深度立体匹配网络的融合代价体及其代价聚合方法

邹正阳，伍云霞†，徐 倩

(中国矿业大学 (北京) 人工智能学院，北京 100083)

摘 要: 深度立体匹配网络使用代价体将三维场景结构编码为双目特征的对应关系, 在机器人定位与避障等场景具有重要应用前景. 然而, 现有代价体方法不能为双目特征建立全面且无冗余的相关信息, 导致视差预测精度不足. 针对该问题, 首次将极线几何约束引入代价体计算流程, 通过多类型代价体优势互补, 提出一种即插即用的融合代价体及其代价聚合方法. 首先, 融合代价体同步计算极线共投影区域内特征向量的全局点积相关信息和局部分组点积相关信息, 保证特征相关的全面性并有效避免了信息冗余; 其次, 在周边信息聚合过程中结合传统聚合方法和融合代价体特性, 提出一种基于深度可分离卷积的自适应加权降维方法, 解决融合代价体在聚合阶段的维度不平衡性和计算效率问题. 将所提方法集成到立体匹配框架并命名为 FusionStereo, 在基准数据集上进行实验验证. 结果表明: FusionStereo 在 KITTI 2015 域内训练后的误匹配率指标 BAD3 为 1.55%, 在 MiddleBurry 跨域测评的误匹配率指标 BAD1 为 17.1%, 明显优于其他类型代价体的对比方法.

关键词: 深度学习；立体匹配；信息冗余；几何约束；融合代价体；代价聚合

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.1461

引用格式: 邹正阳, 伍云霞, 徐倩. 深度立体匹配网络的融合代价体及其代价聚合方法 [J]. 控制与决策, 2025, 40(10): 3145-3154.

Fusion cost volume and cost aggregation method for deep stereo matching network

ZOU Zheng-yang，*WU Yun-xia*†，*XU Qian*

(School of Artificial Intelligence，China University of Mining and Technology-Beijing，Beijing 100083，China)

Abstract: Stereo matching networks encode 3D scene structure into binocular feature correspondences through cost volumes, which has important application in robotic localization and obstacle avoidance. However, the existing cost volume methods fail to establish comprehensive yet non-redundant correlations between binocular features, leading to low disparity accuracy. To address this issue, this research introduces epipolar geometric constraints into the cost volume computation and proposes a plug-and-play fusion cost volume and cost aggregation method by leveraging the complementary advantages of multiple cost volume types. First, the proposed fusion cost volume calculates global dot-product correlations and local grouped dot-product correlations of feature vectors within the epipolar co-projection region simultaneously, which ensures the comprehensiveness of feature correlations information and effectively avoids information redundancy. Then, during information aggregation, this research combines traditional aggregation methods with the characteristics of the fusion cost volume to propose an adaptive weighted dimensionality reduction method based on depthwise separable convolution, which addresses the dimensionality imbalance and computational efficiency issues in the feature aggregation process of the fusion cost volume. The stereo matching framework integrating the proposed method is named FusionStereo, and experimental verification is carried out on the benchmark data sets. Results show that the BAD3 of FusionStereo is 1.55% after training on KITTI 2015, and the BAD1 of Middlebury cross-domain evaluation is 17.1%, which is significantly better than other cost volume methods.

Keywords: deep learning； stereo matching； information redundancy； geometric constraints； fusion cost volume； cost aggregation

收稿日期: 2024-12-18；**录用日期:** 2025-04-21.

基金项目: 国家自然科学基金项目 (52374165).

†通信作者. E-mail: wuyx@cumtb.edu.cn.

本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

0 引 言

双目相机在智能手机、机器人、汽车等领域已广泛应用[1], 使得双目立体视觉在学术界和工业界的关注度日益增加. 双目立体视觉首先通过立体校正[2]对双目视图进行处理, 模拟人类视觉获取双目图像; 然后利用立体匹配算法计算双目图像之间的像素对应关系, 既视差; 最后通过三角映射函数和双目相机的内、外参数, 将视差反投影为观测场景的三维深度. 相比于发展较为成熟的立体校正和深度恢复技术, 求解双目图像像素对应关系的立体匹配算法成为场景深度的估计关键环节[3-5].

传统的立体匹配算法通常依赖于手工特征提取方法, 结合平滑性、遮挡关系和左右一致性等全局优化约束来预测视差值[6-8]. 然而, 这些手工建模的像素特征在处理低纹理区域时存在局限性, 往往导致视差预测精度较低.

随着深度神经网络和人工智能技术的不断发展, 基于代价体概念的深度立体匹配算法在预测精度上取得了显著进展, 逐步成为立体匹配领域的主流算法. 深度立体匹配算法主要由特征提取、代价体计算、代价聚合、视差回归、视差细化等子模块组成. 其中, 特征提取模块能够替代传统手工特征, 主要得益于深度学习在图像理解方面的强大能力以及数据驱动的端到端优化模式[9]. 代价体模块作为立体匹配网络的核心步骤, 通过可微分方式将特征提取模块中的语义特征转化为视差所需的相关信息, 使深度立体匹配算法继承深度学习的语义理解能力和数据驱动模式, 奠定了近 10 年来深度立体匹配算法的发展基础. 因此, 代价体及其相关聚合方法的发展历程基本可以概括为深度立体匹配算法的发展历程.

2014 年, Zbontar 等[10] 首次将传统手工建模特征替换为深度网络特征, 虽然在提升模型性能方面取得了一定成效, 但在后续的代价体处理过程中仍然使用了传统方法. 2015 年, Mayer 等[11] 参考深度光流估计网络提出 DispNetC, 首次将可微分的三维代价体模块引入端到端的立体匹配模型, 实现了深度立体匹配算法的端到端优化. 然而这种方法没有充分利用特征信息, 且忽略了代价体周边信息聚合的步骤. 随后, 研究人员通过建立多尺度代价体[12] 或使用循环神经网络进行特征复用[13-14], 进一步提升了三维代价立体匹配网络的域内训练精度和跨域数据泛化能力. 然而, 三维代价体方法在构建过程中未能有效保留特征向量的内部细节信息, 且存在大量冗余信息, 这导致网络需要繁琐的后处理流程. 为解决这个问题, GWCNet[15] 提出四维代价体, 通过将特征向量拆分为多个子向量组, 并采用滑动方式逐个计算非冗余区域各子向量的局部分组点积相关信息, 从而在代价体中保留了更多的细节信息. 但是, GWCNet 在四维代价体的周边信息聚合阶段依赖大量三维卷积[16], 带来了较高的计算开销. 因此, 研究人员考虑更轻量的代价聚合方法[17-19], 例如, 通过结合传统方法的特征聚合方法[20] 或参考三维代价体引入循环神经网络的特征复用方式[21], 以降低计算资源的消耗.

在近 10 年深度立体匹配算法的其他研究中, 还包括叠加代价体[22]、隐式代价体[23] 等其他方向的发展, 但因预测准确率欠佳而未发展为主流方法. 在最近的研究中, 研究人员考虑将左视图的语义信息集成到四维代价体[24], 或由粗到细的逐步细化代价体[25-26], 虽在预测精度和推理速度上取得突破, 但仍限制在三维代价体或四维代价体概念内, 并未打破两种代价体之间的壁垒. 文献 [27] 使用两个骨干网络分别提取特征, 并将其分别用于建立两种代价体. 这种方法虽然在各项指标中均取得较大增益, 但仍继承了两类代价体的固有缺陷, 不能为立体视图特征建立全面、无冗余的相关信息. 因此如何打破两种代价体之间的壁垒并使二者相互弥补是本文解决的主要问题.

为此, 本文首次将极线几何约束引入代价体计算流程, 实现了多类型代价体的优势互补与协同整合. 具体而言, 代价体仅需保存极线约束下共投影区域内的相关信息. 因此, 本文首先考虑对三维代价体的冗余信息进行后处理 (置 0), 并限制最大视差搜索范围 (MaxDisp) 来建立两类代价体的关联, 如图 1 所示. 然而, 基于对代价体结构的理论分析, 本文证明四维代价体与后处理三维代价体之间存在可转换性, 并基于此提出一种融合代价体方法. 该方法基于维度约简的特征表示, 通过对分组维度的池化操作实现了特征相关性的优化重构[27]. 其次, 在周边信息聚合过程中考虑到两类代价体的维度不平衡性问题, 将转换后的三维代价体视作保存全局信息的组类, 并结合传统聚合方法提出基于深度可分离卷积的自适应加权降维方法, 如图 2 所示. 至此, 本文打破了两类代价体的联合建立壁垒和联合聚合壁垒, 并将其命名为融合代价体及其代价聚合方法.

在实验中将融合代价体及其聚合方法集成至深度立体匹配框架[13], 构建 FusionStereo 模型. 结果表明, FusionStereo 在预测精度上显著优于现有基于代价体的方法, 并且在推理速度上比原基准模型

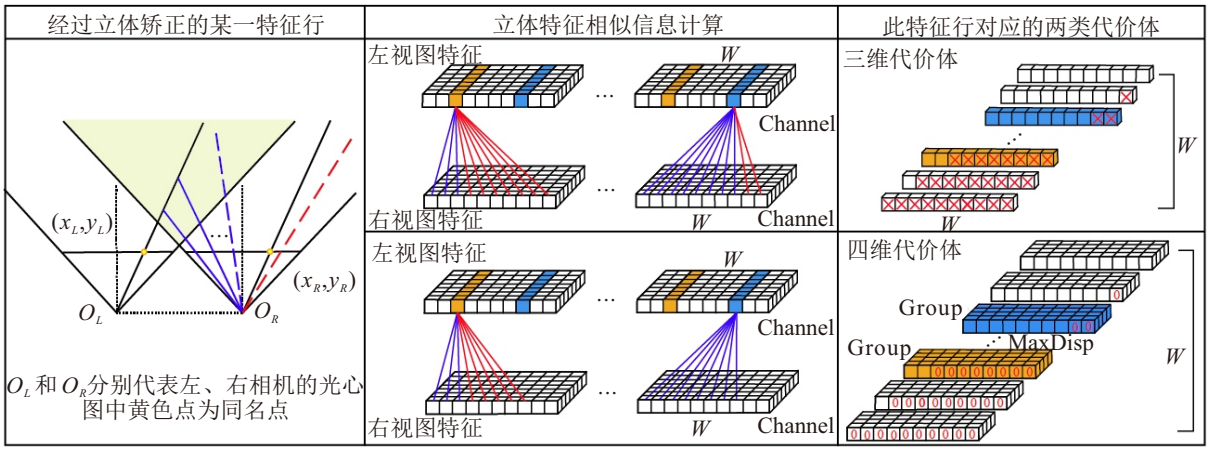

图1　代价体建立流程

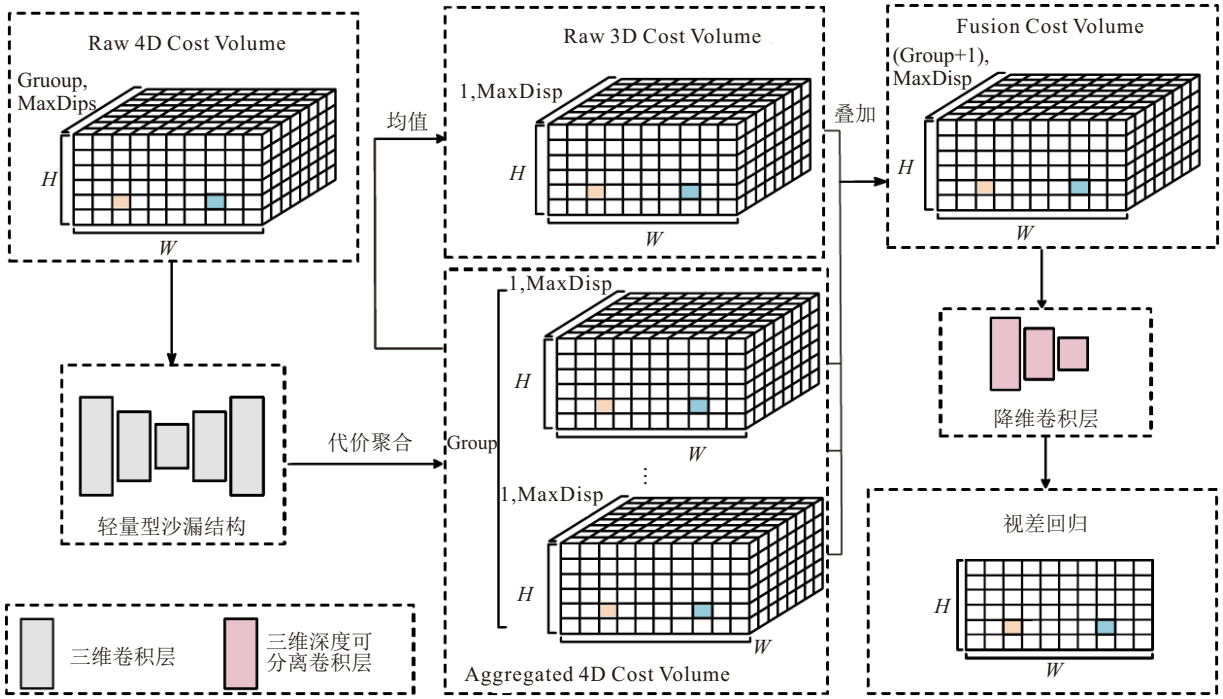

图2　融合代价体及其代价聚合方法流程

RAFT-Stereo[13] 提高了 200%, FusionStereo 能以更低的推理时间实现更高的预测精度, 进一步印证了本文提出的融合代价体为立体视图特征建立全面、无冗余相关信息思路的有效性.

1　先　验

1.1　代 价 体

立体匹配的目的是在右视图中找到所有与左视图对应的像素点, 并返回对应像素之间的横坐标差异, 即视差. 基于深度学习的立体匹配网络以代价体为媒介, 将双目视图特征编码为相关信息. 两类广泛使用的代价体定义为

$$C_{3D}(d,x,y)=\begin{cases}\dfrac{\langle f^l_{xy},f^r_{(x-d)y}\rangle}{\mathrm{Nc}}, & d\leqslant x;\\ \dfrac{\langle f^l_{xy},f^r_{dy}\rangle}{\mathrm{Nc}}, & d>x.\end{cases}\tag{1}$$

$$C_{4D}(g,d,x,y)=\begin{cases}\dfrac{\langle f^l_{xygh},f^r_{(x-d)ygh}\rangle}{\mathrm{Nc/Ng}}, & d\leqslant x;\\ 0, & d>x.\end{cases}\tag{2}$$

其中: C_{3D}和C_{4D}分别为三维和四维代价体, $\langle\ ,\ \rangle$为内积, Nc为特征向量长度, Ng为四维代价体对特征向量的分组数, f^l和f^r分别为左视图特征层和右视图特征层, x和y分别为特征层的横、纵坐标值, g和d分别为分组坐标和距离被搜索点的坐标距离.

1.2　几何约束

经过严格的立体校正, 对应像素点的搜索区域应限制在极线共投影区域内, 此共投影区域如图 1 左侧黄色区域所示. 在共投影区域内, 双目观测点在左图投影点的横坐标永远小于或等于其在右图投影点的横坐标; 当双目观测点位于无穷远处时, 此观测

点与双目光心的连接线近似平行且投影在左、右视图的横坐标相等 (以图中黄色点为例). 因此, 对于双目视图同时观测到的三维空间点$P=(X,Y,Z)$, 在左、右视图中的投影关系为

$$\begin{bmatrix} x_L \\ y_L \\ 1 \end{bmatrix} = K_L \begin{bmatrix} X/Z \\ Y/Z \\ 1 \end{bmatrix},$$
$$\begin{bmatrix} x_R \\ y_R \\ 1 \end{bmatrix} = K_R \begin{bmatrix} X/Z \\ Y/Z \\ 1 \end{bmatrix}. \tag{3}$$

其中: K_L和K_R分别为左、右相机的内参矩阵; x_L, y_L, x_R, y_R分别为空间点P在左、右视图中投影在图像坐标系的横、纵坐标, 且存在$x_R \leqslant x_L$, $y_R = y_L$. 当$x_R = x_L$时, 三维空间点P应位于无穷远处; 当$x_R > x_L$时, 三维空间点不存在且不应考虑在相关信息计算中, 如图 1 红色标识所示, 因此代价体中保存了大量冗余信息.

四维代价体通过对非共投影区域位置提前置 0 (图 1 右侧红色“0”标识) 来缓解冗余信息问题, 而三维代价体则仍然在非共投影区域保存冗余信息 (图 1 右侧红色“叉号”标识). 因此, 如果将三维代价体的冗余信息置 0 且设定视差最大搜索范围 (MaxDisp, 如图 1 右侧下方四维代价体所示), 则三维代价体可以视作四维代价体的一种保存全局相关信息的分组 (Group, 如图 1 左侧下方四维代价体所示).

此外, 四维代价体的聚合和视差回归可以定义为

$$C_{3D}^{\text{aggregation}} = \text{Conv3}d(C_{4D}), \tag{4}$$

$$\text{Disparity} = \sum_{d=1}^{\text{MaxDisp}} d \times \text{Soft}\max(C_{3D}^{\text{aggregation}}). \tag{5}$$

由此可见, 四维代价体首先聚合并降维为三维代价体, 最后通过Softmax函数将三维代价体转变为三维概率体, 并在加权求和后得到近似视差.

综上, 两类代价体的异同主要体现在两方面: 首先, 二者对特征向量和非共投影区域的处理方式不同, 而此差异可以通过后处理建立联系; 其次, 四维代价体在聚合和视差回归过程中使用三维代价体作为处理媒介. 基于以上结论, 本文将在接下来的内容中描述如何利用先验将二者融合并相互发挥作用.

2 融合代价体及其代价聚合方法

首先考虑如何将三维代价体的相关信息计算钳制在有意义的共投影区域内, 并对共投影区域外的冗余信息进行后处理. 由式 (1) 和 (2) 可知, 若三维代价体将冗余信息置 0 并拓展一个分组维度, 则可视作仅保存全局相似信息的一种四维代价体. 沿此思路, 提出一种全新代价体范式, 旨在为立体视图特征提供全面、无冗余的相关信息, 并命名为融合代价体, 如图 2 所示. 然而, 融合代价体对特征向量进行两次代价计算需消耗大量的计算资源, 为缓解该问题, 本文对四维代价体 (Raw 4D Cost Volume) 聚合后的分组维度进行均值池化, 避免了三维代价体的重复聚合, 以简单高效的方式获得包含全局相关信息的三维代价体, 如图 2 Raw 3D Cost Volume 所示.

其次, 在融合代价体的周边信息聚合过程中, 考虑到两类代价体的维度不平衡性, 将融合代价体的全局相似信息视为一个保存全局信息的组类, 并叠加到沙漏型聚合器[16] 聚合后的分组代价体中, 如图 2 所示.

在过去的研究中, 视差回归的操作对象通常为三维代价体, 且传统的代价体通过计算相关信息来区分各候选同名点之间的差异. 因此, 聚合后的融合代价体需要在分组维度自适应加权各组的相关信息, 以降维并用于视差回归函数. 因此, 本研究使用 3D 深度可分离卷积对聚合后的融合代价体进行自适应加权降维, 如图 2 右下所示.

综上所示, 融合代价体、周边聚合及视差回归可以定义为

$$C_{3D}^{\text{raw}}(d,x,y) = \text{MeanPooling}_{\{g\}}(C_{4D}^{\text{Aggregated}}), \tag{6}$$

$$\text{FCV} = \text{Concat}\{\text{Conv3}d^{\text{hourglass}}(C_{4D})_{g=1}^{K}, C_{3D}^{\text{raw}}\}, \tag{7}$$

$$\text{Disparity} = \sum_{d=1}^{\text{MaxDisp}} d \times \text{Soft}\max(\text{Conv3}d(\text{FCV})). \tag{8}$$

其中: C_{3D}^{raw}为对四维代价体均值处理得到的三维代价体; $\text{MeanPooling}_{\{g\}}$为对四维代价体的分组维度进行均值池化; $\text{Conv3}d^{\text{hourglass}}(C_{4D})_{g=1}^{K}$为对四维代价体使用沙漏型聚合器, 然后对分组维度逐个遍历, 且$\text{FCV} \in \mathbb{R}^{(K+1)\times\text{MaxDisp}\times H\times W}$. 在下面的内容中, 本文将介绍以融合代价体及其代价聚合方法为基础的立体匹配网络 FusionStereo.

3 FusionStereo

为评估所提出方法的有效性, 将融合代价体及其代价聚合方法集成到现有网络框架[13], 并命名为 FusionStereo. 在特征提取阶段, FusionStereo 采用与基准方法[15] 类似的 ResNet 结构, 并提取 1/4 原图分辨率特征. 在特征提取的归一化层中, 本文认为立体匹配网络应该将特征层的表征分布限制在特征层上,

因此使用以完整特征层为处理单元的层归一化[28-29]. 网络概述由图 3 所示.

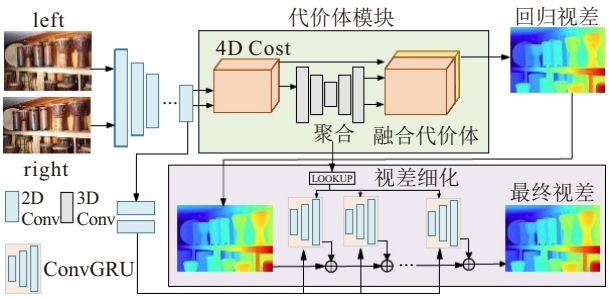

图3　FusionStereo 网络结构

为保证模型实验公平性, 本文按照惯例将代价体建立的最大视差搜索范围限制在 192; 在周边聚合过程中, 对保存更多细节信息的分组代价体使用控制输入输出维度一致的沙漏型结构[16] 处理, 沙漏型结构如式 (7) 的$\mathrm{Conv3}d^{\mathrm{hourglass}}$所示; 使用一层深度可分离三维卷积将融合代价体 FCV 降维成三维代价体, 如式 (8) 中的$\mathrm{Conv3}d(\mathrm{FCV})$所示; 最后对降维后的 FCV 进行视差回归, 并作为后续视差细化阶段的初始视差.

在视差细化过程中, 使用基于 ConvGRU 的循环视差细化方式. 细化过程如下所示:

$$L = \Psi(\mathrm{FCV},\ \mathrm{Disparity}^{t-1}), \tag{9}$$

$$\mathrm{Hidden}_t = \mathrm{ConvGRU}(L, R_t, Z_t, H_t, \mathrm{Hidden}_{t-1}), \tag{10}$$

$$\Delta\mathrm{Disparity} = \mathrm{Conv2}d(\mathrm{HiddenConv}), \tag{11}$$

$$\mathrm{Disparity}^t = \mathrm{Disparity}^{t-1} + \Delta\mathrm{Disparity}. \tag{12}$$

其中: Ψ为使用上一次循环的输出视差对FCV进行信息索引 (FCV 的回归视差作为第 1 次循环优化的视差); L为索引值, 如图 3 中的LOOKUP所示; Hidden_t为隐藏状态; R_t, Z_t, H_t为 3 种状态参数, 二者均来自于特征提取阶段.

将上述内容集成的网络结构如图 3 所示, 紫色区域代表视差细化阶段, 具体设置细节与文献 [13-14] 相同, 其中 L 代表使用当前阶段的视差在融合代价体中采样. 此外, FusionStereo 相比原基准模型 RAFT-Stereo 在代价体相关过程保存了足够的相关信息, 因此省略了原网络中重新提取左图语义信息并作为视差上采样约束的过程.

最后, 使用L_1损失监督 FusionStereo 的训练, 定义为

$$L = \sum_{t=1}^{N} \gamma^{N-t} L_1(d_t - d_{GT}) + L_1(d_0 - d_{GT}). \tag{13}$$

其中: t为循环细化的次数, d_{GT}为监督真值, d_0为对聚合后的融合代价体使用式 (8) 得到的回归视差, 在实验阶段 γ 取 0.9.

4　实验结果及分析

本节介绍实验的具体细节, 并展示融合代价体及其聚合方法的优越性. 所有实验使用 PyTorch 实现并在 Tesla V100-16G GPU 进行训练, 跟随基准数据集[13] 的训练配置, 使用 AdamW 作为优化器并将学习率和衰减系数设置为 0.000 2 和 0.000 01. 在数据增强中, 将预训练数据 SceneFlow[2] 随机裁剪为 256 × 512, 并添加色度变换用于模型的预训练 (亮度、伽玛和对比度变换), 以增强模型的跨域泛化能力[30]. 根据惯例[14,27,30] 将 FusionStereo 的最大视差搜索范围设置为 192 个像素. 本实验将预训练数据集遍历 20 次, 采样批次设置为 8. 视差循环细化过程中, 为保证实验公平性, 在训练过程中采用与现有方法[13,27] 相同的循环细化次数 22.

为验证模型在中低配置平台中的效果, 使用便携计算设备 Lenovo Legion Y9000P2021H i7(11800H @2.30GHZ)-RTX3070(8G) 进行模型测试, 采用流行的测评指标: 推理时间、EPE(端点误差) 和 BAD(N), 其中 EPE 表示预测视差值与真值之间的平均误差, BAD(N) 表示预测值误差大于 N 个像素的百分比.

4.1　代价体消融及对比实验

在本节中, 所有实验均使用 SceneFlow 的 Train 数据集进行训练[11], 并分别将 SceneFlow 的 Test 数据集和自动驾驶数据集 KITTI 用于域内训练测试和跨域测试.

分组维度消融实验. 本文提出一种深度立体网络的融合代价体及其代价聚合方法, 其中融合代价体在代价聚合过程存在维度不平衡性问题. 因此, 如何选择四维代价体的分组维度, 使融合代价体发挥最佳性能, 成为本实验主要考虑的问题. 具体而言, 采用一种维护输入输出尺寸的沙漏型结构[27] 对分组代价体进行信息聚合, 并将保存全局信息的三维代价体叠加在四维代价体聚合后的分组维度, 如表 1 所示. 当分组代价体的分组维度被设置为 16 时, 融合代价体及其聚合方法取得最佳性能. 而更多或更少的分组维度并不能带来更有效的作用, 这是因为更少的分组维度不能为算法带来更多的细节信息, 而更多的分组维度会使三维代价体保存的全局信息不能完全发挥作用.

代价体对比实验. 融合代价体能够为立体特征图提供综合、无冗余的相关信息, 为体现所提出方法的有效性, 将代价体相关的同类方法进行比较. 结果

表1 分组消融及对比实验

模型配置	SceneFlow Test		KITTI 2012		
	EPE	BAD3	EPE	BAD1	BAD3
FCV - 4	1.22	5.28	1.02	19.7	5.23
FCV - 8	1.27	5.40	1.02	19.4	4.79
FCV - 16	**1.12**	**4.90**	**0.93**	**17.0**	**4.60**
FCV - 24	1.13	4.92	0.96	17.5	4.98
FCV - 32	1.15	4.94	0.98	18.1	5.01
三维代价体	1.58	7.04	1.58	42.3	9.91
四维代价体	1.19	5.97	1.01	17.9	5.24
三维代价体-过滤	1.34	5.81	1.33	27.7	7.13
FCV + 代价聚合方法	**1.06**	**4.46**	**0.89**	**14.8**	**4.59**

注: 粗体为最优.

如表 1 所示, 本文所提出的融合代价体及其代价聚合方法相比其他类似方法能够实现更低的误匹配指标.

循环迭代次数对比实验. 如表 2 所示, 实验中使用 KITTI 2012[31] 进行跨域测评. 结果显示, 相比于 IGEV-Stereo[27], 本文所提出的融合代价体及其代价聚合方法在视差回归阶段即可得到准确鲁棒的视差, 进而为循环优化提供准确的视差初值, 进一步体现了所提出方法的有效性.

表2 循环迭代次数对比实验

Methods	循环优化次数(测评指标使用BAD1)					
	1	3	5	8	16	32
IGEV-Stereo[26]	52.0	47.2	44.5	42.4	40.2	39.6
FusionStereo	**19.2**	**16.7**	**14.8**	**14.6**	**14.5**	**14.3**

注: 粗体为最优.

模型参数对比实验. 本节对 FusionStereo 的配置 (卷积核大小、批量归一化数量、归一化函数类型) 进行消融实验. FusionStereo 原配置为 (Kernel Size: 3 × 3, Batch Size: 8, 层归一化). 表 3 结果显示, 更大的卷积核尺寸 (5 × 5) 不会给立体匹配任务带来显著的帮助, 这是因为立体匹配网络的任务在于匹配像素间差异, 而更大的卷积核可能会影响模型对细节的捕捉能力. 此外, 归一化方法的消融实验验证了第 3 节“特征层的表征分布限制在特征层上”这一推论的正确性.

表3 FusionStereo 模型参数消融实验

模型配置	SceneFlow Test		KITTI 2012		
	EPE	BAD3	EPE	BAD1	BAD3
FusionStereo	1.12	**4.90**	**0.93**	17.0	**4.60**
Kernel Size: 5 × 5	1.38	7.53	0.95	**16.1**	5.19
Batch Size: 4	1.16	5.46	0.97	16.9	5.27
Batch Size: 6	1.14	5.15	0.95	17.3	4.99
批量归一化	1.18	6.05	1.96	56.6	7.26
实例归一化	**1.08**	6.10	1.51	27.3	6.43

注: 粗体为最优.

4.2 FusionStereo 的对比实验

KITTI 基准. 本节实验使用 KITTI 2012[31] 和 2015[32] 对 FusionStereo 进行微调以进一步验证 FusionStereo 域内训练后的预测能力. 具体而言, 首先将 KITTI 拆分成训练数据和测试数据[33], 并在训练数据上进行 300 epoch 的微调. 学习率从 0.001 开始, 在 200 个 epoch 后下降到 0.000 1, 在测试数据中进行性能测试. 如表 4 所示, FusionStereo能以较短的推理时间在大多数指标中取得超越 SOTA 方法的性能. 相比于 2024 年专门针对立体匹配精心设计的 SOTA 方法 Selective-RAFT[14] 和 NMRF-Stereo[34], FusionStereo 仍在 KITTI 2015 的测评指标上取得更优秀的结果. 此外, 如表 5 中 KITTI 2015 推理时间的测评结果所示, FusionStereo 相比于最新方法具有更快的推理速度, 体现了本文所提出的融合代价体及其代价聚合方法的有效性. 此外, 所提出的 FusionStereo 相比原基准方法 RAFT-Stereo[13], 能够以两倍的推理速度在细节处取得更连续的结果.

跨域数据泛化能力. 在最近的研究中, 深度立体

表4 KITTI 数据集对比实验

Methods	KITTI2012		KITTI2015		References
	BAD2	BAD3	EPE	BAD3	
GANet-Deep[20]	2.50	1.60	0.5	1.81	CVPR 2020
RAFT-Stereo[13]	2.42	1.66	0.5	1.82	3DV 2021
CREStereo[21]	2.18	1.46	0.5	1.69	CVPR 2022
ACVNet[24]	2.35	1.47	0.7	2.10	CVPR 2022
IGEV-Stereo[27]	2.70	1.44	**0.4**	1.59	CVPR 2023
Selective-RAFT[14]	2.09	1.43	**0.4**	1.63	CVPR 2024
NMRF-Stereo[34]	**2.07**	**1.35**	**0.4**	1.59	CVPR 2024
FusionStereo	2.12	1.43	**0.4**	**1.55**	**ours**

注: 粗体为最优, _为次优.

表5 泛化能力对比实验

代价体	Methods	Middlebury 2014-H		KITTI 2015		
		BAD1	BAD3	EPE	BAD3	Time
3D	RAFT-Stereo[13]	17.3	8.35	**1.13**	5.31	0.36s
	DLNR[35]	48.8	33.9	12.2	30.1	0.37s
	Selective-RAFT[14]	33.6	19.9	1.54	9.26	0.61s
4D	GWCNet[15]	25.5	11.2	2.36	12.1	0.19s
	CFNet[30]	23.1	11.8	26.1	5.88	0.18s
	BGNet+[19]	40.1	15.9	1.46	8.19	**0.04s**
	MS2D[17]	58.6	34.7	2.63	19.7	0.10s
其他	NMRF-Stereo[34]	19.9	7.5	1.14	**5.14**	0.23s
3D&4D	IGEV-Stereo[27]	18.1	8.55	1.45	6.63	0.36s
	Selective-IGEV[14]	50.7	34.6	1.94	13.4	0.40s
	IGEV++[36]	46.8	30.9	2.08	16.1	0.72s
	FusionStereo	**17.1**	**7.24**	1.14	5.22	0.18s

注: 粗体为最优, _为次优.

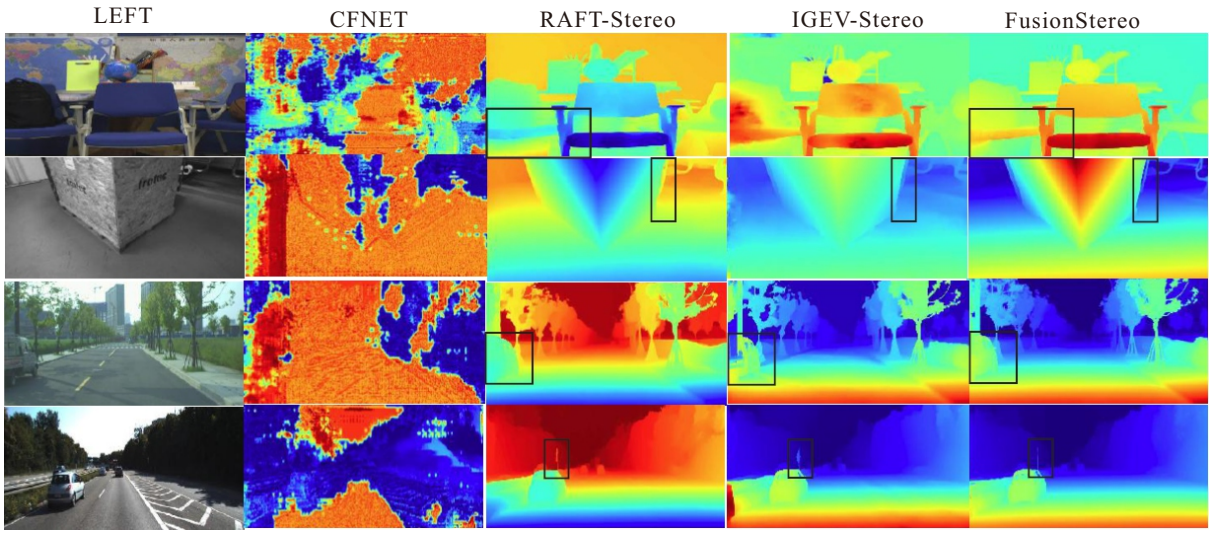

图4　不同模型的可视化结果

匹配网基于数据驱动模型和强大的特征表示能力而表现出瞩目的性能. 然而, 真实场景的深度真值难以获取, 因此研究人员通常在虚拟合成数据上预训练, 并在真实场景中进行跨域数据的泛化能力测试. 此外, 数据驱动的深度立体匹配网络容易将参数优化至虚拟合成数据所在空间, 导致其在真实场景的表现较差, 而跨域数据的泛化能力及其推理速度是衡量深度立体匹配算法能否满足实际应用需求的关键内容[37-38]. 因此, 本节实验将所有方法在 SceneFlow 上进行预训练, 并在室内真实场景和自动驾驶场景进行测试, 结果如表 5 所示. FusionStereo 在 KITTI 2015 跨域测试中的各项指标均具有明显优势. 同时, 强泛化能力基准模型与 FusionStereo 在真实场景的使用效果对比如图 4 所示. 其中 RAFT-Stereo[13]、CFNet[30] 分别基于三维、四维代价体的基准方法, IGEV-Stereo[27] 虽使用两类代价体, 但没有建立两种代价之间的关系. 结果显示, FusionStereo 在室内环境能实现更平滑的效果, 在室外环境能更好地处理细节纹理.

FusionStereo 能够在跨域数据测试中表现优秀主要得益于本研究对代价体相关模块的改进. 具体而言, 深度立体匹配网络的关键在于尽量匹配同名点, 而非利用深度网络的强大语义理解能力[39]. 因此本文提出的融合代价体及其代价聚合方法旨在为双目特征视图建立全面且无冗余的相关信息, 当面临域差异时更关注其特征向量是否匹配准确, 并能够为后续的视差细化提供更稳定、鲁棒的数据基础.

MiddleBurry 基准[39]. MiddleBurry 作为一个真实室内数据集, 包含 15 张不同分辨率及真值的立体图像. 考虑使用 MiddleBurry 训练可能出现的过拟合问题, 本实验同样将其用于跨域数据的泛化能力测试. 实验结果由表 5 所示, 以融合代价体及其代价聚合方法为核心的 FusionStereo, 同样能在室内数据集泛化能力测试的各项对比指标上大幅超过其他类型代价体的对比方法.

Tartan Air 基准[40]. Tartan Air 是一个公开的合成数据集, 模拟多种真实世界的挑战性环境. 该数据集主要作为 SLAM 数据集, 但也提供了 480 × 640 分辨率的立体图像及其视差真值用于立体匹配方法. 本实验同样将其用于多种挑战性环境的鲁棒性测试. 结果如表 6 所示. 研究选择低纹理的室内环境 (office) 和重复纹理的室外环境 (abandoned factory) 作为挑战性环境的测试数据. 结果显示, 得益于本文所提出方法, FusionStereo 依然能够在 Tartan Air 的挑战性环境中取得 SOTA 性能, 这一优势主要得益于融合代价体能够对各类型环境的特征建立全面且准确的相关信息, 从而为后续的视差预测与细化提供了数据基础.

表6　Tartan Air 对比实验

Methods	Office		Abandoned Factory		计算量	访问量
	BAD1	BAD3	EPE	BAD3	FLOPS	BYTE
RAFT-Stereo[13]	46.2	21.1	2.71	21.2	822 G	11.1M
Selective-RAFT[14]	99.8	86.4	37.8	83.6	1.61 T	11.7M
CFNet[30]	45.5	29.1	14.6	29.1	**371 G**	22.2M
IGEV-Stereo[27]	53.7	32.5	3.62	32.5	1.46 T	12.5M
FusionStereo	**14.7**	**9.10**	**1.33**	**6.88**	482 G	**8.31M**

注: 粗体为最优.

4.3　FusionStereo 的实时版本

为进一步验证所提出融合代价体及其代价聚合方法在低分辨率特征层的有效性. 本文在实验中提取更低尺度的 1/8 原图分辨率特征图, 大大降低了算

法的计算复杂度并进一步提高了模型推理速度. 实验保持 FusionStereo 其他超参数不变, 并命名为 FusionStereo-Realtime. 实时性版本的重点在于真实场景的实际部署能力, 因此重点考虑多种场景的泛化能力及其推理速度, 如表 7 所示. 结果显示, 实时性版本能在 KITTI 数据集 (1248 × 384 分辨率) 以 10FPS 以上的帧数运行.

表7 FusionStereo-Realtime 跨域测试实验

数据集	EPE	BAD1	BAD3	time/ms
KITTI2012	0.93	17.0	4.58	90
KITTI2015	1.22	25.8	5.47	90
TarTanAir-Night	1.03	13.4	5.09	60

相比于表 5 的对比实验, 其预测精度虽然在错误匹配指标 BAD3 中保持 SOTA, 但在其余指标中出现了少许下降. 这是因为实时版本提取更低尺度原图分辨率的特征图, 即使融合代价体能够为特征图建立更准确的相关图, 但使用低尺度的特征编码进行视差上采样使视模型缺失了足够的信息来源, 从而使模型不能将视差预测在一个准确的误差范围. 因此 FusionStereo-Realtime 能够在 BAD3 中保持 SOTA, 而在 BAD1 中出现少许的性能下降.

此外, 如表 7 所示, FusionStereo-Realtime 在低光数据集 TarTanAir-Night 的端点误差 (EPE) 仅为 1.03. 因此, 融合代价体及其聚合方法即使在低尺度特征层使用, 仍能在多种场景中实现稳定的立体匹配.

4.4 FusionStereo 的模型复杂度分析

本节实验对比 FusionStereo 与代表性基准模型的计算复杂度和空间复杂度, 结果如表 6 所示. 结果表明, FusionStereo 在显著提升预测精度的同时, 计算复杂度比 SOTA 模型 RAFT-Stereo 降低约 40%, 空间复杂度降低约 28%. 这一性能提升主要归因于融合代价体所实现的以下优化: 首先, 该设计有效简化了计算流程, 通过消除原有方法中对左视图语义特征的提取过程, 在保证模型精度的同时显著降低了计算开销和存储需求; 其次, 融合代价体能够为视差细化阶段提供更加鲁棒的相关信息, 从而能够减少模型所需的迭代优化次数, 进一步提升计算效率.

4.5 FusionStereo 训练过程分析

图 5 展示了 FusionStereo 模型在训练过程中的损失函数变化曲线和验证集精度变化曲线. 其中红色折线 (train) 为使用图像增强的 SceneFlow 进行训练的曲线; 绿色折线 (test) 为 SceneFlow 的原始测试数据曲线; K12(15) 为 KITTI 2012(2015) 的跨域泛化测试曲线.

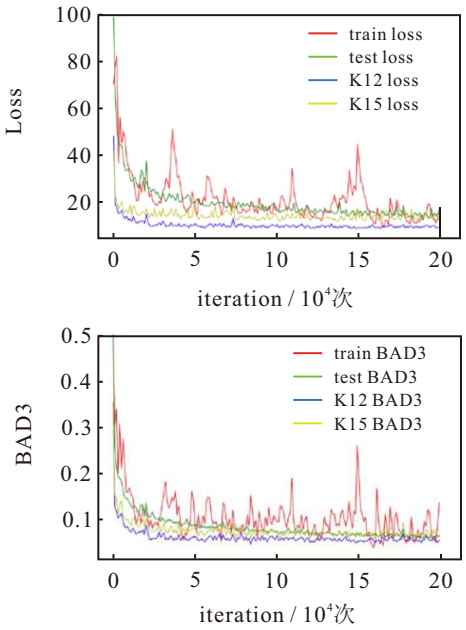

图5 FusionStereo 训练曲线

因为在训练阶段使用了亮度、伽玛和对比度变换的图像增强策略, 所以对原数据添加了较大扰动, 从而导致红色训练数据曲线的波动较大. 而相比于未进行图像增强的测试数据和跨域测试数据, FusionStereo 的收敛则更加稳定.

训练中期的训练曲线虽有较大波动, 但测试数据显示模型训练已逐渐趋于平稳收敛, 未出现明显的过拟合现象, 且跨域泛化测试精度能够保持较高水平.

5 讨 论

本文提出的融合代价体及其代价聚合方法在双目立体匹配任务中展现出显著的性能提升, 但其适用性和潜在改进空间不仅限于双目场景. 以下将讨论该方法在多目立体匹配、视频立体匹配等更复杂场景下的适用性, 并探讨未来改进方向.

多目立体匹配通过引入多视角图像, 能够进一步提升深度估计的精度和鲁棒性. 然而, 不同视角之间的几何一致性约束对同名点的匹配提出了更高的要求. 本文提出的融合代价体及其代价体聚合方法作为一种即插即用的模块, 通过联合利用特征向量的全局信息匹配与局部信息, 能够为多目立体匹配的候选像素相关计算提供思路, 从而提升多目立体匹配的性能.

视频立体匹配不仅需要考虑单帧图像的深度估计, 还需要处理帧与帧之间的时间一致性. 另外, 本文还发现 FusionStereo 虽然能够实现准确的单帧匹配, 但在真实使用过程中存在严重的帧间不稳定性,

尤其在弱光环境及低纹理区域. 因此, 后续将会借助稠密光流预测, 从此特征角度实现跨帧的视差对齐以及特征对齐, 以求能缓解此问题.

6 结 论

为解决现有代价体方法不能为立体视图特征提供全面、无冗余相关信息的问题, 本文结合极限约束对代价体的相关性计算区域进行限定, 提出了一种即插即用的融合代价体及其代价聚合方法. 为验证所提方法的有效性, 将所提出方法与代表性基准网络相集成, 并提出 FusionStereo 及实时版本. 实验中采用多个流行基准数据集进行评估, 结果显示 FusionStereo 在自动驾驶数据集 KITTI 2015 域内训练和跨域测评的误匹配指标 BAD3 上仅有 1.55% 和 5.22%, 在室内数据集 MiddleBurry 跨域测评的错误匹配指标 BAD1 上仅有 17.1%, 优于其他基于代价体方法的深度立体匹配算法, 并且在推理速度上比代表性基准模型 RAFT-Stereo 提高了 200%, 以更快的推理速度实现更高的准确性. 本文所提出实时版本 FusionStereo-Realtime 在 KITTI 数据集以 10 FPS 以上的速度运行, 且在部分指标的表现中仍保持先进性能. 综上所述, 本文所提出的融合代价体及其聚合方法极具优越性, 希望本方法能对基于深度网络的双目深度估计和三维环境感知领域有所帮助.

FusionStereo 虽已满足多种环境的单帧鲁棒立体匹配, 但是帧与帧之间的视差结果存在严重的不稳定问题, 因此在接下来的研究中将考虑借助深度稠密光流预测, 从特征角度建立跨帧的视差对齐以提高帧间稳定性.

参考文献 (References)

[1] 祝志坤, 卢丙举, 李一辰, 等. 基于单双目融合的 AUV 坐落式回收光视觉引导算法[J]. 控制与决策, 2025, 40(1): 28-37.
(Zhu Z K, Lu B J, Li Y C, et al. Light visual guidance algorithm for AUV situated recovery based on monocular and binocular fusion[J]. Control and Decision, 2025, 40(1): 28-37.)

[2] Papadimitriou D V, Dennis T J. Epipolar line estimation and rectification for stereo image pairs[J]. IEEE Transactions on Image Processing, 1996, 5(4): 672-676.

[3] 贾嫣晗, 邹风山, 徐方, 等. 完全在线的双目直接法视觉 SLAM 算法[J]. 控制与决策, 2023, 38(11): 3093-3102.
(Jia Y H, Zou F S, Xu F, et al. Fully online stereo direct vision SLAM algorithm[J]. Control and Decision, 2023, 38(11): 3093-3102.)

[4] 宋海涛, 何文浩, 原魁. 一种基于 SIFT 特征的机器人环境感知双目立体视觉系统[J]. 控制与决策, 2019, 34(7): 1545-1552.
(Song H T, He W H, Yuan K. A stereo vision system based on SIFT feature for robot environment perception[J]. Control and Decision, 2019, 34(7): 1545-1552.)

[5] Poggi M, Tosi F, Batsos K, et al. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5314-5334.

[6] Hirschmuller H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 328-341.

[7] Birchfield S, Tomasi C. Depth discontinuities by pixel-to-pixel stereo[J]. International Journal of Computer Vision, 1999, 35(3): 269-293.

[8] Hirschmüller H, Innocent P R, Garibaldi J. Real-time correlation-based stereo vision with reduced border errors[J]. International Journal of Computer Vision, 2002, 47(1): 229-246.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

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

[11] Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 4040-4048.

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

[13] Lipson L, Teed Z, Deng J. RAFT-stereo: Multilevel recurrent field transforms for stereo matching[C]. 2021 International Conference on 3D Vision. London, 2021: 218-227.

[14] Wang X Q, xu G W, Jia H, et al. Selective-stereo: Adaptive frequency information selection for stereo matching[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 19701-19710.

[15] Guo X Y, Yang K, Yang W K, et al. Group-wise correlation stereo network[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 3268-3277.

[16] Chang J R, Chen Y S. Pyramid stereo matching network[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 5410-5418.

[17] Shamsafar F, Woerz S, Rahim R, et al. MobileStereoNet: Towards lightweight deep networks for stereo matching[C]. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision.

Waikoloa, 2022: 677-686.

[18] Guo X D, Zhang C M, Zhang Y M, et al. LightStereo: Channel boost is all you need for efficient 2D cost aggregation[J/OL]. 2024, arXiv: 2406.19833.

[19] Xu B, Xu Y H, Yang X L, et al. Bilateral grid learning for stereo matching networks[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 12492-12501.

[20] Zhang F H, Prisacariu V, Yang R G, et al. GA-net: Guided aggregation net for end-to-end stereo matching[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 185-194.

[21] Li J K, Wang P S, Xiong P F, et al. Practical stereo matching via cascaded recurrent network with adaptive correlation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 16242-16251.

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

[23] Tankovich V, Hane C, Zhang Y D, et al. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 14357-14367.

[24] Xu G W, Cheng J D, Guo P, et al. Attention concatenation volume for accurate and efficient stereo matching[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 12971-12980.

[25] Shen Z, Dai Y, Song X, et al. Pcw-net: Pyramid combination and warping cost volume for stereo matching[C]. European Conference on Computer Vision. Tel Aviv: Springer, 2022: 280-297.

[26] Feng M J, Cheng J D, Jia H, et al. MC-stereo: Multi-peak lookup and cascade search range for stereo matching[C]. 2024 International Conference on 3D Vision. Davos, 2024: 344-353.

[27] Xu G W, Wang X Q, Ding X H, et al. Iterative geometry encoding volume for stereo matching[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 21919-21928.

[28] Xu J, Sun X, Zhang Z, et al. Understanding and improving layer normalization[J]. Advances in Neural Information Processing Systems, 2019: 32.

[29] Zhang F H, Qi X J, Yang R G, et al. Domain-invariant stereo matching networks[C]. Computer Vision – ECCV 2020. Cham: Springer, 2020: 420-439.

[30] Shen Z L, Dai Y C, Rao Z B. CFNet: Cascade and fused cost volume for robust stereo matching[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 13901-13910.

[31] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, 2012: 3354-3361.

[32] 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, 2: 427-434.

[33] Guo X, Lu J, Zhang C, et al. Openstereo: A comprehensive benchmark for stereo matching and strong baseline[J/OL]. 2024, arXiv: 2312.00343.

[34] Guan T F, Wang C, Liu Y H. Neural Markov random field for stereo matching[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 5459-5469.

[35] Zhao H L, Zhou H Z, Zhang Y J, et al. High-frequency stereo matching network[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 1327-1336.

[36] Xu G, Wang X, Zhang Z, et al. IGEV++: Iterative multi-range geometry encoding volumes for stereo matching[J/OL]. 2025, arXiv: 2409.00638.

[37] Chang T Y, Yang X, Zhang T Z, et al. Domain generalized stereo matching via hierarchical visual transformation[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 9559-9568.

[38] Zhang J W, Li J H, Huang L, et al. Robust synthetic-to-real transfer for stereo matching[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 20247-20257.

[39] Scharstein D, Hirschmüller H, Kitajima Y, et al. High-resolution stereo datasets with subpixel-accurate ground truth[C]. Pattern Recognition. Cham: Springer International Publishing, 2014: 31-42.

[40] Wang W S, Zhu D L, Wang X W, et al. TartanAir: A dataset to push the limits of visual SLAM[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, 2020: 4909-4916.

作者简介

邹正阳 (1997−), 男, 博士生, 主要研究方向为图像处理、三维视觉, E-mail: zhengyang_zou@163.com;

伍云霞 (1967−), 女, 教授, 博士生导师, 主要研究方向为计算机视觉、目标探测与跟踪, E-mail: wuyx@cumtb.edu.cn;

徐倩 (1989−), 女, 博士生, 主要研究方向为计算机视觉、多模态信息融合, E-mail: xuqian_2011@126.com.