**中图法分类号**:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)02-0589-20

**论文引用格式**: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):0589-0608(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):0589-0608)[DOI:10.11834/jig.250081]

# 面向边缘计算设备的轻量级双目立体匹配网络

武忠[1],朱虹[2*],蔺广逢[3],贺丽丽[1]

1. 运城学院山西省智能光电传感应用技术创新中心,运城　044000; 2. 西安理工大学自动化与信息工程学院,西安　710048; 3. 西安理工大学印刷包装与数字媒体学院,西安　710048

**摘　要**:**目的**　现有的高精度双目立体匹配网络因计算开销较高,难以部署到算力受限的边缘计算设备上,极大地限制了双目立体视觉系统的适用场景。针对此问题,提出一种轻量级实时立体匹配网络框架(light-weight binocular stereo matching framework, LBSM)。**方法**　首先,摒弃了高计算开销的"4D代价体+3D卷积"的主流方案,仅采用2D卷积与轻量化通道注意力机制构建了高效的融合注意力(merged attention, MA)代价聚合模块,在降低计算开销的同时减少了信息丢失,使得代价聚合过程更加高效;其次,提出空间自适应视差传播(adaptive disparity propagation, ADP)策略以替代双线性插值,以极低的代价(仅增加约0.03 M参数、0.36 G"乘法—累加"操作数和1.1 ms推理时间)将端点误差和1像素误差分别降低了21.54%和20.73%,揭示了视差上采样策略在轻量级模型中的重要性。在上述方法的基础上,构建了仅依赖2D卷积、无需任何3D卷积层的轻量级模型框架LBSM,通过简单配置即可生成系列模型。**结果**　实验结果表明,所提模型在大型数据集Scene Flow上以更低的计算开销取得了更高的准确性(LBSM-L以68.5%的"乘法—累加"操作数取得了明显高于BGNet(bilateral grid learning network)的准确性),而且在真实道路场景数据集KITTI 2012(Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)和KITTI 2015上分别取得了低至2.41%和2.52%的错误率,显著优于ADCPNet(adaptive disparity candidates prediction network)、P3SNet(parallel pyramid pooling stereo network)等主流轻量级立体匹配模型。此外,在嵌入式AI(artificial intelligence)硬件平台上的部署实测显示,模型处理速度达4.59～20.29帧/s。**结论**　LBSM在计算开销与准确性之间实现了更好的权衡,能够在低算力边缘计算设备上实时完成双目立体匹配任务,在算力受限的实际场景中具有较大的应用潜力。

**关键词**:双目立体视觉;全卷积网络;空间自适应性;轻量级模型;嵌入式智能设备

# Lightweight binocular stereo matching network for edge computing devices

Wu Zhong[1], Zhu Hong[2*], Lin Guangfeng[3], He Lili[1]

*1. Shanxi Province Intelligent Optoelectronic Sensing Application Technology Innovation Center, Yuncheng University, Yuncheng　044000, China; 2. School of Automation and Information Engineering, Xi'an University of Technology, Xi'an　710048, China; 3. Faculty of Printing, Packaging and Digital Media, Xi'an University of Technology, Xi'an　710048, China*

**Abstract**: **Objective**　Computer binocular stereo vision draws inspiration from the fundamental principle of human binocu-

**收稿日期**:2025-03-12;**修回日期**:2025-08-01;**预印本日期**:2025-08-08

*通信作者:朱虹　zhuhong@xaut.edu.cn

**基金项目**:山西省基础研究计划(202403021222304);国家自然科学基金项目(61771386);运城学院应用研究项目(YY-202403);山西省高校思想政治工作质量提升综合改革与精品建设项目(2025年度省级高校数字文物开发项目,序号4);运城学院2025年度博士科研启动项目(YXBQ-202523)

**Supported by**: Fundamental Research Program of Shanxi Province, China(202403021222304); National Natural Science Foundation of China(61771386)

lar perception of object distance. It has significant application potential in numerous fields, including intelligent manufacturing, autonomous driving, robotic visual navigation, aerospace, geographic remote sensing, smart healthcare, and virtual/augmented reality, and is being increasingly widely adopted. In recent years, stereo matching methods based on deep learning have leveraged the powerful learning capabilities of deep neural networks to learn implicit rules of stereo matching from a large number of training samples. These kinds of methods integrate the entire stereo matching process, including feature extraction, cost volume construction, cost aggregation, disparity regression, and disparity refinement, into an end-to-end deep neural network. By leveraging the powerful learning capabilities of deep neural networks, these methods effectively address challenges posed by various complex factors. Currently, the accuracy of deep learning-based stereo matching networks has significantly surpassed that of traditional methods, leading to significant advancements and a substantial improvement in the accuracy of stereo matching. However, as a pixel-level dense matching task, stereo matching inherently involves high computational complexity. Deep learning-based stereo matching models typically require significant computational resources to achieve good disparity accuracy. Existing state-of-the-art stereo matching networks often demand tens to hundreds of giga multiply-accumulate operations to process a pair of stereo images with a resolution of 540 × 960. In many real-world application scenarios, owing to constraints such as hardware costs and power consumption, stereo matching models often need to be deployed on low-power edge devices. This issue imposes stringent requirements on the models: They must not only achieve good disparity accuracy but also exhibit extremely low computational overhead. Present advanced stereo matching networks generally have high computational costs, making their deployment expensive and unsuitable for such practical scenarios. To address this issue, this study proposes a lightweight binocular stereo matching framework named LBSM, which is designed for low-power edge computing devices. **Method** The construction and aggregation of cost volume are two critical steps that significantly influence the overall accuracy and efficiency of the process. Thus, in this study, the popular architecture of "4D cost volume + 3D convolution" with high computational overhead is abandoned. First, the 4D cost volume is directly reshaped into a 3D cost volume by fusing the channel and disparity dimensions, and only 2D convolutions are used for cost aggregation. In this way, the entire cost aggregation process only requires the use of 2D convolution, which significantly reduces computational overhead while minimizing information loss. Second, a lightweight channel attention mechanism is introduced for cost aggregation, avoiding the stacking of many convolutional layers and making the cost aggregation process efficient. Third, a two-stage network following the "coarse-to-fine" architecture is adopted, in which a spatially adaptive disparity upsampling strategy replaces bilinear interpolation, enabling adaptive propagation of low-resolution disparity estimates with minimal computational cost. This upsampling strategy not only significantly enhances the accuracy of lightweight models with minimal computational overhead but also enables us to reduce the number of "coarse-to-fine" iterations. Finally, on the basis of the above methods, the proposed LBSM is constructed without any 3D convolutions. Through straightforward configurations, LBSM can yield a series of models that are capable of achieving real-time stereo matching on low-power edge computing devices. **Result** Ablation and comparative experiments on the large-scale Scene Flow dataset demonstrate that LBSM achieves high disparity accuracy with lower computational overhead. Compared with existing lightweight stereo matching models and even some nonlightweight models, the proposed network framework exhibits clear advantages in accuracy and real-time performance. On the real-world road scene datasets KITTI 2012 and KITTI 2015, the proposed method achieves error rates as low as 2.41% and 2.52%, respectively, significantly outperforming mainstream lightweight stereo matching models such as ADCPNet and P3SNet. Additionally, deployment and testing on embedded AI hardware (i.e., NVIDIA Jetson TX2) platforms show that LBSM achieves much better accuracy with faster processing speed (4.59—20.29 frames per second). **Conclusion** Benefiting from the aforementioned strategies, LBSM relies solely on 2D convolutional layers and exhibits very low computational overhead. It can perform binocular stereo matching tasks in real time on low-power edge computing devices, achieving a better trade-off between computational cost and accuracy than existing models and thus endowing it with significant application potential in real-world scenarios with limited computational resources.

**Key words**: binocular stereo vision; fully convolutional network; spatial adaptivity; lightweight model; embedded intelligent device

## 0 引 言

计算机双目立体视觉借鉴人类双眼感知物体远近的基本原理，依据同一场景、不同视角下采集的两幅二维图像（即参考图像和目标图像，也称左图像和右图像）感知场景的三维信息（Zhang，2023）。作为被动的、拟人化的非接触式三维感知方法，双目视觉具有成本低、易部署、可获得稠密的三维信息和丰富的语义信息等优点，在医学成像（Xia 等，2022）、智能制造（Fu 等，2022；Zhou 等，2023）、三维重建（Ul Haq 等，2019）、自动驾驶（Fan 等，2023；Xie 等，2024）以及机器人导航（Luo 等，2018）等众多领域具有巨大的应用潜力，得到日益广泛的应用。长期以来，作为实现双目立体视觉的核心和难点问题，立体匹配吸引了诸多学者进行持续和深入研究（Scharstein 和 Szeliski，2002）。虽然传统立体匹配方法已取得许多重要进展（程德强 等，2021），但是它们通常依赖人为设定的先验或假设（如局部平滑性假设等），难以同时应对现实场景中存在的噪声、光照、重复纹理、弱纹理和遮挡等诸多复杂因素的影响，因而准确性较低。

基于深度学习的立体匹配方法将特征提取、代价体构建、代价聚合、视差回归和视差精确化等立体匹配流程步骤全部集成到端到端的深度神经网络之中，借助深度神经网络强大的学习能力应对诸多复杂因素带来的挑战（Zhou 等，2020）。目前，深度学习立体匹配网络的准确性大幅度超越了传统方法，已成为立体匹配领域的主流（Laga 等，2022）。

然而，作为像素级稠密匹配任务，立体匹配本身具有较高的计算复杂度，立体匹配网络通常需要以较高的计算开销为代价，才能获得良好的视差准确性。现有的先进立体匹配网络处理一对分辨率为 540 × 960 像素的图像通常也需要几十到几百 G 的“乘法—累加”操作数（Xu 和 Zhang，2020）。在许多现实应用场景中，受功耗、硬件成本等因素的限制，立体匹配模型常常需要部署在嵌入式边缘计算设备上，这对立体匹配模型的计算效率提出严苛要求，不仅要求模型获得较高的视差准确性，而且要求其具有极低的计算开销。因此，研究和实现具有良好视差准确性的轻量级实时网络已成为立体匹配领域亟待解决的重点和热点问题。

StereoNet（stereo network）（Khamis 等，2018）是最早致力于实时立体匹配的网络之一，其在低分辨率下构建 4D 代价体以降低计算开销，再通过视差图的逐级精确化获得全分辨率视差图。然而，其准确性较低且在嵌入式边缘设备上难以实时运行。此后，许多研究致力于设计更轻量级的实时立体匹配网络，其中最具代表性的工作包括 AnyNet（anytime stereo network）（Wang 等，2019）、MADNet（modularly adaptive network）（Tonioni 等，2019）、RTS2Net（real-time semantic stereo network）（Dovesi 等，2020）、ADCPNet（adaptive disparity candidates prediction network）（Dai 等，2022）和 P3SNet（parallel pyramid pooling stereo network）（Emlek 和 Peker，2023）等。受人类视觉系统（Menze 和 Geiger，2015）的启发，轻量级模型的研究工作普遍采用“由粗到细”（coarse-to-fine，CTF）的整体策略以降低计算开销。例如，AnyNet、MADNet 和 RTS2Net 均采用了基于 CTF 策略的网络架构，通过构建低分辨率代价体获得低分辨率的初始视差图，再通过双线性插值提升初始视差图的分辨率，并基于初始视差图和固定视差偏移量获取新的候选视差，从而构建出空间分辨率较高但候选视差个数很少的“窄”代价体。如此经过多次迭代，最终生成全分辨率视差图。虽然上述过程大幅度地降低了计算开销，但是固定的视差偏移量难以适配不同类型的图像区域，而且每次迭代都需要构建代价体并进行代价聚合，多次迭代不可避免地延长了网络的推理时间，因此，三者的准确性和实时性相对较差。为提升候选视差的质量，ADCPNet 使用多层卷积基于初始视差和左图像预测视差偏移量，在实时性和视差准确性上获得了一定的提升。然而，其准确性仍然不足，尤其在大规模数据集上的准确性较低。

虽然 CTF 策略能够显著降低计算开销，是构建轻量级实时立体匹配模型的非常自然的选择，且已成为当前的主流方法，但是现有轻量级实时立体匹配网络的准确性与非实时网络相比，仍然存在巨大差距。值得注意的是，一方面，基于 CTF 策略的网络模型不可避免地需要进行低分辨率视差的上采样，但是现有研究工作很少关注视差上采样环节，而普遍采用简单的双线性插值。然而，双线性插值不可避免地导致模糊的上采样视差，尤其是在物体边缘等视差不连续区域，这使得在全分辨率视差图中恢

复清晰的边缘和精细结构变得更加困难(Deng等,2021)。因此,现有研究工作一定程度上低估了视差上采样环节的重要性。另一方面,在立体匹配流程中,代价体构建和代价聚合两个步骤在很大程度上决定了整个模型的计算开销和准确性。非实时的立体匹配网络通常采用足够多的3D卷积层对4D代价体进行代价聚合(Kendall等,2017),以保证网络模型的准确性。与之相反,因3D卷积层具有较高的计算开销,轻量级实时立体匹配网络不得不大幅度降低3D卷积层的层数,借此减少计算开销,然而这不可避免地降低了网络的准确性。

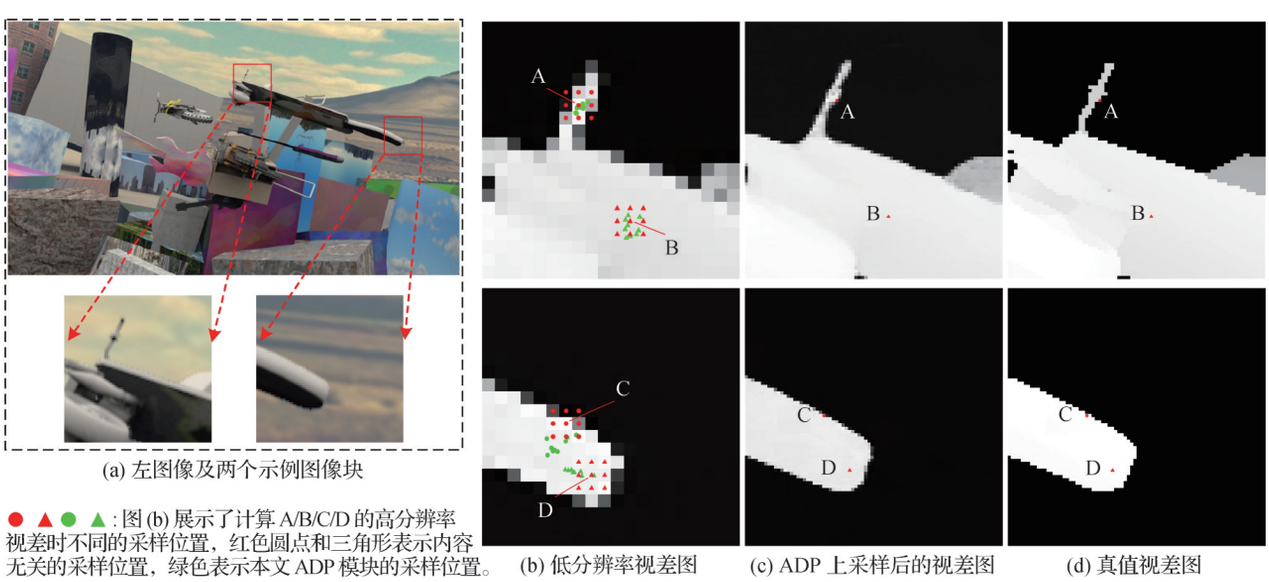
为了克服上述两方面的问题,实现具有良好视差准确性的轻量级实时立体匹配模型,首先,本文提出一种简洁高效的自适应视差传播(adaptive disparity propagation,ADP)策略,并构建了对应的网络模块,以极低的计算开销实现了低分辨率视差的空间自适应上采样,为CTF流程提供高质量的初始视差,从而提升了整个模型的准确性。图1可视化地展示了该上采样策略的实例,图中数据由本文的LBSM-L网络在Scene Flow数据集上得到,ADP以极低的计算开销实现了空间自适应上采样。实验表明,ADP模块以约0.03 M参数量、0.36 G MACs(multiply-accumulate operations per second)和1.1 ms推理时间为代价,将端点误差和1像素误差分别降低了21.54%和20.73%,大幅度地提升了轻量级模型的准确性(详见第2.5节)。

其次,设计了适用于轻量级实时立体匹配的融合注意力(merged attention, MA)代价聚合策略(详见第1.4节),舍弃了高计算开销的"4D代价体+3D卷积"的主流方案,通过融合通道和视差两个维度把4D代价体直接重整为3D代价体,仅使用2D卷积进行代价聚合,在降低计算开销的同时减少信息丢失,保持代价聚合的效果。所提策略使得代价聚合过程仅依赖低计算开销的2D卷积和通道注意力机制,能够在大幅度缩减计算开销的同时保持良好的视差准确性。

以上述方法为核心,本文构建了具有良好视差准确性的轻量级实时立体匹配网络框架(lightweight binocular stereo matching framework, LBSM),并从中抽取出多个可配置的超参数,使得能够依据可用计算资源的多寡,轻松地配置该网络的规模,以满足不同算力的应用场景。实验结果表明,LBSM在视差准确性和推理速度之间获得了更好的权衡,与现有先进轻量级实时立体匹配网络相比,表现出明显优势,其准确性甚至优于许多非实时的立体匹配网络。

## 1 算法描述

本节首先介绍所提网络的设计思路和总体框架




(a) 左图像及两个示例图像块

● ▲ ● ▲:图(b)展示了计算A/B/C/D的高分辨率视差时不同的采样位置,红色圆点和三角形表示内容无关的采样位置,绿色表示本文ADP模块的采样位置。

(b) 低分辨率视差图 (c) ADP上采样后的视差图 (d) 真值视差图

图1 自适应视差传播(ADP)示例

Fig. 1 Adaptive disparity propagation (ADP) examples ((a) left image and two example image patches; (b) low-resolution disparity maps; (c) disparity maps after ADP upsampling; (d) ground-truth disparity maps)

（如图 2 所示）；然后，详细描述实现轻量级实时立体匹配的关键策略和网络模块；最后，介绍训练此框架的损失函数。

### 1.1 设计思路和总体框架

作为像素级稠密匹配任务，立体匹配需要在左、右图像之间建立像素级的稠密对应关系，任务本身具有较高的计算复杂度，并且单幅图像可能包含多种不同特性的区域，如局部细节丰富的区域、细节匮乏的低/弱纹理区域等，立体匹配算法必须同时利用局部细节特征和上下文信息。因此，设计低计算开销、具有良好准确性、可在嵌入式 AI 边缘设备上实时运行的立体匹配网络是一项颇具挑战性的任务。本文从多个研究领域中获取灵感，以应对上述挑战。

一方面，主流的立体匹配神经网络均采用 3D 卷积对基于图像特征的 4D 代价体进行代价聚合，然而，4D 代价体和 3D 卷积带来了大量的计算开销，是整个立体匹配网络的性能瓶颈。针对性地，本文设计了融合注意力（MA）代价聚合策略，通过融合通道和视差两个维度，将 4D 代价体重整为 3D 代价体，再基于通道注意力机制和 2D 卷积以极低的计算开销完成代价聚合（详见第 1.4 节）。此策略不仅大幅度地降低了模型的计算开销，而且其准确性明显优于简单堆叠的 3D 卷积层。

另一方面，受人类视觉系统的启发（Mallot 等，1996；Menz 和 Freeman，2003），计算机视觉领域的许多任务常常采用“由粗到细”的 CTF 策略，鉴于此，LBSM 亦采用两阶段的 CTF 策略来降低计算开销，但是使用本文提出的计算开销极低的自适应视差传播（ADP）策略，保证初始视差的高效传播并避免多次迭代引入的累积误差（详见第 1.5 节）。

得益于上述精心设计的代价聚合模块和自适应视差传播模块，该框架摆脱了高计算开销的 4D 代价体和 3D 卷积，成为仅依赖 2D 卷积的立体匹配网络，从而具有极低的计算开销。所提出的 LBSM 网络框架如图 2 所示，其主要由多尺度特征提取（multiscale feature extraction，MFE）模块、第 1 匹配阶段和第 2 匹配阶段构成。其中，第 1 匹配阶段使用融合注意力（MA）代价聚合模块来聚合低分辨率的代价体，生成低分辨率的视差估计结果，再通过自适应视差传播（ADP）将低分辨率视差提升到更高分辨率，以作为第 2 匹配阶段的初始视差；第 2 匹配阶段以初始视差为基础，从原始图像和重构误差中获取置信度信息，通过视差偏移估计（disparity offset estimation，DOE）学习获得视差偏移量，然后，基于视差偏移量和初始视差得出第 2 阶段的少量候选视差，再构建出空间分辨率较高但候选视差个数很少的“窄”代价体，从而保持较低的计算开销。

### 1.2 多尺度特征提取（MFE）模块

局部细节特征和上下文信息对于立体匹配都非常重要。因此，本文设计了多尺度特征提取（MFE）模块，通过捕获图像的低层级局部细节信息以及高级语义上下文，为每个像素提取鲁棒的特征，从而为

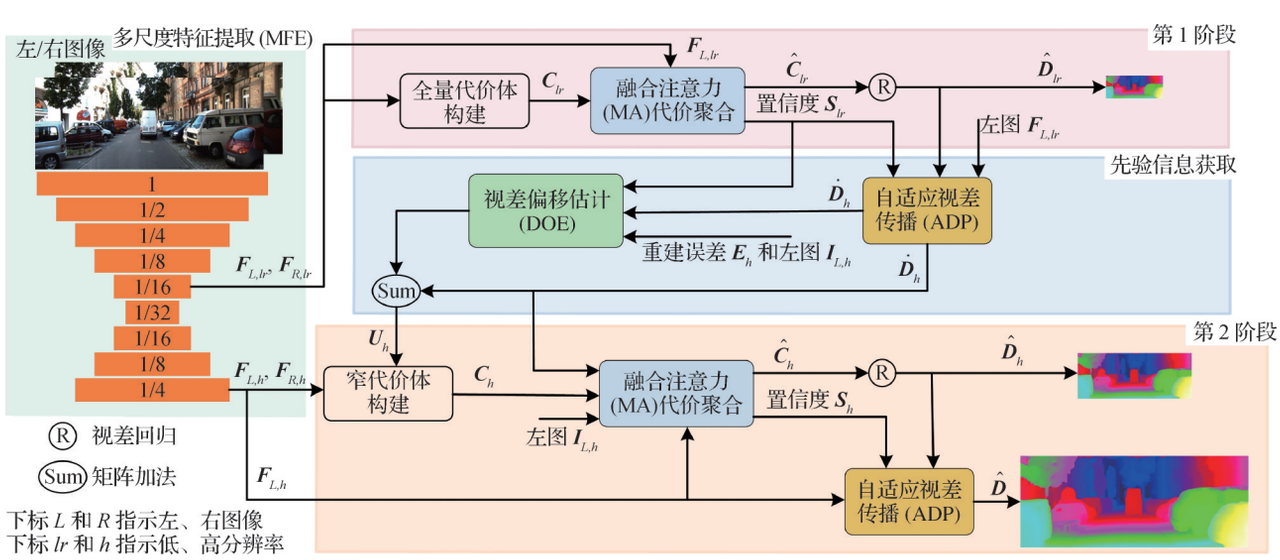



图 2 LBSM 网络的总体框架

Fig. 2 The overall framework of the proposed light-weight binocular stereo matching (LBSM)

构建高质量代价体打下基础。如图3所示，MFE模块具有轻量级的“编码器—解码器”结构，其中的深层浅层特征混合（deep-shallow-mix，DSM）模块和残差模块如图4所示。

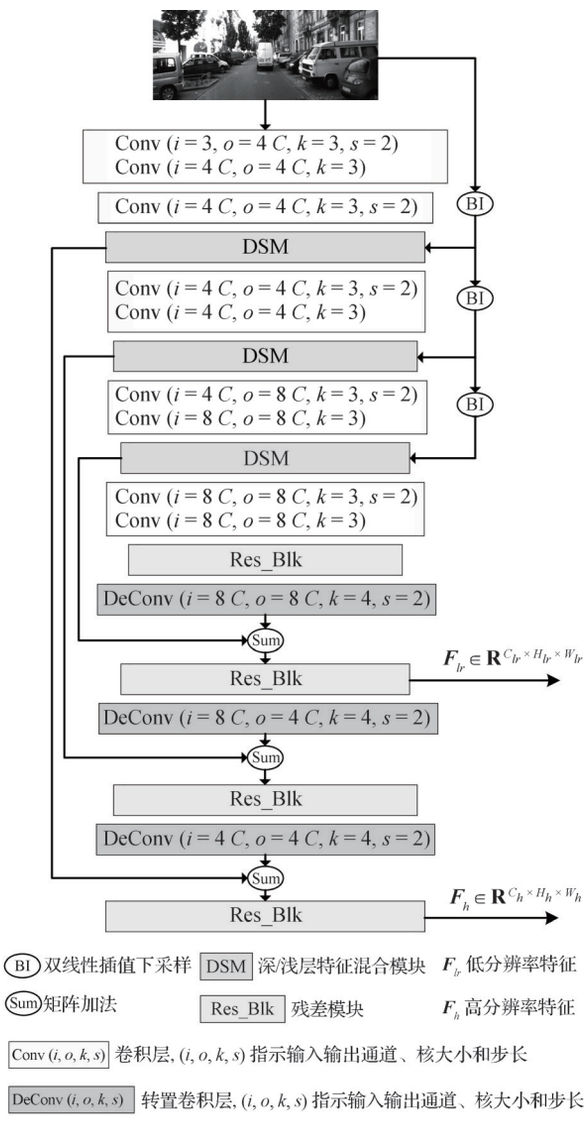
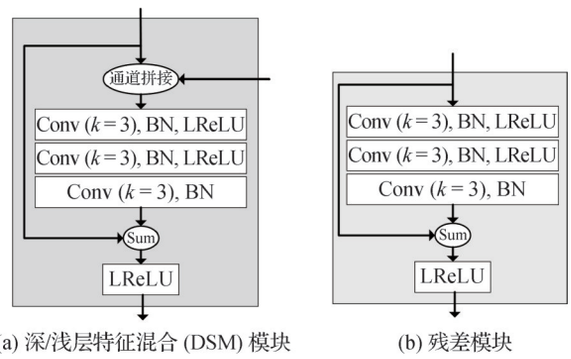



图3　多尺度特征提取（MFE）模块

Fig. 3　Multiscale feature extraction (MFE) module

MFE模块最终输出低分辨率特征图 $\boldsymbol{F}_{lr} \in \mathbf{R}^{C_{lr} \times H_{lr} \times W_{lr}}$ 和高分辨率特征图 $\boldsymbol{F}_h \in \mathbf{R}^{C_h \times H_h \times W_h}$，其中，下标 $lr$ 和 $h$ 分别指示低分辨率和高分辨率，$C_{lr}$ 和 $C_h$ 表示特征图的通道数，$H_{lr} \times W_{lr}$ 和 $H_h \times W_h$ 表示特征图的空间尺寸（亦即分辨率）。记原始输入图像的分辨率为 $H_0 \times W_0$，设置 $H_{lr} = \frac{1}{16}H_0$，$W_{lr} = \frac{1}{16}W_0$，$H_h = \frac{1}{4}H_0$，$W_h = \frac{1}{4}W_0$，$C_{lr} = 8C$，$C_h = 4C$，其中，$C$ 为可配置的超参数。




(a) 深/浅层特征混合 (DSM) 模块　　(b) 残差模块

图4　深/浅层特征混合（DSM）模块和残差模块

Fig. 4　Deep-shallow-mix (DSM) module and residual block ((a) deep-shallow-mix (DSM) module; (b) residual block)

### 1.3　代价体构建

如图2所示，在获得左、右图像的低分辨率和高分辨率特征图后，即可分别构建低分辨率的全量代价体和高分辨率的窄代价体。

在第1匹配阶段，采用分组相关策略（Guo等，2019）构建全量代价体 $\boldsymbol{C}_{lr}$。具体地，先将左、右图像的低分辨率特征图 $\boldsymbol{F}_{L,lr} \in \mathbf{R}^{C_{lr} \times H_{lr} \times W_{lr}}$ 和 $\boldsymbol{F}_{R,lr} \in \mathbf{R}^{C_{lr} \times H_{lr} \times W_{lr}}$ 分别沿通道维度分为 $G$ 组，其中，下标 $L$ 和下标 $R$ 分别指示左、右图像的特征图；然后，计算对应特征组之间的相关性，再将计算结果拼接起来。该过程可具体描述为

$$\boldsymbol{C}_{lr}(g,d,y,x) = \frac{1}{C_{lr}/G}\left\langle \boldsymbol{F}^{g}_{L,lr}(\cdot,y,x), \boldsymbol{F}^{g}_{R,lr}(\cdot,y,x-d) \right\rangle \tag{1}$$

式中，$\langle \cdot, \cdot \rangle$ 表示向量内积，$\boldsymbol{F}^{g}_{L,lr}$ 和 $\boldsymbol{F}^{g}_{R,lr} \in \mathbf{R}^{(C_{lr}/G) \times H_{lr} \times W_{lr}}$ 分别表示左、右特征图的第 $g$ 组特征，$x$ 和 $y$ 分别表示像素的列索引和行索引（亦即像素在宽度和高度两个维度上的索引坐标），$\boldsymbol{F}^{g}_{L,lr}(\cdot,y,x) \in \mathbf{R}^{C_{lr}/G}$ 表示第 $x$ 列第 $y$ 行像素的特征（一维向量），$d$ 表示候选视差，低分辨率代价体 $\boldsymbol{C}_{lr} \in \mathbf{R}^{G \times D_{lr} \times H_{lr} \times W_{lr}}$，$D_{lr}$ 表示低分辨率下候选视差的总数。本文设置 $G = 8$。

在第2匹配阶段，候选视差 $\boldsymbol{U}_h \in \mathbf{R}^{N \times H_h \times W_h}$ 均由网络自身基于先验信息预测得到（详见第1.5节），其中，超参数 $N$ 表示每个像素的候选视差的个数，本文设置 $N = 7$，远远小于高分辨率下所有可能视差的总数，从而大幅度地降低了代价聚合过程的计算开销。此时，使用左、右图像的高分辨率特征图 $\boldsymbol{F}_{L,h}$ 和 $\boldsymbol{F}_{R,h} \in \mathbf{R}^{C_h \times H_h \times W_h}$ 通过对应特征做差的方式构建窄代价体 $\boldsymbol{C}_h$，即

$$\boldsymbol{C}_h(\cdot,d,y,x) = \boldsymbol{F}_{L,h}(\cdot,y,x), -\boldsymbol{F}_{R,h}(\cdot,y,x-d) \tag{2}$$

式中，$\boldsymbol{F}_{L,h}$ 和 $\boldsymbol{F}_{R,h} \in \mathbf{R}^{C_h \times H_h \times W_h}$，$\boldsymbol{C}_h \in \mathbf{R}^{C_h \times N \times H_h \times W_h}$。为了恢复视差图中的细节信息，窄代价体 $\boldsymbol{C}_h$ 需要具备较高的空间分辨率（即 $H_h \times W_h$），但是超参数 $N$ 远远小于高分辨率下所有可能视差的总数，大幅降低了代价体的尺寸以及代价聚合过程所需的计算开销。

### 1.4 融合注意力（MA）代价聚合模块

在主流的立体匹配网络中，代价聚合模块需要使用多层 3D 卷积对 4D 代价体进行卷积操作，是整个立体匹配流程中计算开销最大的部分。由于代价聚合对整个模型的准确性具有重要影响，简单地缩减代价聚合模块的规模会降低匹配代价的准确性，从而难以获得准确的视差估计结果。为了设计适用于轻量级模型的代价聚合模块，本文基于通道注意力机制和 2D 卷积构建了低计算开销的融合注意力（MA）代价聚合模块，在大幅度缩减计算开销的同时保持良好的准确性。

第 1 匹配阶段的 MA 模块结构如图 5 所示。首先，将通道和视差两个维度融合成一个维度，从而将属于 $\mathbf{R}^{G \times D_{lr} \times H_{lr} \times W_{lr}}$ 的 4D 代价体转换为属于 $\mathbf{R}^{C_m \times H_{lr} \times W_{lr}}$ 的 3D 代价体；紧接着，通过一个 2D 卷积层，将 3D 代价体的通道数压缩为 $C_m/2$；然后，从左图像的低分辨率特征图 $\boldsymbol{F}_{L,lr}$ 中学习获得注意力矩阵 $\boldsymbol{A} \in \mathbf{R}^{C_m/2 \times H_{lr} \times W_{lr}}$，用 $\boldsymbol{A}$ 对代价体进行注意力加权；最后，将注意力加权前后的代价体拼接起来，并采用多个 2D 卷积层融合它们。

为了进一步提升性能，再次执行最后两个步骤，如图 5 所示。需要注意的是，常见的通道注意力机制一般设定注意力矩阵属于 $\mathbf{R}^{C_m/2}$，即设定“不同的通道具有不同的重要性，而不考虑位置变化”。与之不同，本文的注意力矩阵 $\boldsymbol{A} \in \mathbf{R}^{C_m/2 \times H_{lr} \times W_{lr}}$，即设定“不同的通道具有不同的重要性，并且同一通道在不同的像素位置也应具有不同的重要性”，这一机制更符合立体匹配任务的实际情况。

在第 2 匹配阶段，采用类似的融合注意力代价聚合策略对窄代价体进行代价聚合，并且额外引入第 1 阶段的视差图 $\dot{\boldsymbol{D}}_h$ 和下采样的左图像 $\boldsymbol{I}_{L,h}$ 作为辅助信息，详细过程如图 6 所示。

### 1.5 自适应视差传播（ADP）和视差偏移估计（DOE）

提高代价体的空间分辨率有利于恢复视差图中的细节信息，进而获得高质量的视差图，因此，第 2 匹配阶段需要构建具有较高空间分辨率的代价体。

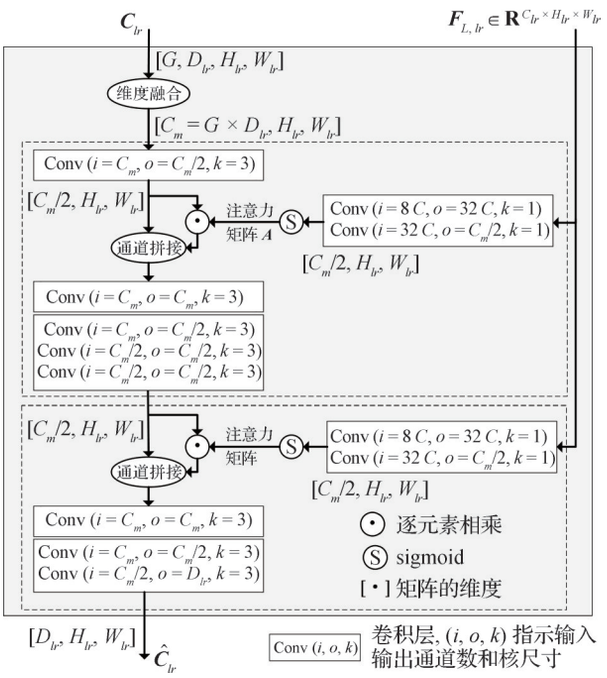



图 5 第 1 匹配阶段的融合注意力（MA）代价聚合策略

Fig. 5 Merged attention (MA) cost aggregation strategy at the first matching stage

然而随着空间分辨率的提高，候选视差范围也会等比例增大，导致高分辨率代价体引入大量的计算开销。鉴于此，本文模型在第 2 匹配阶段放弃构建全量代价体，而仅构建“窄”代价体，即空间分辨率高但候选视差个数（$N$）很小的代价体。于是，为每个像素位置选取 $N$ 个合适的候选视差就成为构建窄代价体的关键。Dai 等人（2022）基于初始视差估计值和下采样后的左图像两种信息预测候选视差；Zhang 等人（2020）的研究表明，代价体的方差一定程度上反映了视差估计值的置信度。本文指出，经过第 1 匹配阶段，实际上已经能够获得关于候选视差的更多先验信息，这些先验信息包括初始视差估计值、左图像、左右图像之间的重构误差以及视差估计值的置信度等。这些先验信息使得估计高质量的候选视差成为可能。基于这些先验知识，本文在第 2 匹配阶段通过简洁高效的自适应视差传播（ADP）模块和视差偏移估计（DOE）模块为每个像素预测 $N$ 个候选视差。

自适应视差传播（ADP）模块的结构如图 7 所示。ADP 模块在低分辨率视差图、视差置信度和图像特征的引导下，借助变形卷积（Zhu 等，2019）以极低的计算开销完成低分辨率视差图的空间自适应上采样。

具体地，ADP 模块首先基于三方面的先验信息

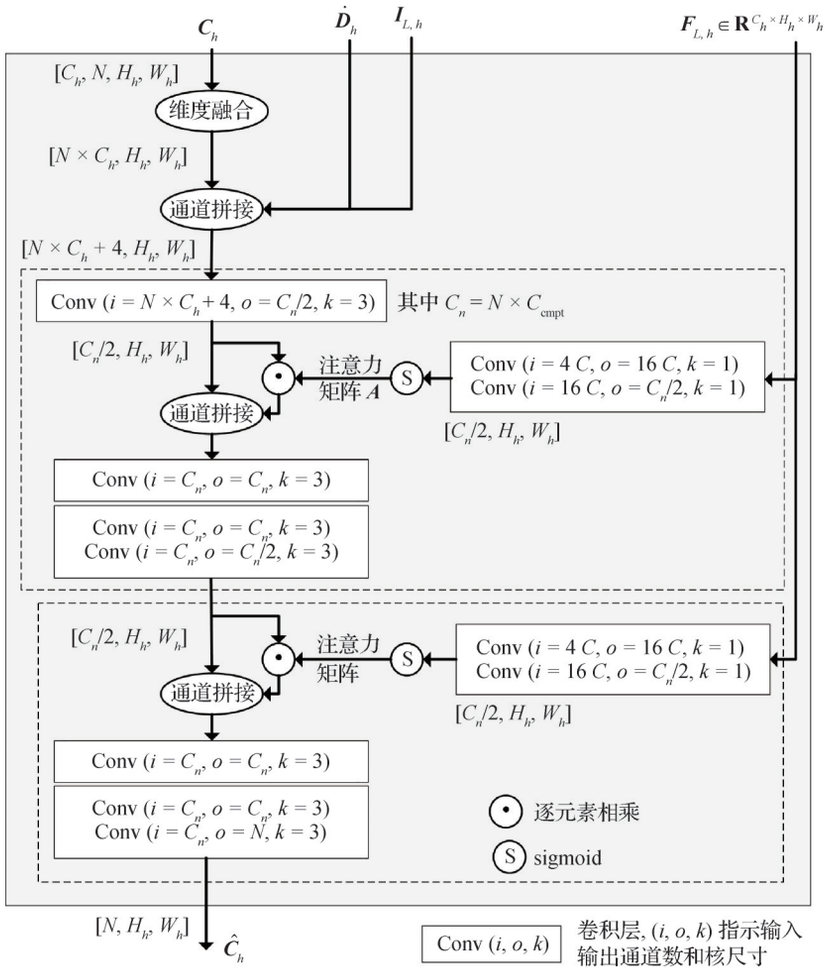



图6　第2匹配阶段的融合注意力(MA)代价聚合策略

Fig. 6　Merged attention (MA) cost aggregation strategy at the second matching stage

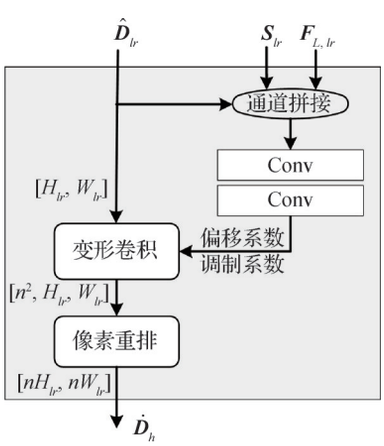



图7　自适应视差传播(ADP)模块

Fig. 7　Adaptive disparity propagation (ADP) module

(包括低分辨率视差图 $\hat{\boldsymbol{D}}_{lr} \in \mathbf{R}^{H_{lr} \times W_{lr}}$、左图像的低分辨率特征图 $\boldsymbol{F}_{L,lr}$ 以及低分辨率视差图的置信度 $\boldsymbol{S}_{lr}$)预测变形卷积的偏移系数和调制系数;接着,通过变形卷积提升低分辨率视差图 $\hat{\boldsymbol{D}}_{lr} \in \mathbf{R}^{H_{lr} \times W_{lr}}$ 通道数至 $n^2$,即获得维度为 $\mathbf{R}^{n^2 \times H_{lr} \times W_{lr}}$ 的多通道视差图,再经过像素重排(pixel-shuffle)即可获得高分辨率的视差图 $\dot{\boldsymbol{D}}_h \in \mathbf{R}^{H_h \times W_h}$。其中,$H_{lr} \times W_{lr}$ 和 $H_h \times W_h$ 表示代价体的空间分辨率,下标 $lr$ 和 $h$ 分别指示低分辨率和高分辨率。上述过程以极低的计算开销实现了低分辨率视差的空间自适应上采样,该过程可具体描述为

$$\boldsymbol{S}_{lr} = \delta\left(\hat{\boldsymbol{C}}_{lr}\right) \tag{3}$$

$$\dot{\boldsymbol{D}}_h = ADP\left(\hat{\boldsymbol{D}}_{lr}, \boldsymbol{F}_{L,lr}, \boldsymbol{S}_{lr}\right) \tag{4}$$

式中,$\hat{\boldsymbol{C}}_{lr} \in \mathbf{R}^{D_{lr} \times H_{lr} \times W_{lr}}$ 表示代价聚合后的低分辨率代价体,$\delta\left(\hat{\boldsymbol{C}}_{lr}\right) \in \mathbf{R}^{H_{lr} \times W_{lr}}$ 表示沿着 $\hat{\boldsymbol{C}}_{lr}$ 的视差维度计算标准差,反映了低分辨率视差估计值的置信度,$ADP(\cdot, \cdot, \cdot)$ 表示ADP模块(如图7所示),$\boldsymbol{F}_{L,lr}$ 表示左图像的低分辨率特征图。

视差偏移估计(DOE)模块的结构如图8所示。

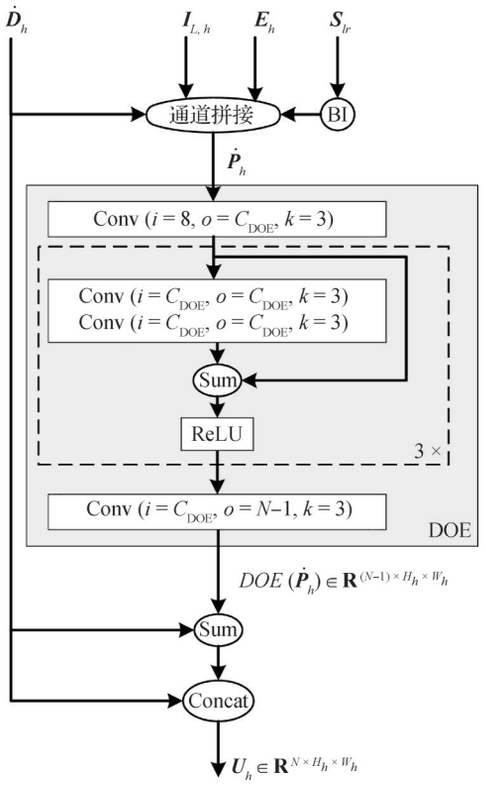



图 8　视差偏移估计（DOE）模块

Fig. 8　Disparity offset estimation (DOE) module

首先，DOE 模块整合已知的先验信息，得到先验信息矩阵 $\dot{\boldsymbol{P}}_h$。具体为

$$\dot{\boldsymbol{P}}_h = Concat\left(\dot{\boldsymbol{D}}_h, \boldsymbol{I}_{L,h}, \boldsymbol{E}_h, BI\left(\boldsymbol{S}_{lr}\right)\right) \tag{5}$$

式中，$Concat(\cdot,\cdot,\cdot,\cdot)$ 表示通道拼接操作，$\boldsymbol{I}_{L,h}$ 表示从全分辨率左图像 $\boldsymbol{I}_L$ 降采样得到的高分辨率左图像，$\boldsymbol{E}_h = \boldsymbol{I}_{L,h} - warp\left(\boldsymbol{I}_{R,h}, \dot{\boldsymbol{D}}_h\right)$ 表示基于右图像 $\boldsymbol{I}_{R,h}$ 和视差图 $\dot{\boldsymbol{D}}_h$ 来重建左图像 $\boldsymbol{I}_{L,h}$ 时的重建误差，$BI(\cdot)$ 表示双线性插值上采样。然后，DOE 模块使用 2D 卷积基于先验信息矩阵 $\dot{\boldsymbol{P}}_h$ 估计出视差偏移量，即可得到高分辨率的候选视差矩阵 $\boldsymbol{U}_h$，具体为

$$\boldsymbol{U}_h = Concat\left(DOE\left(\dot{\boldsymbol{P}}_h\right) + \dot{\boldsymbol{D}}_h, \dot{\boldsymbol{D}}_h\right) \tag{6}$$

式中，$DOE(\cdot)$ 表示 DOE 模块，$\boldsymbol{U}_h \in \mathbf{R}^{N\times H_h \times W_h}$ 为每个像素位置记录了 $N$ 个候选视差。注意到，与使用固定视差偏移量的方法不同，DOE 模块允许不同位置的像素获得不同的候选视差。

基于候选视差矩阵 $\boldsymbol{U}_h$ 和相应分辨率的左、右图像特征图即可构建出高分辨率的窄代价体 $\boldsymbol{C}_h \in \mathbf{R}^{C_h \times N \times H_h \times W_h}$，其中，$C_h$ 表示图像特征的通道数。窄代价体的具体构建过程详见第 1.3 节。

### 1.6　视差回归和损失函数

在所提 LBSM 网络的架构中，第 1 匹配阶段采用流行的 soft argmin 视差回归策略（Kendall 等，2017），从聚合代价体计算出视差估计值。该策略通过 softmax 归一化操作将聚合代价的负值（即 $-c_d$）转化为关于候选视差 $d$ 的概率分布，然后，将候选视差 $d$ 的概率加权均值作为视差估计值 $\hat{d}$，即

$$\hat{d} = \sum_d d \times \mathrm{softmax}\left(-c_d\right) \tag{7}$$

$$\mathrm{softmax}\left(-c_d\right) = \frac{\mathrm{e}^{-c_d}}{\sum_d \mathrm{e}^{-c_d}} \tag{8}$$

第 2 匹配阶段采用类似的视差回归策略，基于聚合后的窄代价体计算出视差估计值，其计算式与式（7）和式（8）类似，不同之处在于每个像素仅有 $N$ 个候选视差。

从图 2 所示的网络架构图可以看出，除了最终的全分辨率视差图，LBSM 网络的中间流程也可以输出低分辨率的视差图。因此，为了更加有效地监督整个网络流程，本文在训练 LBSM 网络的过程中同时监督全分辨率视差图和中间结果视差图，将损失函数定义为

$$L = \sum_{m=1}^{M} \lambda_m L_m \tag{9}$$

$$L_m = \frac{1}{T}\sum_{(x,y)} smooth_{\mathrm{L_1}}\left(\boldsymbol{D}_{\mathrm{gt}}(y,x) - \boldsymbol{D}_{\mathrm{est}}^{m}(y,x)\right) \tag{10}$$

式中，自然数 $M$ 表示需要监督的视差图数量，标量 $\lambda_m$ 用于平衡各损失项的重要程度，$\boldsymbol{D}_{\mathrm{gt}}$ 表示真值视差图，$\boldsymbol{D}_{\mathrm{est}}^{m}$ 表示网络输出的第 $m$ 个视差图，$T$ 表示像素的总数，$(x,y)$ 为像素坐标，平滑的 $\mathrm{L_1}$ 函数 $smooth_{\mathrm{L_1}}(\cdot)$ 定义为

$$smooth_{\mathrm{L_1}}(x) = \begin{cases} \dfrac{1}{2}x^2 & |x| < 1 \\ |x| - 0.5 & \text{其他} \end{cases} \tag{11}$$

## 2　实验与结果分析

本节在立体匹配领域广泛认可的 Scene Flow、KITTI 2015、KITTI 2012 和 ETH3D 数据集上进行实验。

### 2.1　数据集详情

Scene Flow 数据集（Mayer 等，2016）是由德国弗莱堡大学的视觉小组制作发布的大型仿真合成数据

集，共包含39 000多对立体匹配图像，其中训练集和验证集图像共35 454对，测试集图像4 370对，图像分辨率均为540×960像素，并且提供了稠密的真值视差图用于训练和测试。

KITTI数据集由德国卡尔斯鲁厄理工学院（Karlsruhe Institute of Technology，KIT）和芝加哥丰田技术研究院（Toyota Technological Institute at Chicago，TTI-C）使用真实的自动驾驶平台采集制作，为立体视觉、光流估计等任务提供了具有挑战性的、真实场景的测试基准，其中KITTI 2012数据集（Geiger等，2012）和KITTI 2015数据集（Menze和Geiger，2015）是可用于双目立体匹配的两个数据子集。KITTI 2012数据集包含了194对训练图像和195对测试图像；KITTI 2015数据集包含了200对训练图像和200对测试图像，并且包含了更多的动态场景。两个数据集的图像分辨率约为376×1 240像素，为训练图像提供了由激光雷达点云转换而来的、稀疏的真值（ground truth）视差图，平均稠密程度约为50%。

ETH3D（multi-view stereo benchmark of ETH Zurich）（Schops等，2017）是立体匹配领域中广泛认可和使用的小型数据集，它包含了室内外的多种场景。

### 2.2 评价指标

实验采用立体匹配领域常用的定量评价指标，客观地比较不同模型的性能，所用评价指标包括：1）端点误差（end-point error，EPE）：表示预测视差图和真值视差图之间的平均绝对误差；2）$n$像素误差（$n$-pixel error，$n$-px）：是指视差误差大于$n$的像素占有效像素（即具有有效真值视差的像素）的百分比；3）D1错误率：是指视差的绝对误差大于3像素且相对误差大于5%真值视差的像素所占的百分比。特别地，KITTI 2015和KITTI 2012数据集还支持在非遮挡区域（Noc）或全图像区域（All）内分别统计上述指标。

ETH3D数据集上的官方评价指标包括端点误差（EPE）、A95和均方根误差（root mean square error，RMSE）等。其中，A95是指“视差最准确的前95%像素中，视差误差的最大值”，RMSE为视差误差的均方根。

此外，实验采用参数量、“乘法—累加”操作数（MACs）和推理时间（Time）来评估和比较不同模型的计算开销。

### 2.3 实验环境与参数配置

本文采用PyTorch框架编程实现所提出的模型，在单张NVIDIA RTX 3090 GPU上基于Adam优化器（Kingma和Ba，2015）端到端地进行训练。在训练过程中，将图像随机裁剪为256×512像素。设置最大候选视差$d_{\max}=192$，损失权重$\lambda_1=0.25$，$\lambda_2=\lambda_3=0.5$、$\lambda_4=1.0$和$\lambda_5=1.5$。

在Scene Flow数据集上，采用最大学习率为$10^{-3}$的单周期学习率调度器（Smith和Topin，2019），训练本文网络共60轮。由于KITTI 2012和KITTI 2015数据集的训练样本数量较少，不足以充分地训练网络，故使用Scene Flow数据集上的预训练模型在KITTI 2012和KITTI 2015训练集构成的混合训练集上进行800轮训练，设置初始学习率为$5\times10^{-4}$且每200轮减小一半。在ETH3D数据集上采用与KITTI数据集相同的训练策略。

### 2.4 不同配置下的性能测试

通过配置不同的超参数，可以伸缩LBSM网络的规模（参数量），从而能够依据计算资源的多寡，灵活地权衡网络的准确性和计算开销，以使其适用于多样化的算力场景。具体地，本文从LBSM网络的框架中提取出对模型参数量具有重要影响的3个可配置超参数，包括多尺度特征提取（MFE）模块中的通道参数$C$、视差偏移估计（DOE）模块中的通道参数$C_{\mathrm{DOE}}$以及窄代价体代价聚合过程中的通道参数$C_{\mathrm{cmpt}}$。首先，将通道参数$C$、$C_{\mathrm{DOE}}$和$C_{\mathrm{cmpt}}$分别设置为2、4和8，得到本文LBSM网络的计算开销最低的模型变体LBSM-S；然后，将LBSM-S的通道参数分别乘以扩张因子2和4，对应地得到模型LBSM-M和LBSM-L。此外，在LBSM-L的基础上加入视差精确化模块，即得到模型LBSM-L-Refine。

在Scene Flow数据集上分别训练和测试上述模型变体，从而全面地了解所提模型在不同规模下的实际性能，结果如表1所示。首先，最轻量级的模型LBSM-S仅有0.42 M参数和3.36 G MACs，在未使用特殊加速技术的前提下，在单张RTX 3090 GPU上处理一对分辨率为576×960像素的立体图像只需12.7 ms（即78.7帧/s）；其次，随着模型规模的逐步增大，LBSM的端点误差（EPE）大幅度降低，从1.611像素逐步降低到0.948像素，3像素误差（3-px）从8.20%逐步降低到4.86%，表明模型的准确性得到显著提升；注意到，在3种模型变体LBSM-S/M/L中，

计算开销最大的LBSM-L只有1.91 M参数和39.38 G MACs，其处理一对576×960像素的立体图像也仅需16.8 ms（约59.5帧/s）；最后，模型LBSM-L-Refine通过引入视差精确化模块，以增加计算开销和推理延迟为代价，进一步提升了准确性。

上述结果表明，LBSM网络框架具有非常高的计算效率，并且可依据计算资源的多寡配置该网络框架，使其适用于不同算力的应用场景。

**表1　不同规模的LBSM模型在Scene Flow数据集上的性能**

**Table 1　Performance of LBSM models of different scales on the Scene Flow dataset**

| 模型变体 | Scene Flow | | | | 参数量/M | MACs/G | Time/ms |
|---|---|---|---|---|---|---|---|
| | EPE/像素 | 1-px/% | 2-px/% | 3-px/% | | | |
| LBSM-S | 1.611 | 18.18 | 10.95 | 8.20 | **0.42** | **3.36** | **12.7** |
| LBSM-M | 1.197 | 13.59 | 8.23 | 6.20 | 0.72 | 10.71 | 12.9 |
| LBSM-L | 0.948 | 11.08 | 6.54 | 4.86 | 1.91 | 39.38 | 16.8 |
| LBSM-L-Refine | **0.864** | **9.53** | **5.64** | **4.27** | 2.09 | 60.69 | 28.1 |

注：加粗字体表示各列最优指标。Time是在单张RTX 3090 GPU上测得的推理时间，输入图像尺寸被填充为576×960像素。

### 2.5　消融实验

为了评估所提策略或方法的有效性和计算开销，基于LBSM-M模型进行消融实验。具体地，首先构建与LBSM-M架构相同的基线模型，该基线模型分别采用双线性插值和多层3D卷积执行视差图上采样和代价聚合；然后，将本文方法逐步集成到该基线模型中，测试模型性能的变化。实验结果如表2所示，从第1行和第2行数据可以看出，自适应视差传播（ADP）模块以仅仅增加约0.03 M参数量、0.36 G MACs和1.1 ms推理时间为代价，将端点误差（EPE）和1像素误差（1-px）分别降低了21.54%和20.73%，大幅度地提升了基线模型的准确性。该实验结果验证了自适应视差传播策略的有效性和高效性，并且有力地证明了视差上采样策略重要性。现有研究工作普遍采用双线性插值上采样，在一定程度上低估了视差上采样策略的重要性。

继续增加融合注意力（MA）代价聚合模块之后，模型的误差指标进一步降低，准确性得到进一步提升。对比第1、4行数据可以看出，所提方法不仅大幅度地提升了基线模型的准确性，而且降低了模型的计算量（MACs）和推理时间，表明本文方法不仅非常有效，而且具有非常低的计算开销。

### 2.6　对比实验与结果分析

为了验证LBSM网络的实际性能，将其与ADCPNet、CDANet、FBPGNet和BGNet等当前主流的轻量级立体匹配模型进行实验对比。

**表2　在Scene Flow数据集上的消融实验**

**Table 2　Ablation studies on the Scene Flow dataset**

| 序号 | 融合注意力(MA)代价聚合 | | 自适应视差传播 | Scene Flow | | | | 参数量/M | MACs/G | Time/ms |
|---|---|---|---|---|---|---|---|---|---|---|
| | 第1阶段 | 第2阶段 | | EPE/像素 | 1-px/% | 2-px/% | 3-px/% | | | |
| 1 | × | × | × | 1.620 | 17.80 | 10.67 | 7.92 | **0.26** | 11.47 | 14.9 |
| 2 | × | × | √ | 1.271 | 14.11 | 8.56 | 6.46 | 0.29 | 11.83 | 16.0 |
| 3 | √ | × | √ | 1.254 | 13.86 | 8.44 | 6.38 | 0.58 | 11.80 | 15.8 |
| 4 | √ | √ | √ | **1.197** | **13.59** | **8.23** | **6.20** | 0.72 | **10.71** | **12.9** |

注：加粗字体表示各列最优指标。Time是在单张RTX 3090 GPU上测得的推理时间，输入图像尺寸被填充为576×960像素。“√”和“×”分别表示使用和未使用。

2.6.1　Scene Flow数据集上的对比实验

Scene Flow数据集上的实验结果如表3所示。同时，将表3中不同模型的端点误差（EPE）和“乘法—累加”操作数（MACs）绘制到二维坐标系中，以直观地展示和比较不同模型在准确性和模型规模之间的权衡（trade-off），如图9所示。

**表3　在Scene Flow数据集上的实验结果**

**Table 3　Experimental results on the Scene Flow dataset**

| 模型 | EPE/像素 | 参数量/M | MACs/G | 时间/ms | GPU |
|---|---|---|---|---|---|
| ADCPNet-SN7（Dai等，2022） | 2.89 | **0.05** | – | – | – |
| ADCPNet-MN7（Dai等，2022） | 1.79 | 0.4 | – | – | – |
| ADCPNet-LN9（Dai等，2022） | 1.43 | 3.6 | – | – | – |
| CDANet（Liang等，2023b） | 1.42 | – | – | 10 | RTX 2080 Ti |
| FBPGNet（Wen等，2022） | 1.20 | 1.3 | – | 44 | RTX 3090 |
| 2D-MSNet（Shamsafar等，2022） | 1.14 | 2.2 | 136.89 | 103* | RTX 3090 |
| BGNet（Xu等，2021） | 1.17 | 3.0 | 57.46 | 25 | RTX 2080 Ti |
| StereoNet（Khamis等，2018） | 1.10 | 0.6 | 65.21† | 15 | Titan X |
| DeepPruner-Fast（Duggal等，2019） | 0.97 | 7.5 | 219.13 | 48* | RTX 3090 |
| CKDNet（Pan等，2024） | 0.87 | 2.2 | – | 96 | Titan Xp |
| AANet（Xu和Zhang，2020） | 0.87 | 3.9 | 180.01 | 65* | RTX 3090 |
| GFANet（Zhang和Zhang，2023） | **0.82** | – | – | 63 | RTX 3090 |
| LBSM-S（本文） | 1.61 | 0.42 | **3.36** | 12.7 | RTX 3090 |
| LBSM-M（本文） | 1.20 | 0.72 | 10.71 | 12.9 | RTX 3090 |
| LBSM-L（本文） | 0.95 | 1.91 | 39.38 | 16.8 | RTX 3090 |
| LBSM-L-Refine（本文） | 0.86 | 2.09 | 60.69 | 28.1 | RTX 3090 |

注：加粗、下划线字体分别表示各列最优、次优结果。“–”表示原文作者未公布相关指标和源码，“†”表示本文作者实现和测得的数据（因原作者未公布源码），“*”表示本文作者使用官方开源代码测得的数据。

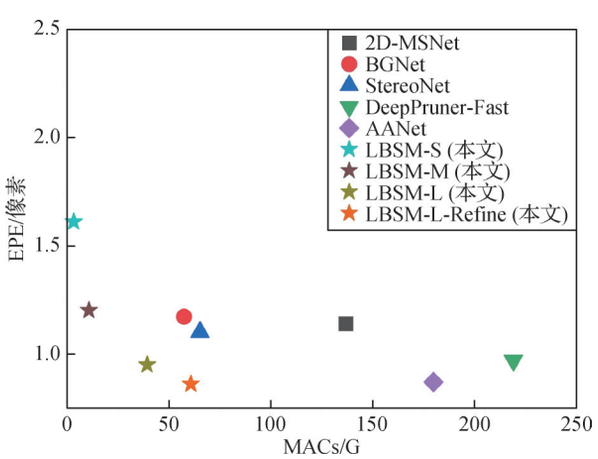



图9　不同模型在Scene Flow数据集上的准确性和计算开销

Fig. 9　Accuracy and computational cost of different models on the Scene Flow dataset

从表3可以看出，轻量级模型ADCPNet-SN7具有最低的参数量（仅0.05 M），但是其端点误差（EPE）高达2.89像素，准确性明显不足；ADCPNet-MN7以0.4 M参数量取得了1.79像素的端点误差，在准确性和参数量之间获得了不错的权衡，与之相比，本文LBSM-S以相近的参数量（0.42 M vs. 0.4 M）取得了更高的准确性（EPE 1.61 vs. 1.79）。轻量级模型CDANet以及经典模型StereoNet都具有非常高的推理速度，但是其端点误差分别为1.42像素和1.10像素，均大于1像素，仍然未达到亚像素级。注意到，StereoNet的参数量和MACs分别为0.6 M和65.21 G，而BGNet的参数量和MACs分别为3.0 M和57.46 G。两者对比说明，参数量小并不意味着

MACs更少。与之相比，本文模型LBSM-L以1.91 M参数量和39.38 G MACs取得了亚像素级的端点误差（EPE为0.95），准确性明显高于StereoNet和BGNet，同时LBSM-L的MACs仅为两者的60.4%和68.5%。

注意到，非轻量级模型DeepPruner-Fast、CKD-Net、AANet和GFANet以较高的计算开销和推理时间获得了亚像素级的端点误差，其准确性明显高于现有的轻量级模型。与之相比，本文的全量模型LBSM-L-Refine能够以大幅度领先的MACs和推理速度取得更优或相当的准确性。例如，LBSM-L-Refine以44.6%的推理时间（28.1 ms vs. 63 ms）取得了与GFANet非常接近的准确性（EPE 0.86 vs. 0.82）；LBSM-L-Refine以33.7%的MACs和43.2%的推理时间取得了与AANet几乎一致的准确性（EPE 0.86 vs. 0.87），同时以27.7%的MACs和58.5%的推理时间取得了显著优于DeepPruner-Fast的准确性（EPE 0.86 vs. 0.97）。

上述结果表明，与现有的轻量级立体匹配模型、甚至非轻量级模型相比，LBSM网络框架在准确性和实时性方面均具有明显优势。特别地，图9直观地展示了不同模型在端点误差和计算开销之间的权衡。可以看出，与现有先进轻量级模型相比，LBSM-M、LBSM-L和LBSM-L-Refine更靠近坐标原点，表明它们在准确性和计算开销之间取得了更好的权衡，使其在计算资源受限的场景中具有较大的应用潜力。

2.6.2 KITTI 2012和KITTI 2015数据集上的对比实验

为了考察所提轻量级模型在真实道路和街景下的有效性，分别在KITTI 2012和KITTI 2015数据集上测试所提模型的性能，并与现有轻量级模型和具有代表性的非轻量级模型进行性能对比，结果如表4所示。可以看出，本文的全量模型LBSM-L- Refine在KITTI 2012和KITTI 2015上均取得了最优的准确性。首先，在6种现有的非轻量级模型中，AANet取得了最高的准确性，与之相比，本文LBSM-L-Refine模型以2.5倍的推理速度取得了更高的准确性。其次，在18种现有的轻量级模型中，ADCPNet-L-Refine取得了最高的准确性，与之相比，本文LBSM-L-Refine模型以相近的推理速度显著地提升了准确性，使得KITTI 2012上的3-px-Noc和3-px-All误差指标分别降低了13.6%和9.4%；在KITTI 2015上的D1-Noc和D1-All错误率分别降低了7.9%和7.0%。特别注意到，本文的LBSM-L模型（未采用视差精确化模块）的推理时间仅为14 ms，其准确性已经非常接近现有模型ADCPNet-L-Refine（采用了视差精确化模块），并且已经显著优于其他17种轻量级模型。最后，现有模型ADCPNet-SN3、GWDRNet-S、HRSNet和CMNet等的推理时间更短，但是准确性明显不足，即便是准确性较高的GWDRNet-S模型，在KITTI 2015数据集上的D1-Noc和D1-All错误率也分别高达3.37%和3.64%，远高于本文的LBSM-M和LBSM-L模型。上述结果表明，所提模型具有显著的性能优势，在准确性和推理速度之间取得了良好权衡，在对实时性和准确性要求较高的场景中具有较大的应用潜力。

图10和图11分别展示了不同模型在KITTI 2012和KITTI 2015数据集上的视差结果实例以及对应的误差图，并在误差图上标注了误差指标的具体数值。误差图中白色表示误差较大的像素，红色表示遮挡区域。可以看出，本文模型不仅在物体边缘处获得了更锋利的视差边缘，而且在路面、墙面等平坦和弱纹理区域也获得了更准确的视差结果。

2.6.3 ETH3D数据集上的对比实验

本文模型和其他快速立体匹配模型在ETH3D数据集上的性能对比如表5所示。为了能与尽量多的现有模型进行公平比较，表5的推理时间均在单张RTX 2080Ti GPU上测得，表中指标数值越低越好。从表5可以看出，ADCPNet-SN3和AnyNet_C1虽然推理速度较快，但准确性较低；DeepPruner-Fast和ADCPNet-L-Refine准确性较高，但推理速度较慢。与DeepPruner-Fast和ADCPNet-L-Refine相比，本文的LBSM-L分别以1.92倍和3.17倍的推理速度取得明显更高的准确性。上述结果表明，本文方法在准确性和计算开销之间取得更好的权衡。

此外，图12展示了不同模型预测的视差图，可以看出，本文LBSM-L-Refine模型在物体边界处生成了更清晰的视差边缘，而且在平滑、平坦和低纹理区域也获得了更平滑的视差结果。

## 3 低算力硬件平台上的部署

许多实际应用场景要求立体匹配模型在低功

表4　在KITTI 2012和KITTI 2015数据集上的实验结果

Table 4　Experimental results on the KITTI 2012 and KITTI 2015 datasets

| 类型 | 模型 | KITTI 2012 | | KITTI 2015 | | 时间/ms |
|---|---|---|---|---|---|---|
| | | 3-px-Noc/% | 3-px-All/% | D1-Noc/% | D1-All/% | |
| 非轻量级模型 | 2D-MSNet(Shamsafar等,2022) | – | – | 2.54 | 2.83 | 97* |
| | MFPSNet(Pan等,2021) | – | – | 2.98 | 3.27 | 87 |
| | AANet(Xu和Zhang,2020) | 1.91 | 2.42 | **2.32** | 2.55 | 57* |
| | FADNet(Wang等,2020) | 2.04 | 2.46 | 2.39 | 2.60 | 50 |
| | DeepPruner-Fast(Duggal等,2019) | 2.26 | 2.73 | 2.35 | 2.59 | 40* |
| | FBPGNet(Wen等,2022) | – | – | 4.04 | 4.30 | 40 |
| 轻量级模型 | ADCPNet-L-Refine(Dai等,2022) | 2.20 | 2.66 | 2.52 | 2.71 | 27 |
| | RST-Stereo(Ye等,2022) | – | – | – | 2.84 | 22 |
| | ADCPNet-LN9(Dai等,2022) | 2.44 | 3.04 | 2.84 | 3.09 | 21 |
| | RTS2Net(Dovesi等,2020) | – | – | 3.22 | 3.56 | 20 |
| | MADNet(Tonioni等,2019) | 5.34 | 6.39 | 4.27 | 4.66 | 20 |
| | RTSNet(Lee和Shin,2019) | 2.43 | 2.90 | – | 3.41 | 20 |
| | RCVNet(Ye等,2024) | 2.41 | 2.88 | – | 2.98 | 16 |
| | StereoNet(Khamis等,2018) | – | – | – | 4.83 | 15 |
| | P3SNet+(Emlek和Peker,2023) | 3.55 | 4.42 | 4.32 | 4.72 | 15 |
| | RTSMNet(Xie等,2021) | – | – | 3.57 | 3.88 | 14 |
| | AnyNet_C32(Wang等,2019) | 3.60 | 4.33 | 4.63 | 4.95 | 13 |
| | P3SNet(Emlek和Peker,2023) | 3.65 | 4.46 | 4.65 | 5.05 | 12 |
| | AnyNet_C1(Wang等,2019) | 5.53 | 6.42 | 6.66 | 7.10 | 8 |
| | CMNet(Liu等,2023) | – | 3.41 | – | 3.84 | 8 |
| | ADCPNet-MN7(Dai等,2022) | 3.14 | 3.84 | 3.69 | 3.98 | 7 |
| | HRSNet(Huang等,2023) | – | – | – | 4.28 | 7 |
| | GWDRNet-S(Liang等,2023a) | 2.65 | 3.17 | 3.37 | 3.64 | 6 |
| | ADCPNet-SN3(Dai等,2022) | 4.86 | 5.70 | 5.33 | 5.70 | 6 |
| | LBSM-S(本文) | 3.07 | 3.70 | 3.48 | 3.77 | 11 |
| | LBSM-M(本文) | 2.52 | 3.12 | 2.88 | 3.20 | 12 |
| | LBSM-L(本文) | 2.21 | 2.77 | 2.58 | 2.81 | 14 |
| | LBSM-L-Refine(本文) | **1.90** | **2.41** | **2.32** | **2.52** | 23 |

注:加粗、下划线字体分别表示各列最优、次优结果。现有模型按推理时间降序排列。“–”表示原文作者未公布相关指标和源码,“*”表示使用官方开源代码在本文硬件平台(RTX 3090 GPU)实测的推理时间。

耗、低算力的嵌入式AI硬件平台上实时运行,因此,在常用的嵌入式AI硬件平台NVIDIA Jetson TX2上部署和测试所提出的LBSM网络,并与其他先进轻量级模型进行性能对比。特别地,轻量级立体匹配模型的相关研究工作常常会采用深度学习推理加速工具包NVIDIA TensorRT(TRT)进一步提升模型的推理速度,本文也测试了该加速工具包对所提模型的实际加速效果。注意到,该工具包会自动综合利用多种加

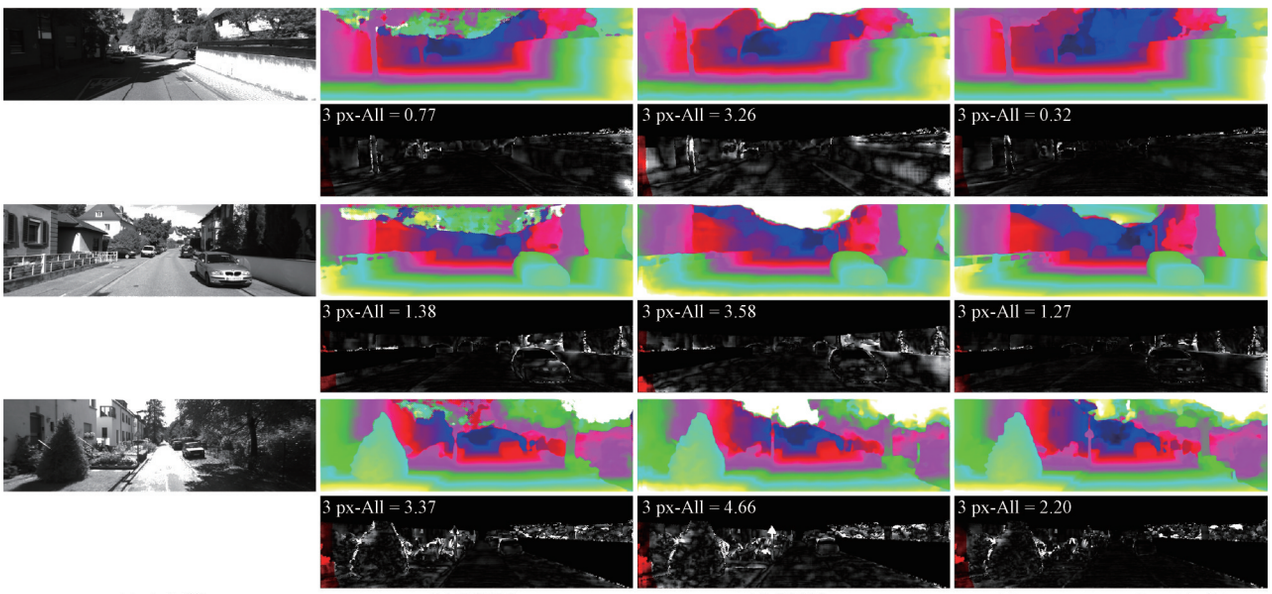



图 10　KITTI 2012 数据集上不同模型预测的视差图及对应的误差图

Fig. 10　The disparity maps predicted by different models and corresponding error maps on KITTI 2012 dataset

((a) left images; (b) RTSNet; (c) P3SNet; (d) LBSM-L-Refine (ours))

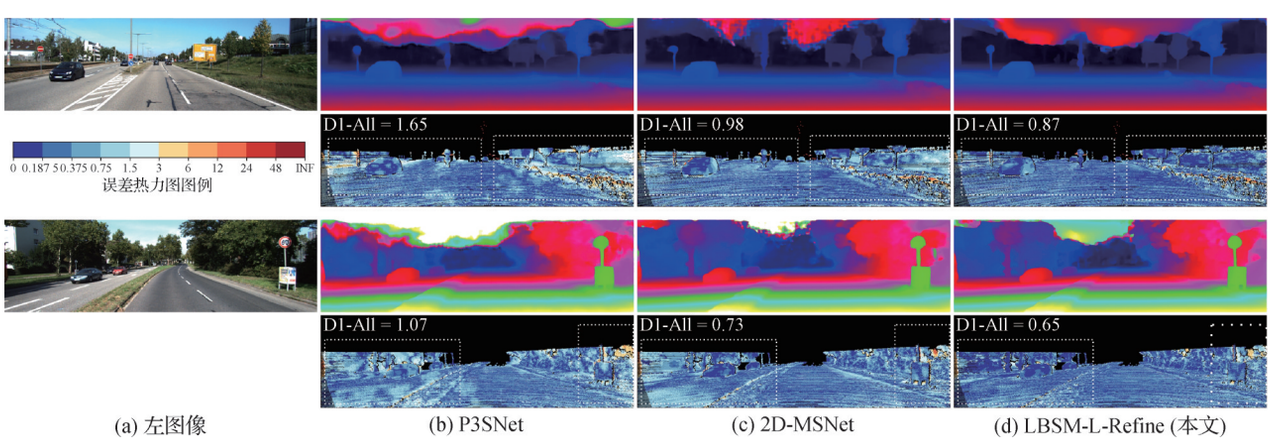



图 11　KITTI 2015 数据集上不同模型预测的视差图及对应的误差图

Fig. 11　The disparity maps predicted by different models and corresponding error maps on KITTI 2015 dataset

((a) left images; (b) P3SNet; (c) 2D-MSNet; (d) LBSM-L-Refine (ours))

速手段(如网络层融合、精度量化、硬件感知优化等)对模型进行推理加速,本文将量化参数设置为“—fp16,—best”,使其自动为每个网络层选择 FP32 和 FP16 两者中性能最优的精度,同时保证精度损失可控。

详细数据如表 6 所示。可以看出,在参与对比的 11 种现有的轻量级模型中,DeepPruner-Fast 和 ADCPNet-L-Refine 取得了较高的准确性,与之相比,本文模型 LBSM-L-Refine 分别以 3.3 倍和 1.2 倍的推理速度(帧率)取得了更高的准确性;现有模型 ADCPNet-SN3 和 AnyNet_C1 在 Jetson TX2 平台上分别取得了 22.32 帧/s 和 10.40 帧/s 的推理速度,然而两者的准确性较低,前者的错误率指标 3-px-All 和 D1-All 高达 5.70%,后者的错误率指标更是达到了 6.42% 和 7.10%。与之相比,模型 LBSM-S 以 10.78 帧/s 的推理速度取得了 3.70% 和 3.77% 的错误率,在速度和准确性之间取得更好的权衡。特别地,在使用 TensorRT 加速后,模型 LBSM-S 的推理速度从 10.78 帧/s 提升到 20.29 帧/s。

图 13 更加直观地展示了不同模型在 Jetson TX2

表5　快速模型在ETH3D数据集上的实验结果

Table 5　Experimental results of fast models on ETH3D dataset

| 模型 | EPE/像素 | A95 | RMSE | 时间/ms |
|---|---|---|---|---|
| AnyNet_C1 | 0.68 | 2.38 | 1.20 | 14* |
| MADNet | 0.62 | 1.80 | 1.11 | 63 |
| ADCPNet-SN3 | 0.62 | 2.35 | 1.14 | **8** |
| AnyNet_C32 | 0.55 | 2.14 | 1.08 | 16* |
| ADCPNet-MN7 | 0.53 | 2.11 | 1.06 | 17 |
| ADCPNet-LN9 | 0.50 | 1.82 | 1.07 | 65 |
| DeepPruner-Fast | 0.46 | 1.68 | 0.92 | 46* |
| ADCPNet-L-Refine | 0.41 | 1.50 | 0.88 | 76 |
| LBSM-S(本文) | 0.52 | 2.00 | 0.99 | 19 |
| LBSM-M(本文) | 0.50 | 1.84 | 0.92 | 23 |
| LBSM-L(本文) | 0.40 | 1.42 | 0.81 | 24 |
| LBSM-L-Refine(本文) | **0.35** | **1.18** | **0.76** | 39 |

注:加粗、下划线字体分别表示各列最优、次优结果。现有模型按EPE降序排列。*表示使用原文开源代码测得的数据。

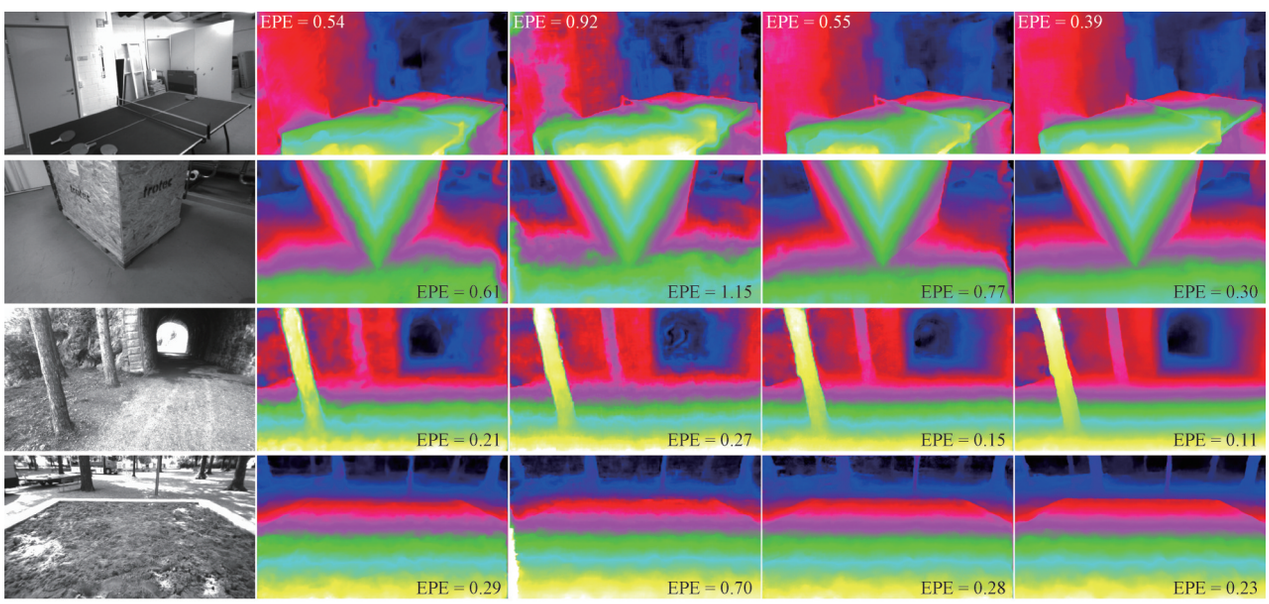



图12　ETH3D数据集上不同模型预测的视差图

Fig. 12　The disparity maps predicted by different models on the ETH3D dataset

((a) left images; (b) AnyNet_C32; (c) MADNet; (d) ADCPNet-LN9 ; (e) LBSM-L-Refine (ours))

硬件平台、KITTI 2015数据集上的错误率和推理速度(帧率)的对比情况。对比现有方法可以看出,在未引入TensorRT加速时,本文模型对应的数据点(五边形图标)已经相对地更加靠近坐标系右下方;在引入TensorRT加速之后,本文模型的推理速度得到进一步提升,对应的数据点(五角星图标)大幅度地向右移动。上述结果表明,所提模型在准确性和推理速度之间取得了更出色的权衡,同时,在实际应用中可依据计算资源的多寡轻松地配置该模型的规模,以满足多样化的算力场景。

表 6　在 Jetson TX2 嵌入式硬件平台上的测试结果

Table 6　Test results on the Jetson TX2 hardware platform

| 模型 | KITTI 2012 | KITTI 2015 | 帧率/(帧/s) |
|---|---|---|---|
| | 3-px-All/% | D1-All/% | |
| StereoNet(Khamis 等,2018) | – | 4.83 | 1.78 |
| DeepPruner-Fast(Duggal 等,2019) | 2.73 | <u>2.59</u> | 0.68 |
| MADNet(Tonioni 等,2019) | – | 4.66 | 3.85 |
| AnyNet_C1(Wang 等,2019) | 6.42 | 7.10 | 10.40 |
| AnyNet_C32(Wang 等,2019) | 4.33 | 4.95 | 3.50 |
| RTS2Net(Dovesi 等,2020) | – | 3.56 | 2.30 |
| RTSMNet+TRT(Xie 等,2021) | – | 3.88 | 3.50 |
| ADCPNet-SN3(Dai 等,2022) | 5.70 | 5.70 | **22.32** |
| ADCPNet-MN7(Dai 等,2022) | 3.84 | 3.98 | 8.93 |
| ADCPNet-LN9(Dai 等,2022) | 3.04 | 3.09 | 2.49 |
| ADCPNet-L-Refine(Dai 等,2022) | <u>2.66</u> | 2.71 | 1.89 |
| LBSM-S(本文) | 3.70 | 3.77 | 10.78 |
| LBSM-M(本文) | 3.12 | 3.20 | 8.07 |
| LBSM-L(本文) | 2.77 | 2.81 | 4.45 |
| LBSM-L-Refine(本文) | **2.41** | **2.52** | 2.26 |
| LBSM-S+TRT(本文) | 3.70 | 3.77 | <u>20.29</u> |
| LBSM-M+TRT(本文) | 3.12 | 3.20 | 13.69 |
| LBSM-L+TRT(本文) | 2.77 | 2.81 | 7.62 |
| LBSM-L-Refine+TRT(本文) | **2.41** | **2.52** | 4.59 |

注:加粗、下划线字体分别表示各列最优、次优结果。“+TRT”表示使用 NVIDIA TensorRT 加速技术。

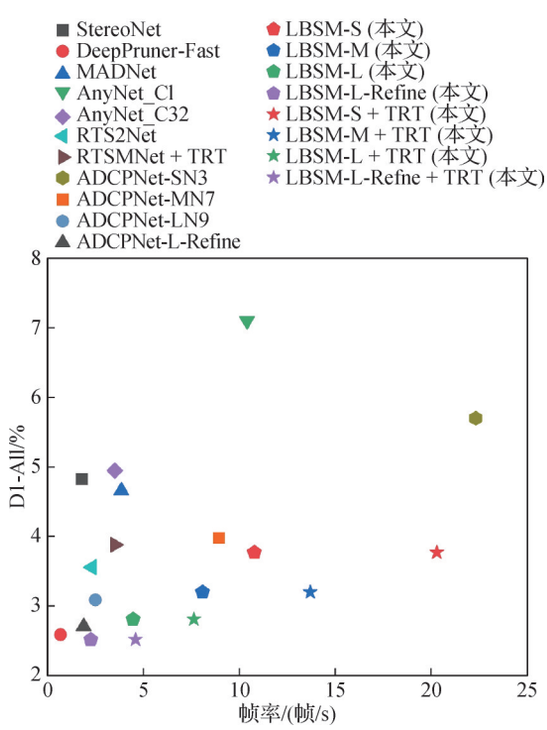



图 13　轻量级模型在 Jetson TX2 硬件平台、KITTI 2015 数据集上的错误率和帧率比较

Fig. 13　Comparison of error rate and frame rate for lightweight models on the Jetson TX2 using the KITTI 2015 dataset

## 4　结　论

针对双目立体匹配网络因计算量较大而难以在低算力边缘计算设备上实时运行的问题,本文提出一种轻量级的双目立体匹配网络框架 LBSM。该框架采用分辨率从低到高的两级级联架构,在低分辨率下构建全量代价体,而在高分辨率下构建窄代价体,避免了高分辨率代价体带来的高计算开销;其次,通过融合代价体的视差和通道两个维度,将 4D 代价体重构为 3D 代价体,再通过轻量化的通道注意力机制完成代价聚合,使得整个代价聚合过程仅依赖 2D 卷积层,从而避免了堆叠大量 3D 卷积层,以更低的计算开销获得了更好的代价聚合效果;最后,设计了计算开销极低的自适应视差上采样策略以实现视差估计值的空间自适应传播,使得低分辨率下的视差估计结果能够更加高效地向高分辨率传播。得益于上述多种措施和方法,LBSM 框架仅依赖低计算开销的 2D 卷积而不需要任何 3D 卷积层。实验结果表明,所提方法在准确性和计算开销之间得到了

更好的权衡，能够在低算力边缘计算设备上以良好的准确性实时地完成双目立体匹配任务。注意到，本文实测和分析了NVIDIA TensorRT对所提模型的加速效果，后续研究工作将深入探索模型剪枝、重参数化以及稀疏化等神经网络加速技术对所提模型的加速潜力。

## 参考文献(References)

Cheng D Q, Li H X, Kou Q Q, Yu Z K, Zhuang H D and Lyu C. 2021. Stereo matching algorithm based on edge preservation and improved cost aggregation. Journal of Image and Graphics, 26(2): 438-451 (程德强, 李海翔, 寇旗旗, 于泽宽, 庄焕东, 吕晨. 2021. 融合边缘保持与改进代价聚合的立体匹配算法. 中国图象图形学报, 26(2): 438-451) [DOI: 10.11834/jig.200041]

Dai H, Zhang X C, Zhao Y L, Sun H B and Zheng N N. 2022. Adaptive disparity candidates prediction network for efficient real-time stereo matching. IEEE Transactions on Circuits and Systems for Video Technology, 32(5): 3099-3110 [DOI: 10.1109/TCSVT.2021.3102109]

Deng Y, Xiao J M, Zhou S Z and Feng J S. 2021. Detail preserving coarse-to-fine matching for stereo matching and optical flow. IEEE Transactions on Image Processing, 30: 5835-5847 [DOI: 10.1109/TIP.2021.3088635]

Dovesi P L, Poggi M, Andraghetti L, Marti M, Kjellstrom H, Pieropan A, et al. 2020. Real-time semantic stereo matching//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 10780-10787 [DOI: 10.1109/ICRA40945.2020.9196784]

Duggal S, Wang S L, 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. Seoul, Korea (South): IEEE: 4383-4392 [DOI: 10.1109/ICCV.2019.00448]

Emlek A and Peker M. 2023. P3SNet: parallel pyramid pooling stereo network. IEEE Transactions on Intelligent Transportation Systems, 24(10): 10433-10444 [DOI: 10.1109/TITS.2023.3276328]

Fan R, Wang L, Junaid Bocus M and Pitas I. 2023. Computer stereo vision for autonomous driving: theory and algorithms//Hosny K M, Salah A, eds. Recent Advances in Computer Vision Applications Using Parallel Processing. Cham, Germany: Springer: 41-70 [DOI: 10.1007/978-3-031-18735-3_3]

Fu J H, Liu H D, He M Q and Zhu D H. 2022. A hand-eye calibration algorithm of binocular stereo vision based on multi-pixel 3D geometric centroid relocalization. Journal of Advanced Manufacturing Science and Technology, 2(1): #2022005 [DOI: 10.51393/j.jamst.2022005]

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. Providence, USA: IEEE: 3354-3361 [DOI: 10.1109/CVPR.2012.6248074]

Guo X Y, Yang K, Yang W K, Wang X G and Li H S. 2019. Group-wise correlation stereo network//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3273-3282 [DOI: 10.1109/CVPR.2019.00339]

Huang J H, Wei W, Liang B F, Liu C, Shang W L and Li J. 2023. Real-time stereo disparity prediction based on patch-embedded extraction and depthwise hierarchical refinement for 3-D sensing of autonomous vehicles on energy-efficient edge computing devices. IEEE Transactions on Green Communications and Networking, 7(1): 424-433 [DOI: 10.1109/TGCN.2023.3233963]

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. 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. Munich, Germany: Springer: 573-590 [DOI: 10.1007/978-3-030-01267-0_35]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]: 1-15

Laga H, Jospin L V, Boussaid F and Bennamoun M. 2022. A survey on deep learning techniques for stereo-based depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(4): 1738-1764 [DOI: 10.1109/TPAMI.2020.3032602]

Lee H and Shin Y. 2019. Real-time stereo matching network with high accuracy//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 4280-4284 [DOI: 10.1109/ICIP.2019.8803514]

Liang B F, Wei W, Huang J H, Liu C, Yang H, Yang R, et al. 2023a. Real-time stereo image depth estimation network with group-wise L1 distance for edge devices towards autonomous driving. IEEE Transactions on Vehicular Technology, 72(11): 13917-13928 [DOI: 10.1109/TVT.2023.3284011]

Liang B F, Yang H, Huang J H, Liu C and Yang R. 2023b. Real-time stereo matching network based on 3D channel and disparity attention for edge devices toward autonomous driving. IEEE Access, 11: 76781-76792 [DOI: 10.1109/ACCESS.2023.3297052]

Liu C, Wei W, Liang B F, Liu X H, Shang W L and Li J. 2023. ConvMLP-mixer based real-time stereo matching network towards autonomous driving. IEEE Transactions on Vehicular Technology,

72(2): 2581-2586 [DOI: 10.1109/TVT.2022.3206612]

Luo C, Yu L J and Ren P. 2018. A vision-aided approach to perching a bioinspired unmanned aerial vehicle. IEEE Transactions on Industrial Electronics, 65(5): 3976-3984 [DOI: 10.1109/TIE.2017.2764849]

Mallot H A, Gillner S and Arndt P A. 1996. Is correspondence search in human stereo vision a coarse-to-fine process? Biological Cybernetics, 74(2): 95-106 [DOI: 10.1007/BF00204198]

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. Las Vegas, USA: IEEE: 4040-4048 [DOI: 10.1109/CVPR.2016.438]

Menz M D and Freeman R D. 2003. Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism. Nature Neuroscience, 6(1): 59-65 [DOI: 10.1038/nn986]

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

Pan B Y, Pang J X and Cheng J. 2024. CKDNet: a cascaded knowledge distillation framework for lightweight stereo matching//Proceedings of the 5th International Conference on Intelligent Computing and Human-Computer Interaction. Nanchang, China: IEEE: 398-403 [DOI: 10.1109/ICHCI63580.2024.10808024]

Pan B Y, Zhang L M and Wang H Z. 2021. Multi-stage feature pyramid stereo network-based disparity estimation approach for two to three-dimensional video conversion. IEEE Transactions on Circuits and Systems for Video Technology, 31(5): 1862-1875 [DOI: 10.1109/TCSVT.2020.3014053]

Scharstein D and Szeliski R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 47(1): 7-42 [DOI: 10.1023/A:1014573219977]

Schops T, Schonberger J L, Galliani S, Sattler T, Schindler K, Pollefey, 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 Computer Vision. Waikoloa, USA: IEEE: 677-686 [DOI: 10.1109/WACV51458.2022.00075]

Smith L N and Topin N. 2019. Super-convergence: very fast training of neural networks using large learning rates//Proceedings of SPIE 11006, Artificial Intelligence and Machine Learning for Multi-domain Operations Applications. Bellingham, USA: SPIE: 369-386 [DOI: 10.1117/12.2520589]

Tonioni A, Tosi F, Poggi M, Mattoccia S and Di Stefano L. 2019. Real-time self-adaptive deep stereo//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 195-204 [DOI: 10.1109/CVPR.2019.00028]

Ul Haq M U, Ashfaque M, Mathavan S, Kamal K and Ahmed A. 2019. Stereo-based 3D reconstruction of potholes by a hybrid, dense matching scheme. IEEE Sensors Journal, 19(10): 3807-3817 [DOI: 10.1109/JSEN.2019.2898375]

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

Wang Y, Lai Z H, 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 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 5893-5900 [DOI: 10.1109/ICRA.2019.8794003]

Wen B, Zhu H, Yang C, Li Z C and Cao R X. 2022. Feature back-projection guided residual refinement for real-time stereo matching network. Signal Processing: Image Communication, 103: #116636 [DOI: 10.1016/j.image.2022.116636]

Xia W Y, Chen E C S, Pautler S and Peters T M. 2022. A robust edge-preserving stereo matching method for laparoscopic images. IEEE Transactions on Medical Imaging, 41(7): 1651-1664 [DOI: 10.1109/TMI.2022.3147414]

Xie Q W, Long Q, Li J P, Zhang L M and Hu X Y. 2024. Application of intelligence binocular vision sensor: mobility solutions for automotive perception system. IEEE Sensors Journal, 24(5): 5578-5592 [DOI: 10.1109/JSEN.2023.3311479]

Xie Y, Zheng S W and Li W H. 2021. Feature-guided spatial attention upsampling for real-time stereo matching network. IEEE MultiMedia, 28(1): 38-47 [DOI: 10.1109/MMUL.2020.3030027]

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

Xu H F and Zhang J Y. 2020. AANet: adaptive aggregation network for efficient stereo matching//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1959-1968 [DOI: 10.1109/CVPR42600.2020.00203]

Ye X Q, Yan B B, Liu B Y, Wang H C, Qi S, Chen D, et al. 2022. Improved real-time three-dimensional stereo matching with local consistency. Image and Vision Computing, 124: #104509. [DOI: 10.1016/j.imavis.2022.104509]

Ye X W, Chen Z J and Deng J W. 2024. Recombination cost volume stereo matching on edge devices//Proceedings of the 6th IEEE International Conference on Power, Intelligent Computing and Systems.

Shenyang, China: IEEE: 359-362 [DOI: 10.1109/ICPICS62053.2024.10796764]

Zhang Y J. 2023. 3-D Computer Vision: Principles, Algorithms and Applications. Singapore, Singapore: Springer [DOI: 10.1007/978-981-19-7580-6]

Zhang Y K and Zhang J H. 2023. GFANet: group fusion aggregation network for real time stereo matching. IEEE Robotics and Automation Letters, 8(7): 4251-4258 [DOI: 10.1109/LRA.2023.3280818]

Zhang Y M, Chen Y M, Bai X, Yu S H J, Yu K, Li Z W, et al. 2020. Adaptive unimodal cost volume filtering for deep stereo matching// Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12926-12934 [DOI: 10.1609/aaai.v34i07.6991]

Zhou K, Meng X X and Cheng B. 2020. Review of stereo matching algorithms based on deep learning. Computational Intelligence and Neuroscience, 2020(1): #8562323 [DOI: 10.1155/2020/8562323]

Zhou L F, Zhang L and Konz N. 2023. Computer vision techniques in manufacturing. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 53(1): 105-117 [DOI: 10.1109/TSMC.2022.3166397]

Zhu X Z, Hu H, Lin S and Dai J F. 2019. Deformable ConvNets V2: more deformable, better results//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9300-9308 [DOI: 10.1109/CVPR.2019.00953]

## 作者简介

武忠，男，博士，讲师，主要研究方向为计算机视觉、双目立体视觉和图像处理。E-mail：wuzhong@ycu.edu.cn

朱虹，通信作者，女，教授，主要研究方向为图像处理、模式识别、计算机视觉、三维感知与重建。

E-mail：zhuhong@xaut.edu.cn

蔺广逢，男，副教授，主要研究方向为视觉信息处理与模式识别。E-mail：lgf78103@xaut.edu.cn

贺丽丽，女，讲师，主要研究方向为图像处理和模式识别。

E-mail：helili@ycu.edu.cn