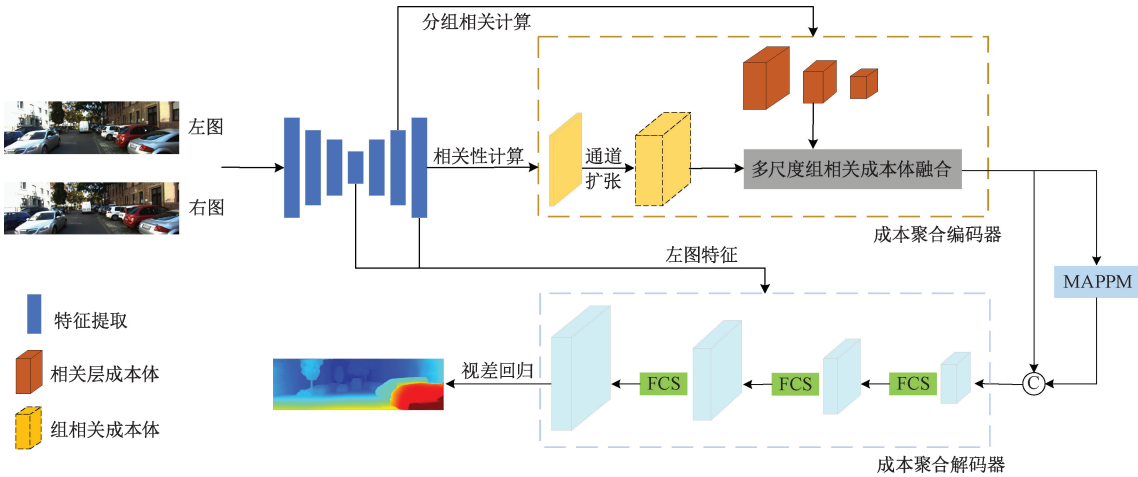


图 1 FCS-Stereo 网络

Fig. 1 FCS-Stereo network architecture

1.1 特征提取

很多模型[12-13,16]在特征提取模块中采用了类似于 ResNet 的结构,使得网络能够提取出丰富的特征,但这往往需要耗费更多的计算资源。相比较而言,MobileNet 不但具有更小的网络体积,还兼具了较好的特征提取能力。为此,为进一步加快模型收敛的速度,本文采用在 imagNet 上预训练后的 MobileNetV2[22] 作为特征提取模块中的下采样骨干网络。进而,采用 2D 转置卷积构建上采样模块,转置卷积核的大小为 4×4,步长为 2。同时,在上采样的过程中,利用 3×3 的 2D 卷积连接特征提取网络中每个尺度上的特征,如式(1)所示。

$$\boldsymbol{F}_i=\begin{cases}\boldsymbol{M}_i, i=\dfrac{1}{32}\\ \boldsymbol{Conv}_{3\times3}(\boldsymbol{Conv}^{\mathrm{T}}_{4\times4}(\boldsymbol{F}_{i/2})\parallel\boldsymbol{M}_i), \text{其他}\end{cases}\tag{1}$$

其中,$i\in\left\{\dfrac{1}{32},\dfrac{1}{16},\dfrac{1}{8},\dfrac{1}{4}\right\}$,$\boldsymbol{F}_i$ 代表第 i 尺度的特征图,$\boldsymbol{M}_i$ 表示 MobileNetV2 骨干网络中获得的第 i 尺度的特征图,$\parallel$ 代表串联操作。经过上述操作,最终得到大小为 $C\times H/4\times W/4$ 的左右特征图,其中 $C=48$。如图 1 所示,特征提取部分获得的多尺度特征 $\boldsymbol{F}_i$,可以后续阶段构建不同的成本体。

1.2 成本体构建

基于 4D 方法的模型利用级联或分组相关的方法构建成本体来提升匹配性能,但也引入了较多的特征通道,显著增加了成本聚合阶段的计算负担。而基于 3D 的相关层仅生成单个特征通道,因此本文构建了多通道的相关层成本体 $\boldsymbol{V}_{correx}$。首先,通过计算 1/4 分辨率处左右特征图的余弦相似度,构建 $D/4\times H/4\times W/4$ 大小的相关层成本体,D 是网络设置的最大视差,值为 192。$\boldsymbol{V}_{corr}(d,x,y)$ 的相关性计算公式如式(2)所示,其中,d 为 (0,48) 之间的视差值,(x,y) 为图像像素的坐标,$\boldsymbol{F}_L$ 和 $\boldsymbol{F}_R$ 是左右特征图,$<,>$ 代表内积操作。

$$\boldsymbol{V}_{corr}(d,x,y)=\frac{\langle\boldsymbol{F}_L(x,y),\boldsymbol{F}_R(x-d,y)\rangle}{\parallel\boldsymbol{F}_L(x,y)\parallel_2\cdot\parallel\boldsymbol{F}_L(x,y)\parallel_2}\tag{2}$$

由于 FCS-Stereo 网络在计算成本体时只采用了单个通道来表示视差信息,左右视图的相关性很容易受到噪声的影响。因此,本文对相关层成本体中的通道采取扩张操作。扩张操作时,采用 3×3 卷积将成本体的通道增加到 8,然后添加批归一化和 Mish 激活函数操作。

同时,借助 GWCNet[13] 提出的方法,本文构建了不同

尺度的组相关成本体 $\boldsymbol{V}_{\mathrm{gwc}}^{i}$。利用从特征提取模块获取的多尺度特征 $\boldsymbol{F}_i$，在大小为 $H/8\times W/8$，$H/16\times W/16$，$H/32\times W/32$ 的左右特征图上分别构建成本体，$\boldsymbol{V}_{\mathrm{gwc}}^{i}(d,x,y,g)$ 的分组相关计算的如式(3)所示。

$$\boldsymbol{V}_{gwc}^{i}(d,x,y,g)=\frac{1}{N_c^i/N_g}\cdot\frac{\langle \boldsymbol{F}_L(x,y),\boldsymbol{F}_R(x-d,y)\rangle}{\|\boldsymbol{F}_L(x,y)\|_2\cdot\|\boldsymbol{F}_L(x,y)\|_2} \tag{3}$$

其中，N_c^i 代表不同尺度的特征层的通道数，g 代表分组的编号，N_g 代表分组数，将 N_g 设置为 8。

1.3　成本聚合

在基于 4D 成本体的立体匹配网络中[12-13]，需要 3 个编码器-解码器结构来实现成本聚合，这导致模型的参数量过多，增加了计算复杂度。因此，本文在成本聚合阶段只利用单个 3D 编码器-解码器完成成本聚合，以达到轻量化的要求。同时，在聚合阶段设计 MGCVF、FCS、MAPPM 等模块来提升匹配精度，下面将详细介绍这 3 个模块。

1）多尺度组相关成本体融合

最近，一些方法证明了多尺度的成本体包含更多的轮廓和区域等高层次信息，在网络中融入这些成本体，可以提升匹配准确率。例如，Shen 等[23]利用提出不同尺度的成本体生成初始成本体，引导网络去关注不同尺度的图像区域。此后，Shen 等[24]又提出了 PCWNet（pyramid combination and warping costvolume for stereo matching），在编码过程中逐步引入多尺度的成本体，辅助聚合网络生成精确的视差图。

本文在融合过程参考 PCWNet[24]的思路，在编码器中设计了 MGCVF 模块。PCWNet 在编码过程中，添加级联和组相关融合的多尺度成本体，每个尺度的成本体包含了太多的信息，需要依靠大量的 3D 卷积来完成正则化。而本文模型是轻量化结构，因此，FCS-Stereo 只使用组相关成本体，来构建简单有效的融合网络。在成本体构建阶段，得到了通道数为 8，分辨率大小为原来分辨率 1/8、1/16 和 1/32 的 $\boldsymbol{V}_{gwc}^{1}$、$\boldsymbol{V}_{gwc}^{2}$、$\boldsymbol{V}_{gwc}^{3}$ 等成本体，融合低分辨率成本体来补充初始成本体丢失的信息。具体的融合步骤如图 2 所示，网络的输入是 1 个经过通道扩张的相关层成本体 $\boldsymbol{V}_{correx}$ 和 3 个低分辨率的组相关成本体 $\boldsymbol{V}_{\mathrm{gwc}}^{i}$。

多尺度成本体融合模块：融合模块主要有两个输入，一个是从高分辨率成本体传递下来的信息，另一个是从特征提取部分获取的各个尺度的组相关成本体。通过引入不同尺度下左右特征的分组相关性，帮助网络理解轮廓和区域信息。具体的每个融合模块 $\boldsymbol{MF}_i$ 如式(4)所示。

$$\boldsymbol{MF}_i=\boldsymbol{Conv}(\boldsymbol{E}_i\,\|\,\boldsymbol{V}_{gwc}^{i}) \tag{4}$$

其中，$\|$ 代表串联操作，$\boldsymbol{Conv}$ 是卷积核大小为 3×3×3 的 3D 卷积操作，$i\in\{1,2,3\}$ 表示不同尺度。E_i 代表对上一个成本体进行降采样的编码块，该编码块使用了两个卷积核大小为 3×3×3 的 3D 卷积，其中第一个 3D 卷积步长为 2。

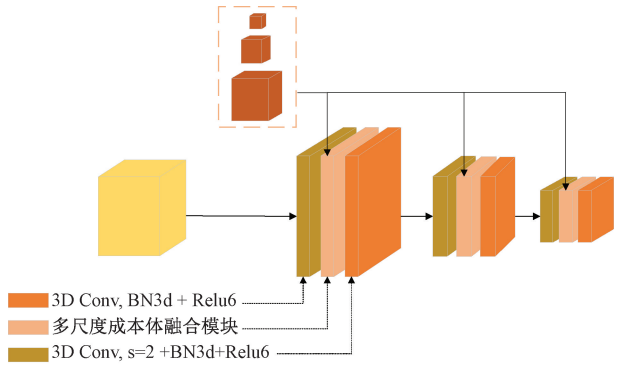


图 2　多尺度成本体融合步骤

Fig. 2　Multi-scale cost volume fusion steps

2）选择性融合上下文信息模块

在成本聚合的编码阶段，本文引入了不同尺度的成本体，覆盖了多尺度特征，此时的网络已经包含了充分的信息。在解码器部分，为了实现高质量的上采样，设计了选择性融合上下文信息模块。

葛兰等[25]利用特征提取部分的浅层特征指导初始视差图进行优化。CoEx 使用特征提取网络的图像特征图作为每个像素的权重，激发成本体相关的几何特征，从而带来更好的性能。因此，本文同样借用参考图像的特征图，利用其丰富的关键性细节信息，设计了 FCS 模块，帮助网络从低分辨率成本体中解码出精确的高分率成本体。如图 3 所示，FCS 模块上分支的输入是高分辨率的左图图像特征，下分支的输入低分辨率的成本体，通过特征图的辅助，最后输出上采样后的高分辨率成本体。

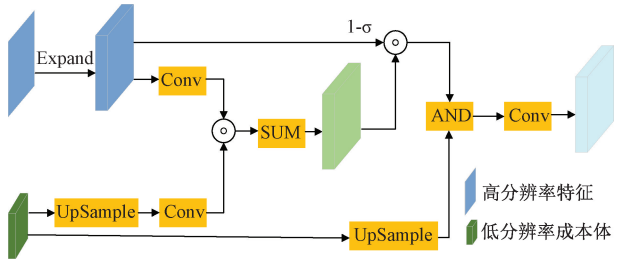


图 3　FCS 模块

Fig. 3　FCS module

对于给定地高分辨率的特征 C，先将其在视差维度扩展，最后形成 $B\times C\times D\times H\times W$ 大小的成本体 $\boldsymbol{C}_{\mathrm{ex}}^{i+1}$，其中 D 为 1。对于从低分辨率成本体传递过来的几何信息 $\boldsymbol{G}^i$，先进行上采样操作，得到高分辨率的成本体。然后通过上下文特征和几何特征的相关关系来生成注意力权重，从而自适应地融合来自图像特征图的有用信息，完成高质量的解码。受到自注意力机制[26]的启发，注意力权重的输出可以表示为：

$$\boldsymbol{\sigma}=Sigmoid\left(f^{3D}(\boldsymbol{C}_{\mathrm{ex}}^{i+1})\times \mathrm{f}^{3D}(\boldsymbol{G}^i)_{up}\right) \tag{5}$$

其中，f^{3D} 表示使用三维的逐点卷积，up 代表上采样操作，使用卷积核大小为 4 的转置卷积。利用像素的特征

来指导几何成本体进行上采样,可以获得更清晰的细节。如式(6)所示,$\boldsymbol{\sigma}$ 代表两个特征之间的关联度,用 $1-\boldsymbol{\sigma}$ 作为 $\boldsymbol{C}_{ex}^{i+1}$ 的权重来辅助模型进行上采样得到 $\boldsymbol{G}^{i+1}$,可以防止上下文信息淹没几何信息。

$$\boldsymbol{G}^{i+1}=\boldsymbol{Conv}((\boldsymbol{G}^{i})_{up}+(1-\boldsymbol{\sigma})\boldsymbol{C}_{ex}^{i+1}) \tag{6}$$

FCS 模块借助于特征中所包含的关键信息,通过注意力模块传递给解码器,可以有选择地融合上下文信息,以实现有效灵活的成本聚合。

3) 多分支聚合的金字塔池化模块

Hong 等[27]在语义分割任务中提出了分割网络 DDRNet,设计了深度聚合的金字塔池化模块(deep aggregation pyramid pooling module,DAPPM),在低分辨率特征处提取多尺度信息,通过级联的方式融合特征,最终表现出优异的性能。在立体匹配任务中,为了聚合不同区域的上下文信息,提高网络理解全局语境的能力,很多模型也使用多尺度卷积来进行视差估计。PSMNet[12]在特征提取阶段使用空间金字塔池化模块,通过不同大小的卷积核得到新的尺寸不一的特征图,然后在通道维度上拼接他们,兼顾了全局语义信息与局部细节信息。

上述模型的金字塔池化模块只在图像的空间维度上进行尺度融合,并没有考虑到视差维度的信息。因此,本文将 DAPPM 中的 2D 卷积替换成 3D 卷积,将其应用在成本聚合阶段,改进后提出了一个多分支聚合的金字塔池化模块,图 4 显示了 MAPPM 的内部结构。

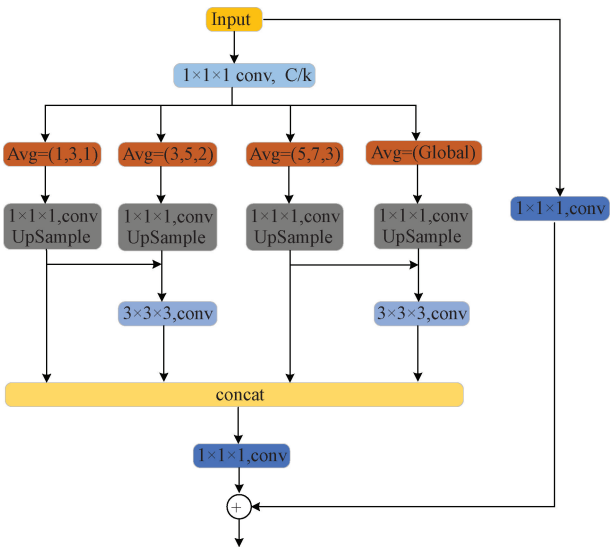


图 4 MAPPM 模块

Fig. 4 MAPPM module

图 4 中 Avg(1,3,1)表示卷积核大小为 1×3×3,步长为 1 的平均池化。Global 表示全局平均池化,C 代表初始通道数,k 为 4。MAPPM 从低分辨率的成本体中提取上下文信息,以 1/32 分辨率大小的成本体作为输入,使用上述的池化操作,通过三线性插值对成本体进行上采样,使其与原始输入的大小相同。

在 DAPPM 中,每个尺度包含的通道太多,而本文的模型使用的 3D 卷积会增加更多的计算量,所以将输入成本体的通道数减少为原来的 1/4。由于本文模型多了视差维度的信息,DAPPM 逐层从低到高的融合策略会使得不同的视差信息混合在一起,导致视差的真实特征被掩盖,进而影响整体的准确性。为了提高网络对不同视差区域的识别能力,本文只用 3×3×3 的卷积融合其中相邻的两个分支。

1.4 视差回归

最终得到的成本聚合体提供了原始图像中的像素在每个视差级别的匹配置信度值。FCS-Stereo 通过 *soft*max 操作将其转换为概率分布,参考 CoEx 提出的视差计算的方法,不使用整个分布的期望值,而是仅仅使用成本聚合量前 k 个值来预测视差。这里设置 k 为 2,预测视差图 $\boldsymbol{D}_0$ 大小为 $B\times1\times H/4\times W/4$,$\boldsymbol{d}_1$,$\boldsymbol{d}_2$ 代表概率值最高的两个视差,视差值计算如下:

$$\boldsymbol{D}_0=soft\max(\boldsymbol{c}_{\boldsymbol{d}_1})\times\boldsymbol{d}_1+soft\max(\boldsymbol{c}_{\boldsymbol{d}_2})\times\boldsymbol{d}_2 \tag{7}$$

最后利用每个像素的超像素权重将视差图 $\boldsymbol{D}_0$ 逐步上采样到原始分辨率大小得到预测视差图 $\boldsymbol{D}_1$。

1.5 损失函数

整个网络是以有监督的端到端的方式进行训练,FCS-Stereo 模型使用的损失函数如式(8)所示。

$$L=\sum_{i=0}^{1}\lambda_i Smooth_{L_1}(\boldsymbol{D}_i-\boldsymbol{D}_{gt}) \tag{8}$$

其中,$\boldsymbol{D}_0$ 是 1/4 分辨率的预测视差图、$\boldsymbol{D}_1$ 是全分辨率的预测视差图,$\boldsymbol{D}_{gt}$ 代表真实视差值,λ_0 和 λ_1 分别为 1 和 0.3。$Smooth_{L_1}$ 公式如下:

$$Smooth_{L_1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{其他}\end{cases} \tag{9}$$

2 实验结果和分析

2.1 数据集与实验指标

SceneFlow 数据集是一个大型的合成立体匹配数据集,由 35 454 对训练图像和 4 370 对测试图像组成,视差范围从 1~468,图像分辨率为 960×540。该数据集提供密集的视差图真实值,使用 SceneFlow 的"finalpass"版本用于训练和评估,通过端点误差 EPE、视差异常值 D1 和大于 n 像素误差等指标来评估模型的性能。EPE 公式如式(10)所示,其中 $\boldsymbol{d}_{est}$ 是预测视差值,$\boldsymbol{d}_{gt}$ 是真实视差值。视差异常值 D1 为预测视差值误差大于 3 个像素或大于 5%的真实视差值,大于 n 像素误差是预测误差大于 n 像素的像素比例。

$$EPE=\frac{1}{N}\sum_{(x,y)\in N}|\boldsymbol{d}_{est}(x,y)-\boldsymbol{d}_{gt}(x,y)| \tag{10}$$

KITTI 数据集包括 KITTI 2012 和 KITTI 2015,是真实世界驾驶场景的数据集。其中 KITTI 2012 由 194 对训练图像和 195 对测试图像组成,KITTI 2015 由 200 对训练图像和 200 对测试图像组成。这两个数据集都是通过

LIDIR 传感器获得的视差图，作者没有对结果做插值，所以生成的是稀疏的真实值。在 KITTI 2012 和 KITTI 2015 上，主要评估指标分为非遮挡区域（Noc）和全部区域（All）。对于 KITTI 2012，本文主要对比各区域的三像素误差、四像素误差和端点误差等指标。对于 KITTI 2015，使用第一帧图像的背景像素预测误差比（D1-bg）、前景像素预测预测误差比（D1-fg）以及全部像素预测误差比（D1-all）作为评价指标。

2.2　实验设置

本研究使用 PyTorch 框架来实现本文的模型，在 RTX 3090 GPU 上对模型进行训练和测试。实验使用 Adam（$\beta_1=0.9$，$\beta_2=0.999$）作为本文的优化器，随机将图像裁剪为 $W=512$、$H=256$ 大小进行训练。

首先在 Sceneflow 数据集上对模型进行 64 轮的训练，初始学习率为 0.001，在第 20、32、44 和 54 轮次后将学习率减半，之后再对模型进行 24 轮的微调。在 KITTI 数据集上，使用在 SceneFlow 数据集上预训练的权重进行迁移训练。本文在 KITTI 2012 和 KITTI 2015 的混合数据集上进行 1 000 轮的训练，初始学习率设置为 0.001，在第 400、600、800 和 900 轮次后减半。最后在分别在 KITTI 2012 和 KITTI 2015 测试集上使用训练好的模型进行测试验证，将得到的视差结果图提交到 KITTI 在线测试网站上进行评估。

2.3　消融实验

本文提出了 FCS-Stereo 模型，在网络中设计了多尺度组相关成本体融合 MGCVF、多分支聚合金字塔池化模块 MAPPM 和 FCS 模块，为了验证模型的有效性，在 Sceneflow 数据集上进行消融实验，使用上述端点误差指标 EPE、视差异常值和大于 n 像素误差等指标来评估模块的性能，如表 1 所示，“√”表示该模块被使用。基线模型是构建一个基于改进相关层方法的成本体，使用通用的编码器-解码器结构对其进行正则化。

表 1　在 SceneFlow 上的消融实验

Table 1　Ablation study on SceneFlow

模型	MGC-CF	FCS	MAPPM	EPE（px）	D1/%	>1px/%	>2px/%	>3px/%	T/ms
基线				0.803	2.98	9.26	5.12	3.66	27
MGCVF	√			0.722	2.72	8.42	4.61	3.35	27
FCS		√		0.654	2.45	7.38	4.07	3.03	32
MGCVF+FCS	√	√		0.628	2.32	7.01	3.96	2.90	32
FCS-Stereo	√	√	√	0.601	2.21	6.76	3.78	2.76	33

从表 1 可以看出，MGCVF 和 FCS 模块均有效提升了网络的性能，FCS-Stereo 网络比基线模型的 EPE 降低了 0.2，其余指标也均有改善，提高了立体匹配的精确度。其中，FCS 模块表现很好，直接将误差指标从 0.80 降低到 0.65，D1 降低了 0.53%，像素误差大于 1 像素、2 像素、3 像素的误差指标分别降低了 1.88%、1.05%、0.63%。加入 MGC-CF 模块后仅用了很少的计算成本，降低了 EPE 和 D1 等误差。最后，添加 MAPPM 模块，总体网络的表现最好。相比于基线模型，本文模型仅增加了很少的时间开销，完成了更好的匹配效果。

为了展示各个模块匹配的效果，本文从 Scene Flow 测试集上选择了包含边缘、遮挡、复杂形状和弱纹理等信息的 3 组图片，FCS-Stereo 的预测部分结果如图 5 所示。从各模块的视差结果图可以看出，FCS-Stereo 算法清晰地还原了物体的细节和轮廓，且在缺乏纹理的结构中也能找到合适的对应点。

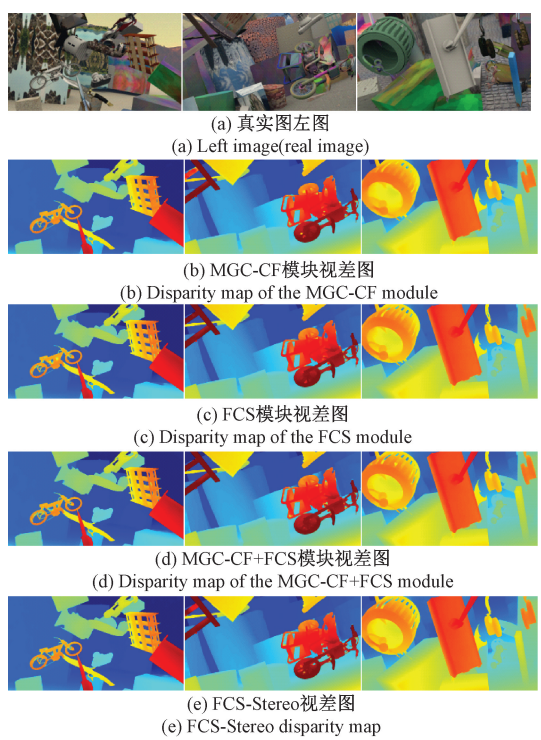


图 5　匹配效果对比

Fig. 5　Comparison of matching results

2.4　对比实验

本文选择算法 EdgeStereo、AANet、GAANET、LMNet、GFAN、LightStereo-M、IGEV-Stereo、CFNet、3D-MobileStereoNet、BGNet＋、CoEx、Fast-ACVNet＋＋与 FCS-Stereo 进行对比，这里，本文根据各算法设计的侧重

点分为上下两栏,上栏是更注重精确性的算法,下栏是更注重实时性的算法。考虑实时性方法的推理时间为≤80 ms,将在 KITTI 在线测试网站上提交的评估结果对各算法进行对比。

1) KITTI2015 对比实验

FCS-Stereo 在 KITTI 2015 测试集上的评估结果如表 2 所示。与其他实时性算法相比,本文算法拥有更高的匹配精度,在全部像素(all)和非遮挡像素(noc)区域中,FCS-Stereo 在背景区域表现更好,背景异常值(D1-bg)比 LightStereo-M 降低了 0.16%,总体像素误差更少,且没有增加太多时间成本。而与高精度算法中表现最好的 IGEV-Stereo 相比,本文的运行时间仅为 IGEV-Stereo 的 1/6,总体来说本模型综合表现最好。

表 2 KITTI2015 各算法对比表

Table 2 Algorithm comparison on KITTI 2015

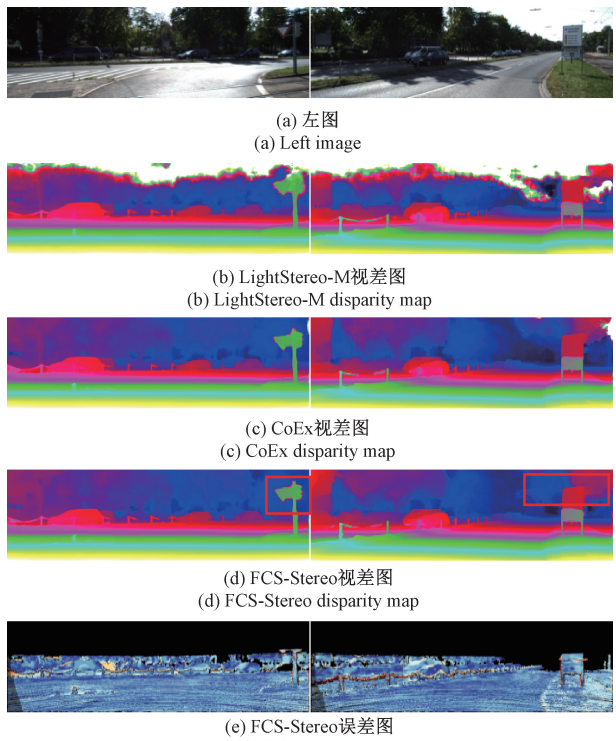
算法	All/%			Noc/%			时间/ms
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	
EdgeStereo[11]	1.84	3.30	2.08	1.69	2.94	1.89	320
IGEV-Stereo[14]	1.38	2.67	1.59	1.27	2.62	1.49	180
CFNet[23]	1.54	3.56	1.88	1.43	3.25	1.73	180
AANet[8]	1.99	5.39	2.55	1.80	4.93	2.32	62
GAANET[9]	1.91	4.25	2.30	1.73	3.99	2.11	80
GFAN[10]	1.78	3.17	2.01	1.95	3.54	2.21	60
3D-MobileStereoNet[17]	1.75	3.87	2.10	1.63	3.50	1.92	72
BGNet+[16]	1.81	4.09	2.19	1.66	3.76	2.01	32
LMNet[18]	2.81	4.39	3.24	—	—	—	15
CoEx[19]	1.79	3.82	2.13	1.65	3.42	1.95	27
Fast-ACVNet++[20]	1.70	3.53	2.01	1.56	3.29	1.85	45
LightStereo-M[21]	1.81	3.22	2.04	1.67	3.06	2.04	23
FCS-Stereo	1.65	3.56	1.97	1.52	3.40	1.83	33

为了对比各算法的匹配结果,本文选择 3D 算法中 LightStereo-M 和 4D 算法中 CoEx 的视差结果图以及本算法视差结果图和误差结果图。如图 6 所示,FCS-Stereo 生成的视差图在背景区域的结果更加平滑,误差更少,能捕捉到物体的边界轮廓,可以还原细小结构。

2) KITTI2012 对比实验

FCS-Stereo 在 KITTI 2012 上的评估结果如表 3 所示。与高精度的算法相比,FCS-Stereo 用了更少的推理时间,而性能接近于 EdgeStereo。而与其他实时性算法相比,FCS-Stereo 在表现上次于 Fast-ACVNet++,但是本文模型在匹配速度上有一定优势。比实时性算法 LightStereo-M 在非遮挡区域的三像素误差降低了 0.06%。

图 7 中展示了各算法在 KITTI 2012 数据集上的部分测试结果,对比其他算法可以看出,在一些如树木等有较多重叠和细节的复杂区域,本文算法能够生成相对稳定的视差值。对于远景和近景中的车辆,FCS-Stereo 也可以在物体轮廓处保持平滑过渡,在光照变化的区域拥有较好的匹配效果,具有较强的适应性。



(a) 左图
(a) Left image

(b) LightStereo-M视差图
(b) LightStereo-M disparity map

(c) CoEx视差图
(c) CoEx disparity map

(d) FCS-Stereo视差图
(d) FCS-Stereo disparity map

(e) FCS-Stereo误差图
(e) FCS-Stereo error map

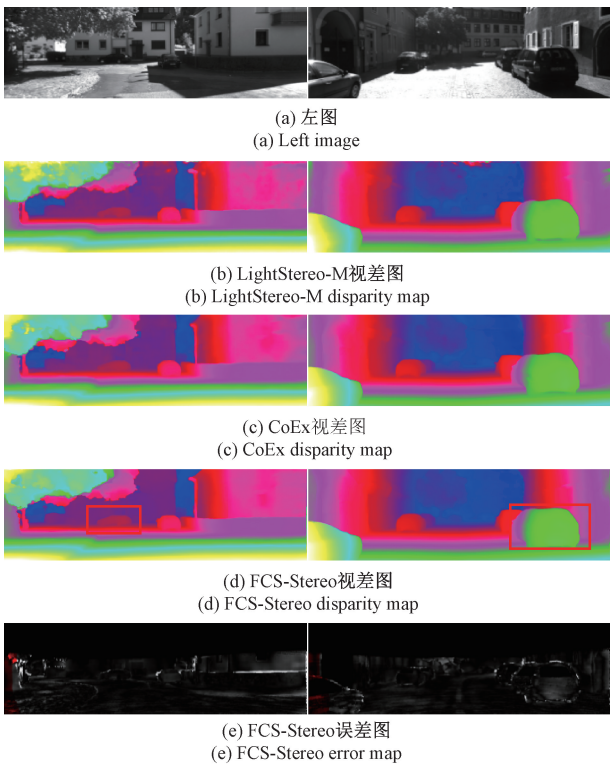
图 6 KITTI2015 测试结果对比

Fig. 6 Comparison of testing results on KITTI 2015

表 3　KITTI2012 各算法对比表

Table 3　Algorithm comparison on KITTI 2012

算法	>3px/%		>4px/%		Avg		时间/ms
	Noc	All	Noc	All	Noc	All	
EdgeStereo[12]	1.46	1.83	1.07	1.24	0.4	0.5	320
IGEV-Stereo[14]	1.12	1.44	0.88	1.12	0.4	0.4	180
CFNet[23]	1.23	1.58	0.92	1.18	0.4	0.5	180
AANet[8]	1.91	2.42	1.46	1.87	0.5	0.6	62
GAANET[9]	1.73	2.22	—	—	0.5	0.6	80
BGNet+[16]	1.62	2.03	1.16	1.48	0.5	0.6	32
CoEx[19]	1.55	1.93	1.15	1.42	0.5	0.5	27
Fast-ACVNet++[20]	1.45	1.85	1.06	1.36	0.5	0.5	45
LightStereo-M[21]	1.56	1.91	1.10	1.46	0.5	0.5	20
FCS-Stereo	1.50	1.89	1.12	1.42	0.5	0.5	33



(a) 左图
(a) Left image

(b) LightStereo-M视差图
(b) LightStereo-M disparity map

(c) CoEx视差图
(c) CoEx disparity map

(d) FCS-Stereo视差图
(d) FCS-Stereo disparity map

(e) FCS-Stereo误差图
(e) FCS-Stereo error map

图 7　KITTI2012 测试结果对比

Fig. 7　Comparison of testing results on KITTI 2012

3　结　论

本文提出了一种选择性融合上下文信息的立体匹配算法 FCS-Stereo，先使用改进的相关层算法和单个编码器-解码器来减少网络的参数量。然后，在成本聚合编码器中设计多尺度组相关成本体融合增强网络全局和结构区域表示，在解码器中设计选择性融合上下文信息模块来增强网络的细节恢复能力。最后，设计多分支聚合的金字塔池化模块，使得网络可以在复杂场景下生成更稳定准确的视差图。在 KITTI2015、KITTI2012 数据集上的实验结果表明，FCS-Stereo 可以在背景区域获得更少的误差，同时在复杂场景如树木、阴影、轮廓等区域生成更稳定、平滑的视差图。与其他算法相比，FCS-Stereo 在匹配精度和匹配速度上实现了较好的平衡。尽管模型在复杂场景中能够保持较好的稳定性，但是对于低光照区域的处理仍然存在不足，且在一些背景区域会出现过度平滑导致信息丢失的问题。该如何根据场景的不同需求来调整平滑和细节保留的权重，仍然是需要研究的一个课题。

参考文献

[1]　郑好，段发阶，白子博，等. 一种基于重投影和 3D-DIC 的曲面变形测量方法[J]. 仪器仪表学报，2024，45(8)：268-285.
ZHENG H，DUAN F J，BAI Z B，et al. A surface deformation measurement method based on reprojection and 3D-DIC [J]. Chinese Journal of Scientific Instrument，2024，45(8)：268-285.

[2]　王晓峰，孙志恒，喻骏，等. 基于细节信息增强的无监督双目立体匹配算法[J]. 电子测量技术，2024，47(5)：94-101.
WANG X F，SUN ZH H，YU J，et al. Unsupervised stereo matching algorithm of binocular based on detail information enhancement[J]. Electronic Measurement Technology，2024，47(5)：94-101.

[3]　余雪飞，顾寄南，黄则栋，等. 基于边缘检测与注意力机制的立体匹配算法[J]. 电子测量技术，2022，45(11)：167-172.
YU X F，GU J N，HUANG Z D，et al. Stereo matching algorithm based on edge detection and attention mechanism [J]. Electronic Measurement Technology，2022，45(11)：167-172.

[4]　覃业宝，孙炜，范诗萌，等. 全距离深度平衡立体匹配网络[J]. 电子测量与仪器学报，2023，37(8)：30-39.
QIN Y B，SUN W，FAN SH M，et al. Full range depth balanced stereo matching network[J]. Journal of

Electronic Measurement and Instrumentation, 2023, 37(8):30-39.

[5] ŽBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of Machine Learning Research, 2016, 17(65):1-32.

[6] PARK H, LEE K M. Look wider to match image patches with convolutional neural networks[J]. IEEE Signal Processing Letters, 2017, 24(12): 1788-1792.

[7] MAYER N, ILG E, HÄUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]. IEEE Conference on Computer Vision and Pattern recognition, 2016: 4040-4048.

[8] XU H F, ZHANG J Y. Aanet: Adaptive aggregation network for efficient stereo matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1959-1968.

[9] 宋昊,毛宽民,朱洲. 基于 GAANET 的立体匹配算法[J]. 计算机科学,2024,51(4):229-235.
SONG H, MAO K M, ZHU ZH. Algorithm of stereo matching based on GAANet[J]. Computer Science, 2024, 51(4):229-235.

[10] ZHANG Y K, ZHANG J H. GFANet: Group fusion aggregation network for real time stereo matching[J]. IEEE Robotics and Automation Letters, 2023, 8(7): 4251-4258.

[11] SONG X, ZHAO X, FANG L J, et al. Edgestereo: An effective multi-task learning network for stereo matching and edge detection[J]. International Journal of Computer Vision, 2020, 128(4): 910-930.

[12] CHANG J R, CHEN Y SH. Pyramid stereo matching network[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5410-5418.

[13] GUO X Y, YANG K, YANG W K, et al. Group-wise correlation stereo network[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3273-3282.

[14] XU G W, WANG X Q, DING X H, et al. Iterative geometry encoding volume for stereo matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21919-21928.

[15] KHAMIS S, FANELLO S, RHEMANN C, et al. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction[C]. European Conference on Computer Vision(ECCV), 2018: 573-590.

[16] XU B, XU Y H, YANG X L, et al. Bilateral grid learning for stereo matching networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12497-12506.

[17] SHAMSAFAR F, WOERZ S, RAHIM R, et al. Mobilestereonet: Towards lightweight deep networks for stereo matching[C]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2417-2426.

[18] XUE Y B, ZHANG D D, LI L D, et al. Lightweight multi-scale convolutional neural network for real time stereo matching[J]. Image and Vision Computing, 2022,124: 104510.

[19] BANGUNHARCANA A, CHO J W, LEE S, et al. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation[C]. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2021: 3542-3548.

[20] XU G W, CHENG J D, GUO P, et al. Attention concatenation volume for accurate and efficient stereo matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12981-12990.

[21] GUO X D, ZHANG CH M, ZHANG Y M, et al. Lightstereo: Channel boost is all your need for efficient 2D cost aggregation[J]. ArXiv preprint arXiv:2406.19833, 2024.

[22] SANDLER M, HOWARD A, ZHU M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510-4520.

[23] SHEN ZH L, DAI Y CH, RAO ZH B. CFNet: Cascade and fused cost volume for robust stereo matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13906-13915.

[24] SHEN ZH L, DAI Y CH, SONG X B, et al. PCW-Net: Pyramid combination and warping cost volume for stereo matching[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 280-297.

[25] 葛兰,贾振堂. 深浅层特征结合的自监督立体匹配[J]. 电子测量技术,2023,46(12):143-149.
GE L, JIA ZH T. Self-supervised stereo matching combining deep features and shallow featuret[J]. Electronic Measurement Technology,2023,46(12):143-149.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural information Processing systems, 2017, 30:5998-6008.

[27] HONG Y D, PAN H H, SUN W CH, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes[J]. ArXiv preprint arXiv: 2101.06085, 2021.

作者简介

宁安琪,硕士研究生,主要研究方向为计算机视觉和图像处理。

E-mail:2120224814@qq.com

於跃成(通信作者),教授,硕士生导师,主要研究方向为计算机视觉和图像处理。

E-mail:zhjyuyuecheng@163.com

杨帆,硕士研究生,主要研究方向为计算机视觉和图像处理。

E-mail:302678032@qq.com

李响,硕士研究生,中级工程师,主要研究方向为边缘云计计算。

E-mail:lixiang@chinamobile.com