黄锦娜, 程阳, 朱家兵. 基于组合代价体的多级迭代立体匹配方法[J]. 智能计算机与应用,2025,15(9):33-40. DOI:10.20169/j.issn.2095-2163.25051202

基于组合代价体的多级迭代立体匹配方法

黄锦娜[1], 程 阳[1], 朱家兵[2,3]

(1 安徽理工大学 计算机科学与工程学院, 安徽 淮南 232001; 2 安徽省北斗卫星导航技术重点实验室, 安徽 淮南 232038; 3 淮南师范学院 电子工程学院, 安徽 淮南 232038)

摘 要: 立体匹配是计算机视觉中的一项重要技术,通过从 2 个或多个视角拍摄的场景图像中恢复深度信息,广泛应用于自动驾驶、机器人和 3D 建模等诸多场景中。针对目前立体匹配方法采用 3D 卷积处理会带来大量的计算和内存成本,以及图像中遮挡和无纹理区域表现不佳的问题,提出一种基于组合代价体的多级迭代端到端立体匹配方法。通过引入组合代价体,增加代价体中的特征信息,在减少迭代次数的同时,不影响预测视差图的效果。并在上下文网络中引入多级池化和上下文注意力机制,进一步提升网络性能。经实验可知,该方法在 Scene Flow、KITTI2012/2015、Middlebury2014 和 ETH3D 数据集上相较于基准模型都取得了良好结果。

关键词: 立体匹配; 组合代价体; 多级池化模块; 上下文注意力机制; 多级迭代

中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0033-08

Multi-level iterative stereo matching method based on combined cost volume

HUANG Jinna[1], CHENG Yang[1], ZHU Jiabing[2,3]

(1 School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China; 2 Key Laboratory of Anhui Provincial Beidou Satellite Navigation Technology, Huainan 232038, Anhui, China; 3 School of Electronic Engineering, Huainan Normal University, Huainan 232038, Anhui, China)

Abstract: Stereo matching is an important technology in computer vision. The technique recovers depth information from scene images taken from two or more perspectives and is widely used in many scenarios such as autonomous driving, robotics, and 3D modeling. In view of the fact that the current stereo matching methods bring a lot of computational and memory costs while using 3D convolution processing, as well as poor performance in occluded and textureless areas in the image, a multi-level iterative end-to-end stereo matching method based on combined cost volume is proposed. By introducing a combined cost volume, the feature information in the cost volume is increased, while the number of iterations is reduced without affecting the effect of the predicted disparity map. Multi-level pooling and contextual attention mechanisms are introduced in the context network to further improve the network performance. Experiments show that this method has achieved good results compared with the baseline model on Scene Flow, KITTI2012/2015, Middlebury2014 and ETH3D datasets.

Key words: stereo matching; mixed cost volume; multi-level pooling module; contextual attention mechanism; multi-level iteration

0 引 言

立体匹配是计算机视觉中常用的一种估计深度信息的方法,因此也可称作双目深度估计。其主要模拟人的双眼视觉系统,通过匹配从不同视角拍摄的同一场景的左右视图,来估计图像中像素点的对应关系,由此得到视差值和深度信息[1]。目前,立体匹配方法应用于智能驾驶中的环境感知[2]、机器人导航避障[3]、工业检测与测量[4]和三维重建[5]等多个领域。因此,立体匹配方法也是最近的热点研究方向之一。

基于深度学习的立体匹配是目前的主流方法,

基金项目: 国家自然科学基金(62172183); 安徽省科技重大专项(202003a05020031); 安徽省重点研发计划(202004a05020023); 安徽省科技创新平台重大科技项目(S202305a12020025)。

作者简介: 黄锦娜(1998—),女,硕士研究生,主要研究方向:计算机视觉; 程 阳(2000—),男,硕士研究生,主要研究方向:医学图像分割。

通信作者: 朱家兵(1970—),男,教授,硕士生导师,主要研究方向:雷达信号处理。Email:zjb3617@163.com。

收稿日期: 2025-05-12

可分为非端到端和端到端两个类别。非端到端方法[6-8]通过将深度神经网络代替传统匹配方法中的某一个或者多个步骤进行优化,故该方法还保留着传统方法的大致框架,仍然需要手工设计的正则化函数或视差后处理步骤,因此意味着其存在计算量大和效率低的缺点。

端到端立体匹配网络,根据构建代价体的操作不同,可以将代价体分为相关体和连接体[9]。Mayer 等学者[10]提出的 Disp-Net,采用从左右视图提取的特征图通过相关性(Correlation)操作生成 3D 相关性代价体,然后运用一系列卷积进行代价聚合,最后回归出视差图。Kendall 等学者[11]提出新型深度学习框架 GC-Net,其创造性地引入了 4D 连接代价体的构造方式,通过将特征图在每个差异水平上进行拼接(Concat),能够得到含有更多图像几何信息和上下文信息的代价体。Chang 等学者[12]为了提高上下文信息的提取和利用,提出了 PSMNet。主要采用空间金字塔池化模块提取多尺度的上下文特征信息,随后将左右特征图进行通道维度上的连接生成连接代价体,最后利用堆叠沙漏结构处理成本体积,进一步学习更多的上下文信息。Guo 等学者[13]指出相关卷只能合成单通道特征,因此丢失过多信息,而连接卷又缺乏相关性度量,故提出 GWC-Net,通过将分组连接体和相关体进行结合,生成精确度更高的视差图。Xu 等学者[14]提出的 ACV-Net,采用多层次自适应块匹配方法计算图像像素之间的相关性,通过使用相关代价体中编码的相似性信息来规范连接体积,这样只需要一个轻量级的聚合网络就可以实现整体的高效率和准确性。

上述几个端到端立体匹配网络基本都采用多层 3D 卷积进行代价聚合,这导致网络在资源存储和计算时间上的成本增加。Khamis 等学者[15]提出实时立体匹配网络 StereoNet,其主要思想是通过多尺度的特征提取来得到低分辨率的特征图,并构建小尺寸的代价体来减少后续处理的计算量。Lipson 等学者[16]提出采用多级卷积 GRU 处理代价体,仅用 2D 卷积对视差图进行优化,大大减轻了 3D 卷积带来的计算量。同时创新性地引入全对相关体,捕捉整个图像之间的相关性,使得网络能够更好地识别低纹理和遮挡区域的深度信息。但由于代价体缺少多通道特征,因此在遮挡和纹理稀疏区域表现不佳。

综合上述可知,相关性代价体和连接代价体有着各自的优点,将其进行有效的结合,能够生成高效组合代价体。此外,利用多级卷积 GRU 结构进行视差图的优化迭代,可以减少采用 3D 卷积代价聚合步骤而带来的计算量。由此,本文以 RAFT-Stereo 为基准模型,提出基于组合代价体的多级迭代立体匹配网络,将注意力连接代价体与全对相关代价体进行组合,由此生成的组合代价体含有相关性信息和多通道的特征信息,这些信息也将有助于迭代次数的减少。此外,为了解决 RAFT-Stereo 在遮挡和无纹理区域表现不佳的问题,引入多级池化模块和上下文注意力机制,通过增强全局和局部上下文信息的捕捉来进一步提升网络估计视差的精度。

1 基于组合代价体的多级迭代立体匹配方法

图 1 是本文提出的总体网络框架 CCV-Stereo,以左右视图为输入,首先通过特征网络提取出多尺度特征,采用 1/4 分辨率的特征图,经过多级自适应块匹配方法 MAPM(Multi-level Adaptive Patch Matching)得出多级相关性匹配代价体 MCMV(Multi-level Correlation Matching Cost Volumn),随后通过轻量级的正则化网络得出注意力权重和初始视差图。然后,用注意力权重去过滤 1/4 分辨率特征图构成的初始连接代价体 ICV(Initial Concat Cost Volumn),除去其中的冗余信息,并使得相关性信息能够在注意力代价体 ACV(Attention Cost Volumn)中得到体现。最后,将全对相关量 APV(All-Paris Correlation Volumn)与 ACV 进行结合,在得到组合代价体 CCV(Combined Cost Volumn)的同时,让网络进一步获取全局相关性信息。通过生成高质量组合代价体,并将经过查询得出的特征信息注入多级卷积 GRU 模块,使得网络在减少迭代次数的情况下也能够估计出高精度的视差图。

此外,利用左视图进行上下文信息提取,并为了更全面地捕捉局部特征,引入多级池化模块 MPM(Multi-level Pooling Module),通过多个最大池化层能够获取多尺度上下文信息之外,能够更好地融合全局和局部信息,增强网络对上下文信息的感知。然后通过上下文注意力机制 CTAM(Contextual Attention Mechanism),对通道和空间两个维度进行特征加权,增强表达能力,聚焦重要信息,提升总体性能。

1.1 特征提取编码器

该部分主要包括 2 个部分:

(1)特征网络。给定左右图像 $I_{l(r)} \in R^{3\times H\times W}$,先用下采样,将图像维度压缩到 1/32,再用上采样块将其还原到 1/8 和 1/4 尺度,并最终输出 $f_{l,4}$ 和 $f_{r,4}$ 特征图。

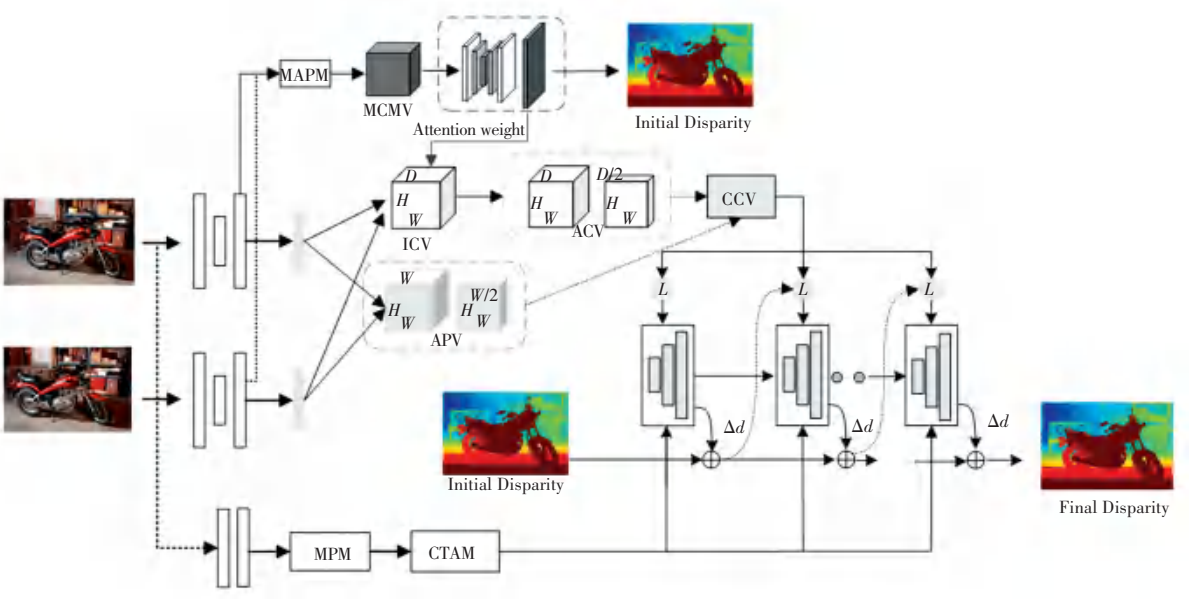

图 1　总体网络框架

Fig. 1　Overall network framework

（2）上下文网络。仅用左视图来提取全局上下文信息，并用一系列残差块和降采样层来得到 1/4 分辨率的上下文特征图，用作后续 MPM 和 CTAM 的输入。利用最后输出的增强上下信息特征初始化 ConvGRU 的更新运算器的隐藏状态，并在每次迭代更新时插入 ConvGRU。

1.2　组合代价体

首先采用分辨率为 1/4 的特征图，让其通过 2 层卷积将通道数压缩到 32，生成 f_l 和 f_r，用于合成初始连接代价体。连接代价体的组成公式如下：

$$C_{\text{concat}} = \text{Concat}\{f_l(x,y), f_r(x-d,y)\} \quad (1)$$

最后得到 C_{concat} 的大小是 $2N_f \times D/4 \times H/4 \times W/4 (N_f = 32)$。

随后，从 I_l 和 I_r 提取左侧特征 $f_{l,4}$ 和右侧特征 $f_{r,4}$，通道数为 N_C，利用多级自适应块匹配方法构建多级匹配相关卷。首先采用 GWC-Net 的分组思想，将特征 $f_{l,4(r,4)}$ 沿通道维度分为 $N_g (N_g = 32)$ 组，则每组通道数为 N_C/N_g。然后，将这 32 组分为 3 个层级 (l_1, l_2, l_3)，其中 l_1 为 8 组，l_2 为 8 组，l_3 为 16 组。对于不同层次中的每个像素，本文利用不同膨胀率的卷积自适应地学习权重，其中用较小膨胀率的卷积去学习特征图中的边缘和细节信息，反之用于涵盖更多的上下文信息。将第 g 个特征组表示为 f_l^g, f_r^g。研究推得的公式如下：

$$C_{\text{patch}}^{l_k}(g,d,x,y) = \frac{1}{N_C/N_g} \sum_{g(i,j)\in\Omega^k} \omega_{ij}^{k,g} \cdot C_{ij}^g(d,x,y) \quad (2)$$

$$C_{ij}^g(d,x,y) = \langle f_l^g(x-i,y-j),\ f_r^g(x-i-d,y-j) \rangle \quad (3)$$

其中，C_{ij}^g 为第 g 特征组的相似性度量；$\langle -, - \rangle$ 表示内积；(x,y) 表示像素的位置；d 表示不同视差等级；$\omega_{ij}^{k,g}$ 表示 k 层特征图上自适应学习的特征块中像素 (i,j) 的权重，对相似度进行加权；$C_{\text{patch}}^{l_k} (k = 1, 2, 3)$ 表示不同层级的匹配成本。然后将所有级别的匹配成本 $C_{\text{patch}}^{l_k}$ 连接起来，得到最终的多级相似性匹配代价体，公式如下：

$$C_{\text{patch}} = \text{Concat}\{C_{\text{patch}}^{l_1}, C_{\text{patch}}^{l_2}, C_{\text{patch}}^{l_3}\} \quad (4)$$

得到的多级相似性匹配卷可表示为 $C_{\text{patch}} \in R^{N_g \times D/4 \times H/4 \times W/4}$，然后应用一个轻量级正则化网络对 C_{patch} 进行处理。该网络主要由 3 个下采样块和 3 个上采样块组成，每个下采样块由 2 个 3×3×3 的卷积组成，上采样块采用了 3×3×3 的转置卷积。最后使用另一个卷积层将通道压缩为 1 得出注意力权重，即 $A \in R^{1 \times D/4 \times H/4 \times W/4}$。

得到注意力权重 A 后，用其来过滤初始连接代价体中的冗余信息，进而增强其表示能力。注意力连接代价体 C_{ACV} 的计算公式为：

$$C_{\text{ACV}} = A \odot C_{\text{concat}} \quad (5)$$

其中，“⊙”表示元素乘积，而权重 A 适用于初始连接体的所有通道。

为了使注入 GRU 的特征更完善，采用 RAFT-Stereo 的方法生成全对相关代价体，可以弥补注意力代价体中所缺少的局部相关性特征。为了增加感

受野,本文使用平均池化的方法对视差维度进行池化,生成注意力连接金字塔和全对相关金字塔,将其进行结合,最后生成组合代价体。

1.3 多级池化模块和上下文注意力机制

传统的池化模块通常都是采用固定的感受野对特征进行聚合,不利于捕捉多尺度上下文信息。而本文多级池化模块主要采用 4 层不同尺度的最大池化操作(32×32,16×16,8×8,4×4),多级池化模块如图 2 所示。首先,采用大尺度的池化层获取更大范围的上下文信息,随后通过由粗到细的方法用较小尺度的池化层弥补缺少的细节信息。最后,通过双线性插值法进行上采样,将所有信息进行有效结合,有助于网络能够在不同的尺度上聚焦更全面的上下文信息和细节,并且提高对不同深度信息的感知能力。

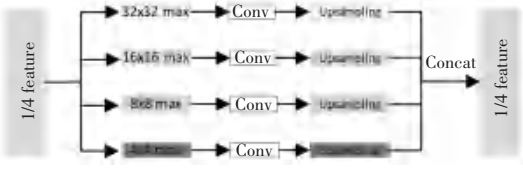

图 2 多级池化模块

Fig. 2 Multi-level pooling module

此外,再引入上下文注意力机制,如图 3 所示,其主要由通道注意力模块和空间注意力模块组成。首先,给定特征图 $F \in R^{C\times H\times W}$,通道注意力模块利用平均池化和最大池化操作对每个通道的空间信息进行全局统计,提取每个通道的关键特征信息得到 $F_{\text{Avg}}, F_{\text{Max}} \in R^{C\times 1\times 1}$。然后,用 2 个连续的 1×1 卷积层对池化结果通过压缩和恢复进行特征学习。随后将 2 个结果进行融合,得到通道注意映射 $M_C(F)$。公式如下:

$$M_C(F) = \text{Conv}(\text{AvgPool}(F)) + \text{Conv}(\text{MaxPool}(F)) \tag{6}$$

空间注意力模块首先采用一个 1×1 卷积对通道数进行压缩,然后使用 2 个 3×3 膨胀卷积捕捉更大范围内的上下文信息,最后再用一个 1×1 卷积得到空间注意映射 $M_S(F)$。公式如下:

$$M_S(F) = f_3^{1\times1}(f_2^{3\times3}(f_1^{3\times3}(f_0^{1\times1}(F)))) \tag{7}$$

在得到 2 个模块的注意力映射后,将其相加并通过 Sigmoid 激活函数得到值为 0~1 之间的上下文注意力权重,将其与输入特征图 F 进行元素相乘,并把相乘结果添加到原始特征图 F 上,最终得到更加精细的特征图 F'。公式如下:

$$M(F) = \sigma(M_C(F) + M_S(F)) \tag{8}$$

$$F' = F + F \otimes M(F) \tag{9}$$

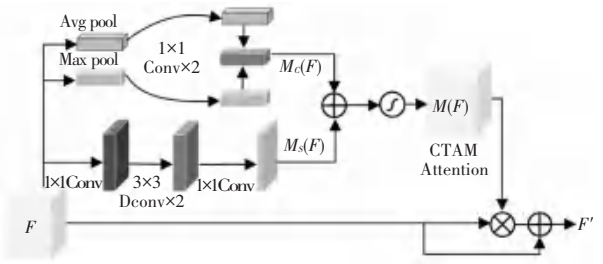

图 3 上下文注意力机制

Fig. 3 Contextual attention mechanism

1.4 损失函数

首先,计算初始视差 d_{init} 的平滑损失 L_1,定义如下:

$$L_{\text{init}} = \text{Smooth}_{L_1}(d_{\text{init}} - d_{\text{gt}}) \tag{10}$$

其中,d_{gt} 表示真实视差值。

随后,计算所有预测视差 $d_i(i = 1,2,\cdots,N)$ 的 L_1 损失。按照 RAFT-Stereo 的方法以指数形式增加权重,总损失定义为:

$$L = L_{\text{init}} + \sum_{i=1}^{N} \gamma^{N-i} \| d_i - d_{\text{gt}} \|_1 \tag{11}$$

其中,N 表示实验迭代总次数;$\gamma = 0.91$;d_{gt} 表示真实视差值。

2 实验

2.1 实验数据集和评价指标

本文实验主要采用的数据集有 Scene Flow[10]、KITTI[17-18]、Middlebury2014[19] 和 ETH3D[20],在这 4 个数据集上对模型进行训练和性能评估。关于各类数据集的具体情况见表 1。

表 1 各类数据集规模大小和应用场景

Table 1 The size and application scenarios of various data sets

Dataset	Train/Test	应用场景
Scene Flow	35 454/4 370	立体匹配和光流估计等视觉任务
KITTI2012	194/195	道路驾驶场景
KITTI2015	200/200	道路驾驶场景
Middlebury2014	15/15	室内纹理较多的场景
ETH3D	27/20	缺乏颜色信息的灰度场景

实验首先在 Scene Flow 数据集上对本文模型进行预训练,重点评估指标有端点误差(EPE)和像素误差百分比($D1$)。随后,在 KITTI 数据集上对得到的预训练模型进行调试,对 KITTI2012 数据集中图像的全部区域(all)和非遮挡区域(noc)进行 EPE、2 像素和 3 像素误差指标的评估,对 KITTI2015 数据集中图像的背景区域(bg)、前景区

域(fg)和全部区域(all)进行($D1$)指标的评估。最后，在 Middlebury2014 和 ETH3D 数据集上进行迁移学习，对 Middlebury2014 主要采用半像素误差率(Half)，0.25 像素误差率(Quarter)进行评估，而 ETH3D 则采用 1 像素误差率。

2.2　实验环境和参数设置

本文模型采用 PyTorch1.11 版本的深度学习框架进行环境搭建，Python 版本是 3.8。采用 2 个 NVIDIA RTX 4090 GPU 进行训练。在所有实验中，本文使用 AdamW 优化器，并在[-1,1]范围内裁剪图像。然后，在 Scene Flow 数据集上将图像随机裁剪为 320×736，并进行 20 k 步的预训练，其中批次处理大小为 8，学习率为 0.000 2，更新迭代次数为 22。预训练完成后，在 KITTI 数据集上对前面得到的 Scene Flow 模型进行 50 k 步的微调，批次处理大小为同样为 8。最后，直接在 ETH3D 和 Middlebury 数据集上进行测试，得到的结果用于验证模型的泛化性能。

2.3　消融实验

本文在 Scene Flow 数据集上进行消融实验，每次实验都采用 32 次的 ConvGRUs 更新迭代轮数，其他实验参数与上述相同。

在 Scene Flow 数据集消融实验结果见表 2。参见表 2 第 2 行，本文提出的组合代价体 CCV 在加入基准模型后，EPE 和 $D1$ 指标都有了明显的降低趋势，其中 EPE 减少了 0.06 像素误差，$D1$ 则降低了 8.4%。最后，测试了整体模型在 Scene Flow 数据集上的性能表现，由表 2 的最后一行数据可知，本文提出的方法较基准模型在 EPE 和 $D1$ 指标上有大幅度的提升，其中 EPE 优化了 16.1%，$D1$ 也降低了 10.9%，验证了本文提出方法的有效性。

表 2　在 Scene Flow 数据集消融实验结果

Table 2　Ablation experiment results in the Scene Flow dataset

Model	EPE/px	>3 px/%	Params/M	Time/s
RAFT-Stereo	0.56	2.85	12.0	0.37
RAFT-Stereo+CCV	0.50	2.61	12.6	0.38
RAFT-Stereo+MPM+CTAM	0.54	2.77	12.4	0.38
RAFT-Stereo+CCV+MPM+CTAM(本文)	0.47	2.54	12.8	0.38

此外，为进一步验证即使减少迭代次数，本文提出的方法也能展现出色的性能效果。减少迭代次数进行预测的结果见表 3。当迭代次数减少到 1、2、4 或 8 次时，仅引入 CCV 就可使 EPE 指标大大优于相同迭代次数的 RAFT-Stereo，比如在 1 次迭代中就比 RAFT-Stereo 高出了 55.09%。最后，迭代次数改为 32 后，CCV-Stereo 达到了最佳性能，比 RAFT-Stereo 高出了 22.95%。

表 3　减少迭代次数进行预测

Table 3　Reducing the number of iterations for prediction

Method	Number of Iterations					
	1	2	4	8	16	32
RAFT-Stereo	2.16	1.21	0.82	0.66	0.63	0.61
R+CCV	0.97	0.83	0.72	0.63	0.56	0.52
R+CCV+MPM+CTAM	0.71	0.67	0.61	0.54	0.51	0.49
Full model	0.68	0.65	0.59	0.53	0.49	0.47

2.4　与其他方法对比

为了展示 CCV-Stereo 在驾驶场景的性能，本文将其与其他已发布的方法在 Scene Flow，KITTI2012 和 KITTI2015 数据集上进行比较，测试结果见表 4。在 Scene Flow 测试集上，本文提出模型的 EPE 指标达到 0.47，比经典的 PSMNet 方法要高出 56.88%，且较基准模型 RAFT-Stereo 精确度提高了 16.07%。

表 4　在 Scene Flow 数据集进行 EPE 参数对比

Table 4　Comparison of EPE parameters in the Scene Flow dataset

Method	PSMNet	CSPN	LEAStereo	GwcNet	GC-Net	RAFT-Stereo	ACVNet	本文
EPE/px	1.09	0.78	0.78	0.76	0.72	0.56	0.48	0.47

此外,在 KITTI2012 和 KITTI2015 的测试集上对本文模型进行了评估,见表 5。在 KITTI2012 数据集上,CCV-Stereo 在各项指标上都有着突出表现。其中,2-noc 降低 8.85%,2-all 降低 6.20%,而 3-all 指标降得最多、达到了 12.05%。随后还在 KITTI2015 数据集进行了测验,且从表 5 中可知,本文模型在 D1-fg 和 D1-all 指标上都取得不错的结果。综合上述可知,CCV-Stereo 在 KITTI2012、2015 数据集中的表现都要明显优于 RAFT-Stereo,并且处理遮挡区域和复杂环境时有着更好的精确性和鲁棒性,由此也验证了本文提出模型的有效性和可行性。

表 5 在 KITTI2012 和 KITTI2015 数据集上进行参数对比

Table 5 Parameter comparison in the KITTI2012 and KITTI2015 datasets

Methods	KITTI2012						KITTI2015		
	2-noc	2-all	3-noc	3-all	EPE-noc	EPE-all	D1-bg	D1-fg	D1-all
GC-Net	2.71	3.46	1.77	2.30	0.6	0.7	2.21	6.16	2.87
PSMNet	2.44	3.01	1.49	1.89	0.5	0.6	1.86	4.62	2.32
GWCNet	2.16	2.71	1.32	1.70	0.5	0.5	1.74	3.93	2.11
HITNet	2.00	2.65	1.41	1.89	0.4	0.5	1.74	3.20	1.98
AcfNet	1.83	2.35	1.17	1.54	0.5	0.5	1.51	3.80	1.89
ACVNet	1.83	2.35	**1.13**	1.47	0.4	0.5	**1.37**	3.07	1.65
RAFT-Stereo	1.92	2.42	1.30	1.66	0.4	0.5	1.58	3.05	1.82
CCV-Stereo(本文)	**1.75**	**2.27**	1.21	**1.46**	**0.4**	**0.5**	1.41	**2.74**	**1.63**

为了评估 CCV-Stereo 的泛化能力,在 ETH3D 和 Middlebury2014 数据集上进行试验。结果见表 6。由表 6 可知,本文提出的模型在低纹理区域表现更加出色,但在 ETH3D 的灰度图像上表现不佳。

表 6 在 Middlebury2014 和 ETH3D 数据集上进行评估

Table 6 Evaluated on Middlebury2014 and ETH3D datasets

Model	Middlebury2014		ETH3D
	Half	Quater	
SGM	25.2	10.7	12.9
PSMNet	15.8	9.8	10.2
GANet	13.5	8.5	6.5
DSMNet	13.8	8.1	6.2
CFNet	15.3	9.8	5.8
RAFT-Stereo	8.7	7.3	3.2
CCV-Stereo(本文)	7.5	6.4	3.4

2.5 可视化结果对比与分析

图 4 展示了在 KITTI 数据集中遮挡区域的视差图对比,可以看到基准模型对图 4 中的遮挡区域处理效果不佳。例如对互遮挡栏杆的处理,CCV-Stereo 可以很好地显示出栏杆的大致形状,但基准模型的结果却模糊不清。在树支撑杆的处理上,本文模型展示的效果也更加理想,得到的结果清晰且连贯,处理有一定的稳定性。

图 5 展示了在 KITTI 数据集中弱纹理区域的视差图对比。显而易见地,本文提出的模型相较于基准模型更具有说服力,其在围栏和栏杆区域的表现效果更为显著,捕捉的细节信息也更多。

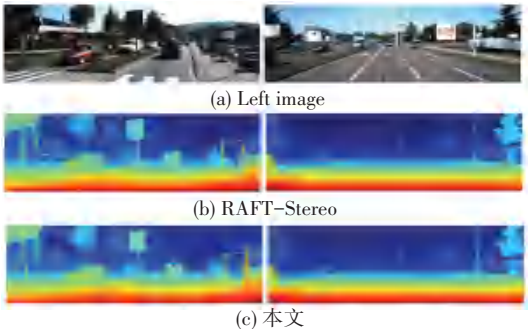

图 4 在 KITTI 数据集中图像的遮挡区域视差图对比

Fig. 4 Comparison of disparity maps of occluded areas in images in the KITTI dataset

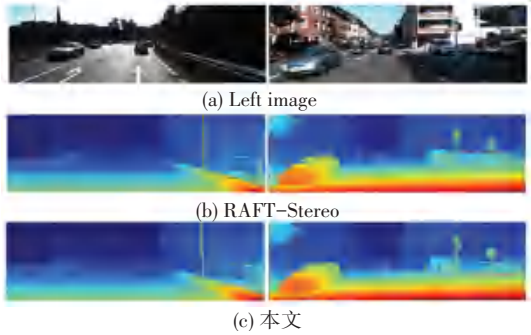

图 5 在 KITTI 数据集中图像的弱纹理区域视差图对比

Fig. 5 Comparison of disparity maps of weak texture areas in images in the KITTI dataset

图 6 展示的是在 Middlebury2014 和 ETH3D 数据集上进行预测的视差图，从图 6 中可以看出，与基准方法相比，CCV-Stereo 可以保留更多细节信息。例如在一些枝干衔接处，本文模型预测出的视差图更加流畅且连贯。CCV-Stereo 相较于基准模型的预测效果有显著提升。

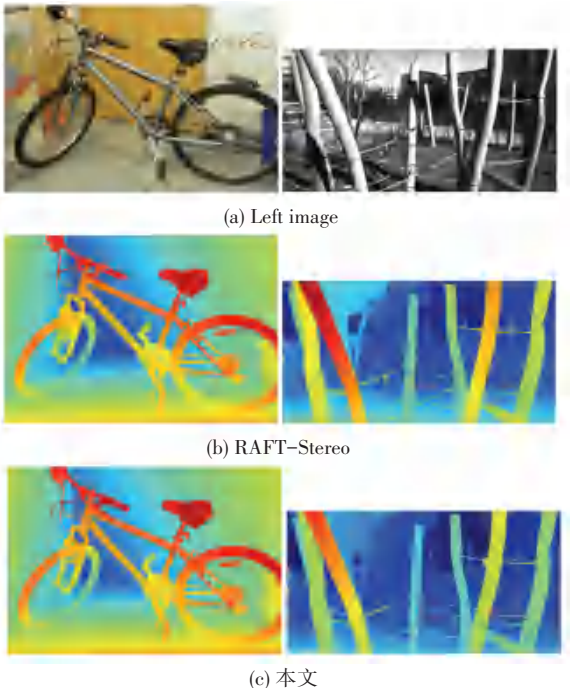

图 6　在 Middlebury2014 和 ETH3D 数据集上预测的视差图对比

Fig. 6　Comparison of disparity maps predicted on Middlebury2014 and ETH3D datasets

3　结束语

为了解决目前立体匹配方法在处理代价体时计算耗能高、且在弱纹理和遮挡区域处理不佳的问题，本文提出一种基于组合代价体的多级迭代立体方法，通过将全对相关体和注意力代价进行组合获得含有丰富特征信息的组合代价体，来帮助模型在减少迭代次数情况下仍可以得到高精度的视差图。此外，再引入多级池化模块和上下文注意力机制使得模型更加关注重点特征，进一步提高模型生成视差图的精确度。最后，在各类数据集上进行训练和测试，将得到的各项指标与其他现有方法进行比较，可知本文所提出的方法在误差精度和对图像中弱纹理与遮挡区域的处理都有着良好表现。

参考文献

[1] 尹晨阳，职恒辉，李慧斌. 基于深度学习的双目立体匹配方法综述[J]. 计算机工程，2022，48(10)：1-12.

[2] LI Zhiqi，YU Zhiding，LAN Shiyi，et al. Is ego status all you need for open-loop end-to-end autonomous driving? [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2024：14864-14873.

[3] JING Junpeng，MAO Ye，MIKOLAJCZYK K. Match-stereo-videos：Bidirectional alignment for consistent dynamic stereo matching [C]//Proceedings of the European Conference on Computer Vision. Cham：Springer，2024：415-432.

[4] KRUMPEK O，SCHLÜTER M，HÜGLE J. Investigation on stereo-ToF data fusion for the inspection of used industrial parts [J]. AIP Conference Proceedings，2024，2989(1)：030008.

[5] 吴昊，李成斌，陈彦良，等. 基于双目立体匹配的三维重建系统研究[J]. 现代计算机，2024，30(13)：78-81.

[6] SHAKED A，WOLF L. Improved stereo matching with constant highway networks and reflective confidence learning [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2017：4641-4650.

[7] YE Xiaoqing，LI Jiamao，WANG Han，et al. Efficient stereo matching leveraging deep local and context information[J]. IEEE Access，2017，5：18745-18755.

[8] ZHAO Shiyu，ZHAO Long，ZHANG Zhixing，et al. Global matching with overlapping attention for optical flow estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2022：17592-17601.

[9] SHEN Zhelun，DAI Yuchao，SONG Xibin，et al. PCW-Net：Pyramid combination and warping cost volume for stereo matching [C]//Proceedings of the European Conference on Computer Vision. Cham：Springer，2022：280-297.

[10] MAYER N，ILG E，HAUSSER P，et al. A large dataset to train convolutional networks for disparity，optical flow，and scene flow estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2016：4040-4048.

[11] KENDALL A，MARTIROSYAN H，DASGUPTA S，et al. End-to-end learning of geometry and context for deep stereo regression [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE，2017：66-75.

[12] CHANG Jiaren，CHEN Yongsheng. Pyramid stereo matching network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2018：5410-5418.

[13] GUO Xiaoyang，YANG Kai，YANG Wukui，et al. Group-wise correlation stereo network [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2019：3273-3282.

[14] XU Gangwei，CHENG Junda，GUO Peng，et al. Attention concatenation volume for accurate and efficient stereo matching [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2022：12981-12990.

[15] KHAMIS S，FANELLO S，RHEMANN C，et al. Stereonet：Guided hierarchical refinement for real-time edge-aware depth prediction [C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham：Springer，2018：573-590.

[16] LIPSON L，TEED Z，DENG Jia. Raft-stereo：Multilevel recurrent field transforms for stereo matching[C]//Proceedings of 2021 International Conference on 3D Vision (3DV). Piscataway，

NJ:IEEE, 2021: 218-227.
[17] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite [C]// Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE, 2012: 3354-3361.
[18] MENZE M, HEIPKE C, GEIGER A. Joint 3d estimation of vehicles and scene flow [J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, 2: 427-434.
[19] SCHARSTEIN D, HIRSCHMÜLLER H, KITAJIMA Y, et al. High-resolution stereo datasets with subpixel-accurate ground truth [C]//Proceedings of the 36th German Conference on Pattern Recognition (GCPR 2014). Cham:Springer, 2014: 31-34.
[20] SCHOPS T, SCHONBERGER J L, GALLIANI S, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE, 2017: 3260-3269.