基于几何先验知识约束的双目视觉深度估计方法

张泽辉[1],王 阳[1],陈博洋[4],张浩轩[1],徐晓滨[1*],吴富龙[1],程胜龙[2],邵海滨[3],李 昊[1]
(1. 杭州电子科技大学,浙江杭州 310018;2. 中汽信息科技(天津)有限公司,天津 300399;
3. 上海交通大学,上海 200240;4. 宁夏石化银骏安全技术咨询有限公司,宁夏银川 750000)

摘 要: 近年来,随着自动驾驶、机器人导航及三维重建等领域的迅速发展,深度估计技术作为感知环境三维结构的关键手段,受到广泛关注。然而,现有基于监督学习的深度估计方法虽然在特定数据集上表现优异,但其泛化能力较弱,且依赖大规模、高质量的标注数据,这严重限制了其在真实工业场景中的应用。因此,本研究提出一种基于几何先验知识约束的双目视觉深度估计方法。首先,组合残差卷积与上下文编码器,从图像数据中提取多尺度特征。接下来,利用特征金字塔结构捕捉不同尺度匹配信息,并保留图像边缘结构细节。然后,设计多级门控制循环(Gated Recurrent Unit,GRU)单元结合不同尺度特征信息对特征匹配参数进行更新,优化视差匹配结果,实现双目视觉深度估计。特别地,本文构建了一种结合监督信号与物理先验的混合损失函数。该函数在传统监督损失的基础上,引入了源自自监督学习范式的几何约束作为正则化项,具体包括左右视差一致性约束和视差结构一致性约束。其中,左右一致性约束通过强制左右视图预测视差满足几何对应关系,以增强模型的几何理解并缓解遮挡区域的误匹配,而结构一致性约束则通过引导视差图在纹理平坦区域保持平滑、在物体边缘处保持清晰,进而提升深度图的结构完整性与视觉质量,以实现增强双目视觉深度估计模型的泛化能力。为验证所提方法的有效性,本文在KITTI 2015和Middlebury等公开数据集上进行训练与评估,并利用SceneFlow数据集进行跨数据集泛化性能测试。实验结果表明,引入几何先验约束后,基线模型的性能得到稳定提升,在KITTI数据集上,端点误差(End-Point Error,EPE)降低了3%~5%,综合误匹配率(D1-all)降低了5%~8%。同时,在Middlebury数据集上的结果进一步证实了该方法在不同场景下的良好泛化性与鲁棒性。消融实验验证了各模块的贡献,超参数敏感性实验确定了损失函数权重的最优配置。此外,迁移实验表明,本文提出的几何先验约束机制具有良好的可移植性,能够适配于多种主流深度估计网络架构,并普遍带来性能增益。

关键词: 深度估计;立体匹配;先验知识;深度学习;几何约束;混合监督学习
基金项目: 国家自然科学基金(No.52401376);浙江省自然科学基金(No.LTGG24F030004);浙江省尖兵领雁科技项目(No.2025C04005);衢州市科技计划项目(No.2024K154)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0195-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250504

Binocular Vision Depth Estimation Method Based on Geometric Prior Knowledge Constraints

ZHANG Zehui[1], WANG Yang[1], CHEN Boyang[4], ZHANG Haoxuan[1], XU Xiaobin[1*], WU Fulong[1], CHENG Shenglong[2], SHAO Haibin[3], LI Hao[1]
(1. *Hangzhou Dianzi University*, *Hangzhou*, *Zhejiang* 310018, *China*;
2. *China Auto Information Technology* (*Tianjin*) *Co.*, *Ltd.*, *Tianjin* 300399, *China*;
3. *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*;
4. *Ningxia Petrochemical Yinjun Safety Technology Consulting Co.*, *Ltd.*, *Yinchuan*, *Ningxia* 750000, *China*)

Abstract: In recent years, with the rapid development of fields such as autonomous driving, robot navigation, and 3D reconstruction, depth estimation technology, as a key means of perceiving the three-dimensional structure of the environment, has garnered widespread attention. However, although the existing deep estimation methods based on supervised learning perform well on specific datasets, their generalization ability is weak and they rely on large-scale, high-quality labeled data, which severely limits their application in real industrial scenarios. Hence, this study proposes a binocular vision depth estimation method based on geometric prior knowledge constraints. First, this study combines residual convolution with the context encoder to extract multi-scale features from image data, and utilizes the feature pyramid structure to capture matching information at different scales for retaining the edge structure details of the image. Then, a multi-level gated recurrent unit (GRU) unit is designed to update the feature matching parameters in combination with feature information of dif-

收稿日期:2025-06-12;录用日期:2026-01-12;责任编辑:覃怀银
*通信作者:徐晓滨

ferent scales, optimize the disparity matching results, and achieve binocular vision depth estimation. Notably, this paper constructs a hybrid loss function that combines supervised signals with physical priors. Based on the traditional supervised loss, this function introduces geometric constraints derived from the self-supervised learning paradigm as regularization terms, specifically including the left-right disparity consistency constraint and the disparity structure consistency constraint. The left-right consistency constraint enforces geometric correspondence between the predicted disparities of the left and right views, enhancing the model geometric understanding and mitigating mismatches in occluded areas. The structural consistency constraint guides the disparity map to remain smooth in texture-flat regions and sharp at object edges, thereby improving the structural integrity and visual quality of the depth map, ultimately enhancing the generalization capability of the binocular vision depth estimation model. To verify the effectiveness of the proposed method, this paper conducts training and evaluation on public datasets such as KITTI 2015 and Middlebury, and uses the SceneFlow dataset for cross-dataset generalization performance. Experimental results show that after introducing geometric prior constraints, the baseline model's performance is consistently improved: on the KITTI dataset, the endpoint error (EPE) is reduced by 3% to 5%, and the overall mismatch rate (D1-all) is reduced by 5% to 8%. Simultaneously, results on the Middlebury dataset further confirm the method's good generalization and robustness across different scenarios. Ablation experiments verify the contributions of each module, while hyperparameter sensitivity experiments determine the optimal configuration for the loss function weights. Additionally, transfer experiments demonstrate that the proposed geometric prior constraint mechanism exhibits good portability, adapting to various mainstream depth estimation network architectures and generally providing performance gains.

Keywords: depth estimation; stereo matching; prior knowledge; deep learning; geometric constraints; hybrid supervised learning

Foundation Item(s): National Natural Science Foundation of China (No. 52401376); Zhejiang Provincial Natural Science Foundation (No.LTGG24F030004); "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No.2025C04005); Science and Technology Project of Quzhou (No.2024K154)

0 引言

深度估计是计算机视觉领域中的核心课题之一，其目标是从单目或双目图像中恢复场景的三维几何信息。随着深度学习技术的快速发展，基于数据驱动的深度估计方法取得了突破性进展，在KITTI、Middlebury等基准数据集上实现了亚像素级精度的深度预测，广泛应用于自动驾驶、机器人导航、三维重建等领域[1-4]。在众多方法中，基于监督学习的双目深度估计方法通过端到端的训练，能够从大量标注数据中学习复杂的特征表示和匹配规律，在特定数据集上往往能达到最佳性能[5-6]。

然而，基于数据驱动监督学习的方法存在其固有的局限性。首先，其性能严重依赖于大规模、高质量的视差标注数据，而获取密集且精确的真实视差标签成本极高，这在很大程度上限制了模型的推广应用。其次，在一个数据集上训练得到的模型，当迁移到其他具有不同域分布的数据集时，性能往往会出现显著下降，即泛化能力较弱[7-8]。尽管现有研究通过引入更复杂的网络架构[6,9]、注意力机制[10]或高效聚合模块[11-12]来提升模型在特定基准上的精度，但模型对标注数据的依赖及其泛化能力不足的问题仍未得到根本解决。

与此同时，自监督学习范式为解决上述问题提供了有价值的思路。该类方法不依赖于真实的视差标签，而是利用图像自身固有的几何和光度一致性作为监督信号。例如，Godard等人[13]在2017年的开创性工作中，提出了一套完整的损失函数框架，包括左右视图的光度重投影损失、左右视差一致性约束(left-right Consistency)和边缘感知的平滑度损失(edge-aware smoothness Loss)。这些基于物理先验的约束条件虽不能完全替代真实标签，但为模型提供了强大的正则化，有效提升了模型的泛化能力，并缓解了对标注数据的依赖。

受此启发，本文将这些经过验证的、源自自监督领域的几何先验知识，引入到深度估计模型训练框架中，以实现监督学习的高精度与自监督学习的强泛化。以性能优异的RaftStereo[14]为基线模型，对其结构进行改进，并构建了一种结合监督信号与物理先验的混合损失函数。本文的主要贡献如下：

(1)构建了以残差卷积与上下文编码器的协同作用为初始模块，以金字塔结构与多级门控循环单元为核心架构的深度估计神经网络模型，在估计精度与运算效率之间取得了良好平衡，完整实现了基于双目视觉的深度估计流程。

(2)提出了一种结合监督信号与物理先验的混合损失函数，本文将经典的自监督几何约束(左右视差一致性约束和视差结构一致性约束)[13]以正则化项的形式引入到监督式双目深度估计模型训练过程中，在不显著增加模型复杂度的前提下，有效降低了模型

对标注数据的过拟合风险,增强了其泛化能力。

(3)在 KITTI 2015 和 Middlebury 公开数据集以及 SceneFlow 数据集的迁移实验上验证了所提框架的有效性。实验结果表明,模型增强了对双目图像间细微差异的理解能力,深度估计精度有一定提升,相比于原模型,几何先验知识约束后其端点误差降低 3%~5%,综合误匹配率降低 5%~8%。

1　相关工作

1.1　深度估计

深度估计技术作为计算机视觉领域的重要分支,经历了从传统立体匹配到基于深度学习方法的显著演变。传统方法依赖于手工设计的特征(如 SIFT、Census)和基于几何原理的代价计算、聚合与优化流程(其一般步骤如图 1 所示)[15-18]。这类方法流程严谨,计算资源需求较低,在纹理丰富的场景中仍具有应用价值。然而,其性能高度依赖于相机参数标定和环境光照的稳定性,在遮挡、弱纹理或反射区域难以获得鲁棒且精确的结果。

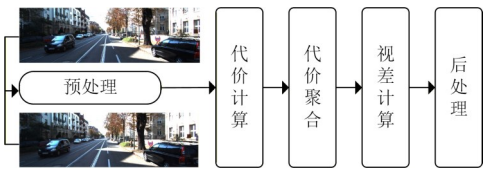

图 1　传统立体匹配深度估计方法一般步骤

Figure 1　General steps of traditional depth estimation method

为了提高深度估计的精度,基于数据驱动的深度估计方法得以提出。这种方法不再完全依赖于手工设计的几何约束和匹配规则,而是让神经网络从大量标注数据中学习如何有效地提取特征、匹配像素,并计算出准确的深度图,能够更灵活地处理各种复杂的场景和纹理,提高了对遮挡、弱纹理等挑战情况的鲁棒性。

全监督方法利用大量真实视差标签进行训练,在特定数据集上往往能达到最高精度。从早期的端到端视差回归[5]到多尺度特征融合[6],逐步解决了传统方法中特征匹配模糊和计算效率低的问题。为进一步提升精度,Zhang 等人[19]引导深度估计聚合网络通过可变形卷积,动态优化了匹配代价体的构建过程,提升了视差估计的匹配精度。Duggal 等人[20]结合神经网络以及可微分块匹配,减小了匹配代价聚合动态搜索的范围,显著降低了匹配过程的计算复杂度。随着注意力机制的普及,Xu 等人[9]引入稀疏点表示的尺度内代价聚合方法和神经网络层近似的跨尺度代价聚合算法,在 Scene Flow 和 KITTI 数据集上实现了快速且精确的深度估计。Xu 等人[21]利用相关性卷积生成注意力权重来过滤拼接卷积中的冗余信息,从而显著提升了立体匹配的准确性和效率。谢昭等人[22]提出一种针对上下采样过程的汇集网络模型,增强了数据之间的关联性,有效改善了复杂物体边界等深度估计的干扰情况。陈震等人[23]构建了由稀疏到稠密的大位移运动光流估计方法,在不同尺度上进行计算,获得了更准确的光流计算精度。面向实时性需求,Tankovich 等人[11]提出层级迭代优化策略,在保持高精度的同时实现每秒 50 帧的实时推理。而 Li 等人[12]通过级联循环网络自适应调整相关性搜索范围,兼顾了模型的计算效率与动态场景适应性,实现了高精度的深度估计。

自监督与无监督方法正是为了缓解上述问题而兴起。它们不依赖于真实的视差标签,而是利用图像自身固有的几何和光度一致性(photometric consistency)作为监督信号[24]。这类方法通过构建图像重投影误差、施加左右一致性约束和平滑性约束来训练网络,虽在绝对精度上可能不及全监督方法,但其摆脱了对标注数据的依赖,展现了更好的泛化潜力。本文工作旨在将全监督方法的高精度与自监督方法的强泛化能力相结合。这些方法共同推动了双目深度估计在自动驾驶、三维重建等领域的实用化进程。

1.2　自监督与几何约束在立体匹配中的应用

自监督学习的核心思想是利用数据本身内在的、无需人工标注的自我监督信号来驱动模型学习。在立体匹配任务中,这种信号主要来源于双目系统固有的几何约束(图 2)。

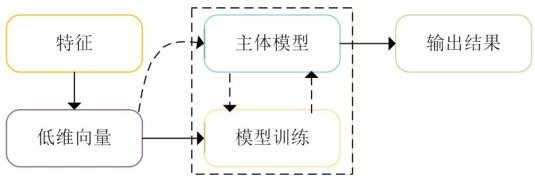

图 2　几何先验知识约束的一般流程

Figure 2　General process of geometric prior knowledge constraints

该类方法的奠基性工作之一是 Godard 等人[13]提出的框架。他们为单目深度估计设计了一套完整的自监督损失函数,其后被广泛借鉴并应用于双目任务中。该框架主要包含三个核心组件。

光度重投影损失:基于预测的视差图,将右视图扭曲至左视图,并计算扭曲后的图像与原始图像之间的光度误差。该损失迫使网络学习到能够保持外观一致性的视差。

左右视差一致性损失:鼓励左视图预测的视差图与右视图预测的视差图在几何上保持一致。该约束有助于处理遮挡问题,并提升预测视差的几何合理性。

边缘感知平滑损失:鼓励视差在纹理平坦区域保

持平滑，同时在图像梯度较大的区域（如物体边缘）允许视差不连续。这与本文使用的视差结构一致性损失目标一致。

这些基于几何先验的约束条件为模型提供了强大的正则化，有效缓解了对标注数据的过度依赖，增强了模型的泛化能力。自此，这套范式成为了自监督深度估计领域的标准实践，并被后续众多研究[25]所沿用和改进。

本文将自监督领域中的此类几何先验约束（左右一致性与结构平滑性）作为一种正则化手段，引入到以 RAFT-Stereo 为代表的先进全监督学习框架中。我们旨在构建一种混合监督范式，探究这种结合能否兼得监督学习的高精度与自监督正则化的强泛化能力，从而提升基础模型在多种场景下的鲁棒性与性能。

2 神经网络深度估计方法

2.1 工作流程

本文构建的深度估计神经网络模型启发于基于光流的神经网络 RaftStereo[14]，其网络结构如图 3 所示，具体工作流程如下。首先，深度估计模型的输入是一对经过校正的立体图像（左视图和右视图），残差卷积特征编码器提取左图和右图的特征用于生成密集的特征图。然后，左右图像的特征图用于构建三维相关体积，通过计算特征点积，生成相似性矩阵，表征像素的匹配代价，之后对相关体积在视差维度上逐级进行 1D 平均池化，形成包含 4 个层级的相关体积。与此同时，上下文编码器仅处理左图输出全局上下文特征。这些特征不仅用于初始化后续多级 GRU 的隐藏状态，还在每次迭代中与匹配特征融合，为优化过程注入场景语义信息。最后，网络通过多级 GRU（Gated Recurrent Unit）迭代更新逐步优化视差场。初始视差场为零，隐藏状态由上下文特征初始化。每次迭代包含相关特征检索、特征融合、多级 GRU 更新和视差增量预测四个关键步骤。经过多次迭代后，网络输出低分辨率视差场，并通过凸组合上采样恢复至原图分辨率，即得到最终输出的视差图。

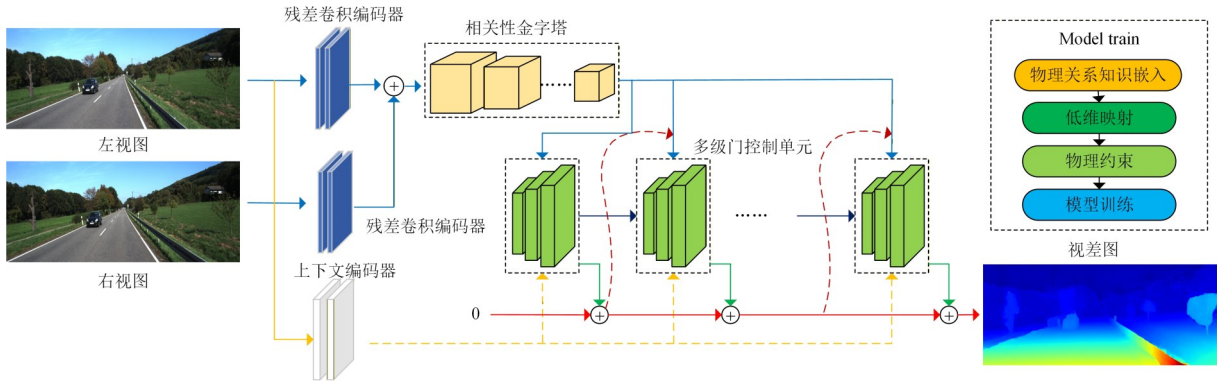

图 3 基于几何先验知识约束的双目视觉深度估计方法整体框架

Figure 3 Overall framework of stereo vision depth estimation method based on geometric prior knowledge constraints

2.2 模型结构

深度估计模型主要包括残差卷积特征编码器、上下文编码器、相关性金字塔以及多级门控制循环单元，各模块的具体结构以及输入输出的介绍如下。

2.2.1 残差卷积特征编码器与上下文编码器

残差卷积特征编码器的输入是一对经过校正的且尺寸为 $H\times W\times 3$ 的立体图像（左视图和右视图），模块的输出是左右图像的密集特征图。该模块结构如图 4 左侧所示，由 7×7 的卷积层、实例归一化层、ReLU 激活函数层、两个残差块和 1×1 的卷积层组成，其中 7×7 的卷积层、实例归一化层和 ReLU 激活函数层用于提取初步特征，然后这些特征通过连续两个残差块和 1×1 的卷积分别提取出左图的密集特征图 $f_l\in \mathbf{R}^{H/4\times W/4\times 256}$ 和右图的密集特征图 $f_r\in \mathbf{R}^{H/4\times W/4\times 256}$。

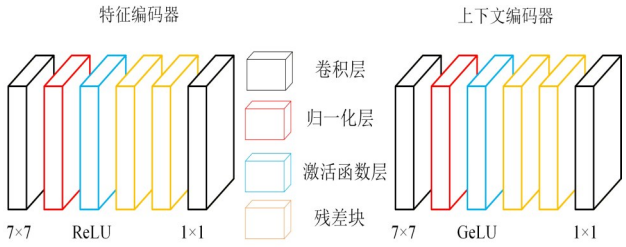

图 4 残差卷积特征编码器和上下文编码器

Figure 4 Residual convolutional feature rncoder and context encoder

上下文编码器的输入是经过校正的左视图，输出是左视图的上下文特征。上下文编码器的结构类似于残差卷积特征编码器，结构如图 4 右侧所示，由 7×7 的卷积层、批归一化层、GeLU 激活函数层、两个残差块和 1×1 的卷积层组成，最终的输出结果是左视图的上下文特征 $c\in \mathbf{R}^{H/4\times W/4\times 256}$。

2.2.2 相关性金字塔

相关性金字塔的输入是左右视图的密集特征图,输出是4层三维相关体积。该结构通过对同y坐标下的每个点进行点积运算计算左图特征与右图特征的点积相似性,获得三维相关体积$C\in\mathbf{R}^{H/4\times W/4\times D/4}$,计算公式如下:

$$C_{xyz}=\sum_{D} f_{l-xyD}\cdot f_{r-xzD} \tag{1}$$

其中,D表示特征向量的维度索引;x和y表示特征图中的行和列索引;z表示在计算相关性时考虑的右图中可能的列索引;f_l和f_r分别表示左右图的密集特征图。之后对三维相关体积的D维度进行核为2步长为2的平均池化,输出4层三维相关体积$C_0\in\mathbf{R}^{H/4\times W/4\times D/4}$、$C_1\in\mathbf{R}^{H/4\times W/4\times D/4}$、$C_2\in\mathbf{R}^{H/4\times W/4\times D/4}$和$C_3\in\mathbf{R}^{H/4\times W/4\times D/4}$,捕捉不同尺度的匹配信息的同时保留精细的结构细节。

2.2.3 多级门控制循环单元

多级门控制单元的输入是4层三维相关体积和左图的上下文特征,输出是最终的视差图。初始化视差场$d_0=0\in\mathbf{R}^{H/4\times W/4\times 1}$,且根据输入的左视图上下文特征初始化隐藏状态$h_0\in\mathbf{R}^{H/4\times W/4\times 64}$。输入数据中的4层三维相关体积,首先需要经过相关特征的检索,根据当前视差估计d_t,从相关金字塔的每一级采样特征,对每个层级,以当前视差值为中心,生成偏移网格,然后进行双线性插值,从每一层提取特征,拼接组合为多尺度相关特征$F_0\in\mathbf{R}^{H/4\times W/4\times 81}$,将检索到的多尺度特征、经过卷积的当前视差估计$d_t$以及上下文特征进行特征融合拼接,从而得到融合特征$F\in\mathbf{R}^{H/4\times W/4\times(81+64+256)}$,融合特征以及不同分辨率下的隐藏状态$h_0$被输入到多级GRU中,具体结构如图5所示。

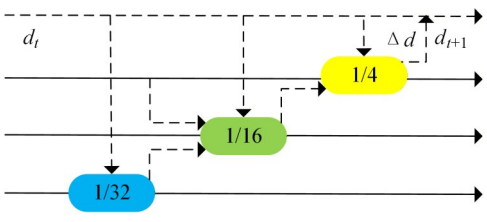

图5　多级门控制循环单元

Figure 5　Multi-level GRU

多级GRU的相邻分辨率视差之间交叉连接,它们之间的隐藏状态通过上下采样传递信息,网络利用上采样层将低分辨率视差图上采样到全分辨率,用于捕获全局场景结构,融合全局与局部信息以及预测视差增量,最后输出更新后的分辨率隐藏状态,其中最高分辨率的隐藏状态被用于预测视差增量Δd,根据$d_{t+1}=\Delta d+d_t$得到d_{t+1},一步步迭代得到最终的视差场,然后经过9×9的凸组合上采样至原图分辨率,获得最后视差图。

2.3 基于监督信号与物理先验的混合损失函数

为解决监督学习模型泛化能力弱、过度依赖标注数据的问题,本文将自监督立体匹配领域中经过验证的几何先验约束作为正则化项,引入到监督学习框架中,构建了一种结合监督信号与物理先验的混合损失函数。该函数受自监督学习工作的启发,旨在兼顾监督学习的高精度与自监督正则化的强泛化能力,通过利用左右视差一致性和视差结构一致性,将其转化为损失函数正则化项以融入模型训练过程中。

左右视差一致性与视差结构一致性是双目立体视觉中的核心几何约束,对提升深度估计的精度和鲁棒性具有重要作用。左右视差一致性指的是,在双目立体视觉系统中,以左视图为参考图像计算得到的视差图与以右视图为参考图像计算得到的视差图在几何关系上应满足一一对应性。这一约束本质上是基于双目视觉的极线几何原理,确保左右视图的深度估计在物理空间中的一致性。

左右视差一致性遵守如下公式:

$$L_{lr}=\frac{1}{N}\sum_{x}\left|d_{\text{left}}(x)-d_{\text{right}}\left(x-d_{\text{left}}(x)\right)\right| \tag{2}$$

其中,$d_{\text{left}}(x)$是左视图中点x的视差,对应的右视图点为$x-d_{\text{left}}(x)$,在训练的中间过程中通过预测视差$d_{\text{pred}}(x)$与$d_{\text{right}}(x)$来求取左右视差一致性,通过左右视差一致性检验,可以检测并修正因光照变化、纹理缺失或算法误差导致的错误视差值,提高深度估计的精度,并增强鲁棒性。

视差结构一致性则强调视差图的全局空间连续性。真实场景中的物体表面通常是平滑连续的,因此投影到图像平面上的视差值也应满足局部平滑性和全局结构性。该约束通过计算预测视差图与输入图像梯度之间的关系,鼓励视差在纹理平坦区域保持平滑,同时在图像梯度较大的区域允许视差不连续。

视差结构一致性的获取如下所示:

$$\begin{aligned}L_{\text{struct}}=\frac{1}{N}\sum_{x,y}\Big(&\left|\nabla_x d(x,y)+\nabla_x I(x,y)\right|\\&+\left|\nabla_y d(x,y)+\nabla_y I(x,y)\right|\Big)\end{aligned} \tag{3}$$

其中,$\nabla_x d$和$\nabla_y d$分别是视差的水平和垂直梯度;$\nabla_x I$和$\nabla_y I$分别是原图的水平和垂直梯度。通过引入视差平滑损失,网络在训练过程中会倾向于生成连续的视差图,避免局部突变,保留边缘信息,从而优化深度图质量。

左右视差一致性与视差结构一致性有助于输出

结果结构性优化，但无法学习深度，需要引入L1范数反映真实结果差异，其公式如下：

$$L_1 = 1/N \sum \left| d_{\text{pred}}(x) - d_{\text{left}}(x) \right| \tag{4}$$

其中，$d_{\text{left}}(x)$是左视图中点x的视差；$d_{\text{pred}}(x)$是点x的预测视差。

最终的损失函数为

$$L_t = \lambda_1 L_1 + \lambda_2 L_{lr} + \lambda_3 L_{\text{struct}} \tag{5}$$

在本研究中，通过将左右视差一致性与视差结构一致性约束作为约束项引入深度估计网络，我们利用物理先验约束了模型的优化方向。左右一致性损失有效提升了左右视图预测的一致性，减少了模型对特定场景数据的过拟合风险；而视差结构一致性损失则在遮挡边界处保持了视差图的平滑过渡，引导网络生成更符合物理规律的预测结果。两者共同作用，显著增强了模型在未知场景下的泛化性能。

2.4 模型训练

本节是基于几何先验知识约束的双目视觉深度估计方法的整体训练流程，具体如算法1所示。

算法1 深度估计模型训练

输入：训练数据集D，迭代次数T

输出：深度估计模型M

初始化深度估计模型
S←训练数据集D的大小
for $t=1$至T do
  //从训练数据集中读取数据
  for 批量数据$d=1$至S do
    I_l, I_r, d_{left}, d_{right}←批量数据
    d_{pred}←中间模型预测
    L_1←视差结果差异(式4)
    L_{lr}←视差一致性计算(式2)
    L_{struct}←结构一致性计算(式3)
    L_t←$L_1+L_{lr}+L_{\text{struct}}$
      模型训练
    if 训练步数为1 000的倍数 then
      模型保存
    end if
  end for
end for

3 实验与分析

3.1 实验环境及实现

实验环境为Windows 11，Python 3.10，PyTorch 2.2.0和CUDA 12.1.实验数据集为公开数据集KITTI-15、Middlebury和SceneFlow，其中KITTI-15包含灰度影像和彩色影像，平均大小为1 242×375像素，涵盖200对训练图像、200对测试图像，以及左右图像真实深度以及光流图像等。真实深度值由旋转激光扫描仪记录，点云密度约为影像像素的30%。在整体训练过程中，学习率设置为0.001，总训练轮次为300轮，合计30 000步。损失函数权重参数分别为主差异损失权重$\lambda_1=0.7$、左右一致性损失权重$\lambda_2=0.1$、视差结构一致性损失权重$\lambda_3=0.2$。模型使用AdamW优化器配合学习率调度器，以动态调整学习率。实验主要采用端点误差（EPE）以及综合误匹配率（D1-all）作为评价指标，其中EPE直接计算预测视差与真实视差的平均绝对误差，反映整体误差的绝对值大小，计算公式如下：

$$\text{EPE} = \frac{1}{N} \sum_x \left| d_{\text{pred}}(x) - d_{\text{gt}}(x) \right| \tag{6}$$

其中，$d_{\text{pred}}(x)$表示点x的预测视差；$d_{\text{gt}}(x)$表示点x的真实视差。

D1-all是KITTI官方评价指标，强调实际应用中对障碍物检测的可靠性，它统计所有像素中视差误差超过3像素精度或超过真实视差值5%的像素比例，计算公式如下：

$$\text{D1-all} = \frac{1}{N} \sum_x \delta\left(\left| d_{\text{pred}}(x) - d_{\text{gt}}(x) \right| > \max\left(3, 0.05 \cdot d_{\text{gt}}(x)\right) \right) \times 100\% \tag{7}$$

其中，δ为指示函数，条件满足时为1，不满足为0，当$\left| d_{\text{pred}}(x) - d_{\text{gt}}(x) \right| > 3$即绝对误差超过3像素或者$\left| d_{\text{pred}}(x) - d_{\text{gt}}(x) \right| > 0.05 \cdot d_{\text{gt}}(x)$即绝对误差超过5%时，判定为误匹配。

3.2 消融实验

消融实验意在探究不同模块对深度神经网络表征能力的影响机制，通过对比几何先验知识约束前后模型在KITTI数据集中的评估指标，对该方法在深度估计任务中的性能差异展开研究。

本实验在保持网络架构、超参数及训练策略完全一致的前提下，仅改变模型是否加入了部分结构。同时为了进一步评估几何先验知识约束对网络表征能力的影响，计算模型的1像素精度和5像素精度，通过不同阈值下的正确率，判断模型在精度与鲁棒性之间的平衡，1像素精度和5像素精度的计算公式如下：

$$\text{Acc}_{1\text{px}} = 1 - \frac{1}{N} \sum_x \delta\left(\left| d_{\text{pred}}(x) - d_{\text{gt}}(x) \right| \right) < 1 \tag{8}$$

$$\text{Acc}_{5\text{px}} = 1 - \frac{1}{N} \sum_x \delta\left(\left| d_{\text{pred}}(x) - d_{\text{gt}}(x) \right| \right) < 5 \tag{9}$$

3.2.1 金字塔模块消融实验

金字塔模块消融实验的结果如表1所示，表1中所有带有“(*)”符号的均表示没有使用金字塔模块的模型。

表1　金字塔模块消融实验

Table 1　Pyramid module ablation experiment

模型	EPE	D1-all/%	1px	5px
stereo^{24}(*)	1.278	5.162	0.732	0.913
stereo^{24}	0.647	2.303	0.871	0.983
stereo^{24}-GC(*)	1.163	4.312	0.791	0.921
stereo^{24}-GC	0.639	2.248	0.876	0.987

上述模型中stereo^{24}的24表示每一个轮次训练过程中多级门控制单元参数的迭代更新次数,带有GC后缀的模型表示模型进行了几何先验知识约束,表1中,没有使用金字塔模块的模型指标要低于使用了金字塔模块的模型,使用后,模型的EPE和D1-all均有明显改善,上述各个模型的预测示例结果如图6所示。

由于缺少了金字塔模块捕捉不同尺度的匹配信息,导致细节出现匹配错误情况,如stereo^{24}(*)模型的天空出现大量匹配错误,道路匹配也与其他模型存在较大差异,stereo^{24}-GC(*)模型的天空也出现了细节丢失以及错误匹配的情况。

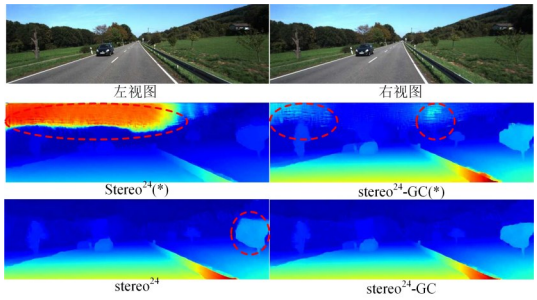

图6　金字塔模块消融实验模型预测结果

Figure 6　Prediction results of the pyramid model ablation experiment

3.2.2　GRU及几何先验知识约束消融实验

GRU模块消融实验旨在探究模块参数更新次数对预测结果的影响,几何先验知识约束消融实验则是为了验证本文方法的有效性,实验的具体结果如表2所示。

表2中stereo^{16}表示模型的多级GRU模块在单轮次训练过程中有16次更新,stereo^{24}类似,其中带有GC后缀的模型表示进行了几何先验知识约束训练之后的模型。实验结果显示,stereo^{16}的D1-all由原先的2.711%降至2.533%,EPE也从0.672降低至0.655,stereo^{24}的D1-all由原先的2.303%降至2.248%,EPE也从0.647降低至0.639,当去除多级GRU模块后,模型的精度断崖式下降。模型的1像素精度与5像素精度均在进行了几何先验知识约束后有略微提升,由此可知,几何先验知识约束对本文提出的网络模型的精度与鲁棒性均有一定提升。上述模型的预测结果示例如图7所示。

表2　GRU模块及几何先验知识约束消融实验

Table 2　Ablation experiment of GRU module and geometric prior knowledge constraints

模型	EPE	D1-all/%	1px	5px
stereo8	0.728	2.936	0.822	0.973
stereo16	0.672	2.711	0.826	0.976
stereo24	0.647	2.303	0.871	0.983
stereo8-GC	0.714	2.708	0.831	0.975
stereo16-GC	0.655	2.533	0.838	0.981
stereo24-GC	0.621	2.118	0.876	0.987

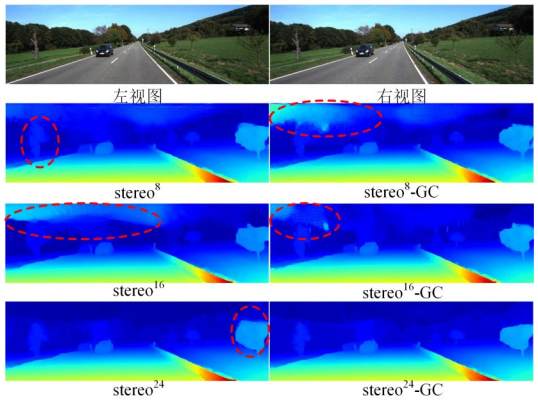

图7　GRU模块及几何先验知识约束消融实验模型预测结果

Figure 7　Prediction results of the ablation experiment model with GRU module and geometric prior knowledge constraints

从上述结果可以看出,经过了几何先验知识约束的模型输出结果边缘更加明显,验证了几何先验知识约束方法在本文提出的深度估计模型中的有效性,并且几何先验知识约束后物体内部变化更加平滑,能提升模型的泛化性和鲁棒性。

3.2.3　效率分析

效率分析是为了计算不同模型的运行速度,保证运行速度与计算量的平衡,同时根据模型效果获得最优选择,部分模型的效率如表3所示。

表3　效率分析表

Table 3　Efficiency analysis table

模型	参数量/M	Flops/T	FPS
Raft-stereo	11.117	3.156	18.5
stereo24(*)	11.224	1.987	23.7
stereo24	11.226	2.378	21.3
stereo24-GC	11.226	2.378	21.4

从表3可知,当去除金字塔模块后,模型的运行速度有略微提升,参数也有少量减少,几何先验知识约束前后模型的参数量没有变化,运行速度也相差无几。但结合模型精度考虑可知,几何先验知识约束后

的模型更优。

3.3 超参数敏感性实验

超参数敏感性实验的核心目标在于系统探究不同权重参数配置对深度神经网络表征能力的多维度影响。为确保实验的客观性与可比性，实验将基于统一的网络架构（stereo24-GC）和训练流程，使用同样的基线模型构建多组权重参数差异化模型，通过对比不同权重的模型在KITTI数据集中的评估指标，显式表示不同参数对模型性能的影响，实验的具体结果如表4所示。实验结果表明，从模型stereo24-GC在KITTI数据集上的超参数敏感性实验数据对比分析可知，当损失函数的三项超参数分别设置为0.7、0.1、0.2时，该模型的综合评估指标达到最优水平，0.7为主差异损失，0.1为左右一致性损失，0.2为视差结构一致性损失，三者加权平衡了特征拟合精度与模型鲁棒性。

表4 stereo^{24}-GC的超参数敏感性试验

Table 4 Hyperparameter sensitivity experiment of model stereo^{24}-GC

λ_1	λ_2	λ_3	EPE	D1-all/%
0.5	0	0.5	0.766	2.793
0.5	0.1	0.4	0.759	2.748
0.5	0.2	0.3	0.772	2.802
0.5	0.3	0.2	0.824	3.161
0.5	0.4	0.1	1.023	3.362
0.5	0.5	0	1.108	3.421
0.6	0	0.4	0.674	2.533
0.6	0.1	0.3	0.653	2.416
0.6	0.2	0.2	0.696	2.443
0.6	0.3	0.1	0.728	2.653
0.6	0.4	0	0.736	2.684
0.7	0	0.3	0.658	2.347
0.7	0.1	0.2	0.621	2.118
0.7	0.2	0.1	0.653	2.292
0.7	0.3	0	0.662	2.443

3.4 迁移实验

迁移实验旨在讨论本文几何先验知识约束方法的可移植性，通过多模型对比开展研究。

本实验选取了3种不同的深度估计神经网络，在保持原有网络核心拓扑结构的前提下，统一嵌入本文所设计的几何先验知识约束组件，3个深度估计神经网络分别是结合了金字塔池化模块以及多尺度上下文聚合模块获得深度的PSMnet[6]、通过分组相关代价聚合和3D卷积神经网络聚合特征获得深度的GWCnet[24]以及自适应聚合与注意力拼接体积增强特征相关性增强图像结构性计算深度的AAnet[9]. 通过对比嵌入前后各模型在KITTI数据集上的表现，量化分析几何先验知识约束方法对不同架构的适配特性以及影响。表5详细展示了几何先验知识约束前后各网络的指标。

表5 几何先验知识约束前后模型性能对比

Table 5 Comparison of model performance before and after geometric constraint regularization

模型	EPE	D1-all/%
PSMnet[6]	0.833	2.622
PSMnet-GC	0.794	2.568
GWCnet[24]	0.669	1.875
GWCnet-GC	0.658	1.787
AAnet[9]	0.714	2.452
AAnet-GC	0.708	2.433

实验结果表明，上述几种网络架构在进行了几何先验知识约束后，模型的EPE和D1-all指标均得到了不同程度的改善。这些改进过后得到优化的模型证明了本文提出的几何先验知识约束方法具有良好的可移植性和广泛的适用性，能够在一定程度上提升不同结构深度估计神经网络的整体性能。

3.5 对比实验及泛化性实验

对比实验和泛化性实验旨在讨论本文几何先验知识约束方法的有效性。本实验对近年来的多个模型分别在KITTI-15和Middlebury上进行对比。同时使用SceneFlow数据集训练模型，在Middlebury和KITTI-15数据集上进行泛化性实验，实验的结果如表6、表7和图8所示。从KITTI-15数据集以及Middlebury数据集的对比实验结果可知，几何约束模块有效地减少了模型在物体边界的误匹配，使得视差图整体更加平滑和精确，使得基线模型的整体误差有小幅度下降。然而在KITTI数据集对离群异常点更为敏感的D1-all指标上，本文方法与MoCha-Stereo相比仍存在一定差距。需明确的是，这一差距的核心并非本文方法的优化能力局限，而是两者在架构定位与设计目标上的本质取舍差异。本文方法基于结构简洁的Raft-Stereo基础架构，核心设计目标是在控制参数量、保障实时部署可行性的前提下，通过混合损失函数提升基础模型的泛化性与实用精度；而MoCha-Stereo作为

表6 KITTI-15数据集对比实验

Table 6 Comparison experiment of the KITTI-15

模型	EPE	D1-all/%
PSMnet[6]	0.833	2.622
GWCnet[24]	0.669	1.875
AAnet[9]	0.714	2.452
P3Snet[26]	1.073	4.320
MoCha-Stereo[27]	—	1.534
stereo24	0.647	2.303
stereo24-GC	0.639	2.248

表 7　Middlebury 数据集对比实验

Table 7　Comparison experiment of the Middlebury

模型	AvgErr	Bad(2px)
LEAstereo[28]	1.434	7.157
GMStereo[29]	1.310	7.140
NS-RAFT-Stereo[30]	—	9.670
Raftstereo[14]	1.273	4.744
stereo24-GC	1.212	5.184

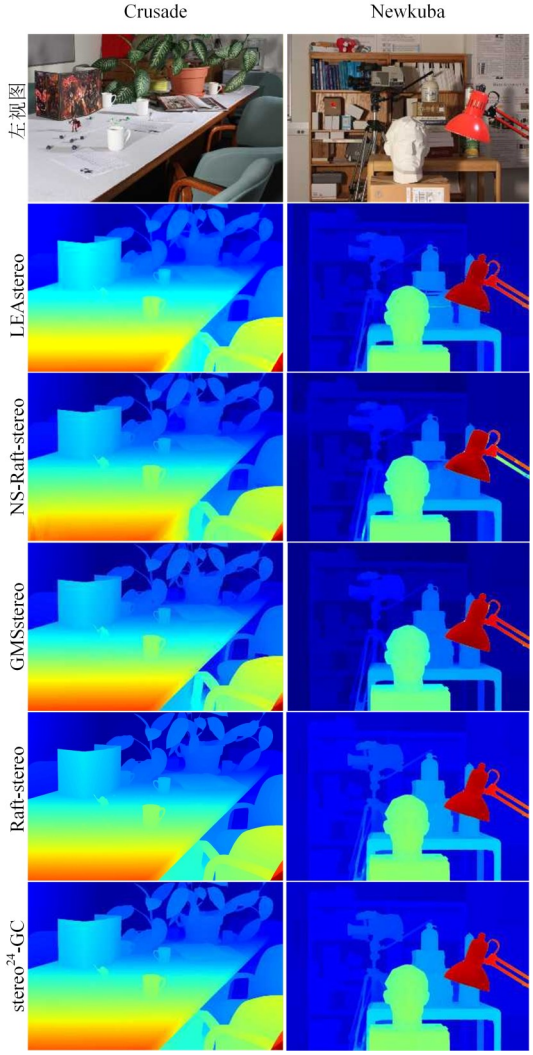

图 8　Middlebury 数据集模型预测结果

Figure 8　Model prediction results on the Middlebury dataset

后续迭代的模型,通过引入大量复杂优化模块、依赖大规模预训练数据与更长训练周期,以架构简洁性为代价,实现了对极致精度的专项突破。尽管纹理稀疏等困难区域仍存在少量离散误差,但本文方法在控制参数量、保障实时部署的前提下,已通过混合损失函数与正则化策略实现了满足工业场景等实用场景的匹配精度,契合核心设计目标。

表 8 和表 9 的模型均在 Sceneflow 数据集进行训练并分别在 Middlebury 和 KITTI-15 测试跨数据集的泛化性能,通过在模型中引入几何先验知识约束,stereo24-GC 在 Middlebury 数据集上的泛化性能不仅显著超越了基线模型,甚至优于 Raft-Stereo,证明了该模块对模型的泛化性有促进作用。

表 8　跨数据集(Middlebury)泛化性实验

Table 8　Middlebury generalization experiment

模型	Bad(2px)
Raft-stereo[14]	12.59
stereo24	13.74
stereo24-GC	12.43

表 9　跨数据集(KITTI-15)泛化性实验

Table 9　KITTI-15 generalization experiment

模型	EPE	D1-all/%
Raft-stereo[14]	1.174	5.687
stereo24	1.198	5.803
stereo24-GC	1.123	5.378

4　结束语

本文在基于神经网络的深度估计方法基础上,通过深度神经网络中的几何先验知识约束技术,显著提升了深度估计在场景中的精度和鲁棒性。聚焦于左右视差一致性和视差结构一致性这两个关键要素,本文将相关的专业知识和规则以几何约束的形式嵌入到深度估计神经网络中。实验结果表明,几何先验知识约束不仅增强了模型对双目图像间细微差异的理解能力,还提升了在复杂场景中准确捕捉和解析视差信息的能力,具体表现为端点误差降低 3%~5%,综合误匹配率降低 5%~8%。此外,几何先验知识约束还增强了模型的泛化能力,在不同架构的深度估计神经网络中均表现出良好的适配性。

未来工作将进一步探索几何先验知识约束技术在其他深度学习任务中的应用,以拓宽其适用范围。同时,针对现实场景中可能出现的更多挑战,计划研发更加先进的知识表示和约束方法,以进一步提升深度估计模型的性能和稳定性。

参考文献

[1] Wofk D, Ma Fangchang, Yang T J, et al. FastDepth: Fast monocular depth estimation on embedded systems[C]// 2019 International Conference on Robotics and Automation. Piscataway: IEEE, 2019: 6101-6108.

[2] He Qingdong, Wang Zhengning, Zeng Hao, et al. Stereo RGB and deeper LIDAR-based network for 3D object detection in autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(1): 152-162.

[3] 曲熠, 陈莹. 基于尺度线索增强的无监督单目深度估计[J]. 电子学报, 2024, 52(9): 3217-3227.
Qu Yi, Chen Ying. Unsupervised monocular depth estimation based on scale clue enhancement[J]. Acta Electronica Sinica, 2024, 52(9): 3217-3227. (in Chinese)

[4] 周晓清, 王翔, 郑锦, 等. 基于自适应空间稀疏化的高效多视图立体匹配[J]. 电子学报, 2023, 51(11): 3079-3091.
Zhou Xiaoqing, Wang Xiang, Zheng Jin, et al. Adaptive spatial sparsification for efficient multi-view stereo matching[J]. Acta Electronica Sinica, 2023, 51(11): 3079-3091. (in Chinese)

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

[6] Chang Jiaren, Chen Yongsheng. Pyramid stereo matching network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5410-5418.

[7] Laga H, Jospin L V, Boussaid F, et al. A survey on deep learning techniques for stereo-based depth estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1738-1764.

[8] 张振宇, 杨健. 基于元学习的双目深度估计在线适应算法[J]. 自动化学报, 2023, 49(7): 1446-1455.
Zhang Zhenyu, Yang Jian. Online adaptation through meta-learning for stereo depth estimation[J]. Acta Automatica Sinica, 2023, 49(7): 1446-1455. (in Chinese)

[9] Xu Haofei, Zhang Juyong. AANet: Adaptive aggregation network for efficient stereo matching[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1956-1965.

[10] Yang Guanglei, Rota P, Alameda-Pineda X, et al. Variational structured attention networks for deep visual representation learning[J]. IEEE Transactions on Image Processing, 2024: 3137647.

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

[12] Li Jiankun, Wang Peisen, Xiong Pengfei, et al. Practical stereo matching via cascaded recurrent network with adaptive correlation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 16242-16251.

[13] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6602-6611.

[14] Lipson L, Teed Z, Deng Jia. RAFT-stereo: Multilevel recurrent field transforms for stereo matching[C]//2021 International Conference on 3D Vision. Piscataway: IEEE, 2021: 218-227.

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

[16] Ma Tao, Zhu Hangbiao, Huang Weijian, et al. Stereo image dense matching based on SGM constrained by feature matching[C]//2023 9th International Conference on Computer and Communications. Piscataway: IEEE, 2023: 1911-1915.

[17] 王笛, 胡辽林. 基于双目视觉的改进特征立体匹配方法[J]. 电子学报, 2022, 50(1): 157-166.
Wang Di, Hu Liaolin. Improved feature stereo matching method based on binocular vision[J]. Acta Electronica Sinica, 2022, 50(1): 157-166. (in Chinese)

[18] 狄红卫, 柴颖, 李逵. 一种快速双目视觉立体匹配算法[J]. 光学学报, 2009, 29(8): 2180-2184.
Di Hongwei, Chai Ying, Li Kui. A fast binocular vision stereo matching algorithm[J]. Acta Optica Sinica, 2009, 29(8): 2180-2184. (in Chinese)

[19] Zhang Feihu, Prisacariu V, Yang Ruigang, et al. GA-net: Guided aggregation net for end-to-end stereo matching[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 185-194.

[20] Duggal S, Wang Shenlong, Ma W C, et al. DeepPruner: Learning efficient stereo matching via differentiable PatchMatch[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 4383-4392.

[21] Xu Gangwei Cheng Junda, Guo Peng, et al. Attention concatenation volume for accurate and efficient stereo matching[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 12971-12980.

[22] 谢昭, 马海龙, 吴克伟, 等. 基于采样汇集网络的场景深度估计[J]. 自动化学报, 2020, 46(3): 600-612.
Xie Zhao, Ma Hailong, Wu Kewei, et al. Sampling aggregate network for scene depth estimation[J]. Acta Automatica Sinica, 2020, 46(3): 600-612. (in Chinese)

[23] 陈震, 张道文, 张聪炫, 等. 基于深度匹配的由稀疏到稠

密大位移运动光流估计[J]. 自动化学报, 2022, 48(9): 2316-2326.
Chen Zhen, Zhang Daowen, Zhang Congxuan, et al. Sparse-to-dense large displacement motion optical flow estimation based on deep matching[J]. Acta Automatica Sinica, 2022, 48(9): 2316-2326. (in Chinese)

[24] Guo Xiaoyang, Yang Kai, Yang Wukui, et al. Group-wise correlation stereo network[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3268-3277.

[25] Yang Delong, Luo Zhaohui, Shang Peng, et al. Unsupervised deep learning of depth, ego-motion, and optical flow from stereo images[C]//2021 9th International Conference on Traffic and Logistic Engineering. Piscataway: IEEE, 2021: 51-56.

[26] Emlek A, Peker M. P3SNet: Parallel pyramid pooling stereo network[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(10): 10433-10444.

[27] Chen Ziyang, Long Wei, Yao He, et al. MoCha-stereo: Motif channel attention network for stereo matching[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 27768-27777.

[28] Cheng Xuelian, Zhong Yiran, Harandi M, et al. Hierarchical neural architecture search for deep stereo matching[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 22158-22169.

[29] Xu H F, Zhang J, Cai J F, et al. Unifying flow, stereo and depth estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13941-13958.

[30] Tosi F, Tonioni A, De Gregorio D, et al. NeRF-supervised deep stereo[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 855-866.

作者简介

张泽辉　男,1994年9月生,湖南衡阳人。杭州电子科技大学副研究员、硕士生导师、浙江省科协青年托举人才。主要研究方向为计算机视觉、故障诊断以及人工智能方法在工业领域的应用。
E-mail: zhangtianxia918@163.com

王　阳　男,2001年5月生,湖北荆州人。现为杭州电子科技大学硕士研究生。主要研究方向为机器学习、深度学习、计算机视觉。
E-mail: 232060349@hdu.edu.cn

徐晓滨　男,1980年3月生,河南郑州人。杭州电子科技大学博士生导师,浙江省杰出青年基金获得者。主要研究方向为故障诊断与智能运维。
E-mail: xuxiaobin1980@hdu.edu.cn