中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-17

论文引用格式:Zou Yaqi, Zhu Lvwei, Gao Ying, Wang Xinjing, Li Jinyu. Fast Iterative Stereo Matching with Low-Dimensional Cost Volume Feedback Correction[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260431. (邹亚琦, 朱旅潍, 高颖, 王馨静, 李金玉. 低维代价体反馈修正的快速迭代立体匹配方法[J/OL]. 中国图象图形学报,XXXX:1-17. DOI: 10.11834/jig.260431.)[DOI:10.11834/jig.260431]

低维代价体反馈修正的快速迭代立体匹配方法

邹亚琦,朱旅潍,高颖,王馨静,李金玉

青岛科技大学数据科学学院,青岛　266100

摘　要: **目的**　高效立体匹配需要兼顾匹配精度与推理效率,而低维代价体压缩容易削弱复杂区域的匹配判别能力。针对弱纹理、遮挡和目标边界等区域的视差估计不稳定问题,本文提出一种基于低维代价体反馈修正的快速迭代立体匹配方法FLISNet(Fast Iterative Stereo Matching with Low-Dimensional Cost Volume Feedback Correction,FLISNet),以提升匹配稳定性与细节恢复能力。**方法**　首先,采用共享权重的特征提取网络获得多尺度匹配特征和上下文特征。随后,通过低维特征投影,通道注意力加权和分组相关计算构建紧凑代价表征,在降低代价体规模的同时保留有效匹配响应。在此基础上,设计反馈递归修正机制,将中间视差估计编码为空间反馈特征,并以残差调制方式对低维代价体进行自适应更新,使当前预测结果参与后续代价修正。其中,FLISNet和FLISNet+分别采用8次和18次反馈递归修正,对应标准设置和精度增强设置。最后,构建多尺度结构补偿模块,融合多尺度上下文特征与门控残差信息,采用由粗到细的级联补偿策略逐步恢复高分辨率视差细节。**结果**　在多类公开立体匹配数据集上的实验结果表明,FLISNet在保持较低推理耗时的同时获得了具有竞争力的匹配精度。标准版本FLISNet在Scene Flow上取得0.68px的EPE和32ms的推理耗时,达到与CoEx相近的精度水平。进一步增加反馈递归修正次数后,FLISNet+将Scene Flow上的EPE降至0.56 px,较Fast-ACVNet+和LightStereo分别降低5.1%和6.7%。在KITTI真实道路场景中,FLISNet+相比LBSM-L-Refine分别降低26.6%和26.2%的主要误差指标,说明增强版本能够进一步改善道路场景中的视差估计效果。在ETH3D数据集上,FLISNet+的Bad2.0误差为0.54%,低于所列对比方法。零样本实验结果表明,FLISNet在未进行目标域微调的情况下仍能保持较稳定的匹配性能,具有一定的跨域适应能力。**结论**　本文方法通过紧凑代价表征,反馈递归修正和多尺度结构补偿,缓解了低维代价体条件下匹配判别能力不足和复杂区域细节恢复不足的问题,在保持较低推理耗时的同时获得了具有竞争力的匹配精度,为快速立体匹配任务提供了一种有效方案。

关键词: 双目立体匹配;轻量化网络;低维代价体;迭代优化;视差细化

Fast Iterative Stereo Matching with Low-Dimensional Cost Volume Feedback Correction

Zou Yaqi, Zhu Lvwei, Gao Ying, Wang Xinjing, Li Jinyu

School of Data Science, Qingdao University of Science and Technology, Qingdao 266100

收稿日期:2026-07-15;**修回日期**:2026-10-22

* **通信作者**:高颖　gaoying@qust.edu.cn

基金项目:国家自然科学基金(62401310);山东省重点研发计划(重大科技创新工程)(2025CXGC010105);山东省高等学校青创科技支持计划(2024KJG053)

Supported by: National Natural Science Foundation of China (62401310); Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (2025CXGC010105); Shandong Provincial Higher Education Youth Innovation Technology Support Program (2024KJG053)

© 中国图象图形学报版权所有

Abstract: **Objective** Accurate and efficient disparity estimation is essential for three-dimensional scene understanding, autonomous driving, robotic perception, and image-based reconstruction. Stereo matching estimates dense correspondences between rectified left and right images, but reliable disparity estimation remains difficult when local appearance provides weak or ambiguous evidence. High-accuracy stereo networks often rely on high-dimensional cost volumes, computationally intensive aggregation, or large recurrent states. These designs improve the representation of candidate matches but also increase memory use and inference time. Compact stereo networks reduce computation by lowering feature dimensionality and simplifying cost aggregation. Excessive compression, however, can weaken the discrimination of similar disparity hypotheses, especially in weakly textured regions, occlusions, and areas near object boundaries. Moreover, when an iterative estimator repeatedly consults an unchanged matching representation, early disparity errors can affect later updates. To balance estimation quality and computational efficiency, this paper proposes FLISNet, a fast iterative stereo matching network with low-dimensional cost volume feedback correction. The method combines compact matching evidence, disparity-guided recursive correction, and multi-scale structural compensation so that intermediate predictions can contribute to subsequent cost refinement. **Method** FLISNet takes a rectified stereo pair as input and uses a shared MobileNetV2 backbone with a U-Net-style reconstruction structure to extract multi-scale matching features. A separate context branch provides features at full, half, and quarter input resolutions. The compact cost representation combines low-dimensional feature projection, channel weighting, and group-wise correlation. Specifically, the left and right matching features are projected into a lower-dimensional space, and learnable channel-wise weights shared across spatial positions emphasize informative responses before correlation is computed within feature groups. This procedure limits cost volume size while retaining discriminative correspondence cues. Recursive feedback correction on the compact cost representation operates at one-quarter of the input resolution. The initial disparity at this resolution is set to zero. A convolutional gated recurrent unit predicts successive disparity residuals, and each intermediate estimate is encoded into a spatial feedback feature. A gate generated from this feature modulates the low-dimensional cost responses through a residual connection. The updated cost responses are then used for subsequent recurrent updates, allowing the current disparity estimate to influence the matching representation itself. The standard FLISNet performs eight feedback updates, while FLISNet+ performs eighteen updates for higher accuracy. Following the final update, the gating features are resized to match the spatial resolutions of the structural compensation stages. The network combines the final quarter-resolution disparity estimate with context and gating features, then progressively restores details at one-quarter, one-half, and full resolution relative to the input image. Three output stages are jointly supervised with loss weights of 0.5, 0.7, and 1.0. This coarse-to-fine compensation process is designed to improve local structures and disparity discontinuities without introducing a large cost volume. **Result** Experiments on Scene Flow, KITTI 2012, KITTI 2015, ETH3D, and Middlebury examine matching accuracy, inference efficiency, and performance across different scene types. On Scene Flow, the standard FLISNet achieves an endpoint error (EPE) of 0.68 pixels with an inference time of 32 ms. CoEx obtains 0.69 pixels and 33 ms in the same comparison. FLISNet+ reduces the Scene Flow EPE to 0.56 pixels with a runtime of 56 ms. This EPE is 5.1% and 6.7% lower than those of Fast-ACVNet+ and LightStereo, respectively, in the same comparison. Increasing the recursive update count from six to twenty reduces EPE from 0.73 to 0.55 pixels, while inference time increases from 27 to 62 ms. The additional EPE reduction from eighteen to twenty updates is only 0.01 pixels, supporting eighteen updates as the accuracy-enhanced setting. On KITTI 2012, FLISNet+ obtains a 3-all error of 1.77% (all pixels with valid ground truth, 3-pixel threshold). Compared with the 2.41% reported for LBSM-L-Refine in the same table, this represents a relative reduction of 26.6%. On ETH3D, FLISNet+ achieves a Bad 2.0 error of 0.54%, the lowest among the methods listed in the comparison. In the zero-shot setting, the standard FLISNet is trained only on Scene Flow and evaluated directly on KITTI 2012, KITTI 2015, and Middlebury. The zero-shot comparison indicates that FLISNet can be applied to unseen datasets without target-domain fine-tuning and still produce relatively stable disparity estimates. Ablation experiments evaluate the role of each component under the same training and evaluation conditions. Replacing the compact cost representation with a standard correlation cost volume increases Scene Flow EPE from 0.68 to 0.72 pixels. Replacing residual feedback modulation with a two-dimensional lookup raises the EPE to 0.74 pixels, while replacing multi-scale compensation with direct upsampling raises it to 0.79 pixels. Using lower-resolution update scales increases the EPE to 0.78 pixels and the inference time to 37 ms. Qualitative

© 中国图象图形学报版权所有

comparisons show clearer object boundaries and improved continuity in weakly textured regions. Inference times in this study were measured on a single NVIDIA RTX 3090 GPU with batch size 1, 384 × 1248 inputs, and FP32 computation. Each measured value is the average over 200 runs after 50 warm-up runs; runtimes quoted from original papers are marked separately. **Conclusion** FLISNet combines low-dimensional matching features with disparity-guided residual modulation and multi-scale structural compensation. The experiments show that the standard configuration maintains competitive disparity accuracy with modest inference time, while increasing the update count improves accuracy when a higher runtime is acceptable. The ablation results support the contributions of compact cost representation, feedback correction, and coarse-to-fine compensation. Future work will investigate model compression, knowledge distillation, and low-cost feature adaptation, along with adaptive matching methods for visually degraded environments.

Key words: stereo matching; lightweight network; low-dimensional cost volume; iterative optimization; disparity refinement

论文引用格式：Zou Yaqi, Zhu Lvwei, Gao Ying, Wang Xinjing, Li Jinyu. Fast Iterative Stereo Matching with Low-Dimensional Cost Volume Feedback Correction[J/OL]. Journal of Image and Graphics. DOI:10.11834/jig.260431.(邹亚琦,朱旅潍,高颖,王馨静,李金玉. 低维代价体反馈修正的快速迭代立体匹配方法[J/OL]. 中国图象图形学报. DOI:10.11834/jig.260431.)

本文代码已开源至ScienceDB,链接:https://doi.org/10.57760/sciencedb.j00240.0009n

The source code is available on ScienceDB: https://doi.org/10.57760/sciencedb.j00240.0009n

0 引 言

双目立体匹配是计算机视觉领域的重要研究方向,其目标是从经过极线校正的双目图像对中建立像素级对应关系,进而恢复场景的三维几何结构。作为连接二维图像感知与三维场景理解的重要桥梁,该技术已广泛应用于自动驾驶,机器人感知,三维重建以及复杂环境视觉感知等任务中。典型的双目立体匹配流程通常包括特征提取,代价体构建,代价聚合和视差回归等步骤。其中,特征表示的判别能力,代价体的表达形式以及后续视差优化策略,直接影响最终视差估计的精度与效率。

早期立体匹配方法主要依赖人工设计的匹配代价和显式优化策略。半全局匹配方法通过沿多个方向进行路径代价聚合,在局部匹配和全局优化之间取得了较好的平衡(Hirschmuller,2008)。部分方法将Hopfield神经网络引入视差空间,以增强边界区域的匹配约束(Xu等,2013)。图像块几何约束与视差概率建模被用于构建分割块关系,进而提升区域级匹配的可靠性(Zhai等,2010)。针对弱纹理区域的代价表达,融合引导图滤波和快速加权中值滤波的方案得到了有效应用(Han和Yang,2018)。这类方法具有较强的可解释性,但其性能表现高度依赖人工设计特征和局部平滑假设,在复杂光照,遮挡和视差不连续区域中的鲁棒性有限。

随着深度学习的发展,端到端立体匹配方法通过学习型特征表示和代价体建模提升了复杂区域的视差估计精度。Mayer等构建的大规模合成立体匹配数据集为深度网络训练提供了基础(Mayer等,2016)。GCNet通过四维代价体和三维卷积正则化实现端到端视差估计(Kendall等,2017)。PSMNet进一步引入空间金字塔池化和级联沙漏结构,以增强多尺度上下文建模能力(Chang和Chen,2018)。随后,AANet,BGNet和ACVNet分别从自适应聚合,双边网格学习和注意力代价体构建等角度优化代价聚合过程(Xu和Zhang,2020;Xu等,2021;Xu等,2022)。这些方法在公开数据集上取得了较高精度,但通常需要构建高维代价体或复杂聚合网络,导致计算和存储开销较大。

Transformer结构和递归迭代机制进一步推动了高精度立体匹配的发展。STTR将立体匹配建模为序列到序列问题,利用注意力机制建立全局对应关系(Li等,2021)。贾迪等人针对弱纹理目标匹配困难问题引入Transformer结构,以增强全局上下文表达能力(Jia等,2024)。在迭代优化方面,RAFT-Stereo利用递归更新机制持续修正视差结果(Lipson等,2021)。HITNet采用分层迭代细化结构处理高分辨率图像对(Tankovich等,2021)。IGEV通过几何编码代价体融合局部匹配与非局部几何信息。

© 中国图象图形学报版权所有

IGEV++进一步构建多范围几何编码体,并结合自适应块匹配和选择性几何特征融合,同时建模大视差区域的粗粒度几何信息与细节区域的细粒度匹配信息(Xu等,2023;Xu等,2025)。这类方法侧重通过更充分的几何信息建模与递归细化提升匹配精度,也进一步推动了立体匹配向更强表征能力方向发展。

面向移动端与边缘计算设备的实时部署需求,高效立体匹配方法通常通过降低网络复杂度、压缩代价体规模和简化聚合结构来减少计算开销。StereoNet利用低分辨率视差估计和引导式层次细化实现快速边缘感知深度预测(Khamis等,2018)。AnyNet采用可随时退出的粗到细结构,使模型能够根据计算资源动态输出视差结果(Wang等,2019)。MobileStereoNet则面向移动端设计轻量化匹配网络(Shamshad等,2022)。这些方法主要从网络结构和部署效率角度提升实时性。

在兼顾匹配精度与计算效率的研究中,紧凑代价建模和高效聚合成为另一重要方向。CoEx通过通道激励增强匹配表达。Fast-ACVNet利用低分辨率相关线索生成高置信视差假设及注意力权重,以缩减视差搜索空间。LightStereo通过增强三维代价体中的视差通道实现高效二维聚合。ADStereo利用跨尺度视差对齐改善多尺度匹配信息。LBSM则采用二维卷积和轻量化通道注意力降低代价聚合开销(Hong等,2021;Xu等,2024;Guo等,2025;Wang等,2025;Wu等,2026)。

部分方法进一步从递归更新和几何信息传播角度增强视差优化能力。CREStereo利用级联递归与自适应相关持续更新局部匹配信息。GIP-Stereo通过多尺度特征聚合、传播增强相关体和几何感知交互细化,使几何信息贯穿特征提取、代价构建和视差细化过程。MoCha-Stereo则通过多层级特征交互增强匹配表征(Li等,2022;Zhao等,2025;Chen等,2024)。此外,Zhang等(2026)将几何先验约束引入双目深度估计,通过多尺度特征、多级GRU更新以及左右视差一致性和视差结构一致性约束强化几何信息引导。Yuan等(2026)则从更广泛的多视角立体匹配与三维重建角度,对图像投影和几何推理等技术路线进行了系统梳理。

针对上述问题,本文提出一种基于低维代价体反馈修正的快速迭代立体匹配方法(Fast Iterative Stereo Matching with Low-Dimensional Cost Volume Feedback Correction,FLISNet)。该方法通过紧凑代价表征降低匹配计算开销,并将视差反馈调制和多尺度结构补偿融入递归更新与视差细化过程,在保持较低推理耗时的同时改善低维代价体条件下的匹配稳定性和复杂区域视差恢复质量。主要贡献如下:

1)提出一种紧凑代价表征。通过低维特征投影、通道注意力和分组相关构建紧凑匹配空间,在降低代价建模开销的同时保留有效匹配响应,并为后续中间视差驱动的反馈修正提供匹配表征基础。

2)设计反馈递归修正机制。该机制将中间视差估计编码为空间反馈特征,以残差调制的方式对低维代价体进行自适应更新,使当前视差参与后续匹配表征更新,并持续引导视差残差预测。

3)构建多尺度结构补偿模块。该模块融合多尺度上下文特征与门控残差信息,采用由粗到细的级联补偿策略逐步恢复高分辨率视差细节,并保留目标边界与局部结构信息,生成全分辨率视差图。

4)在Scene Flow、KITTI 2012、KITTI 2015和ETH3D等数据集上进行实验验证。结果表明,本文方法在保持较低推理耗时的同时取得了具有竞争力的匹配精度,实现了精度与效率的良好折中。

1 方 法

本文提出的FLISNet整体框架如图1所示。给定输入的左右图像,网络首先通过共享权重的特征提取器获得左右图像的多尺度匹配特征,并通过参考图像的上下文分支提取多尺度上下文特征。随后利用低维特征投影,通道注意力加权和分组相关计算构建紧凑注意力代价表征,以降低代价体规模和计算开销。在此基础上,反馈递归修正模块将中间视差估计作为空间引导信息,对低维代价体进行残差调制,并逐步修正视差预测误差。最后,多尺度结构补偿模块结合上下文特征与门控残差信息,采用由粗到细的级联补偿策略恢复全分辨率视差图。

1.1 特征提取

给定输入双目图像对$\mathbf{I}_l,\mathbf{I}_r\in\mathbf{R}^{3\times H\times W}$,本文采用共享权重的MobileNetV2-100作为主干特征编码器提取左右图像的初始匹配特征。为融合局部细节与高层上下文信息,进一步引入U-Net风格的特征重

© 中国图象图形学报版权所有

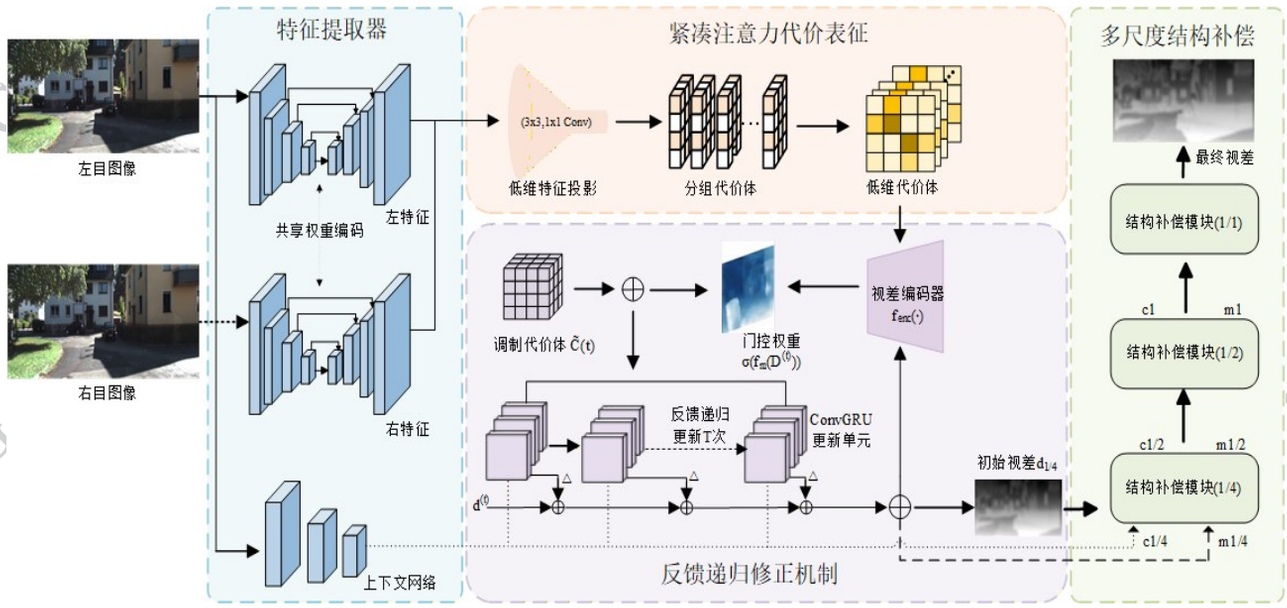


图1　FLISNet整体框架

Fig. 1　Overall framework of FLISNet

构模块。通过自顶向下的上采样和跨尺度跳跃连接实现多尺度特征融合，得到不同分辨率下的左右匹配特征：

$$\{\mathbf{F}_l^s,\mathbf{F}_r^s\}\in\mathbf{R}^{C_s\times\frac{H}{s}\times\frac{W}{s}},\quad s\in\{2,4,8,16,32\},,\qquad(1)$$

C_s表示下采样倍率为s时特征图的通道数，H和W分别表示输入图像的高度和宽度。不同尺度的匹配特征分别侧重局部纹理，边缘结构和高层语义信息。

本文还构建了由残差块与下采样层组成的上下文特征提取分支，用于从参考图像中提取原始分辨率，1/2和1/4分辨率下的多尺度上下文特征：

$$Q_s\in\mathbb{R}^{16\times\frac{H}{s}\times\frac{W}{s}},\quad s\in\{1,2,4\},\qquad(2)$$

Q_s表示尺度s下的上下文特征，通道数统一设置为16，以保证不同尺度特征在后续融合过程中的维度一致。上下文特征主要用于递归更新和视差细化阶段。

1.2　紧凑注意力代价表征

紧凑注意力代价表征由低维特征投影、通道注意力加权和分组相关计算组成，用于控制代价体规模并形成分组匹配响应。

首先，通过3×3卷积和1×1卷积组成的通道压缩映射，将原始特征$\mathbf{F}_l$, $\mathbf{F}_r$投影到一个低维嵌入空间中：

$$\mathbf{E}_l=\phi(\mathbf{F}_l),\quad \mathbf{E}_r=\phi(\mathbf{F}_r),\qquad(3)$$

式中，$\phi(\cdot)$表示特征嵌入函数，$\mathbf{E}_l,\mathbf{E}_r\in\mathbb{R}^{C'\times H_s\times W_s}$，$C'$表示低维嵌入特征的通道数，且$C'\ll C$。该低维投影用于减少冗余通道，降低后续代价体构建开销。

为增强对视差判别更敏感特征通道的响应，本文引入一个可学习的通道注意力权重$\mathrm{W}\in\mathrm{R}^{C'}$，并通过Sigmoid函数对其进行归一化处理：

$$\tilde{\mathrm{W}}=\sigma(\mathrm{W}),\qquad(4)$$

$\tilde{\mathrm{W}}$为空间共享的通道注意力权重。随后，对低维特征进行通道加权：

$$\hat{\mathrm{E}}_l=\tilde{\mathrm{W}}\odot\mathrm{E}_l,\quad \hat{\mathrm{E}}_r=\tilde{\mathrm{W}}\odot\mathrm{E}_r,\qquad(5)$$

式中，$\odot$表示逐元素乘法。

为了降低特征幅值差异对匹配相关计算的影响，本文进一步对加权后的低维特征在通道维度上进行L_2归一化：

$$\overline{\mathrm{E}}_{\mathrm{l}}(\mathrm{x},\mathrm{y})=\frac{\hat{\mathrm{E}}_{\mathrm{l}}(\mathrm{x},\mathrm{y})}{\|\hat{\mathrm{E}}_{\mathrm{l}}(\mathrm{x},\mathrm{y})\|_{\mathrm{L}_2}},\overline{\mathrm{E}}_{\mathrm{r}}(\mathrm{x},\mathrm{y})=\frac{\hat{\mathrm{E}}_{\mathrm{r}}(\mathrm{x},\mathrm{y})}{\|\hat{\mathrm{E}}_{\mathrm{r}}(\mathrm{x},\mathrm{y})\|_{\mathrm{L}_2}},\qquad(6)$$

接下来，将低维特征通道划分为G个分组，每个分组包含C'/G个通道。给定视差假设d，第g个通道组的匹配响应定义为：

$$C_g(\mathrm{x},\mathrm{y},\mathrm{d})=\sum_{c=1}^{C'/G}\overline{\mathrm{E}}_{\mathrm{l},g}^{c}(\mathrm{x},\mathrm{y})\cdot\overline{\mathrm{E}}_{\mathrm{r},g}^{c}(\mathrm{x}-\mathrm{d},\mathrm{y}),\qquad(7)$$

式中，$\mathrm{d}\in[0,\mathrm{D}_s]$，$\mathrm{D}_s$表示当前尺度下的最大视差搜索范围。$\overline{\mathrm{E}}_{\mathrm{l},g}^{c}$和$\overline{\mathrm{E}}_{\mathrm{r},g}^{c}$分别表示第g个通道组中的第c个通道特征。

最后，将所有通道组的匹配响应进行聚合，得到

© 中国图象图形学报版权所有

尺度s下的低维注意力代价体：

$$C(x,y,d)=\sum_{g=1}^{G}C_g(x,y,d)。\tag{8}$$

Fast-ACVNet利用低分辨率相关线索生成视差候选及注意力权重，并据此缩小视差搜索范围。LightStereo通过增强三维代价体中的视差通道进行二维代价聚合。本文将左右匹配特征投影至低维空间，在分组相关计算前利用空间共享的通道权重调整特征响应，形成紧凑代价。该代价用于反馈递归修正，中间视差状态被编码为空间反馈特征，对低维代价响应进行调制，使当前预测参与下一次匹配表征更新。

1.3 反馈递归修正机制

反馈递归修正在1/4分辨率下进行。开始前，将该尺度的初始视差场设置为零，即$d^{(0)}=0$。首次递归更新由紧凑代价表征与对应尺度的上下文特征共同驱动。随后，各次迭代基于当前视差状态生成空间反馈信息，并持续更新匹配表征。

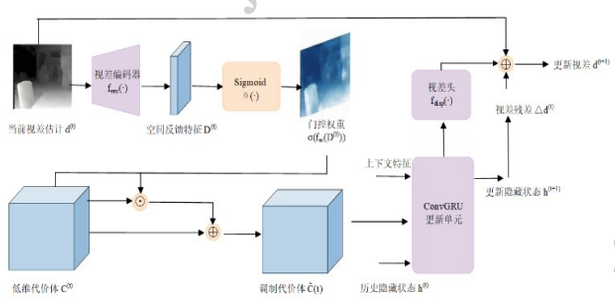


图2 视差反馈递归修正过程

Fig. 2 Disparity-feedback recursive correction process

如图2所示，在第t次迭代过程中，当前视差估计记为$d^{(t)}$，本文首先通过视差编码网络$f_{\mathrm{enc}}(\cdot)$对$d^{(t)}$进行特征编码，将其映射为空间反馈特征：

$$D^{(t)}=f_{\mathrm{enc}}\left(d^{(t)}\right),\tag{9}$$

$f_{\mathrm{enc}}(\cdot)$由若干3×3卷积层组成，用于提取当前视差估计中的局部空间结构信息。随后将空间反馈特征映射为门控权重，并采用残差调制方式更新低维代价体：

$$M_s^{(t)}=\sigma\left(f_m\left(D^{(t)}\right)\right)。\tag{10}$$

$$\hat{C}_s^{(t)}=C_s^{(t)}+C_s^{(t)}\odot M_s^{(t)},\tag{11}$$

式中，$f_{\mathrm{m}}(\cdot)$表示门控映射函数，用于将空间反馈特征映射到与代价体一致的维度。$C_s^{(t)}$表示第t次迭代时尺度s下的低维注意力代价体，$\hat{C}_s^{(t)}$为调制后的代价体，$M_s^{(t)}$为对应尺度下的门控权重。残差式调制在保留原始代价响应的基础上，根据当前视差估计调整匹配响应，供后续递归更新使用。

调制后的代价体特征$\hat{C}_s^{(t)}$与对应尺度的上下文特征Q_s首先拼接形成当前更新输入：

$$X_s^{(t)}=\left[\hat{C}_s^{(t)},Q_s\right],\tag{12}$$

随后，当前更新输入$X_s^{(t)}$与历史隐藏状态$h_s^{(t)}$共同送入对应尺度的更新单元：

$$h_s^{(t+1)}=G_s\left(h_s^{(t)},X_s^{(t)}\right),\tag{13}$$

式中，$h_s^{(t)}$表示第t次迭代时尺度s下的隐藏状态。$G_s(\cdot)$表示对应尺度下的卷积门控循环单元(convolutional gated recurrent unit，ConvGRU)。

在反馈递归修正阶段，利用1/4分辨率下更新后的隐藏状态预测当前迭代的视差残差：

$$\Delta d^{(t)}=f_{\mathrm{disp}}\left(h_{1/4}^{(t+1)}\right),\tag{14}$$

式中，$f_{\mathrm{disp}}(\cdot)$表示由卷积层构成的视差预测头，用于将更新后的隐藏状态映射为视差残差。$h_{1/4}^{(t+1)}$表示第$t+1$次迭代后在1/4分辨率尺度下的隐藏状态。$\Delta d^{(t)}$表示当前迭代预测得到的视差增量。

随后，将该视差残差累加到当前视差估计结果中，完成一次递归更新：

$$d^{(t+1)}=d^{(t)}+\Delta d^{(t)}。\tag{15}$$

递归更新过程中产生的门控信息进一步传递至多尺度结构补偿阶段。完成T次递归更新后，保留最终递归状态对应的门控表示，并根据后续结构补偿层的空间尺度进行尺度变换，得到$\{m_{1/4},m_{1/2},m_1\}$，分别用于1/4、1/2和原始分辨率下的结构细化。

RAFT-Stereo利用相关金字塔查询匹配信息，CREStereo通过自适应相关与级联递归更新视差。本文将当前视差估计编码为空间反馈特征，以门控残差方式调制低维代价响应；调制后的代价特征与上下文特征共同送入ConvGRU，预测下一次视差增量。中间视差状态由此参与后续匹配表征更新。

本文将递归更新过程展开为固定迭代次数，其中标准版本设置为T=8，FLISNet+设置为T=18，以在推理效率和视差精度之间取得平衡。

1.4 多尺度结构补偿模块

在完成T次反馈递归修正后，网络获得1/4分辨率视差估计$d^{(T)}$，同时将最终递归状态产生的门控信息转换为与后续结构补偿尺度对应的$\{m_{1/4},m_{1/2},m_1\}$。多尺度结构补偿模块以$d^{(T)}$为初始

© 中国图象图形学报版权所有

输入，并结合各尺度上下文特征$\{c_{1/4}, c_{1/2}, c_1\}$与对应门控信息，采用由粗到细的级联残差补偿策略逐步恢复高分辨率视差细节。

首先，在1/4分辨率层级，将反馈递归得到的视差$d^{(T)}$、门控信息$\mathrm{m}_{1/4}$和对应的上下文特征$\mathrm{c}_{1/4}$沿通道维度拼接，并通过视差细化模块预测残差：

$$\Delta d_{1/4} = R_{1/4}(\mathrm{Concat}[d^{(T)}, m_{1/4}, c_{1/4}]), \tag{16}$$

式中，Concat(·)表示通道拼接操作，$R_{1/4}(\cdot)$表示1/4分辨率下的视差细化模块。随后，将预测得到的残差与递归阶段输出的视差相加，得到：

$$d_{1/4} = d^{(T)} + \Delta d_{1/4}。 \tag{17}$$

在此基础上，将$d_{1/4}$双线性上采样至1/2分辨率，并按照尺度变化校正视差幅值，得到1/2尺度的输入视差$\hat{d}_{1/2}$。随后，将$\hat{d}_{1/2}$、门控信息$m_{1/2}$和对应上下文特征$c_{1/2}$进行融合，由1/2分辨率细化模块预测视差残差，并将其与输入视差相加得到该尺度的细化结果。

最后，将1/2分辨率下的细化结果进一步恢复至原始分辨率。在全分辨率层级，将当前视差、门控信息m_1与上下文特征c_1进行融合，并通过全分辨率细化模块完成最终残差修正，得到全分辨率视差输出。

1.5 损失函数

在多尺度级联视差细化模块中，网络在一次前向传播过程中生成三阶段视差预测$\{d_1, d_2, d_3\}$，分别对应由粗到细的逐级视差恢复结果。为约束不同阶段的预测结果，本文对多阶段输出进行联合监督。

对于低分辨率阶段的预测结果，先将其上采样并校正至原始分辨率，再与真实视差d^{gt}计算损失。整体训练损失定义为：

$$L_{\mathrm{seq}} = \sum_{i=1}^{3} w_i \left\| \hat{d}_i - d^{\mathrm{gt}} \right\|_1, \tag{18}$$

式中，$\hat{d}_i$表示第i个阶段上采样并校正至原始分辨率后的预测视差，$\{w_1, w_2, w_3\} = \{0.5, 0.7, 1.0\}$表示不同阶段的监督权重。该损失用于对由粗到细的多阶段视差预测结果进行联合约束。

2 实　验

2.1 数据集

本文在Scene Flow、KITTI 2012、KITTI 2015、ETH3D和Middlebury公开立体匹配数据集上对所提出方法进行实验验证。上述数据集覆盖了合成场景、真实自动驾驶场景、高分辨率室内场景以及复杂室内外环境，能够从多个角度评估模型性能。

Scene Flow（Mayer等，2016）是大规模合成数据集，包含35454对训练图像和4370对测试图像，图像分辨率为540×960，并提供稠密视差真值。该数据集场景变化丰富，视差标注完整，常用于立体匹配模型的预训练和基础精度评估。本文选用其Cleanpass子集用于预训练和测试。KITTI 2012（Geiger等，2012）和KITTI 2015（Menze和Geiger，2015）为真实车载双目数据集，图像中包含道路、车辆、行人及复杂光照变化，视差真值由激光雷达点云辅助生成，具有稀疏标注特点。与合成数据相比，KITTI数据集更接近自动驾驶场景中的真实成像条件。本文将这两个数据集用于评估模型在真实道路场景下的匹配精度和实际应用性能。ETH3D（Schöps等，2017）包含多种真实室内外场景，涵盖纹理突变、光照交替和深度跳跃等挑战，能够反映模型在非道路真实场景中的适应能力。本文使用ETH3D评估模型在真实复杂场景中的匹配稳定性。Middlebury（Scharstein等，2014）为高分辨率室内数据集，包含复杂几何结构，丰富纹理变化和精细物体边界，对视差细节恢复能力要求较高。本文主要在零样本设置下使用该数据集，测试模型在未见高分辨率室内场景中的泛化能力。

2.2 评价指标

端点误差（EPE）是立体匹配任务中最常用的评价指标之一，主要用于衡量预测视差与真实视差之间的平均绝对误差。其数学定义如下：

$$\mathrm{EPE} = \left(\frac{1}{N}\right)\sum_{i=1}^{N} \left| d_i^{\mathrm{pred}} - d_i^{\mathrm{gt}} \right|, \tag{19}$$

d_i^{pred}和d_i^{gt}分别表示第i个像素点的预测视差和真实视差，N为参与计算的有效像素数量。EPE指标能够直观反映模型在整体视差估计上的平均误差水平。EPE数值越小，表示预测视差与真实视差越接近。

错误像素比例是一类基于误差阈值的评价指标，用于统计视差误差超过给定阈值的像素所占比例。该指标的数学定义如下：

$$\mathrm{Bad}x = \frac{1}{N}\sum_{i=1}^{N} I\left(\left| d_i^{\mathrm{pred}} - d_i^{\mathrm{gt}} \right| > x\right), \tag{20}$$

式中，$I(\cdot)$为指示函数，x表示误差阈值。本文采用

© 中国图象图形学报版权所有

Bad1.0、Bad2.0和Bad3.0等指标评价不同误差阈值下的匹配性能。

在KITTI数据集中，通常采用D1指标对算法性能进行评估。当像素的视差误差同时满足以下条件之一时，该像素被判定为错误匹配像素：

$$\left|d_i^{\text{pred}}-d_i^{\text{gt}}\right|>3 \text{或} \frac{\left|d_i^{\text{pred}}-d_i^{\text{gt}}\right|}{d_i^{\text{gt}}}>5\%。 \tag{21}$$

KITTI数据集进一步细化了评价区域D1-noc和D1-all分别表示非遮挡区域和全图的错误像素比例。此外，该指标体系还划分了前景与背景，并重点关注反光区域的匹配误差。

除了视差估计精度之外，本文还采用推理耗时（Runtime）评价模型的计算效率，单位为毫秒（ms）。

2.3 实验设置

本文提出的FLISNet网络基于PyTorch框架实现，并采用AdamW优化器进行参数更新。训练过程分为预训练和微调两个阶段。首先，对网络进行500,000次迭代的预训练，初始学习率设置为0.001。为增强模型对不同成像条件的适应能力，预训练过程中以20%的概率对输入双目图像进行数据增强，主要包括颜色抖动、亮度增强和对比度调整等操作，以模拟真实场景中可能出现的曝光变化。

在完成预训练后，本文进一步对模型进行参数微调。其中，真实道路场景下的微调过程共训练60个epoch，学习率设置为0.0002。复杂室内外场景下的微调过程共训练68个epoch，学习率同样设置为0.0002。除常规微调实验外，本文还设计了跨数据集泛化测试。在该测试中，模型不进行任何针对目标数据分布的额外训练，直接利用已有参数完成前向推理，以评估其在未知场景中的泛化能力。

推理耗时分为原论文报告结果和本文统一平台实测结果。对于具有官方开源实现且能够统一复现的方法，本文采用单张NVIDIA RTX 3090 GPU进行统一测试。其余方法保留原论文报告的运行时间，并在表中明确标注来源。考虑到不同文献在GPU型号、输入分辨率及具体测试设置等方面存在差异，原论文报告的运行时间仅作为效率量级参考，不用于严格的直接速度排序。统一实测采用batch size=1，输入张量尺寸固定为384×1248，并保持一致的FP32计算设置。测试前进行50次前向预热，随后连续执行200次推理并统计平均运行时间。

2.4 反馈递归修正次数分析

为分析反馈递归修正机制对视差估计精度的影响，并分析迭代更新过程带来的推理开销，本文首先在Scene Flow数据集上进行递归次数T的参数分析。在保持网络结构、模型参数量、代价体维度和训练设置不变的条件下，仅改变递归更新次数T，评估不同迭代次数下FLISNet的性能变化，结果如表1所示。

表1 反馈递归修正次数实验

Table 1 Effect of Feedback Recursive Correction Iterations

修正次数	6	8	10	12	16	18	20
EPE(px)	0.73	0.68	0.65	0.61	0.59	0.56	0.55
耗时(ms)	27	32	37	42	49	56	62

随递归更新次数增加，模型在Scene Flow数据集上的EPE整体呈下降趋势。当迭代次数由6次增加至12次时，EPE从0.73 px降至0.61 px，误差下降约16.4%，表明在该阶段递归更新能够利用低维代价表征和中间视差反馈优化匹配结果。继续增加迭代次数后，EPE仍进一步下降，但边际收益逐渐减小。当迭代次数由18次增加至20次时，EPE仅由0.56 px降至0.55 px，说明模型的视差估计逐渐接近收敛。

与此同时，推理耗时随递归次数增加近似线性增长，由6次迭代时的27 ms增加至20次迭代时的62 ms。综合考虑精度增益与推理开销，本文将T=8设为标准FLISNet的默认配置，此时EPE为0.68 px、推理耗时为32 ms。将T=18设为FLISNet+的精度增强配置，此时EPE进一步降至0.56 px，而继续增加至20次仅带来0.01 px的额外改善。由此，T=8与T=18分别对应效率优先和精度增强的两种配置。

2.5 消融实验

为验证FLISNet中各关键设计对视差估计精度和推理效率的影响，本文在相同训练和测试条件下进行消融实验。完整FLISNet作为基线模型，其端点误差为0.68px，推理耗时为32ms。实验分别对紧凑注意力代价表征，反馈递归修正机制，多尺度结构补偿模块以及递归更新尺度进行替换或调整，以分析各部分对模型性能的独立贡献。

由表2可知，将紧凑注意力代价表征替换为普通相关代价体后，推理耗时由32ms降至30ms，但

© 中国图象图形学报版权所有

EPE由0.68px升至0.72px，表明普通相关代价体虽然计算更简单，但匹配判别能力不足。将反馈递归修正机制中的残差调制更新替换为二维查找方式后，EPE升至0.74px，耗时增加至34ms，说明二维查找不仅未带来效率优势，也削弱了模型对早期视差误差的逐步修正能力。将多尺度结构补偿模块替换为直接上采样后，EPE升至0.79px，性能下降最为明显，表明简单插值难以恢复目标边界和细小结构信息。将更新尺度调整为[1/16，1/8，1/4]后，EPE升至0.78px且耗时增至37ms，说明过低分辨率下的迭代更新会削弱空间结构信息的保留，并可能在后续视差恢复过程中放大误差。总体来看，各项模块替换均导致EPE不同程度上升，说明FLISNet中的紧凑代价表征，反馈递归修正和多尺度结构补偿设计对最终视差估计具有积极作用。

表2　FLISNet关键模块消融实验

Table 2　Ablation Study on Key Modules of FLISNet

核心组件	基线配置	变体配置	端点误差EPE变化	推理耗时(ms)变化
代价表征	紧凑注意力代价表征	普通相关代价体	0.68→0.72	32→30
递归修正	残差调制更新	二维查找	0.68→0.74	32→34
结构补偿	多尺度结构补偿	直接上采样	0.68→0.79	32→30
更新尺度	[1/4,1/2,1]	[1/16,1/8,1/4]	0.68→0.78	32→37

为分析多阶段联合监督中损失权重的选择依据及模型对权重变化的敏感性，本文进一步开展损失权重敏感性实验。三阶段预测按照由粗到细的方式逐步恢复视差，其中最终阶段直接对应高分辨率输出，因此固定$W_3 = 1.0$，并通过调整前两阶段权重W_1和W_2比较不同监督强度的影响。实验保持网络结构、训练数据、优化器、学习率以及其他训练设置不变，仅调整$\{W_1, W_2, W_3\}$，并在Scene Flow数据集上进行评估。

表3　多阶段损失权重敏感性实验

Table 3　Sensitivity Analysis of Multi-stage Loss Weights

分组	W_1	W_2	W_3	EPE(px)
A	1.0	1.0	1.0	0.75
B	0.3	0.5	1.0	0.70
C	0.5	0.7	1.0	0.68
D	0.7	0.9	1.0	0.71

由表3可见，不同多阶段损失权重设置会对视差估计精度产生一定影响。A组采用等权监督{1.0，1.0，1.0}，EPE为0.75 px。B组减弱前两阶段监督，将权重设置为{0.3，0.5，1.0}后，EPE下降至0.70 px。C组采用{0.5，0.7，1.0}的递增式监督权重，EPE进一步降低至0.68 px，取得最优结果。D组继续提高前两阶段权重至{0.7，0.9，1.0}时，EPE回升至0.71 px。结果表明，适度提高后续阶段的监督权重有利于兼顾中间视差约束与最终高分辨率预测，其中C组在各组设置中获得最低EPE，因此本文采用{0.5，0.7，1.0}作为默认损失权重配置。同时，B、C和D三组递增权重设置下的EPE变化幅度较小，说明模型对前两阶段监督权重的适度调整具有较稳定的性能表现。

2.6　对比实验

本实验选取多种代表性立体匹配方法，在Scene Flow、KITTI 2012、KITTI 2015、ETH3D和Middlebury五个公开数据集上进行评估。这些数据集涵盖合成场景、真实道路场景、真实室内外场景及高分辨率室内场景，可用于观察不同成像条件下的匹配性能。实验同时统计视差估计精度与推理耗时，并通过不进行目标域微调的零样本测试分析模型的跨域适应能力。此外，本文展示KITTI 2012、KITTI 2015和ETH3D上的视差图及局部细节，结合定量指标分析目标边界、弱纹理等区域的匹配效果。

由表4可以看出，不同高效立体匹配方法在Scene Flow数据集上的精度和推理耗时之间存在明显差异。LBSM面向边缘计算设备，通过二维卷积和轻量通道注意力降低代价聚合开销，其设计更侧重于低延迟部署。本文则进一步关注低维代价表征条件下的匹配精度保持与递归修正能力。LBSM-L原论文报告的推理耗时为17 ms，其EPE为0.95 px，

© 中国图象图形学报版权所有

仍存在一定精度损失。相比之下,本文提出的FLISNet在32 ms推理耗时下取得0.68px的EPE,相比LBSM-L误差降低约28.4%,说明标准版本在保持较低推理开销的同时具有更好的匹配精度。

进一步地,FLISNet+将EPE降低至0.56px,取得表中最优结果。相比Fast-ACVNet+和LightStereo,误差分别降低约5.1%和6.7%,相比LBSM-L-Refine,误差降低约34.9%。虽然FLISNet+的推理耗时增加至56ms,但仍低于DeepPrunerFast和AANet,并获得了更高的视差估计精度。结果表明,FLISNet标准版本在精度和推理耗时之间取得了较好折中,而FLISNet+通过增加递归更新强度进一步提升了匹配精度,适合对视差估计质量要求更高的任务。

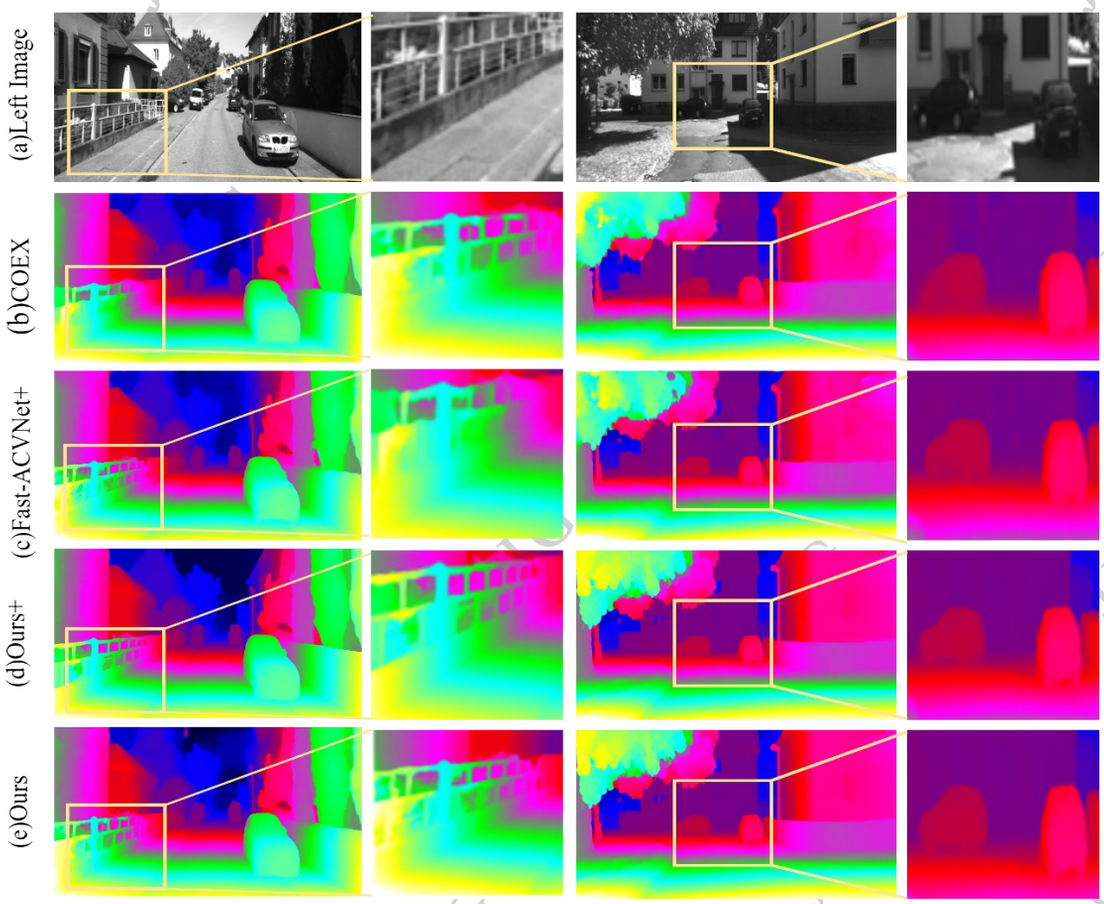
表5给出了本文方法与现有高精度立体匹配方法和高效立体匹配方法在KITTI 2012与KITTI 2015数据集上的对比结果。可以看出,高精度方法在部

表4 Scene Flow数据集实验结果

Table 4 Experimental Results on the Scene Flow Dataset

方法	EPE(px)	推理耗时(ms)
DeepPrunerFast(Duggal等,2019)	0.97	63*
AANet(Xu和Zhang,2020)	0.87	65*
BGNet(Xu等,2021)	1.17	28*
CoEx(Hong等,2021)	0.69	33*
Fast-ACVNet(Xu等,2024)	0.64	39*
Fast-ACVNet+(Xu等,2024)	<u>0.59</u>	45*
LightStereo(Guo等,2025)	0.60	37*
LBSM-L(Wu等,2026)	0.95	17†
LBSM-L-Refine(Wu等2026)	0.86	28†
FLISNet(Ours)	0.68	32*
FLISNet+(Ours+)	**0.56**	56*

注:加粗、下划线字体分别表示各列最优、次优结果。"*"表示本文在RTX 3090平台实测。"†"表示原论文报告的数据。



((a)Left Image;(b)COEX;(c)Fast-ACVNet+;(d)Ours+;(e)Ours)

图3 KITTI 2012数据集上的可视化结果

Fig. 3 Visualization Results on the KITTI 2012 Dataset

© 中国图象图形学报版权所有

分指标上仍具有较强优势。如LEAStereo和ACVNet在KITTI 2012与KITTI 2015的多个指标上取得较低误差，但其推理耗时难以满足实时应用需求。

可以看出，LBSM-L具有较低推理耗时，但其在KITTI 2015上的D1-all误差为2.81%，而FLISNet在32ms推理耗时下将D1-all误差降低至2.20%。在KITTI 2012上，FLISNet取得1.65%的3-noc误差和2.16%的3-all误差，整体精度优于LBSM-L。该结果表明，本文方法在增加有限推理开销的情况下，能够有效提升真实道路场景中的匹配精度。

进一步增加反馈递归修正次数后，FLISNet+在多项精度指标上取得同组最优结果，并在KITTI 2015的D1-fg指标上达到全表最优，表明增强版本能够进一步改善复杂区域的视差估计效果。与同为精度增强设置的LBSM-L-Refine相比，FLISNet+将KITTI 2012的3-all误差由2.41%降低至1.77%，将KITTI 2015的D1-all误差由2.52%降低至1.86%，分别降低约26.6%和26.2%。从整体数值来看，FLISNet和FLISNet+在保持较低推理耗时的同时，在KITTI 2012和KITTI 2015的主要误差指标上均优于对应的高效基线方法，体现出本文方法对真实道路场景的适应能力。上述结果表明，本文方法在真实道路场景中能够取得较好的精度与效率折中。

图3和图4分别给出了KITTI 2012和KITTI 2015数据集上的可视化结果。可以看出，CoEx和Fast-ACVNet+在道路等大面积区域能够获得较为连续的视差估计，但在车辆边界和远处弱纹理区域仍存在一定误匹配。相比之下，本文方法在目标轮廓和局部结构处具有更清晰的视差边界。

ETH3D数据集包含大量真实室内外场景，对模型的跨场景适应能力和局部细节恢复能力提出了较高要求。本实验采用精度增强配置FLISNet+进行测试。从表6可以看出，FLISNet+在Bad 2.0指标上取得0.54%的误差，为表中所列方法的最低值，说明该配置在ETH3D场景中能够保持较低的大误差像素比例。在Bad 1.0指标上，FLISNet+的误差为3.18%。综合来看，本文精度增强配置在真实室内外场景中展现了对较大视差误差的控制能力。

图5展示了ETH3D数据集上的可视化结果。可以看出，在弱纹理区域，物体边界以及深度变化较明显的区域，部分对比方法容易出现视差不连续、边缘模糊或局部误匹配现象。相比之下，本文方法生成的视差图具有更好的区域连续性和边界保持能力，能够减少局部噪声和异常视差点，提升真实室内外场景中的视差估计稳定性。

为评估标准迭代配置下的跨域泛化能力，本实验选用FLISNet进行零样本测试。FLISNet仅在Scene Flow数据集上训练，无需任何目标域微调，直接在KITTI 2012、KITTI 2015与Middlebury数据集上进行推理，结果如表7所示。其中，FLISNet在KITTI 2012和KITTI 2015上的D1-all分别为11.7%和10.2%，在Middlebury上的Bad 2.0为19.68%，指标均为表中最低值。结果表明，标准配置能够将从合成数据中学习的匹配表征应用于真实道路和高分辨率室内场景，并保持较稳定的视差估计性能。

3 结 论

针对快速立体匹配中计算开销与匹配精度难以兼顾的问题，本文提出一种基于低维代价体反馈修正的快速迭代立体匹配方法FLISNet。该方法通过紧凑代价表征构建低维匹配空间，利用中间视差反馈对代价体进行递归调制，并结合多尺度结构补偿模块逐级恢复高分辨率视差细节。实验结果表明，本文方法在多个公开基准数据集上取得了较稳定的匹配结果。其中，FLISNet在保持较低推理延迟的同时获得了具有竞争力的视差精度，增强版本FLISNet+通过增加递归次数进一步提升了复杂区域的匹配质量。本文方法仍存在一定局限性。增强版本在提升精度的同时会相应增加推理开销，且低维代价体在极端弱纹理，严重遮挡和大视差变化区域中的细粒度表达能力仍有进一步提升空间。未来工作将进一步研究模型压缩，知识蒸馏和低开销特征适配策略，以降低模型部署成本。同时，将探索复杂退化环境的自适应匹配方法，以提升模型在真实复杂场景中的适应能力。

© 中国图象图形学报版权所有

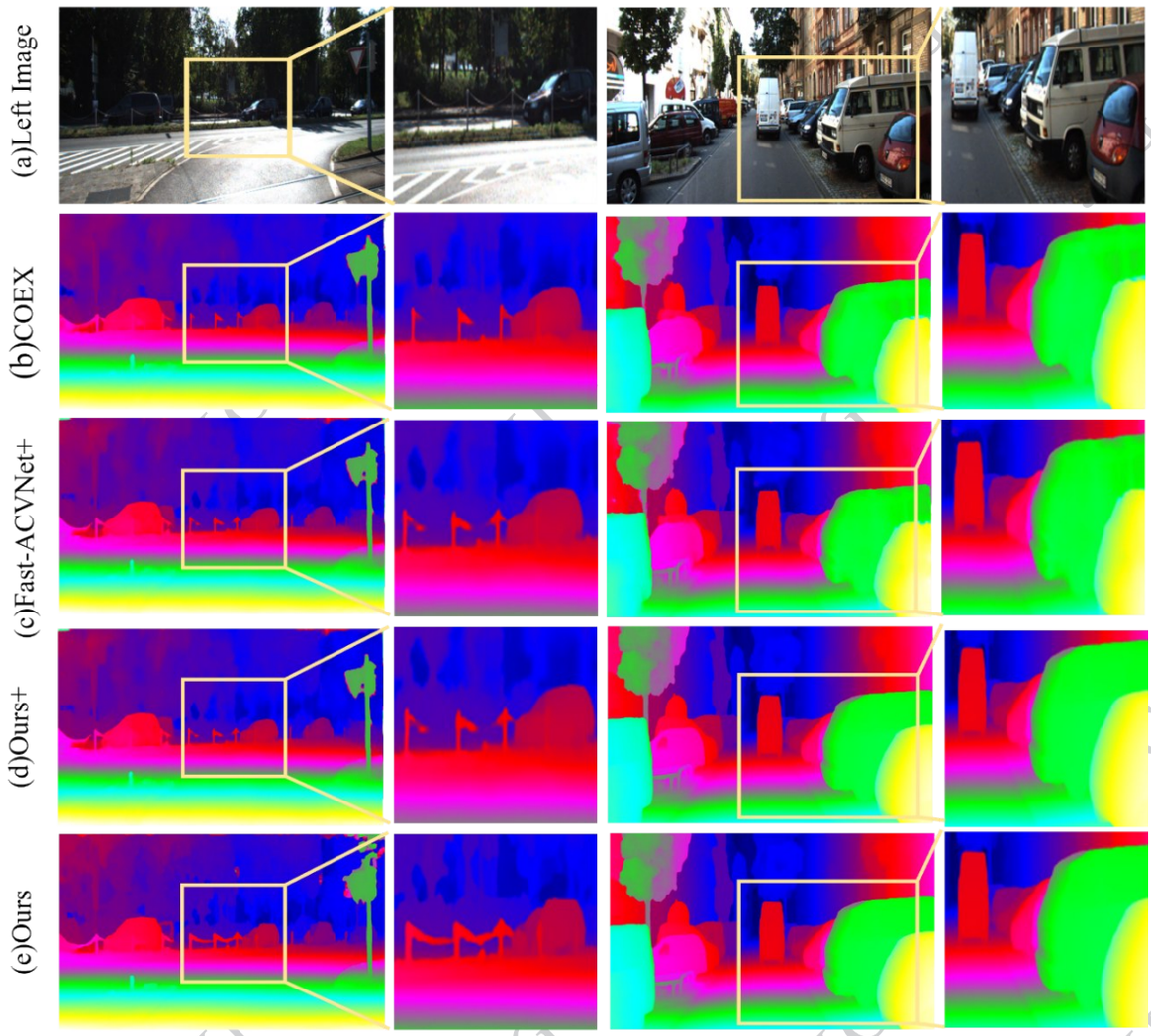
表5　KITTI2012与KITTI2015实验结果对比

Table 5　Comparison of Experimental Results on KITTI 2012 and KITTI 2015

类型	方法	KITTI 2012				KITTI 2015			Runtime (ms)
		3-noc	3-all	4-noc	4-all	D1-bg	D1-fg	D1-all	
精度	GCNet(Kendall等,2017)	1.77	2.30	1.36	1.77	2.21	6.16	2.87	900†
	PSMNet(Chang和Chen,2018)	1.49	1.89	1.12	1.42	1.86	4.62	2.32	310*
	GwcNet(Guo等,2019)	1.32	1.70	0.99	1.27	1.74	3.93	2.11	180*
	GANet(Zhang等,2019)	1.19	1.60	0.91	1.23	1.48	3.46	1.81	180†
	CFNet(Shen等,2021)	1.23	1.58	0.92	1.18	1.54	3.56	1.88	180†
	RAFT-Stereo(Lipson等,2021)	–	–	–	–	1.75	**2.89**	1.96	380*
	LEAStereo(Cheng等,2020)	**1.13**	**1.45**	**0.83**	**1.08**	1.40	2.91	**1.65**	300†
	ACVNet(Xu等,2022)	**1.13**	1.47	0.86	1.12	**1.37**	3.07	**1.65**	200*
效率	DispNetC(Mayer等,2016)	4.11	4.65	2.77	3.20	4.32	4.41	4.34	60†
	DeepPrunerFast(Duggal等,2019)	–	–	–	–	2.32	3.91	2.59	50*
	AANet(Xu和Zhang,2020)	1.91	2.42	1.46	1.87	1.99	5.39	2.55	62*
	DecNet(Yao等,2021)	–	–	–	–	2.07	3.87	2.37	50†
	BGNet(Xu等,2021)	1.77	2.15	–	–	2.07	4.74	2.51	28*
	BGNet+(Xu等,2021)	1.62	2.03	1.16	1.48	1.81	4.09	2.19	35*
	CoEx(Hong等,2021)	1.55	1.93	1.15	1.42	1.79	3.82	2.13	33*
	HITNet(Tankovich等,2021)	1.41	1.89	1.14	1.53	1.74	3.20	1.98	54†
	Fast-ACVNet(Xu等,2024)	1.68	2.13	1.23	1.56	1.82	3.93	2.17	39*
	Fast-ACVNet+(Xu等,2024)	1.45	1.85	1.06	1.36	1.70	3.53	2.01	45*
	LightStereo(Guo等,2025)	1.39	1.82	1.02	1.33	**1.64**	3.01	1.89	48*
	LBSM-L(Wu等,2026)	2.21	2.77	–	–	–	–	2.81	14†
	LBSM-L-Refine(Wu等,2026)	1.90	2.41	–	–	–	–	2.52	23†
	FLISNet(Ours)	1.65	2.16	1.17	1.55	1.96	3.42	2.20	32*
	FLISNet+(Ours+)	**1.36**	**1.77**	**0.99**	**1.30**	1.69	**2.73**	**1.86**	56*

注:表中方法按精度方法和效率方法进行分组。加粗、下划线字体分别表示各列最优、次优结果。“*”表示本文作者使用官方开源代码在统一RTX 3090平台下实测的推理时间。“†”表示原论文报告的推理时间。“–”表示原论文未提供相关结果。

© 中国图象图形学报版权所有



((a)Left Image;(b)COEX;(c)Fast-ACVNet+;(d)Ours+;(e)Ours)

图4　KITTI 2015数据集上的可视化结果

Fig. 4　Visualization Results on the KITTI 2015 Dataset

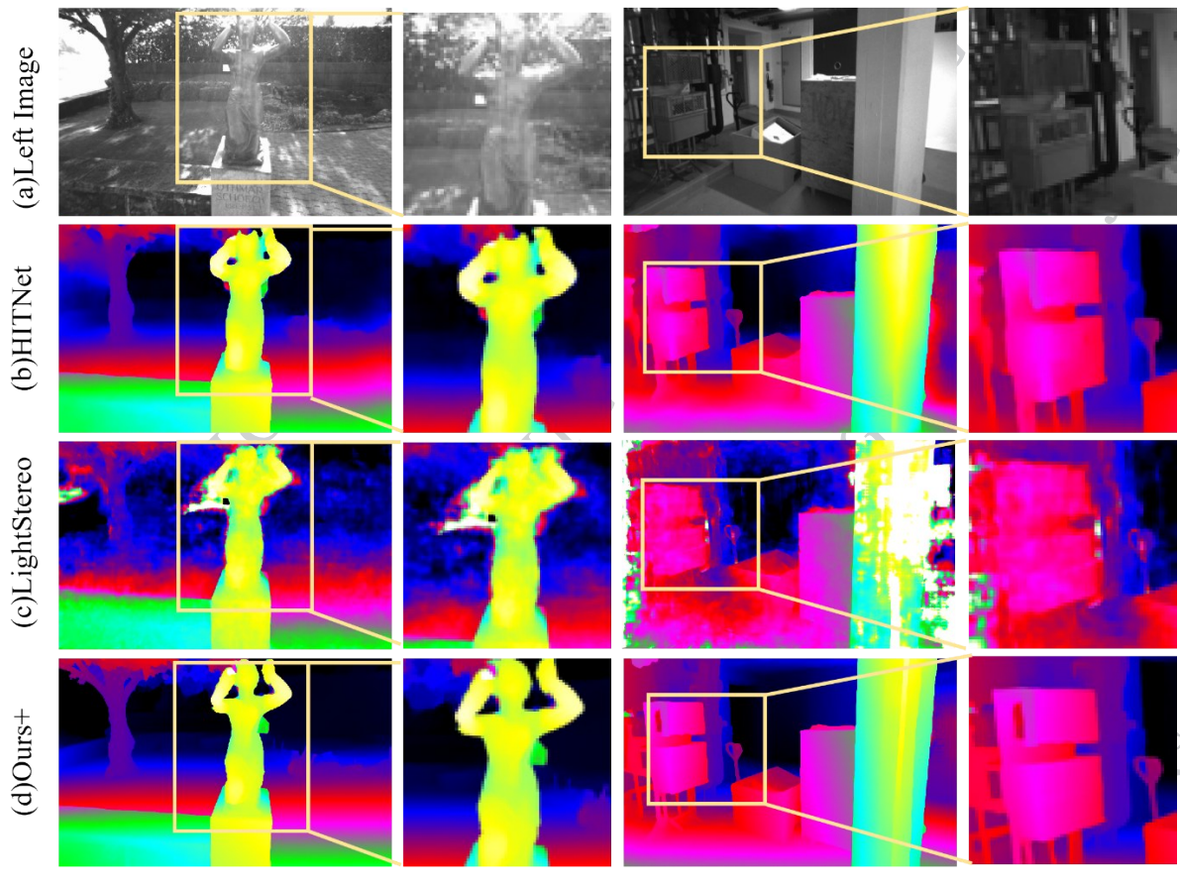
表6　ETH3D 数据集实验结果

Table 6　Experimental Results on the ETH3D Dataset

方法	Bad 1.0(%)	Bad 2.0(%)
PSMNet(Chang和Chen,2018)	5.02	1.09
GANet(Zhang等,2019)	6.56	1.10
CFNet(Shen等,2021)	3.31	0.77
HITNet(Tankovich等,2021)	<u>2.79</u>	0.80
ACVNet(Xu等,2022)	**2.58**	<u>0.57</u>
FLISNet+(Ours+)	3.18	**0.54**

注:加粗、下划线字体分别表示各列最优、次优结果。

© 中国图象图形学报版权所有



((a)Left Image;(b)HITNet;(c)LightStereo;(d)Ours+)

图5　ETH3D数据集可视化结果

Fig. 5　Visualization Results on the ETH3D Dataset

表7　KITTI 2012、KITTI 2015与Middlebury数据集上的零样本泛化实验

Table 7　Zero-Shot Generalization Experiments on the KITTI 2012, KITTI 2015, and Middlebury Datasets

方法	KITTI2012 D1-all(%)	KITTI2015 D1-all(%)	Middlebury Bad 2.0(%)
PSMNet(Chang和Chen,2018)	15.1	16.3	30.25
DeepPrunerFast(Duggal等,2019)	16.8	15.9	30.83
BGNet(Xu等,2021)	24.8	20.1	37.00
CoEx(Hong等,2021)	13.5	11.6	25.51
Fast-ACVNet(Xu等,2024)	<u>12.4</u>	<u>10.6</u>	<u>20.13</u>
FLISNet(Ours)	**11.7**	**10.2**	**19.68**

注:加粗、下划线字体分别表示各列最优、次优结果。

参考文献(References)

Bangunharcana A, Cho J W, Lee S, Kweon I S, Kim K S and Kim S. 2021. Correlate-and-Excite: real-time stereo matching via guided cost volume excitation// Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE: 3542-3548 [DOI: 10.1109/IROS51168.2021.9635909]

Chang J R and Chen Y S. 2018. Pyramid stereo matching network// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 5410-5418 [DOI: 10.1109/CVPR.2018.00567]

Chen Z, Long W, Yao H, Zhang Y J, Wang B S, Qin Y B, et al. 2024. MoCha-Stereo: motif channel attention network for stereo matching// Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE:

© 中国图象图形学报版权所有

27768-27777[DOI:10.1109/CVPR52733.2024.02623]

Cheng X, Zhong Y, Harandi M, Dai Y, Chang X, Drummond T, et al. 2020. Hierarchical neural architecture search for deep stereo matching// Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, Inc.: 22158-22169[DOI:10.5555/3495724.3497613]

Duggal S, Wang S, Ma W C, Hu R and Urtasun R. 2019. DeepPruner: learning efficient stereo matching via differentiable PatchMatch// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE: 4383-4392[DOI:10.1109/ICCV.2019.00448]

Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite// Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE: 3354-3361[DOI: 10.1109/CVPR.2012.6248074]

Guo X, Yang K, Yang W, Wang X and Li H. 2019. Group-wise correlation stereo network// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 3268-3277[DOI:10.1109/CVPR.2019.00339]

Guo X, Zhang C, Zhang Y, Zheng W, Nie D, Poggi M, et al. 2025. LightStereo: channel boost is all you need for efficient 2D cost aggregation// Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 8738-8744[DOI:10.1109/ICRA55743.2025.11127711]

Han X J and Yang H Y. 2018. Fusion of gray scale cost aggregation for stereo matching. Journal of Software, 29(S2): 44-53 (韩先君, 杨红雨. 2018. 融合灰度尺度的代价聚合的立体匹配. 软件学报, 29(S2): 44-53)

Hirschmüller H. 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2): 328-341[DOI: 10.1109/TPAMI.2007.1166]

Jia D, Cai P, Wu S, Wang Q and Song H L. 2024. Transformer network for stereo matching of weak texture objects. Journal of Image and Graphics, 29(8): 2413-2425 (贾迪, 蔡鹏, 吴思, 王骞, 宋慧伦. 2024. 面向弱纹理目标立体匹配的Transformer网络. 中国图象图形学报, 29(8): 2413-2425)[DOI:10.11834/jig.230575]

Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, et al. 2017. End-to-end learning of geometry and context for deep stereo regression// Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 66-75[DOI:10.1109/ICCV.2017.17]

Khamis S, Fanello S, Rhemann C, Kowdle A, Valentin J and Izadi S. 2018. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction// Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 596-613[DOI:10.1007/978-3-030-01267-0_35]

Li J, Wang P, Xiong P, Cai T, Yan Z, Yang L, et al. 2022. Practical stereo matching via cascaded recurrent network with adaptive correlation// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 16242-16251[DOI:10.1109/CVPR52688.2022.01578]

Li T, Ma W, Xu S B and Zhang X P. 2020. Task-adaptive end-to-end networks for stereo matching. Journal of Computer Research and Development, 57(7): 1531-1538 (李曈, 马伟, 徐士彪, 张晓鹏. 2020. 适应立体匹配任务的端到端深度网络. 计算机研究与发展, 57(7): 1531-1538)[DOI: 10.7544/issn1000-1239.2020.20190478]

Li Z, Liu X, Drenkow N, Ding A, Creighton F X, Taylor R H, et al. 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 6177-6186[DOI: 10.1109/ICCV48922.2021.00614]

Lipson L, Teed Z and Deng J. 2021. RAFT-Stereo: multilevel recurrent field transforms for stereo matching// Proceedings of 2021 International Conference on 3D Vision (3DV). London, UK: IEEE: 218-227[DOI:10.1109/3DV53792.2021.00032]

Lou J, Liu W, Chen Z, Liu F and Cheng J. 2023. ELFNet: evidential local-global fusion for stereo matching// Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 17738-17747[DOI: 10.1109/ICCV51070.2023.01630]

Mayer N, Ilg E, Häusser P, Fischer P, Cremers D, Dosovitskiy A, et al. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 4040-4048[DOI: 10.1109/CVPR.2016.438]

Menze M and Geiger A. 2015. Object scene flow for autonomous vehicles// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 3061-3070[DOI:10.1109/CVPR.2015.7298925]

Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X, et al. 2014. High-resolution stereo datasets with subpixel-accurate ground truth// Proceedings of the 36th German Conference on Pattern Recognition (GCPR). Münster, Germany: Springer: 31-42[DOI:10.1007/978-3-319-11752-2_3]

Schöps T, Schönberger J L, Galliani S, Sattler T, Schindler K, Pollefeys M, et al. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 2538-2547[DOI: 10.1109/CVPR.2017.272]

Shamsafar F, Woerz S, Rahim R and Zell A. 2022. MobileStereoNet: towards lightweight deep networks for stereo matching// Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Com-

© 中国图象图形学报版权所有

puter Vision (WACV). Waikoloa, USA: IEEE: 677-686[DOI:10.1109/WACV51458.2022.00075]

Shen Z, Dai Y and Rao Z. 2021. CFNet: cascade and fused cost volume for robust stereo matching// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 13901-13910 [DOI: 10.1109/CVPR46437.2021.01369]

Tankovich V, Häne C, Zhang Y, Kowdle A, Fanello S and Bouaziz S. 2021. HITNet: hierarchical iterative tile refinement network for real-time stereo matching// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 14357-14367 [DOI: 10.1109/CVPR46437.2021.01413]

Wang D and Hu L L. 2022. Improved feature stereo matching method based on binocular vision. Acta Electronica Sinica, 50(1): 157-166 (王笛, 胡辽林. 2022. 基于双目视觉的改进特征立体匹配方法. 电子学报, 50(1): 157-166)[DOI: 10.12263/DZXB.20200919]

Wang Q, Shi S, Zheng S, Zhao K and Chu X. 2020. FADNet: a fast and accurate network for disparity estimation// Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE: 101-107 [DOI: 10.1109/ICRA40945.2020.9197031]

Wang Y, Lai Z, Huang G, Wang B H, van der Maaten L, Campbell M, et al. 2019. Anytime stereo image depth estimation on mobile devices// Proceedings of 2019 IEEE International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE: 5893-5900[DOI:10.1109/ICRA.2019.8794003]

Wang Y, Li K, Wang L, Hu J, Wu D O and Guo Y. 2025. ADStereo: efficient stereo matching with adaptive downsampling and disparity alignment. IEEE Transactions on Image Processing, 34: 1204-1218 [DOI:10.1109/TIP.2025.3540282]

Wu Z, Zhu H, Lin G F and He L L. 2026. Lightweight binocular stereo matching network for edge computing devices. Journal of Image and Graphics, 31(2): 589-608 (武忠, 朱虹, 蔺广逢, 贺丽丽. 2026. 面向边缘计算设备的轻量级双目立体匹配网络. 中国图象图形学报, 31(2): 589-608)[DOI:10.11834/jig.250081]

Xu B, Xu Y, Yang X, Jia W and Guo Y. 2021. Bilateral grid learning for stereo matching networks// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 12492-12501[DOI:10.1109/CVPR46437.2021.01231]

Xu G, Cheng J, Guo P and Yang X. 2022. Attention concatenation volume for accurate and efficient stereo matching// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 12971-12980[DOI:10.1109/CVPR52688.2022.01264]

Xu G, Wang X, Ding X and Yang X. 2023. Iterative geometry encoding volume for stereo matching// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 21919-21928[DOI:10.1109/CVPR52729.2023.02099]

Xu G, Wang Y, Cheng J, Tang J and Yang X. 2024. Accurate and efficient stereo matching via attention concatenation volume. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4): 2461-2474[DOI:10.1109/TPAMI.2023.3335480]

Xu G, Wang X, Zhang Z, Cheng J, Liao C and Yang X. 2025. IGEV++: iterative multi-range geometry encoding volumes for stereo matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(8): 7108-7122 [DOI: 10.1109/TPAMI.2025.3569218]

Xu H and Zhang J. 2020. AANet: adaptive aggregation network for efficient stereo matching// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 1956-1965[DOI:10.1109/CVPR42600.2020.00203]

Xu S, Ye N, Zhu F, Xu S S and Zhou L L. 2013. Solving the stereo matching in disparity image space using the Hopfield neural network. Journal of Computer Research and Development, 50(5): 1021-1029 (徐昇, 业宁, 朱发, 徐姗姗, 周溜溜. 2013. Hopfield网络在视差空间上的立体匹配求解. 计算机研究与发展, 50(5): 1021-1029)

Yao C, Jia Y, Di H, Li P and Wu Y. 2021. A decomposition model for stereo matching// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 6087-6096[DOI:10.1109/CVPR46437.2021.00603]

Yuan Z L, Li Z H, Chen K H, Mao T L, Jiang H and Wang Z Q. 2026. Multiview stereo 3D reconstruction research: a survey. Journal of Image and Graphics, 31(3): 657-685 (袁祯泷, 李泽昊, 陈科桦, 毛天露, 蒋浩, 王兆其. 2026. 多视角立体匹配三维重建研究综述. 中国图象图形学报, 31(3): 657-685)[DOI:10.11834/jig.250348]

Zhai Z G, Lu Y and Zhao H. 2010. Stereo matching algorithm using segment geometric constraint and disparity probability. Journal of Software, 21(11): 2985-2998 (翟振刚, 陆耀, 赵红. 2010. 利用块几何约束及视差概率的立体匹配算法. 软件学报, 21(11): 2985-2998)[DOI:10.3724/SP.J.1001.2010.03695]

Zhang F, Prisacariu V A, Yang R and Torr P H S. 2019. GA-Net: guided aggregation net for end-to-end stereo matching// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 185-194[DOI:10.1109/CVPR.2019.00027]

Zhang W, Shao X T, Yang W, Guo M K and Jing N Z. 2020. An efficient and accurate stereo matching algorithm based on convolutional neural network. Journal of Computer-Aided Design & Computer Graphics, 32(1): 45-53 (张文, 邵小桃, 杨维, 郭鸣坤, 景年昭. 2020. 基于卷积神经网络的高效精准立体匹配算法. 计算机辅助设计与图形学学报, 32(1): 45-53)[DOI:10.3724/SP.J.1089.2020.17823]

© 中国图象图形学报版权所有

Zhang Z H, Wang Y, Chen B Y, Zhang H X, Xu X B, Wu F L, et al. 2026. Binocular vision depth estimation method based on geometric prior knowledge constraints. Acta Electronica Sinica, 54(1): 195-205 (张泽辉, 王阳, 陈博洋, 张浩轩, 徐晓滨, 吴富龙, 等. 2026. 基于几何先验知识约束的双目视觉深度估计方法. 电子学报, 54(1): 195-205)[DOI:10.12263/DZXB.20250504]

Zhang Z Y and Yang J. 2023. Online adaptation through meta-learning for stereo depth estimation. Acta Automatica Sinica, 49(7): 1446-1455 (张振宇, 杨健. 2023. 基于元学习的双目深度估计在线适应算法. 自动化学报, 49(7): 1446-1455)[DOI:10.16383/j.aas.c200286]

Zhao Y, Chen Z, He J, Li W, Xiao Y, Tian C, et al. 2025. GIP-Stereo: geometry-aware information propagation network for stereo matching. Knowledge-Based Systems, 327: 114062[DOI:10.1016/j.knosys.2025.114062]

作者简介

邹亚琦,女,硕士研究生。主要研究方向为双目立体匹配。E-mail:4025181035@mails.qust.edu.cn

朱旅潍,女,硕士研究生。主要研究方向为三维感知与重建。E-mail:4023112073@mails.qust.edu.cn

高颖,女,博士,副教授。主要研究方向为机器学习和计算机视觉。E-mail:gaoying@qust.edu.cn

王馨静,女,硕士研究生。主要研究方向为语义分割和目标跟踪。E-mail:e402418004@mails.qust.edu.cn

李金玉,女,本科生。主要研究方向为图像处理。E-mail:lijjinyu@mails.qust.edu.cn

© 中国图象图形学报版权所有