DOI:10.19651/j.cnki.emt.2520067

融合多核注意力解码器的立体匹配算法*

苏志豪 于红绯 曹 杨
(辽宁石油化工大学人工智能与软件学院 抚顺 113001)

摘 要:针对在反射区域匹配精度下降的问题,提出了一种基于多核注意力解码器和代价体动态增强机制的立体匹配算法。首先在特征提取网络中使用多核注意力解码器,该解码器采用并行深度卷积捕捉不同分辨率的空间细节,强化算法局部细节和全局信息的融合能力,提升在反射区域的视差预测精度。接着算法采用代价体动态增强机制来优化代价体,根据全局上下文信息选择性融合不同尺度的代价体,避免冗余信息叠加,从而增强代价体对反射区域的建模能力,缓解在反射区域匹配精度下降的问题。在 Scene Flow 数据集实验结果表明,所提方法在 EPE 和 D1 指标上分别达到 0.45 和 2.40%,在 KITTI 数据集反射区域上 3-All 指标达到 4.59%,与基线方法相比降低了 8.2%。表明所提方法能有效提升在反射区域的匹配精度,并且模型参数量也有所下降。

关键词:双目立体匹配;端到端视差估计;反射区域;深度卷积;代价体优化
中图分类号:TP391.41;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.6040

Stereo matching algorithm with multi-kernel attention decoder

Su Zhihao Yu Hongfei Cao Yang
(College of Artificial Intelligence and Software, Liaoning Petrochemical University, Fushun 113001, China)

Abstract: To address the issue of degraded matching accuracy in reflective regions, a stereo matching algorithm is proposed based on a multi-kernel attention decoder and a cost volume dynamic enhancement mechanism. First, the feature extraction network employs a multi-kernel attention decoder, which utilizes parallel depth-wise convolutions to capture spatial details at different resolutions. This enhances the algorithm′s ability to fuse local details and global information, thereby improving disparity prediction accuracy in reflective regions. Next, the algorithm adopts a cost volume dynamic enhancement mechanism to optimize the cost volume. By selectively fusing multi-scale cost vol-umes based on global contextual information, it avoids redundant information accumulation, thus strengthening the cost volume′s modeling capability in reflective regions and mitigating the accuracy degradation problem. Experi-mental results on the Scene Flow dataset show that the proposed method achieves 0.45 EPE and 2.40% D1, while on reflective regions of the KITTI dataset, it attains 4.59% 3-All error, representing an 8.2% reduction com-pared to baseline methods. These results demonstrate that the proposed approach effectively improves matching accuracy in reflective regions while also reducing model parameters.

Keywords: binocular stereo matching; end-to-end disparity estimation; reflective regions; depth-wise convolution; cost volume optimization

0 引 言

立体匹配作为计算机视觉领域的核心任务之一,旨在通过分析左右视图的像素对应关系计算视差图,进而实现场景的深度感知,在自动驾驶、机器人导航和三维重建等领域具有重要应用价值[1-2]。传统立体匹配方法通常分为代价计算、代价聚合、视差计算和视差优化 4 个步骤[3],并基于优化范围的不同分为全局匹配[4]、半全局匹配[5]和局部匹配[6]等方法。尽管传统方法在早期展现出较好的性能,但其依赖手工设计的特征和启发式规则[7-8],在反射区域中面临计算复杂度高、泛化能力差等问题。

基于深度学习的立体匹配方法显著提升了匹配精度和鲁棒性。早期工作如 MC-CNN[9] 采用卷积神经网络(convolutional neural network, CNN)替代传统代价计算,

收稿日期:2025-10-20
* 基金项目:辽宁省教育厅高校基本科研项目(LJ212510148003)资助

通过数据驱动学习特征相似性,但需独立训练各模块,难以全局优化。随后,端到端方法(如 DispNet[10])通过单一网络直接从图像对预测视差图实现了联合优化。PSMNet[11]进一步引入金字塔池化模块和 3D 卷积,增强了对上下文信息的利用,但高分辨率的 3D 代价体积导致计算和内存开销剧增。为提升效率,GWC-Net[12]通过分组相关(group-wise correlation)优化代价体积计算,平衡了精度与速度。虽然基于 CNN 的方法显著提升了立体匹配任务的性能,但由于网络难以充分提取上下文特征和全局信息,在反射区域下进行视差预测仍面临挑战。

近年来,基于迭代优化的方法成为研究热点。例如,RAFT-Stereo[13]提出利用从全对相关性中检索的局部代价值来循环更新视差场,然而由于全对相关性缺乏全局信息,在处理反射区域时仍存在局限性。孟凡娜等[14]提出在上下文网络中引入自适应空间卷积块,通过加权聚合多个卷积核的响应,增强对复杂场景中病态区域的特征提取能力,但是由于上下文网络并不用于构建代价体,无法为代价体在反射区域提供更好的特征,所以在反射区域的匹配精度有一定的限制。IGEV-Stereo[15]为解决病态区域的局部模糊性问题,进一步设计几何编码体积模块,融合非局部几何信息,但其性能高度依赖初始特征提取的准确性,若全局信息不足容易使误差累积,从而影响在反射区域的匹配精度。

因此,为解决在反射区域的匹配精度下降问题,本文提出了一种基于多核注意力的立体匹配算法。本文的主要贡献为:1)提出一种面向立体匹配任务的多核注意力解码器,通过构建长距离分组门控融合器和双注意力引导特征增强模块,在保持网络参数量精简的同时实现局部与全局特征的有效聚合;2)设计了代价体动态增强机制,该机制通过建立代价体特征间的长程依赖关系,根据全局上下文信息选择性融合不同尺度的代价体,提升代价体与预测视差的相关性。

1 融合多核注意力解码器的立体匹配网络

针对目前立体匹配算法在反射区域匹配精度不足的问题,本文提出了一种基于多核注意力解码器和代价体动态增强机制的立体匹配算法。考虑到 IGEV-Stereo[15]网络在立体匹配任务中展现出高效迭代优化机制的优势,本研究以其为基准框架进行改进。通过使用多核注意力解码器和代价体动态增强机制,本算法能够更有效地处理反射区域下的匹配难题,在保持计算效率的同时显著提升了匹配精度。

1.1 本文网络架构

首先,选用轻量级的 MobileNetV2 作为特征提取主干网络。该选择主要基于两点考虑:其一,其倒残差结构与线性瓶颈设计能在保持较强特征表示能力的同时,显著减少模型参数与计算量,这符合构建高效立体匹配模型的初衷;其二,其多尺度、高分辨率的特征输出为后续设计的多核注意力解码器提供了丰富的空间细节信息,有利于反射区域等细节部分的特征增强。接着引入多核注意力解码器,通过构建长距离分组门控融合器和双注意力引导特征增强模块,在保持网络参数量精简的同时实现局部和全局特征的有效聚合。同时通过独立的上下文网络生成场景理解特征,为后续处理奠定基础。随后开始构建代价体(cost volume):先将左右图像特征按通道分组进行相关性计算,形成基础的匹配代价表示,再通过代价体动态增强机制(cost volume dynamic enhancement, CVDEM)正则化后得到几何编码体积(geometry encoding volume, GEV),利用代价体动态增强机制能够自适应地校准特征响应,从而提升代价体和视差的相关性。接着对 GEV 进行"Soft argmin"操作得到初始视差图,为后续优化提供高质量起点。最后,沿用了 IGEV-Stereo 的 ConvGRU 迭代优化过程,每步从代价体中检索多尺度匹配特征,结合当前视差估计和上下文信息进行渐进式更新,并通过空间上采样机制将低分辨率预测视差逐步细化到原图尺寸。本文算法架构如图 1 所示。

1.2 多核注意力解码器

IGEV-Stereo 的特征提取网络主要依赖传统卷积操作,可能对局部细节和全局信息的融合能力有限。尤其在纹理缺失和反射区域,局部细节的差异性较小。为此,本文提出了多核注意力解码器(multi-kernel attention decoder, MKAD),该多核注意力解码器主要由长距离分组门控融合器和双注意力引导特征增强模块构成,采用并行深度卷积捕捉不同分辨率的空间细节,有效缓解纹理缺失部分的匹配歧义,并抑制镜面反射导致的代价体噪声。其次,MKAD 通过多核卷积扩大感受野,聚合周围区域的上下文信息,为反射区域提供补充性特征表示,弥补了反射区域局部细节差异性不足的缺点,从而维持结构的连续性。并且与基准网络相比,本网络模型参数规模有所降低。该多核注意力解码器提高了对局部细节和全局信息的融合能力,并且提升了反射区域的匹配精度。MKAD 解码器整体结构如图 2 所示。

上文提到的长距离分组门控融合器(long-range grouped gating fuser, LGGF)如图 3(a)所示。该模块通过分组卷积捕获局部结构信息,并利用通道自适应的注意力权重重新校准特征重要性,从而提升对于立体匹配任务的敏感度。该模块以解码器输出的特征图 $\boldsymbol{F} \in \mathbf{R}^{C\times H\times W}$ 作为输入,首先通过分组卷积(Group=8,核大小 3×3)提取局部空间上下文特征,随后经过批归一化(BN)和 ReLU 激活函数增强非线性表达能力。接着利用 1×1 卷积压缩通道维度并通过 Sigmoid 函数生成通道注意力权重 $\boldsymbol{A} \in [0,1]^{C\times H\times W}$,最终与原始特征进行逐元素相乘。

作为前文所述多核注意力解码器的核心组件,双注意力引导特征增强模块(dual-attention guided feature enhancer, DGFE)如图 3(b)所示。具体而言,该模块通过

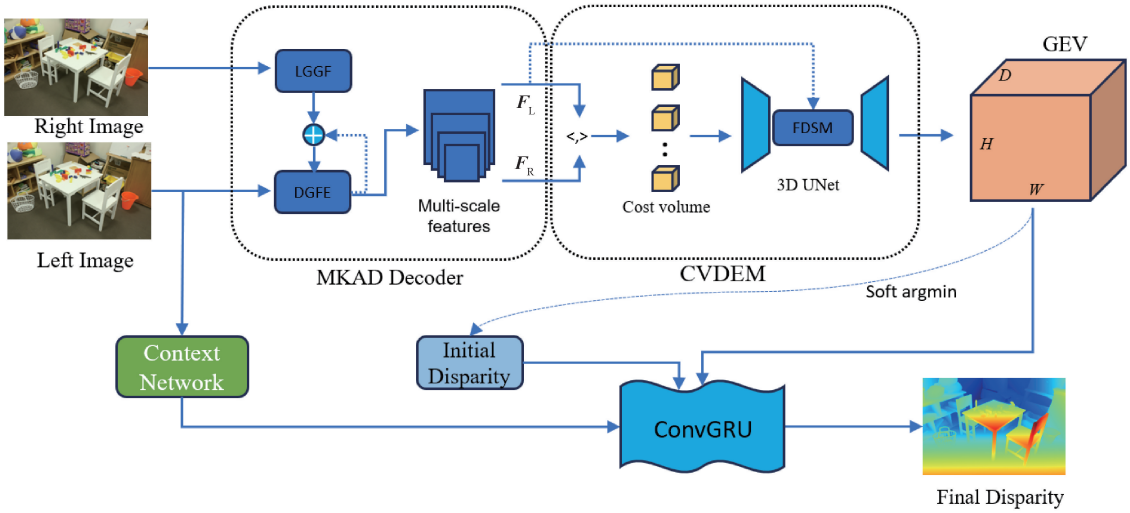

图 1　本文算法架构

Fig. 1　Architecture of the proposed algorithm

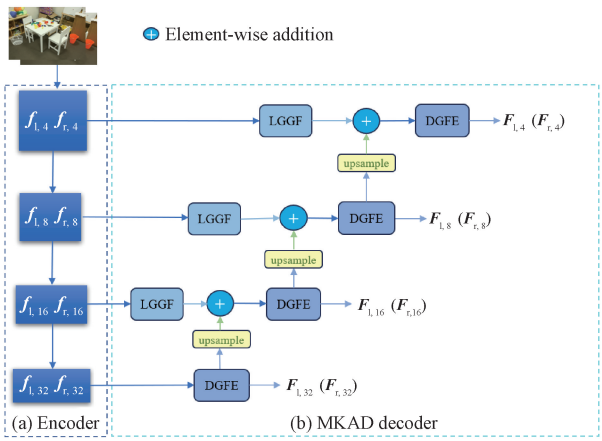

图 2　多核注意力解码器

Fig. 2　Multi-kernel attention decoder

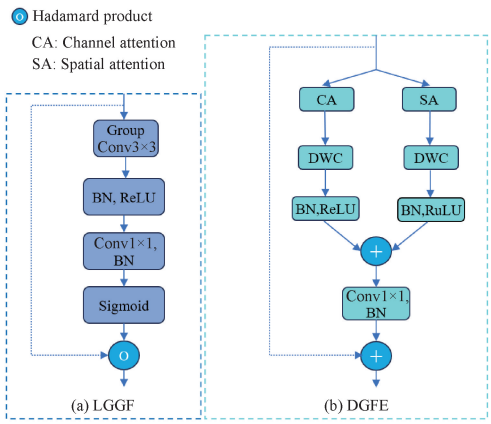

图 3　LGGF 和 DGFE 结构

Fig. 3　LGGF and DGFE architecture

并行的通道注意力(chamel attention,CA)和空间注意力(spatial attention,SA)分支来分别捕捉关键特征通道和重要空间区域的特征信息,其中 CA 分支采用 3×3 深度可分离卷积处理通道加权特征,SA 分支使用 5×5 深度可分离卷积处理空间加权特征,两个分支经 BN 和 ReLU 激活后特征相加,再通过 1×1 卷积和 BN 层进行特征融合,最后与原始输入特征进行残差连接。这种设计既保留了原始特征信息,又通过双注意力机制和多尺度卷积增强了特征表达能力,有效提升了模型在复杂场景下的匹配精度。其结构如图 3(b)所示。

多核注意力解码器整体的执行步骤如下:给定左图和右图 $\boldsymbol{I}_{l(r)} \in \mathbf{R}^{3\times H\times W}$,首先应用在 ImageNet 上预训练的 MobileNetV2 将 $\boldsymbol{I}_{l(r)}$ 下采样至原始尺寸的 1/32,得到特征 $\boldsymbol{f}_{l,i}(\boldsymbol{f}_{r,i}) \in \mathbf{R}^{C_i\times\frac{H}{i}\times\frac{W}{i}}(i=4,8,16,32)$。接着将经过编码器后得到的多尺度特征 $\boldsymbol{f}_{l,i}(\boldsymbol{f}_{r,i})$ 输入到 MKAD,将其恢复至 1/4 尺度,得到特征 $\boldsymbol{F}_{l,i}(\boldsymbol{F}_{r,i}) \in \mathbf{R}^{C_i\times\frac{H}{i}\times\frac{W}{i}}(i=4,8,16,32)$,如图 2 所示。具体的,首先将 $\boldsymbol{f}_{l,32}(\boldsymbol{f}_{r,32})$ 作为输入送到 DGFE 模块,直接输出特征 $\boldsymbol{F}_{l,32}(\boldsymbol{F}_{r,32})$,并且对它进行上采样,得到 $\boldsymbol{F}^{up}_{l,32}(\boldsymbol{F}^{up}_{r,32})$。接着将 $\boldsymbol{f}_{l,16}(\boldsymbol{f}_{r,16})$ 送入到 LGGF 模块,将其输出与 $\boldsymbol{F}^{up}_{l,32}(\boldsymbol{F}^{up}_{r,32})$ 相加,然后再经过 DGFE 模块得到 $\boldsymbol{F}_{l,16}(\boldsymbol{F}_{r,16})$。同理,可得到 $\boldsymbol{F}_{l,8}(\boldsymbol{F}_{r,8})$ 和 $\boldsymbol{F}_{l,4}(\boldsymbol{F}_{r,4})$。

1.3　代价体动态增强机制

在立体匹配任务中,代价体(cost volume)作为匹配代价的密集表征结构,其构建与优化对最终视差估计的精度至关重要。代价体通过构建一个三维的匹配代价空间(高度×宽度×视差),评估了左右图像中每个像素在所有可能视差下的匹配相似度。它通过枚举所有可能的视差假设,构建了一个全局搜索空间,确保每个像素的匹配代价

均被显式计算。尽管单个像素的匹配代价可能不可靠,但代价体保留了所有候选视差的代价分布,为后续优化提供了充分的信息基础。然而,原始代价体往往受到多种因素的干扰,这些干扰使得代价体在局部区域可能出现噪声,进而导致视差估计的偏差。正则化技术的引入成为优化代价体的关键环节。正则化的核心思想是通过引入先验约束或上下文信息,对原始代价体进行优化,从而抑制异常值并增强匹配代价的可靠性。随着深度学习的发展,基于 3D 卷积网络的正则化方法展现出显著优势。这类方法将代价体视为三维特征空间,通过堆叠的 3D 卷积层进行多层次的特征融合。但是这类方法通常仅依赖局部卷积操作,难以有效建模全局结构,且在反射和弱纹理区域易产生模糊性。

因此,本文提出了一种 CVDEM 机制,将 3DUnet 和本文设计的特征动态选择模块(feature dynamic selection module,FDSM)相结合,增强代价体的全局建模能力。该机制与 IGEV-Stereo 原始正则化模块的核心区别在于:原始方法主要依赖堆叠的 3D 卷积进行局部特征融合,而 CVDEM 通过引入 FDSM,建立了代价体特征间的长程依赖关系,实现了基于全局上下文信息的多尺度代价体自适应加权融合。这种设计能够根据图像内容动态校准特征响应,避免冗余信息叠加,从而显著提升代价体在反射区域的特征质量及其与预测视差的相关性。FDSM 结构如图 4 所示。

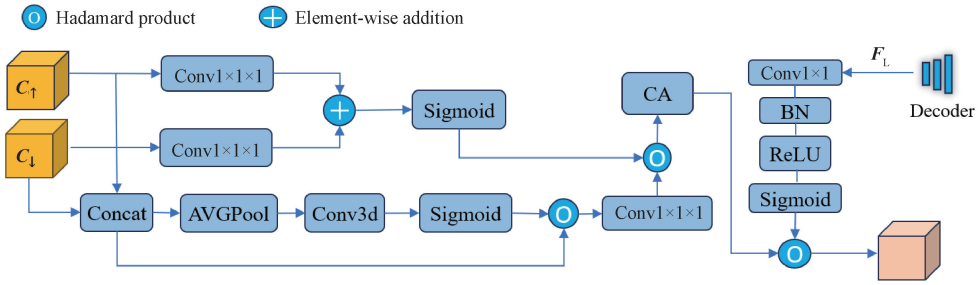

图 4 特征动态选择模块

Fig. 4 Feature dynamic selection module

使用特征 $\boldsymbol{F}_{\mathrm{l}}$ 和 $\boldsymbol{F}_{\mathrm{r}}$ 计算内积得到相似特征代价体 $\boldsymbol{C}\in\mathbf{R}^{C\times D\times\frac{H}{4}\times\frac{W}{4}}$,然后使用轻量级 3D 正则化网络进一步处理 $\boldsymbol{C}$,3D 正则化网络基于轻量级 3D UNet,该 UNet 由 3 个下采样块和 3 个上采样块组成。每个下采样块由两个 3×3×3 的 3D 卷积组成,每个上采样块由一个 4×4×4 的 3D 转置卷积和两个 3×3×3 的 3D 卷积组成。利用经过下采样和上采样分别得到的代价体 $\boldsymbol{C}_{\downarrow}$ 和 $\boldsymbol{C}_{\uparrow}$,作为代价体动态增强机制的输入。将 $\boldsymbol{C}_{\downarrow}$ 和 $\boldsymbol{C}_{\uparrow}$ 沿通道维度拼接得到 $\boldsymbol{C}_{\mathrm{c}}\in\mathbf{R}^{2C\times D\times\frac{H}{4}\times\frac{W}{4}}$,通过全局平均池化压缩空间和视差维度,生成全局统计量 $\mathrm{AvgPool}(\boldsymbol{C}_{\mathrm{c}})\in\mathbf{R}^{2C\times1\times1\times1}$,然后使用 1×1×1 卷积和 Sigmoid 激活生成通道注意力权重 $\boldsymbol{W}_1$。将注意力权重应用于拼接后的特征,通过 1×1×1 卷积将通道数从 $2C$ 降为 C 得到 $\boldsymbol{C}_{\boldsymbol{W}1}$。公式表示为式(1)和(2)。

$$\boldsymbol{W}_1=\sigma(\mathrm{Conv3d}(\mathrm{AvgPool}(\boldsymbol{C}_{\mathrm{c}})))\in\mathbf{R}^{2C\times1\times1\times1} \tag{1}$$

$$\boldsymbol{C}_{\boldsymbol{W}1}=\mathrm{Conv3d}(\boldsymbol{C}_{\mathrm{c}}\odot\boldsymbol{W}_1)\in\mathbf{R}^{C\times D\times\frac{H}{4}\times\frac{W}{4}} \tag{2}$$

另外,对输入代价体 $\boldsymbol{C}_{\downarrow}$ 和 $\boldsymbol{C}_{\uparrow}$ 分别使用 1×1×1 卷积生成单通道权重 $\boldsymbol{W}_2$,将权重 $\boldsymbol{W}_2$ 与降维后的特征 $\boldsymbol{C}_{\boldsymbol{W}1}$ 逐元素相乘,得到经过特征动态增强的代价体 $\boldsymbol{C}_{\boldsymbol{W}2}$。其公式表示为式(3)和(4)。最后通过原始特征的权重来引导激励代价体,得到最终的结果。

$$\boldsymbol{W}_2=\sigma(\mathrm{Conv3d}(\boldsymbol{C}_{\downarrow})+\mathrm{Conv3d}(\boldsymbol{C}_{\uparrow}))\in\mathbf{R}^{1\times D\times\frac{H}{4}\times\frac{W}{4}} \tag{3}$$

$$\boldsymbol{C}_{\boldsymbol{W}2}=\boldsymbol{C}_{\boldsymbol{W}1}\odot\boldsymbol{W}_2\in\mathbf{R}^{C\times D\times\frac{H}{4}\times\frac{W}{4}} \tag{4}$$

1.4 损失函数

本文沿用 IGEV-Stereo 设计损失函数,该函数计算从 GEV 回归的初始视差 d_0 上的平滑 L1 损失[5]:

$$\mathcal{L}_{\mathrm{init}}=\mathrm{Smooth}_{\mathrm{L}_1}(d_0-d_{\mathrm{gt}}) \tag{5}$$

其中,d_{gt} 表示视差真实值。计算所有预测视差 $\{d_i\}_{i=1}^{N}$ 的 L1 损失,并且遵循以指数方式增加权重,总损失定义为:

$$\mathcal{L}_{\mathrm{stereo}}=\mathcal{L}_{\mathrm{init}}+\sum_{i=1}^{N}\gamma^{N-i}\parallel d_i-d_{gt}\parallel_1 \tag{6}$$

其中,$\gamma=0.9$,并且 d_{gt} 表示真实视差值。

2 实验与分析

2.1 数据集与评价指标

本研究在两个立体匹配基准数据集 Scene Flow[8] 和 KITTI[16-17] 上进行实验验证。Scene Flow[8] 大规模合成数据集,包含 35 454 个训练样本和 4 370 个测试样本,提供密集视差真值。选用 Finalpass 版本(含运动模糊和离焦效果)以增强模型在真实场景的泛化能力。KITTI 真实驾驶场景数据集,其中又分为 KITTI2012[16] 和 KITTI2015[17],共包含 392 组训练样本和 392 组测试样本。本文的定量评价指标采用端点误差(EPE)和误匹配率(D1),在 KITTI 数据集中根据区域的不同分别评价背景区域(bg)、前景区域(fg)、非遮挡区域(noc)以及全部区域(all)。

2.2　实验环境与参数配置

本文模型基于 PyTorch 深度学习框架实现，使用 NVIDIA RTX 3090 GPU 进行训练。对于所有训练，本文使用 AdamW[18]优化器，并将梯度裁剪到[－1，1]。首先在 Scene Flow 数据集上做预训练。训练参数如表 1 所示。

表 1　预训练模型参数

Table 1　Parameters of the pre-trained model

参数	值
训练步数	400 k
学习率	0.000 2
迭代轮次	22
批量大小	4

其次本文使用 KITTI2012 和 KITTI2015 混合数据集在预训练模型的基础上做微调训练，训练 100 k 步。随机裁剪图像到 320×736，并使用数据增强进行训练。对于所有实验，使用单周期学习率调度，学习率为 0.000 2，并在训练过程中使用 22 个更新迭代。

2.3　消融实验与分析

为验证 MKAD 与 CVDEM 在本文网络的整体效果，本文在 KITTI 数据集上进行了系统性的消融实验。所有实验模型 steps 设置为 50 000。以 IGEV-Stereo 为基准模型，分别使用 MKAD 与 CVDEM 进行性能验证。

首先，将特征提取网络替换为本文设计的 MKAD，其余结构保持不变，添加 MKAD 之后，EPE 减小了 0.07，D1 指标降低了 1.0%。随后，仅添加 CVDEM。最后在同时添加 MKAD 和 CVDEM 后，EPE 减小了 0.08，D1 指标降低了 1.5%。实验结果如表 2 所示，表 2 中 IGEV-Stereo 表示完整的基准网络架构，IGEV-Stereo＋MKAD 表示将 IGEV-Stereo 的特征提取网络替换为 MKAD，IGEV-Stereo＋CVDEM 表示在 IGEV-Stereo 中引入 CVDEM。

表 2　消融实验

Table 2　Ablation study

消融模型	EPE/pixel	D1/%	Param/M
IGEV-Stereo	0.32	6.3	12.60
IGEV-Stereo＋MKAD	0.25	5.3	12.55
IGEV-Stereo＋CVDEM	0.26	5.1	12.71
IGEV-Stereo＋MKAD＋CVDEM(本研究)	0.24	4.8	12.51

为了深入分析 MKAD 的内部工作机制，本研究对其核心组件进行了系统的消融实验研究，如表 3 所示。实验结果表明，单独使用 LGGF 模块时，模型在 EPE 指标上达到 0.29，D1 为 5.8%，这一结果验证了 LGGF 能有效增强解码器对局部特征的感知能力并抑制噪声响应。而仅采用 DGFE 模块时，EPE 和 D1 分别降至 0.28 和 5.6%，说明双注意力引导特征增强模块能有效聚合周围的上下文信息，为反射区域提供补充性特征表示。值得注意的是，当完整集成 LGGF 和 DGFE 模块构成最终 MKAD 结构时，模型性能获得进一步显著提升，EPE 和 D1 分别优化至 0.25 和 5.3%，这一结果不仅明显优于单一模块配置，更充分证明了两种机制在特征表示优化方面具有显著的协同增强效应，能够更全面地捕捉和利用多层次特征信息。

表 3　仅改变 MKAD 结构的消融分析

Table 3　Ablation analysis of modifying only the MKAD

模型	EPE/pixel	D1/%
LGGF	0.29	5.8
DGFE	0.28	5.6
LGGF＋DGFE	0.25	5.3

为了验证所提出的 CVDEM 的有效性，本文以 IGEV-Stereo 作为基准模型，对比了不同的代价体正则化方案：原始 IGEV-Stereo 的 3D CNN 正则化模块、加入非局部注意力模块的变体(IGEV-Stereo＋Non-local)和以及本文提出的 CVDEM 方法。如表 4 所示，实验结果表明，与 3D CNN 正则化模块和非局部注意力模块相比，本文的 CVDEM 方法在 EPE 和 D1 指标上取得了更好的结果。

表 4　对 CVDEM 的有效性分析

Table 4　Analysis of CVDEM′s effectiveness

模型	EPE/pixel	D1/%
IGEV-Stereo	0.32	6.3
IGEV-Stereo＋Non-local	0.30	5.8
IGEV-Stereo＋CVDEM	0.26	5.1

2.4　方法对比

为测试所提方法在真实驾驶环境中的实际效果，本研究利用 Scene Flow、KITTI2012 和 KITTI2015 基准数据集进行了实验评估。根据表 5 的实验数据显示，本方法在 Scene Flow 数据上 EPE 指标达到了 0.45，比基准模型下降了 4.2%。另外，在 KITTI2015 测试集的整体区域(all)及非遮挡区域(noc)均取得优异表现，且两个区域的性能差异较小，这说明了算法对遮挡情况也具有良好的适应能力。

如表 6 所示，在 KITTI2015 测试集上的表现进一步验证了本方法的优越性。在整体区域(all)评估中，本方法 D1-all 指标达到 1.56%，优于 RAFT-Stereo(1.82%)和 CRE-Stereo(1.69%)。在非遮挡区域(noc)的 D1-all 指标更是达到 1.46%，与最近的方法 MDA(1.48%)相比仍有提升。特别值得关注的是，本方法在前景区域(D1-fg)的表

表 5 在 Scene Flow 测试集上与主流立体匹配方法的定量对比结果

Table 5 Quantitative comparison results with mainstream stereo matching methods on the Scene Flow test set

方法	EPE/pixel
PSMNet[11]	1.09
GwcNet[12]	0.76
GANet[19]	0.84
CSPN[20]	0.78
ACVNet[21]	0.48
IGEV-Stereo[15]	0.47
本研究	**0.45**

现尤为突出(2.62%),表明其前景区域的视差预测能力更强。

表 7 展示了在 KITTI2012 测试集整体区域(all)的详细对比结果。本方法在 6 个评价指标中全部取得最优,特别是在 3 pixel 误差阈值下的 noc 指标(1.06%)和 all 指标(1.40%)均领先于对比方法。EPE 指标与最优方法持平,本方法展现出更稳定的性能表现。

在反射区域(Reflective),如表 8 所示,本方法展现出优势。在 2 pixel 误差阈值下的 noc 指标(6.81%)和 all 指标(7.99%)均领先于对比方法,较次优的 ASC-Stereo 分别提升 4.5%和 0.2%。在 3 pixel 误差阈值下的表现更为突出,noc 指标(3.83%)较 IGEV-Stereo 提升 11.9%。平均

表 6 在 KITTI2015 测试集上与主流立体匹配方法的定量对比结果

Table 6 Quantitative comparison results with mainstream stereo matching methods on the KITTI2015 test set

方法	KITTI-2015(all)			KITTI-2015(noc)			Runtime/s
	D1-bg/%	D1-fg/%	D1-all/%	D1-bg/%	D1-fg/%	D1-all/%	
RAFT-Stereo[13]	1.58	3.05	1.82	1.45	2.94	1.69	0.38
CRE-Stereo[22]	1.45	2.86	1.69	1.33	2.60	1.54	0.41
IGEV-Stereo[15]	1.38	2.67	1.59	1.27	2.62	1.49	**0.18**
Los[23]	1.42	2.81	1.65	1.29	2.66	1.52	0.19
CroCo-Stereo[24]	1.38	2.65	1.59	1.3	2.56	1.51	0.93
MDA[25]	1.37	2.64	1.58	1.26	2.58	1.48	0.32
ASC-Stereo[26]	1.42	2.69	1.58	1.24	2.64	1.47	0.44
本研究	**1.35**	**2.62**	**1.56**	**1.24**	**2.58**	**1.46**	0.29

表 7 在 KITTI2012 测试集上与主流立体匹配方法的定量对比结果(all)

Table 7 Quantitative comparison results with mainstream stereo matching methods on the KITTI2012 test set (all) %

方法	KITTI-2012(all)					
	2-noc	2-all	3-noc	3-all	EPE-noc	EPE-all
RAFT-Stereo[13]	1.92	2.42	1.30	1.66	0.4	0.5
CRE-Stereo[22]	1.72	2.18	1.14	1.46	0.4	0.5
IGEV-Stereo[15]	1.71	2.17	1.12	1.44	0.4	0.4
MS-ACV[26]	1.75	2.30	1.11	1.44	0.4	0.5
DVANet[27]	1.78	2.39	1.09	1.52	0.4	0.5
HCR[28]	1.69	2.18	1.09	1.42	0.4	0.4
MDA[25]	1.76	2.26	1.09	1.43	0.4	0.5
本研究	**1.68**	**2.11**	**1.06**	**1.40**	**0.4**	**0.4**

误差(Avg)指标 0.9 更是所有方法中的最优结果,这充分证明本方法在强反射等复杂场景下的独特优势,有效克服了反射干扰问题。

2.5 可视化结果与分析

LGGF 门控机制定量分析。LGGF 模块的核心是通过 Sigmoid 函数生成通道注意力权重(即门控信号),并通过与原始特征相乘来重新校准特征重要性。为探究门控阈值的敏感性及其对反射区域的影响,本文定义了一个反射区域平均置信度指标:在 KITTI 数据集的反射区域掩码内,统计每张图像预测视差正确的像素占反射区域内所有像素的比例,然后求平均。该值越高说明模型对反射区域中所预测正确的像素越多。通过在后处理中引入一个阈值来模拟不同的门控严格程度:将 LGGF 输出的门控权重中低于阈值的部分置零。随后,绘制不同阈值下模型在数

表 8　在 KITTI2012 测试集上与其他方法的定量对比结果(Reflective)

Table 8　The quantitative comparison results with other methods on the KITTI 2012 test set (Reflective)　%

方法	KITTI-2012(Reflective)					
	2-noc	2-all	3-noc	3-all	Avg-noc	Avg-all
RAFT-Stereo[1]	8.41	9.87	5.40	6.48	—	—
CRE-Stereo[22]	9.71	11.26	6.27	7.27	1.4	1.4
IGEV-Stereo[15]	7.57	8.80	4.35	5.00	1.0	1.1
DVANet[27]	9.63	11.99	5.68	7.48	—	—
HCR[28]	9.78	11.93	6.01	7.68	1.2	1.3
MDA[25]	9.79	11.89	5.64	7.22	—	—
ASC-Stereo[26]	7.13	8.01	4.39	4.62	—	—
本研究	6.81	7.99	3.83	4.59	0.9	0.9

据集上的整体 D1 误差指标与在反射区域的平均置信度的变化曲线，如图 5 所示。

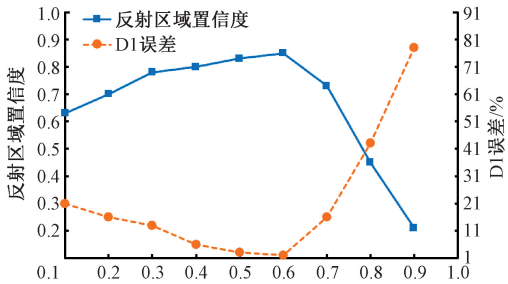

图 5　LGGF 门控阈值对模型性能的影响

Fig. 5　The impact of LGGF gating threshold on model performance

从图 5 中可以观察到：当阈值从 0 升至约 0.3 时，反射区域平均置信度显著上升，而整体 D1-all 误差稳步下降。这表明适度的门控可以有效地抑制低置信度的噪声特征，同时增强了对正确匹配有贡献的高置信度特征。这与 LGGF 模块的设计初衷完全一致。在阈值处于 0.3～0.6 时，模型同时保持了较高的反射区域置信度和较低的整体误差，说明门控机制在此范围内达到了一个稳定且高效的工作状态。最终模型采用的 Sigmoid 激活函数等效于使用了一个接近 0.6 的软阈值，正好位于此最优区间内。当阈值超过 0.6 后，整体误差开始回升。这是因为过高的阈值过度修剪了门控信号，导致一些虽有噪声但包含有效信息的特征通道也被完全关闭，引入了不必要的信息损失，反而损害了匹配性能。该定量分析强有力地证明了 LGGF 中的门控机制并非一个无关紧要的组件，而是一个关键的设计。它有针对性地抑制了不可靠的噪声影响，从而为提升在反射区域的匹配精度做出了直接贡献。

为深入探究 DGFE 模块在反射区域的有效性，本文将其内部的通道与空间注意力权重进行了可视化，结果如图 6 所示。

图 6　DGFE 模块注意力可视化

Fig. 6　Visualization of DGFE module attention

图 6 第 2 列展示了通道注意力的相应情况。可以观察到，DGFE 模块所侧重的特征通道能够将更高的权重分配给物体的边缘、轮廓以及受反射影响但仍有部分纹理信息的区域，而非均匀地处理整个反射区域。这使得模型能够从反射区域的周边或内部未被完全覆盖的纹理中提取有效的匹配线索，从而弥补了反射区域局部特征缺失的问题。图 6 第 3 列展示了空间注意力的分布。可以观察到，在存在高光反射区域的特征信息被显著抑制，而另一些对几何结构和纹理更敏感的局部区域则被增强。这表明 DGFE 的空间注意力机制能够自动甄别并削弱那些因反射而产生的、不稳定的特征，转而依赖那些受干扰较小、更能反映真实场景结构的特征信息。

DGFE 模块通过空间注意力过滤掉受高光污染的局部噪声特征，并通过通道注意力提取全局的边缘轮廓和纹理信息，并且包括了受反射影响但仍有部分纹理信息的区域。两者的协同工作，实现了对反射区域特征的自适应增强与净化，有效抑制了反射干扰，为后续的代价体构建提供了更加丰富的特征表示。

CVDEM 动态融合机制分析:为更加直观的展示 CVDEM 能够随图像反射强度的变化来提取全局上下文信息并选择性融合不同尺度的代价体,本文提供了其融合权重的可视化分析。选取 KITTI 数据集中包含多个反射区域的场景,将 CVDEM 为代价体生成的权重图上采样至输入图像分辨率,得到空间权重热力图。

如图 7 所示,其展示了 3 个场景的输入图像与对应的权重热力图。可以观察到,在反射强度较低的区域,热力图值普遍较高(颜色偏红),这表明网络更倾向于依赖来自感受野更大的上下文信息。而在反射强度较高的区域(如建筑物的反光、路面上的高光),热力图值显著降低(颜色偏蓝),这表明网络动态地降低了粗尺度特征的权重,相应地增强了来自细尺度的、包含更多细节信息的特征在融合中的贡献,以避免粗尺度特征中可能存在的模糊性对预测准确视差造成干扰。

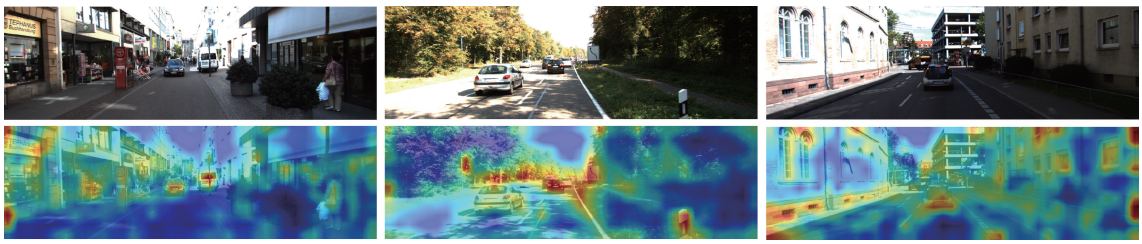

图 7 CVDEM 融合权重随反射强度的变化分析

Fig. 7 Analysis of CVDEM fusion weight variation with reflectance intensity

图 8 和 9 对比展示了本文算法与基准方法在 KITTI 数据集的视差预测结果。在每组图像中,第 1 列为左视图,第 2 列为基准算法结果,第 3 列为本文算法结果。红色矩形框标识关键对比区域。实验结果表明,基准算法 IGEV-Stereo 在强光照和反射表面等挑战性场景下表现欠佳,预测结果存在偏差。相较而言,本文算法展现出以下优势:1)在玻璃等复杂反射区域的细节保持能力显著提升;2)对树枝和远距离车辆等小目标的视差估计误差明显降低;3)整体视差图的连续性和一致性得到有效改善,在视差不连续区域能够保持相对清晰的边界。这些结果验证了本文算法在复杂光照条件下的鲁棒性和场景适应能力。

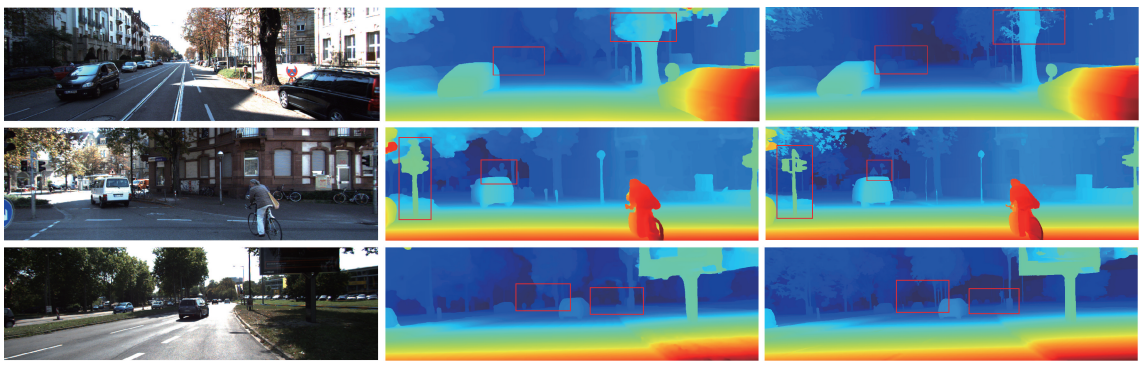

图 8 KITTI 数据集遮挡区域视差图对比

Fig. 8 Comparison of disparity maps of occluded areas in the KITTI dataset

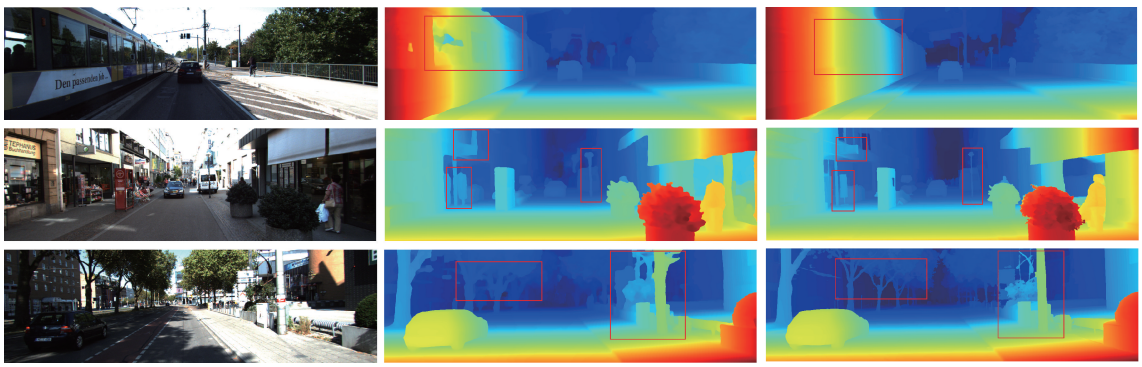

图 9 KITTI 数据集反射区域视差图对比

Fig. 9 Comparison of disparity maps of reflection areas in the KITTI dataset

图 8 展示了在 KITTI 数据集中对于较难预测视差的细节区域的视差图对比。通过观察可以发现，基准算法在树枝和物体轮廓等细节区域的表现比较模糊，而本文算法对于树枝、交通信号灯和远处的汽车等细节区域具有更好的预测效果。

图 9 展示了在 KITTI 数据集中反射区域的视差图对比。通过观察可以发现，本文算法对于玻璃的镜面反射等区域有更好的推理结果，受光照强度的影响较小，与基准方法相比能够预测出更加合理的视差值。

为深入探究本算法的局限性并明确其性能边界，本节选取了两种极端场景进行失败案例分析：夜间强反射与密集细微物体。并提供了这些场景下基准模型和本文所提模型的视差预测图。如图 10 所示，第 2 列是基准模型的预测视差图，第 3 列是本文模型的预测视差图。

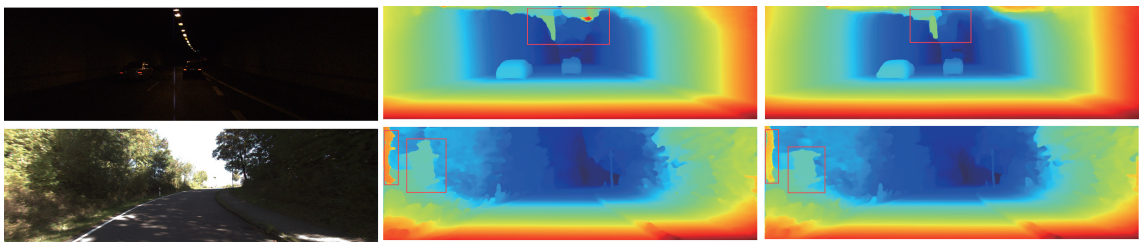

图 10　失败案例分析

Fig. 10　Analysis of failed cases

1）夜间强反射场景分析。如图 10 第 1 行所示，由于数据集中没有夜间拍摄场景，本文利用隧道场景来模拟夜间的环境。图中强烈的路灯与黑暗的环境形成巨大反差。在高光区域中心（如图红色框内），由于感光元件饱和导致局部纹理信息完全丢失，无论是本文方法还是基准方法均出现一定范围的预测失败。这表明当反射强度超出一定范围，导致图像信息严重损失时，仅从 RGB 图像中恢复几何信息是极其困难的，可能需要引入额外的光源先验或物理模型。

2）密集细微物体场景分析。如图 10 第 2 行所示，对于前景中密集、稀疏的树叶丛，无论是本文方法还是基准方法预测的视差图都出现了明显的噪声和不连续性。失败原因主要有两点：（1）固有歧义性：树叶丛本身具有复杂的、非刚性的几何结构以及大量的遮挡，其视差本身就不连续，属于一个病态问题；（2）感受野与细节的权衡：虽然 MKAD 模块通过多核卷积捕捉了多尺度特征，但对于单个树叶级别的细微结构，其特征响应仍然不足。

上述分析表明，本文算法在夜间极端反射和具有高频细节的复杂非刚性结构上仍面临挑战。这为未来的工作指明了方向，可能需要引入额外的传感器信息或先验知识[29-30]。

3　结　　论

本研究针对立体匹配算法在反射区域匹配精度下降的问题，提出了一种融合多核注意力解码器与代价体动态增强机制的改进算法。首先，MKAD 的核心思想在于通过并行的深度可分离卷积与双注意力机制，构建了一个高效的特征解码器。它不仅在多个分辨率上捕捉空间细节以缓解匹配歧义，更通过注意力权重重校准了特征响应，有效抑制了反射区域引入的噪声，为代价体构建提供了更具判别力的输入特征。其次，CVDEM 的核心思想是引入一种基于全局上下文感知的动态选择机制，以替代传统的静态代价体正则化。它通过建立代价体特征间的长程依赖，自适应地确定来自不同尺度代价体特征的融合比例，从而增强代价体对反射区域的建模能力，并提升其与最终视差预测的相关性。MKAD 与 CVDEM 的协同作用机制在于它们分别从“前端特征”与“后端代价”两个层面共同解决了反射区域的挑战。MKAD 在前端提供了对反射区域更鲁棒的特征表示，从源头上降低了代价体构建的模糊性；CVDEM 则在后端根据包含反射区域在内的全局上下文信息，对可能已受污染的代价体进行智能化清洗与增强。二者形成前后联动的处理流，共同确保了从特征到视差整个链路在反射场景下预测视差的准确性。

然而，本文方法仍存在一定的局限性。首先，尽管对反射区域有显著改善，但其性能在很大程度上依赖于初始特征提取的准确性，在夜间极端光照或密集细微物体场景下，基础特征的退化可能限制后续模块的效能。其次，模型中引入的 3D 卷积与迭代优化机制，在追求高精度的同时，其推理速度尚无法满足实时性要求极高的应用场景。基于以上分析，未来的研究工作将围绕以下几个方向展开：探索更轻量、更高效的代价体构建与正则化架构，以在精度与速度间取得更优平衡；研究将时序信息或多模态传感数据引入立体匹配框架，以应对极端场景下的挑战。

参考文献

[1]　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.

[2]　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.

[3] 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.

[4] ZHANG K, LU J B, LAFRUIT G. Cross-based local stereo matching using orthogonal integral images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(7): 1073-1079.

[5] HIRSCHMULLER H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 328-341.

[6] 周秀芝,文贡坚,王润生.自适应窗口快速立体匹配[J].计算机学报,2006(3):473-479.
ZHOU X ZH, WEN G J, WANG R SH. Fast stereo matching using adaptive window[J]. Chinese Journal of Computers, 2006(3): 473-479.

[7] BIRCHFIELD S, TOMASI C. A pixel dissimilarity measure that is insensitive to image sampling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(4): 401-406.

[8] MEI X, SUN X, ZHOU M C, et al. On building an accurate stereo matching system on graphics hardware[C]. IEEE International Conference on Computer Vision Workshops, 2012: 467-474.

[9] ŽBONTAR J, LECUN Y. Computing the stereo matching cost with a convolutional neural network[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1592-1599.

[10] 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.

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

[12] 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.

[13] LIPSON L, TEED Z, DENG J. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching[C]. International Conference on 3D Vision(3DV), 2022: 218-227.

[14] 孟凡娜,邹永佳,曹杨,等. 基于自适应空间卷积的立体匹配算法[J].激光与光电子学进展,2025,62(8):246-254.
MENG F N, ZOU Y J, CAO Y, et al. Stereo matching algorithm based on adaptive spatial convolution[J]. Laser & Optoelectronics Progress, 2025,62(8):246-254.

[15] 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.

[16] 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, 2012: 3354-3361.

[17] 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(1): 427-434.

[18] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. ArXiv preprint arXiv: 1711.05101, 2017.

[19] ZHANG F H, PRISACARIU V, YANG R G, et al. GA-Net: Guided aggregation net for end-to-end stereo matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 185-194.

[20] CHENG X J, WANG P, YANG R G. Learning depth with convolutional spatial propagation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019, 42(10):2361-2379.

[21] 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(CVPR), 2022: 12971-12980.

[22] LI J K, WANG P S, XIONG P F, et al. Practical stereo matching via cascaded recurrent network with adaptive correlation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16263-16272.

[23] LI K H, WANG L G, ZHANG Y, et al. LoS: Local structure-guided stereo matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024:19746-19756.

[24] WEINZAEPFEL P, LUCAS T, LEROY V, et al. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow[C]. IEEE/CVF International Conference on Computer Vision, 2023: 17969-17980.

[25] ZHOU J L, HUANG W Q, LIAO Q M, et al. Multi-dimensional attention on cost volume for stereo matching[C]. International Joint Conference on Neural Networks(IJCNN), 2024:1-8.

[26] 秦伦明, 余斌, 崔昊杨,等. 基于改进 ACV 模型的视差估计方法[J]. 激光与光电子学进展, 2024, 61(24): 141-150.
QIN L M, YU B, CUI H Y, et al. Parallax estimation method based on improved ACV model[J]. Laser & Optoelectronics Progress, 2024, 61(24): 141-150.

[27] ZHAO T, DING M Y, ZHAN W, et al. Depth-aware volume attention for texture-less stereo matching[J]. ArXiv preprint arXiv:2402.08931, 2024.

[28] YUAN T M, HU J CH, OU SH J, et al. Hourglass cascaded recurrent stereo matching network[J]. Image and Vision Computing, 2024, 147: 105074.

[29] 赖东阳,邱志斌,杨泽鼎,等. 基于 YOLOv9-SOEP 与双目立体视觉的输电线路山火火焰高度测量[J]. 电子测量与仪器学报,2025, 39(8): 258-268.
LAI D Y, QIU ZH B, YANG Z D, et al. Flame height measurement of transmission lines wildfire based on YOLOv9-SOEP and binocular stereo vision[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(8):258-268.

[30] 刘秋骅,徐晓苏. 基于双目相机深度估计的相机-LiDAR 端到端外参标定方法[J]. 仪器仪表学报, 2025,46(5):214-225.
LIU Q H, XU X S. End-to-end camera-LiDAR extrinsic calibration method based on stereo camera depth estimation[J]. Chinese Journal of Scientific Instrument,2025,46(5):214-225.

作者简介

苏志豪,硕士研究生,主要研究方向为计算机视觉、立体匹配。

E-mail:suzhihao1019@163.com

于红绯(通信作者),副教授,主要研究方向为计算机视觉、人工智能、自动驾驶。

E-mail:yuhfln@163.com

曹杨,副教授,主要研究方向为计算机视觉、机器博弈。

E-mail:caoyang821213@163.com