DOI:10.19651/j.cnki.emt.2518961

基于 3D 代价体注意力的实时立体匹配网络*

李 冰[1,2] 严熠萌[1,2] 张鑫磊[1,2] 邵宝文[1,2] 翟永杰[1,2]
(1.华北电力大学自动化系 保定 071003; 2.保定市电力系统智能机器人感知与控制重点实验室 保定 071003)

摘 要: 信息丰富且计算高效的代价体对于高精度高效率的立体匹配至关重要,为构建信息丰富且计算高效的代价体和实现高精度高效率的立体匹配,在 Fast-ACVNet 的基础上,提出一种轻量化网络 Efficient-ACVNet,提升立体匹配在代价体构建阶段的效率。首先,使用计算量更低的 3D 代价体作为代价体注意力,引入逆瓶颈残差块堆叠对称沙漏结构对 3D 代价体进行代价聚合,并提出多尺度视差通道注意力模块来进一步提升代价聚合效果,聚合后的 3D 代价体作为代价体注意力将用于构建和滤波信息冗余的 4D 代价体,提升 4D 代价体的信息量和计算效率,最后引入伪 3D 残差块和设计伪 3D 下采样模块对 4D 代价体代价聚合,进一步降低网络复杂度。实验结果表明,与基准方法相比,所提算法在 SceneFlow 数据集上的端点误差(EPE)降低了 9.375%,在 KITTI15 数据集上的前景区域(D1-fg)异常值百分比下降 19%,并且网络运行时间从 39ms 下降到 25ms。

关键词: 立体匹配;深度学习;代价体注意力;逆瓶颈残差块;伪 3D 卷积
中图分类号: TP391.4;TN911.73 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Real-time stereo matching network based on 3D cost volume attention

Li Bing[1,2] Yan Yimeng[1,2] Zhang Xinlei[1,2] Shao Baowen[1,2] Zhai Yongjie[1,2]
(1. Department of Automation, North China Electric Power University, Baoding 071003, China;
2. Baoding Key Laboratory of Intelligent Robot Perception and Control for Power Systems, Baoding 071003, China)

Abstract: An information-rich and computationally efficient cost volume is crucial for high-precision and high-efficiency stereo matching. To construct such a cost volume and achieve accurate stereo matching, a lightweight network, Efficient-ACVNet, was proposed based on Fast-ACVNet to improve the efficiency of cost volume construction in stereo matching. First, a computationally less intensive 3D cost volume was used as cost volume attention. Inverse bottleneck residual blocks were employed to stack a symmetric hourglass structure for cost aggregation of the 3D cost volume, and multi-scale disparity channel attention modules was introduced to further enhance the aggregation effect. The aggregated 3D cost volume served as cost volume attention to construct and filter the information-redundant 4D cost volume, improving its information content and computational efficiency. Finally, pseudo-3D residual blocks were introduced and pseudo-3D downsampling modules was designed for cost aggregation of the 4D cost volume, further reducing network complexity. Experimental results showed that compared to the baseline method, the proposed algorithm reduced the endpoint error (EPE) by 9.375% on the SceneFlow dataset, decreased the outlier percentage (D1-fg) in the foreground region by 19% on the KITTI15 dataset, and reduced network runtime from 39 ms to 25 ms.
Keywords: stereo matching; deep learning; cost volume attention; inverted bottleneck residual block; pseudo-3D convolution

0 引 言

立体匹配作为三维重建[1]和自动驾驶[2]等领域的核心技术,旨在通过计算左右图像对应像素点的水平位移(即视差)[3],为深度估计、环境感知和场景理解提供关键数据支持。

传统的立体匹配方法在 Scharstein 等[4]的研究中被划分为 4 个关键步骤:代价计算、代价聚合、视差计算和视差

收稿日期:2025-05-28
*基金项目:国家自然科学基金面上项目(62373151)、国家自然科学基金联合项目(U21A20486)、河北省自然科学基金面上项目(F2023502010)、中央高校基本科研业务费专项资金(20237488)资助

优化。传统立体匹配算法在立体匹配初期有着很好的性能，但后续研究发现传统立体匹配方法在无纹理区域的表现并不理想[5]。Zbontar 和 LeCun 等[6]提出的 MC-CNN 是一项突破性工作，首次将深度学习算法应用到传统立体匹配流程中，该非端到端方法使得计算精度得到提升。此后，Mayer 等[7]提出首个端到端立体匹配网络 DispNetC，该网络构建 3D 代价体，使用编码-解码的 2D 卷积神经网络结构直接输出视差图。

近年来，随着深度学习立体匹配网络的快速发展，如何构建信息丰富且计算高效的代价体是深度学习的立体匹配的关键所在[8]。Kendall 等[9]通过拼接特征图构建出信息丰富的 4D 代价体，并引入 3D 卷积层进行代价聚合，该网络结构为后续大多数网络运用。Guo 等[10]将多通道左右特征图沿着通道维度进行分组，通过组相关的计算构建 4D 代价体，并对相关性代价体实现计算优化，提升了立体匹配性能和精度。Shen 等[11]提出以粗到细的方式构建多尺度级联 4D 代价体，通过融合多尺度代价体获取不同尺度的信息，网络信息量得以提升，不过网络在每阶段都进行大量的 3D 卷积，需要巨大的计算成本。Tankovich 等[12]通过多分辨率初始化、可微的传播过程与 warp 机制来实现视差预测，该网络避免了传统 3D 卷积，显著提升了推理速度。Xu 等[8]认为分组相关性构建的代价体有丰富的相关性信息缺少内容信息，而拼接构建的代价体拥有内容信息却缺少相关性信息。为此，Xu 等[8]提出 Fast-ACVNet，利用 Fine-to-Important(F2I)策略，将分组相关性构建的 4D 代价体作为注意力，去滤波通过拼接特征图构建的信息冗余的 4D 代价体，使滤波后的代价体变得信息丰富且计算高效，但低分辨率的 4D 代价体注意力的精度不高且计算复杂度高。

在轻量级立体匹配网络的设计上，Khamis 等[13]使用原始图像 1/8 分辨率的左、右特征图计算出两者之间的差异，构建出 4D 差异代价体，然后使用左图颜色指导完成视差优化，但低分辨率代价体的预测精度有限。Xu 等[14]利用卷积层构建并行多分尺度 3D 代价体并提出自适应聚合网络，提升了匹配精度和效率，然而多层的卷积层增大了网络的复杂度。Guo 等[15]构建出 3D 代价体，并使用 MobileNet V2[16]的倒置残差块来进行高效代价聚合并实现实时立体匹配，然而 3D 代价体所含信息不足，网络精度并不高。在立体匹配网络的轻量化设计中，大多数网络都在减少或避免 3D 卷积层，然而这也会造成精度上的牺牲[5]。

针对以上问题，为构建信息丰富且计算更加高效的代价体，实现高精度高效率的立体匹配，本文对 Fast-ACVNet 网络进行改进。首先将基准网络中用于生成代价体注意力权重的低分辨率 4D 代价体替换为简洁的高分辨率 3D 代价体，降低网络计算成本，提升匹配精度；然后，为提升 3D 代价体的聚合性能，引入 ConvNeXt V2[17]的逆瓶颈残差块堆叠对称沙漏结构对 3D 代价体进行代价聚合，并提出多尺度视差通道注意力模块增强代价聚合；聚合后的代价体会用于生成代价体注意力，用于滤波 4D 代价体的冗余信息，实现信息丰富且计算高效的 4D 代价体；对于 4D 代价体的代价聚合，设计伪 3D 下采样模块和引入伪 3D 残差块[18]堆叠沙漏结构进行代价聚合，降低网络计算成本。最后进行视差回归，得到最终的预测视差图。

1　基于 3D 代价体注意力的立体匹配方法

基于基准网络中 Fine-to-Important(F2I)策略的良好效果，在 Fast-ACVNet 网络的基础上进行改进，得到更高效的 Efficient-ACVNet。Efficient-ACVNet 结构如图 1 所示，由四部分组成：特征提取网络，用于提取左右图像的多尺度特征；3D 代价体构建与 2D 代价聚合，用于生成代价体注意力；高似然视差假设与代价体注意力生成，用于构建 4D 代价体和滤波；4D 代价体的构建与 3D 代价聚合以及视差回归，用于生成最终预测视差图。通过 3D 代价体注意力构建出信息丰富且计算高效的代价体，在保证匹配精度的同时，显著提升计算效率，实现效率与性能的均衡。

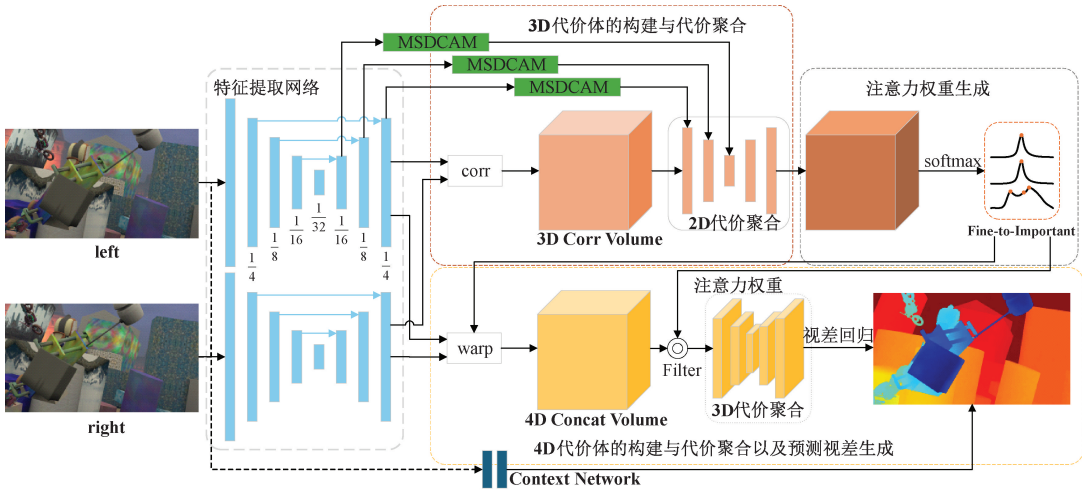

图 1　Efficient-ACVNet 整体结构

Fig. 1　Overall architecture of Efficient-ACVNet

1.1　特征提取网络

特征提取网络采用与Fast-ACVNet同样的特征提取网络,即已在ImageNet上完成预训练的MobileNet V2[16]的特征提取网络,实现对4个不同尺度进行特征图提取,有效地将分辨率降低至初始大小的1/4、1/8、1/16和1/32。随后,使用带有跳跃连接的反卷积模块对特征图进行恢复,特征提取网络将原始输入图像的1/4,1/8,1/16和1/32分辨率的多尺度特征图作为输出。基准网络通过特征通道分组构建的4D相关性代价体[10],需要特征提取网络输出的通道数分别为(96, 96, 96, 160),本文网络利用特征图构建的3D相关性代价体将避免对特征通道分组的操作,特征提取网络中各尺度特征图的输出通道数减少为(24, 48, 96, 160)。

1.2　构建3D代价体与代价聚合

为构建信息丰富且计算更加高效的代价体,提出利用原始图像1/4分辨率下的左、右特征图之间的相关性构建3D代价体,减少运行时间。在代价聚合阶段引入逆瓶颈残差块堆叠的对称沙漏结构进行代价聚合,并提出多尺度视差通道注意力模块增强代价聚合,丰富代价体信息,提升匹配精度。

1) 构建3D代价体

通过特征提取网络输出的原始图像1/4分辨率的左、右特征图,构建3D相关性代价体。当视差 d 的范围在0到 $D-1$(D 为最大视差值)内时,通过计算左特征图与右特征图在水平方向上平移 d 个像素后的相似度来构建代价体,构建代价体 $\boldsymbol{C}_{corr}$ 的计算公式如下:

$$\boldsymbol{Ccorr}(d,h,w)=\frac{1}{C}\sum_{C=1}^{C}\boldsymbol{f}_{l,4}(h,w)\cdot\boldsymbol{f}_{r,4}(h,w-d) \tag{1}$$

式中:C 代表视差通道数,h 和 w 为图像的高度和宽度,$\boldsymbol{f}_{l,4}$ 和 $\boldsymbol{f}_{r,4}$ 为特征提取层得到的原始图像1/4分辨率的左、右特征图。当 d 为0时,代价体在所有通道上对左特征图与右特征图在相同空间位置的特征向量进行逐元素乘积,并计算其平均值。

2) 逆瓶颈残差块组成的2D代价聚合

Guo等[15]使用MobileNet V2[16]的倒置残差块堆叠沙漏结构进行代价聚合展现出了良好的性能,为进一步提升代价聚合的效果,引入ConvNeXt V2[17]的逆瓶颈残差块。逆瓶颈残差块具有更大的感受野和逆瓶颈操作,能够提升特征表达能力并提升计算效率,逆瓶颈残差块结构如图2所示。代价聚合时,逆瓶颈残差块将代价体 $\boldsymbol{C}_{corr}$ 作为输入($Input$),$\boldsymbol{C}_{corr}$ 先通过一个7×7的具有大感受野的深度可分离卷积,该卷积会对每个视差通道独立操作以捕获空间特征得到代价体 $\boldsymbol{C}_m$,代价体 $\boldsymbol{C}_m$ 计算公式如下:

$$\boldsymbol{C}_m=LayerNorm(\boldsymbol{W}_{depthwise}*\boldsymbol{C}_{corr}) \tag{2}$$

式中:$\boldsymbol{W}_{depthwise}$ 表示深度可分离卷积的权重,$*$ 表示卷积运算,$LayerNorm$ 表示层归一化。接着,扩展通道后的代价体 $\boldsymbol{C}_m$ 会进行一次逐点卷积,该逐点卷积由线性激活函数等价实现,用于提升视差通道的通道数得到代价体 $\boldsymbol{C}_n$,代价体 $\boldsymbol{C}_n$ 计算公式如下:

$$\boldsymbol{C}_n=GRN\{GELU[Linear(\boldsymbol{C}_m)]\} \tag{3}$$

式中:$Linear$ 代表线性激活函数,$GELU$ 代表 $GELU$ 激活函数,GRN 代表全局响应归一化层。代价体提升的通道数将由扩展因子(扩展通道数时的乘积因子)决定。然后,代价体 $\boldsymbol{C}_n$ 再通过另一个逐点卷积恢复到原始的视差通道数代价体 out,计算公式如下:

$$\boldsymbol{out}=Linear(\boldsymbol{C}_n) \tag{4}$$

式中:$Linear$ 代表线性激活函数。最后,添加残差连接来优化梯度传播:

$$\boldsymbol{out}=\boldsymbol{out}+\boldsymbol{In\,put} \tag{5}$$

代价聚合下采样模块采用MobileNet V2的倒置卷积块,下采样模块将代价体 $\boldsymbol{C}_{corr}$ 作为输入($Input$),$\boldsymbol{C}_{corr}$ 先通过一个1×1的通道拓展卷积,该卷积会提升视差通道数,并进行下采样得到代价体 $\boldsymbol{C}_m$,代价体 $\boldsymbol{C}_m$ 计算公式如下:

$$\boldsymbol{C}_m=ReLU6(\boldsymbol{W}_{expand}*\boldsymbol{C}_{corr}) \tag{6}$$

式中:$\boldsymbol{W}_{expand}$ 表示拓展卷积的权重,$*$ 表示卷积运算,$ReLU6$ 表示 $ReLU6$ 激活函数,代价体提升的通道数将由扩展因子(扩展通道数时的乘积因子)决定。接着,扩展通道后的代价体 $\boldsymbol{C}_m$ 会进行一次3×3的深度可分离卷积,该卷积会对每个视差通道独立操作以捕获空间特征得到代价体 $\boldsymbol{C}_n$,代价体 $\boldsymbol{C}_n$ 计算公式如下:

$$\boldsymbol{C}_n=ReLU6(\boldsymbol{W}_{depthwise}*\boldsymbol{C}_m) \tag{7}$$

式中:$\boldsymbol{W}_{depthwise}$ 代表深度可分离卷积的权重。最后,代价体 $\boldsymbol{C}_n$ 再通过另一个1×1的投影卷积恢复到原始的视差通道数代价体 out,计算公式如下:

$$\boldsymbol{out}=\boldsymbol{W}_{project}*\boldsymbol{C}_n \tag{8}$$

式中:$\boldsymbol{W}_{project}$ 代表投影卷积的权重。

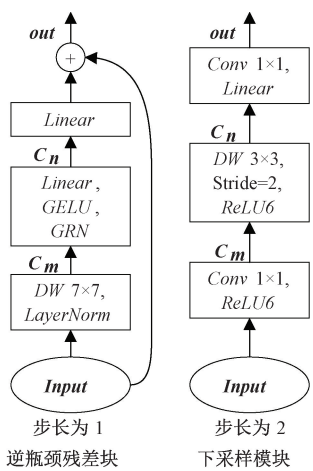

图2　2D代价聚合卷积块结构图

Fig. 2　2D cost aggregation convolution block architecture diagram

为进一步提升代价聚合的效果,提出对称沙漏结构的代价聚合网络,代价聚合网络由下采样模块和逆瓶颈残差

块组成，代价聚合网络的结构如图 3 所示，该结构使用下采样模块将代价体于原始图像 1/4 分辨率下采样至 1/8 和 1/16 分辨率，随后通过带有跳跃连接的反卷积上采样逐步恢复至 1/4 分辨率的代价体。在每一分辨率层级下，逆瓶颈残差块的数量按照(1，2，4，2，1)进行配置，并将逆瓶颈残差块和下采样模块的扩展因子设置为 4。网络采用编码器-解码器结构，具有对称的下采样和上采样路径，能够同时获取多尺度上下文信息并保持精细的空间细节。

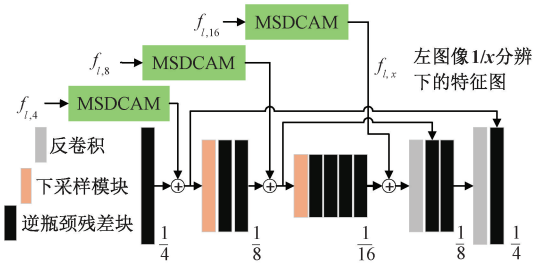

图 3　2D 代价聚合结构图

Fig. 3　2D cost aggregation architecture diagram

3) 多尺度视差通道注意力模块

根据大核卷积和条形卷积作为卷积注意力在增强特征表达能力方面展现出显著的有效性[19-21]。为提升 2D 代价聚合效果，丰富代价体的信息。在 Guo 等[15]提出的多尺度注意力模块的基础上提出一种多尺度视差通道注意力模块(multi-scale disparity channel attention module，MSDCAM)，对特征图提取多尺度特征用于增强 2D 代价聚合。

模块分为 3 个用于不同分辨率的注意力小模块，用于适应不同分辨率大小，每个小模块分别使用不同核大小的深度可分离卷积。如图 4 所示，在原始图像 1/4 分辨率下，卷积核组合为 1×1、7×1、1×7、11×1、1×11、21×1 和 1×21；在 1/8 分辨率下，采用 1×1、5×1、1×5、13×1、1×13、19×1 和 1×19 的卷积核组合；在 1/16 分辨率下，则使用 1×1、5×1、1×5、11×1、1×11、17×1 和 1×17 的卷积核组合。利用非对称条形卷积捕捉各特征图的水平和垂直条形特征信息，这些信息指导网络更好的识别图像中的结构特征，适合立体匹配任务。MSDCAM 模块利用多尺度图像结构特征以及语义信息来指导代价聚合过程，增强代价聚合。

将特征提取网络输出的多尺度左特征图在各自尺度下通过 MSDCAM 模块，提取水平和垂直方向的条形特征。首先，特征图通过 1×1 卷积扩大通道数，再经过条形卷积提取多尺度条形特征，然后对条形特征进行拼接形成全面的多尺度特征表示，最后通过一个 1×1 卷积进行通道融合，将多尺度条形特征进行融合并对特征通道重新校准，使网络能够获取不同尺度下的相关信息。最终输出与正在聚合的代价体通过逐元素相乘的方式相结合，增强代价聚合的效果。

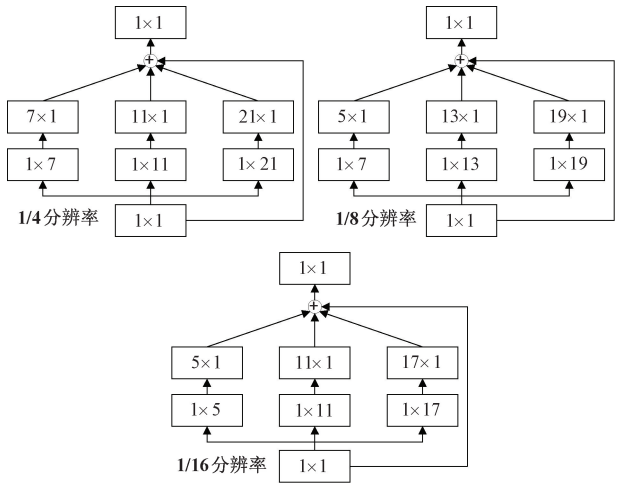

图 4　多尺度视差通道注意力模块

Fig. 4　Multi-scale disparity channel attention module

1.3　代价体注意力生成

生成代价体注意力的方式沿用基准网络 Fast-ACVNet 的 Fine-to-Important(F2I)策略。

F2I 策略将聚合后的 3D 代价体通过 $softmax$ 得到概率体 $\boldsymbol{P}$，Fine-to-Important(F2I)策略在每个像素点处选择 $\boldsymbol{P}$ 中概率值最高的 K 个值作为注意力权重 $\boldsymbol{A}^F$，计算公式如下：

$$\boldsymbol{A}^F = max^K\{\boldsymbol{P}\} \tag{9}$$

以及将概率体 $\boldsymbol{P}$ 进行 $argmax$ 视差回归作为高似然性假设 $\boldsymbol{D}^{hyp}$，计算公式如下：

$$\boldsymbol{D}^{hyp} = argmax\{\boldsymbol{P}\} \tag{10}$$

1.4　4D 代价体的构建与代价聚合以及视差回归

1) 4D 代价体构建

利用高似然假设 $\boldsymbol{D}^{hyp}$ 拼接特征提取网络输出的左、右特征图构建 4D 代价体 $\boldsymbol{C}_{concat}$，$\boldsymbol{C}_{concat}$ 构建方式如下：

$$\boldsymbol{C}_{concat}(\cdot,\boldsymbol{D}^{hyp},h,w) = Concat\{\boldsymbol{f}_l(h,w),\boldsymbol{f}_r(h-\boldsymbol{D}^{hyp},w)\} \tag{11}$$

式中：h 和 w 为图像的高度和宽度，$\boldsymbol{f}_l$ 和 $\boldsymbol{f}_r$ 为特征提取网络输出原始图像 1/4 分辨率的左、右特征图，$Concat$ 表示拼接操作。最后，使用代价体注意力权重 $\boldsymbol{A}^F$ 对 4D 代价体 $\boldsymbol{C}_{concat}$ 进行滤波操作：

$$\boldsymbol{C}_{concat} = \boldsymbol{A}^F * \boldsymbol{C}_{concat} \tag{12}$$

2) 伪 3D 残差块和伪 3D 下采样模块

从伪 3D 残差块[18]获得启发，引入伪 3D 残差块并设计伪 3D 下采样模块来代替 3D 卷积，用于保证网络的轻量化，提升网络的精度和稳定性。伪 3D 残差块和伪 3D 下采样模块如图 5 所示。伪 3D 模块通过解耦 3×3×3 卷积为 1×3×3 和 3×1×1 两个卷积来降低计算量，并在前后添加 1×1×1 逐点卷积增强特征融合能力，在每一层卷积后，都添加批归一化和 $ReLU6$ 激活函数，以提升网络稳定性。伪 3D 卷积不但保持与标准 3D 卷积相当的性能，还减

少计算复杂度。下采样时,4D 代价体会先经过 1×1×1 的逐点卷积,逐点卷积步长为 2,用于下采样和提升通道数;然后,代价体会经过 1×3×3 的 3D 卷积对空间维度进行聚合;随后,代价体会经过 3×1×1 的 3D 卷积对视差维度进行聚合;最后,代价体再经过 1×1×1 的逐点卷积进行特征信息融合。

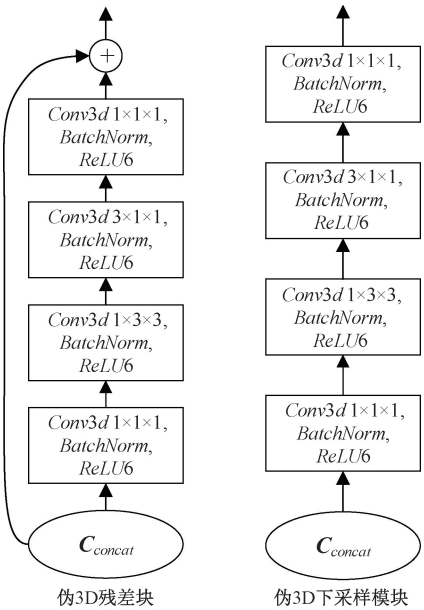

图 5　伪 3D 卷积结构图

Fig. 5　Pseudo-3D convolution architecture diagram

伪 3D 残差块在代价聚合时,代价体会先经过 1×1×1 的逐点卷积,用于对空间和视差维度进行变换,以适应后续卷积对空间和视差维度的聚合;然后,代价体会经过 1×3×3 和 3×1×1 的 3D 卷积对空间维度和视差维度分别进行聚合;随后,代价体再经过 1×1×1 的逐点卷积进行特征融合;最后,添加残差连接来优化梯度传播。

3) 3D 代价聚合

3D 代价聚合网络的沙漏结构仅由 2 个伪 3D 下采样模块、4 个伪 3D 残差块和两个反卷积组成,如图 6 所示。该结构首先会将代价体从原始图像 1/4 分辨率下采样至 1/8 和 1/16 分辨率,随后利用 3D 反卷积上采样恢复到初始代价体分辨率大小。并借鉴 CoEx[22] 的方法,利用左图像提取的特征图引导代价聚合,增强代价聚合的效果。

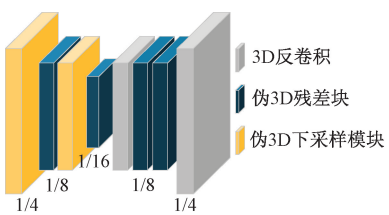

图 6　3D 代价聚合结构图

Fig. 6　3D cost aggregation architecture diagram

4) 视差回归

根据 CoEx[22] 利用 top-k 进行视差回归的良好效果,使用 top-k 进行视差回归,将聚合代价体积中每个像素处的前 2 个值进行 softmax 运算,以计算期望的视差值。最后,依照全卷积分割[23]的方法,利用围绕每个像素的"超像素"权重,将输出的视差预测结果上采样至原始输入图像的分辨率。

2　实验结果与分析

2.1　数据集与评估方法

SceneFlow 数据集[7]是合成的立体图像数据集,该数据集旨在提供一个大规模、多样化且具有精确标注的场景流数据,以支持相关领域的研究和算法开发。数据集包含超过 35 454 对训练图像和 4 370 对测试图像,数据集里包括 FlyingThings3D、Driving 和 Monkaa 子数据集。网络使用 SceneFlow 的"finalpass"数据集用于训练和评估,并使用端点误差(EPE)作为评估指标。

KITTI 数据集包含 KITTI2012[24] 和 KITTI2015[25]。KITTI2012 和 KITTI2015 是真实世界数据集,分别有 194/195 对和 200/200 对图像用于训练和测试。对于这些数据集,使用激光雷达(LiDAR)系统的收集了稀疏的真实视差数据。评估方式依据 KITTI 官方的在线基准测试。在线基准测试中对于 KITTI2012,采用非遮挡区域(noc)和所有(all)像素的错误像素百分比还有平均端点误差(EPE)作为评估结果。对于 KITTI2015,采用评估背景区域(D1-bg)、前景区域(D1-fg)和所有像素(D1-all)的视差异常值百分比(D1)作为评估结果。在泛化性评估中,网络使用仅在 SceneFlow 数据集上训练 90 个周期的模型,以 KITTI2012 和 KITTI2015 的测试数据集作为泛化性评估对象,并以视差异常值百分比(D1)作为评估指标。

Middlebury2014 数据集[26]包含 15 组训练图像和 15 组测试图像。网络使用仅在 SceneFlow 数据集上训练 90 个周期的模型,以每组图像的 1/2 分辨率版本作为泛化性评估对象,并以误差大于 2 像素的像素百分比作为评估指标。

2.2　实验过程

Efficient-ACVNet 的实验在服务器上进行,服务器的操作系统为 Ubuntu 20.04.6,GPU 为 NVIDIA RTX 3090,深度学习框架为 Pytorch 2.1.1,Torchvison 0.16.2,CUDA12.2,Python 3.10。网络的最大视差值设置为 192。训练时,损失函数采用 *Smooth L*1,损失权重采用(1.0, 0.3, 0.5, 0.3)。对于 SceneFlow 数据集,Efficient-ACVNet 的批量大小设置为 24,并使用 AdamW 优化器结合 OneCycleLR 调整学习率,其中最大学习率设置为 0.000 5。在最终评估和消融实验中,Efficient-ACVNet 网络采用 SceneFlow 数据集并训练 90 个周期。训练时还使用了随机裁剪对数据集进行数据增强。对于

KITTI 数据集，网络在 SceneFlow 数据集上预训练的基础上，使用由 KITTI2012 和 KITTI2015 训练数据集组成混合训练集，对网络进行 500 个周期的微调。KITTI 数据集采用的批量大小为 2，并使用 AdamW 优化器结合 OneCycleLR 调整学习率，最大学习率为 0.000 2。

伪 3D 卷积模块(pseudo-3D convolution，P3D)的有效性实验在同样配置的服务器上进行。训练时 StereoNet-P3D 的损失权重采用(0.3，0.5，0.7，1)，并使用 RMSprop 优化器结合 MultiStepLR(gamma＝0.5)调整学习率，网络最大学习率为 0.01；Fast-ACVNet-P3D 的损失权重采用(1.0，0.3，0.5，0.3)，并使用 AdamW 优化器结合 OneCycleLR 调整学习率，其中最大学习率设置为 0.000 5。两个网络损失函数采用 *Smooth L*1，批量大小设置为 24，最终在 SceneFlow 数据集上训练 90 个周期，以端点误差(EPE)作为评估指标。

2.3　实验结果

表 1 将 Efficient-ACVNet 与 SceneFlow 数据集上的其他几种先进实时立体匹配方法进行了对比。Efficient-ACVNet 在参数量上牺牲少量性能，从而实现精度和运行时间的大幅提升。在精度方面，Efficient-ACVNet 的端点误差(EPE)达到 0.58，相比基准网络 Fast-ACVNet 下降 9.375%，表明网络通过 3D 代价体注意力以及引入逆瓶颈残差块使网络代价体的信息更加丰富。在速度方面，Efficient-ACVNet 的运行时间为 25 ms，比基准网络 Fast-ACVNet 的 39 ms 快 14 ms，表明网络通过 3D 代价体注意力以及伪 3D 卷积模块使网络的代价体更加简洁高效。

表 1　在 SceneFlow 数据集上对比其他主流实时立体匹配网络

Table 1　Comparison with state-of-the-art real-time stereo matching networks on SceneFlow dataset

算法	FLOPs/G	参数量/M	EPE	耗时/ms
StereoNet	85.93	0.40	1.10	20
CoEx	53.39	2.72	0.67	36
AANet	152.86	2.97	0.87	93
HITNet	50.23	0.42	**0.55**	36
LightStereo-S	22.71	3.44	0.73	**17**
LightStereo-M	36.36	7.64	0.62	23
Fast-ACVNet	79.34	3.08	0.64	39
Fast-ACVNet＋	93.08	3.20	0.59	45
本文算法	76.94	4.01	0.58	25

图 7 展示了本文算法和基准网络在 SceneFlow 数据集上的部分结果对比，所提算法有着比基准网络更为丰富的细节。

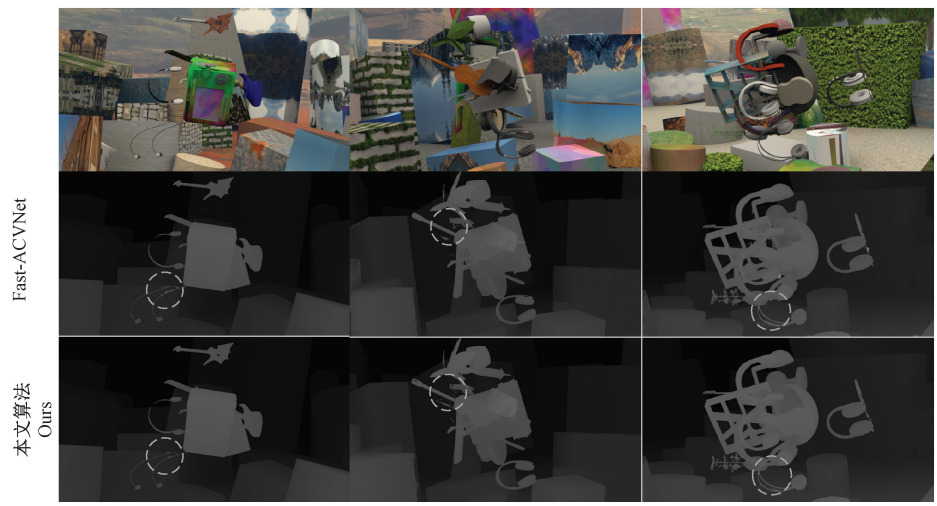

图 7　在 SceneFlow 数据集上的部分表现

Fig. 7　Partial performance evaluation on SceneFlow dataset

表 2 为 Efficient-ACVNet 与其他几种先进实时立体匹配方法在 KITTI2012 和 KITTI2015 在线基准测试的对比结果。在 KITTI2012 的 3 像素和 4 像素的非遮挡区域(noc)和所有(all)像素的错误像素百分比指标中，Efficient-ACVNet 相比基准网络分别下降 9.5%、9.8%、6.5%和 9.6%。在 KITTI2015 的各项评估指标中，Efficient-ACVNet 相比于基准网络，对于背景区域(D1-bg)、前景区域(D1-fg)和所有像素(D1-all)的视差异常值百分比的指标分别下降 4.3%、19%和 8.3%，并且在 D1-fg 指标上取得了最佳成绩。在 KITTI 数据集中，网络的运行时间比基准网络快 14 ms。网络提出的 3D 代价体注意力、逆瓶颈残差块和伪 3D 卷积的网络架构，在 KITTI 数据集基准测试中，相比基准网络同样展现出精度和计算效率的提升。

图 8 展示本文算法和优化网络 Fast-ACVNet＋在 KITTI 数据集上的部分结果对比，所提算法有相比优化网络在前景区域的表现更加平滑且对于场景中物体的结构

表2　在KITTI数据集上对比其他主流实时立体匹配网络

Table 2　Comparison with state-of-the-art real-time stereo matching networks on KITTI dataset

算法	KITTI2012						KITTI2015			耗时
	3-noc	3-all	4-noc	4-all	EPE-noc	EPE-all	D1-bg	D1-fg	D1-all	/ms
StereoNet	—	—	—	—	0.8	0.9	4.30	7.45	4.83	22
CoEx	1.55	1.93	1.15	1.42	0.5	0.5	1.79	3.82	2.13	33
AANet	1.91	2.42	1.46	1.87	0.5	0.6	1.99	5.39	2.55	62
HITNet	1.41	1.89	1.14	1.53	0.4	0.5	1.73	3.20	1.98	54
LightStereo-S	1.88	2.34	1.30	1.65	0.6	0.6	2.00	3.80	2.30	17
LightStereo-M	1.56	1.91	1.10	1.36	0.5	0.5	1.81	3.22	2.04	23
Fast-ACVNet	1.68	2.13	1.23	1.56	0.5	0.6	1.82	3.93	2.17	39
Fast-ACVNet+	1.45	1.85	1.06	1.36	0.5	0.5	1.70	3.53	2.01	45
本文算法	1.52	1.92	1.15	1.41	0.5	0.6	1.74	3.18	1.99	25

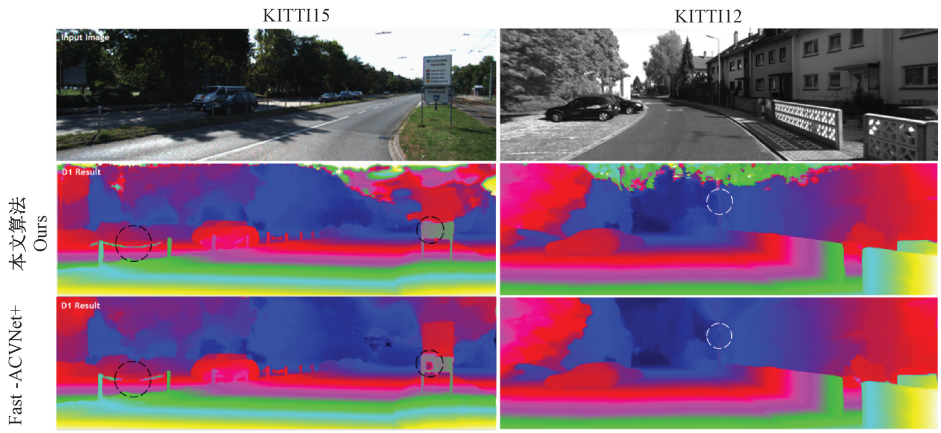

图8　在KITTI数据集上的部分表现

Fig. 8　Partial performance evaluation on KITTI dataset

细节更为丰富。

表3展示本文算法的泛化性结果,相较于基准网络,本文算法在KITTI2012数据集上的视差异常值百分比(D1)下降10.5%,在KITTI2015数据集上的视差异常值百分比(D1)下降14.7%,在Middlebury2014数据集上的误差大于2像素的像素百分比下降3.9%。并且对比其他几种实时立体匹配算法中,本文算法的泛化性在KITTI2和Middlebury2014数据集取得领先。

表3　网络泛化性表现

Table 3　Network generalization performance

算法	KITTI12 D1/%	KITTI15 D1/%	Middlebury 2014>2 px
CoEx	13.5	11.6	25.51
LightStereo-S	11.57	9.01	19.63
Fast-ACVNet	12.4	10.6	20.13
本文算法	11.10	9.04	19.34

2.4　消融实验

1) 2D代价聚合分析

如表4所示,其中实验a～c分析逆瓶颈残差块里不同大小通道扩展因子对匹配精度的影响;实验d～f分析不同数量逆瓶颈残差块配置对匹配精度的影响;实验g和h分析逆瓶颈残差块的有效性;实验g和i分析多尺度视差通过注意力模块的有效性。

表4中实验a～c分析了当固定扩展因子为4时,逆瓶颈残差块的块配置分别设置为(1, 1, 1, 1, 1)、(1, 2, 4, 2, 1)和(2, 4, 8, 4, 2)时对网络精度的影响。结果表明,当块配置最小为(1, 1, 1, 1, 1)的实验a,其端点误差(EPE)最高,为0.650 7。当块配置最大为(2, 4, 8, 4, 2)的实验c,其端点误差最低,为0.570 6。结果表明,逆瓶颈残差块在每层分辨率上的配置增多时,会显著提升网络的精度,但同时也会增加网络的运行时间和计算量,运行时间从24.52 ms增加到28.36 ms,每秒浮点运算次数(FLOPs)从73.79 G增加到83.34 G。最终,为了平衡精度和效率,本文选择块数量配置为(1,2,4,2,1)。

表4中实验d～f分析了当逆瓶颈残差块固定数量配

表 4　消融实验

Table 4　Ablation experiments

序号	块配置	拓展因子	ConvNeXtV2	MSDCAM	P3D	EPE	FLOPs/G	参数量/M	耗时/ms
a	(1 1 1 1 1)	(4 4 4 4 4)	√	√	√	0.650 7	73.79	2.94	24.52
b	(1 2 4 2 1)	(4 4 4 4 4)	√	√	√	0.584 4	76.94	4.01	25.34
c	(2 4 8 4 2)	(4 4 4 4 4)	√	√	√	0.570 6	83.34	5.59	28.36
d	(1 2 4 2 1)	(2 2 2 2 2)	√	√	√	0.591 4	76.45	3.94	23.21
e	(1 2 4 2 1)	(4 4 4 4 4)	√	√	√	0.584 4	76.94	4.01	25.34
f	(1 2 4 2 1)	(8 8 8 8 8)	√	√	√	0.582 3	77.91	4.16	28.27
g	(1 2 4 2 1)	(4 4 4 4 4)	√	√	√	0.584 4	76.94	4.01	25.34
h	(1 2 4 2 1)	(4 4 4 4 4)	—	√	√	0.616 7	77.35	3.99	24.36
i	(1 2 4 2 1)	(4 4 4 4 4)	√	—	√	0.612 6	76.40	3.91	24.22
j	(1 2 4 2 1)	(4 4 4 4 4)	√	√	—	0.588 7	79.23	4.15	25.54

置为(1，2，4，2，1)时，不同大小通道扩展因子的逆瓶颈残差块对网络精度的影响。结果表明，随着通道扩展因子由 2 增加到 4 再增加到 8 时，端点误差(EPE)从 0.591 4 下降到 0.584 4 再下降到 0.582 3，网络的精度得到提高。说明在代价聚合过程中，视差通道的通道数会影响网络的精度。然而，精度提升会造成运行时间和计算量的增加，每秒浮点运算次数(FLOPs)从 76.45 G 增加到 76.94 G 再增加到 77.91 G，运行时间也从 23.21 ms 增加到 25.34 ms 再增加到 28.27 ms。当拓展因子由 4 增加到 8 时，网络的精度提升并不大，为此，设定网络的拓展因子为 4，实现网络在精度和效率的平衡。

表 4 中实验 g 和 h 分析引入逆瓶颈残差块的有效性。实验 g 使用逆瓶颈残差块拓展因子为 4、块数量配置为(1，2，4，2，1)的网络为基准，其端点误差(EPE)为 0.584 4。作为对比，使用 Guo 等[15]的代价聚合方法，即使用 MobileNet V2[16]的倒置残差块进行代价聚合，倒置残差块的拓展因子为 4，块数量配置为(1，2，4，2，1)。实验表明，在使用倒置残差块时，端点误差为 0.616 7，引入逆瓶颈残差块后，端点误差下降到 0.584 4。这表明逆瓶颈残差块的大感受野和逆瓶颈操作能够使 2D 代价聚合性能更佳。

表 4 中实验 g 和 i 分析多尺度视差通过注意力模块的有效性。实验 i 使用拓展因子为 4、块数量配置为(1，2，4，2，1)的网络为基准，其端点误差(EPE)为 0.612 6。在加入多尺度视差通道注意力模块(MSDCAM)后，如实验 g，端点误差进一步降至 0.584 4，且每秒浮点运算次数(FLOPs)仅从 76.40 G 增加到 76.94 G，运行时间仅从 24.22 ms 增加到 25.34 ms。这表明 MSDCAM 在对计算效率影响较小的情况下，能够利用多尺度图像特征中蕴含的结构语义信息来指导和增强代价聚合过程，提升网络精度。

2) 3D 代价聚合分析

表 4 中实验 g 和 j 分析伪 3D 卷积块的有效性。实验 g 使用伪 3D 残差块和伪 3D 下采样模块堆叠的沙漏结构为基准，其端点误差(EPE)为 0.584 4。使用基准网络 Fast-ACVNet 的 3D 代价聚合网络作为对比，其网络由 6 个 3D 卷积和 2 个 3D 反卷积组成。实验表明，伪 3D 卷积块替换 3D 卷积后，端点误差由 0.588 7 下降到 0.584 4，计算浮点数由 79.23 G 下降到 76.94 G，参数量也从 4.15 M 下降到 4.01 M。结果表明，伪 3D 残差块和伪 3D 下采样模块相比标准 3D 卷积在提升网络精度的情况下，还降低了网络的复杂度和参数量。

3) 伪 3D 卷积模块有效性分析

表 5 中分析伪 3D 模块在替换 StereoNet[13]和 Fast-ACVNet[8]的 3D 卷积层后的有效性。在伪 3D 加入 StereoNet 后，网络复杂度由 85.93 G 下降到 75.08 G，参数量由 0.4 M 下降到 0.35 M，在 EPE 上提高 0.03，运行时间增加 1 ms。在伪 3D 加入 Fast-ACVNet 后，网络复杂度由 79.34 G 下降到 76.50 G，参数量由 3.08 M 下降到 2.91 M，网络在 EPE 没有提升，运行时间降低 14 ms。结果表明，在不同网络结构下，伪 3D 卷积模块在保持着标准 3D 卷积层精度的同时，可以降低网络的复杂度和参数量。

表 5　伪 3D 卷积的有效性

Table 5　Effectiveness of pseudo-3D convolution(P3D)

算法	FLOPs/G	参数量/M	EPE	耗时/ms
StereoNet-P3D	75.08	0.35	1.13	21
Fast-ACVNet-P3D	76.50	2.91	0.64	25
本文算法	76.94	4.01	0.58	25

3　结　　论

为构建信息丰富且计算高效的代价体和实现高精度高效率的立体匹配，在 Fast-ACVNet 网络的基础上进行改进提出精度更高、运行时间更快的 Efficient-ACVNet。实验结果表明，Efficient-ACVNet 通过 3D 代价体生成的代

价体注意力对4D代价体进行构建和滤波,实现代价体信息丰富且简洁高效,提高了网络效率。为了提升3D代价体注意力的性能,引入逆瓶颈残差块和多尺度视差通道注意力模块增强3D代价体的聚合性能,提升对4D代价体的滤波效果;为保证网络的轻量性,引入伪3D残差块并设计伪3D下采样模块,降低网络的计算成本和复杂度,实现高效的实时立体匹配。然而,3D代价体的内容信息相对有限,且对于训练无标签区域的数据集时,预测视差在无标签区域处表现糟糕。对于目前,标签数据获取成本太高,未来将研究无监督的立体匹配算法。

参考文献

[1] 朱代先,巩若琳,孔浩然,等. 基于注意力机制的多视图立体重建算法[J]. 电子测量技术, 2024, 47(16): 130-138.
ZHU D X, GONG R L, KONG H R, et al. Multi-view stereo reconstruction algorithm based on attention mechanism [J]. Electronic Measurement Technology, 2024, 47(16): 130-138.

[2] 周志伟,周建江,王佳宾,等. 基于雷达和视觉多级信息融合的目标检测网络[J]. 电子测量技术, 2024, 47(24): 110-117.
ZHOU ZH W, ZHOU J J, WANG J B, et al. Target detection network based on multi level information fusion of radar and vision[J]. Electronic Measurement Technology, 2024, 47(24): 110-117.

[3] 李冰,王豪伟,韩宇辰,等. 基于双目视觉的拖车钩检测与定位方法研究[J]. 电子测量技术, 2024, 47(3): 1-8.
LI B, WANG H W, HAN Y CH, et al. Research on tow hook detection and location method based on binocular vision [J]. Electronic Measurement Technology, 2024, 47(3): 1-8.

[4] SCHARSTEIN D, SZELISKI R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms [J]. International Journal of Computer Vision, 2002, 47(1): 7-42.

[5] 尹晨阳,职恒辉,李慧斌. 基于深度学习的双目立体匹配方法综述[J]. 计算机工程, 2022, 48(10): 1-12.
YIN CH Y, ZHI H H, LI H B. Survey of binocular stereo-matching methods based on deep learning[J]. Computer Engineering, 2022, 48(10): 1-12.

[6] ZBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches [J]. Journal of Machine Learning Research, 2016, 17(1): 2287-2318.

[7] MAYER N, ILG E, HAUSSER 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. Las Vegas, NV, USA. New York: IEEE Press, 2016: 4040-4048.

[8] XU G W, WANG Y, CHEN J D, et al. Accurate and efficient stereo matching via attention concatenation volume[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2461-2474.

[9] KENDALL A, MARTIROSYAN H, DASGUPTA S, et al. End-to end learning of geometry and context for deep stereo regression [C]. IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 66-75.

[10] 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. Long Beach, CA, USA: IEEE Press, 2020: 3268-3277.

[11] 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. Nashville, TN, USA: IEEE, 2021: 13901-13910.

[12] TANKOVICH V, HANE C, ZHANG Y D, et al. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021: 14357-14367.

[13] 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. Munich, Germany: Springer, 2018: 596-613.

[14] XU H F, ZHANG J Y. AANet: Adaptive aggregation network for efficient stereo matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 1956-1965.

[15] GUO X D, ZHANG CH M, ZHANG Y M, et al. Lightstereo: Channel boost is all your need for efficient 2d cost aggregation[EB/OL]. (2024-06-28) [2025-02-13]. https://arxiv.org/abs/2406.19833.

[16] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, 2018: 4510-4520.

[17] WOO S, DEBNATH S, HU R H, et al. ConvNeXt

V2: Co-designing and scaling convnets with masked autoencoders [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE, 2023: 16133-16142.

[18] QIU ZH F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]. IEEE Conference on Computer Vision. Venice, Italy: IEEE, 2017: 5534-5542.

[19] PENG CH, ZHANG X Y, YU G, et al. Large kernel matters—improve semantic segmentation by global convolutional network [C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 1743-1751.

[20] HOU Q B, ZHANG L, CHENG M M, et al. Strip pooling: Rethinking spatial pooling for scene parsing[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 4002-4011.

[21] GUO M H, LU CH Z, HOU Q B, et al. SegNeXt: Rethinking convolutional attention design for semantic segmentation[C]. 36th International Conference on Neural Information Processing Systems. New Orleans, LA, USA: ACM, 2022: 1140-1156.

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

[23] WANG Y H, WU ZH J, DAI J, et al. Evaluating robotic-assisted partial nephrectomy surgeons with fully convolutional segmentation and multi-task attention networks[J]. Journal of Robotic Surgery, 2023, 17(5):2323-2330.

[24] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012: 3354-3361.

[25] MENZE M, HEIPKE C, GEIGER A. Joint 3D estimation of vehicles and scene flow[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, II-3/W5: 427-434.

[26] SCHARSTEIN D, HIRSCHMULLER H, KITAJIMA Y, et al. High-resolution stereo datasets with subpixel-accurate ground truth [C]. German Conference on Pattern Recognition. Munster, Germany: Springer, 2014: 31-42.

作者简介

李冰，博士，副教授，主要研究方向为模式识别与电力视觉。

E-mail: li_bing_hb@126.com

严熠萌，硕士研究生，主要研究方向为模式识别与计算机视觉。

E-mail: 1437319613@qq.com

张鑫磊，硕士研究生，主要研究方向为模式识别与计算机视觉。

E-mail: 2398812564@qq.com

邵宝文，硕士研究生，主要研究方向为模式识别与计算机视觉。

E-mail: 704755757@qq.com

翟永杰(通信作者)，博士，教授，主要研究方向为模式识别与电力视觉。

E-mail: zhaiyongjie@ncepu.edu.cn