基于引导信息的双目立体匹配算法

魏 东,何 雪*
(沈阳工业大学 信息科学与工程学院,辽宁 沈阳 110870)

摘 要:针对现有立体匹配算法在边缘、遮挡、视差不连续、弱纹理等区域匹配误差较大的问题,提出一种在利用视差注意力机制的基础上引入边缘和语义信息的立体匹配算法。在利用视差注意力机制进行代价计算和代价聚合中引入边缘细节信息改善边缘和遮挡区域匹配误差较大的问题,并对引入边缘信息时与特征提取过程中得到的不同尺度特征图融合的时机进行了讨论,确定浅层大尺度特征图引入边缘信息可以提高匹配精度;在视差优化中引入语义信息改善视差不连续和弱纹理区域匹配精度不高的问题,并对不同尺度特征图求取的语义信息对匹配精度的影响进行讨论,利用深层小尺度特征图提取语义信息可以提高匹配精度。提出的方法在 SceneFlow 数据集上进行了测试,将基准网络 PASMNet 的误差降低了 49.05%,并与其他算法进行对比分析。实验结果表明,边缘和语义等引导信息的引入有针对性地改善了现有算法在边缘、遮挡、视差不连续和弱纹理区域的视差精度,从而提高了整体预测精度。

关键词:立体匹配;双目视觉;边缘信息;语义信息;视差注意力机制
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2022)12-0159-06
doi:10.3969/j.issn.1673-629X.2022.12.024

Binocular Stereo Matching Algorithm Based on Guidance Information

WEI Dong, HE Xue*
(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: Aiming at the problem that the existing stereo matching algorithms have large matching errors in areas such as edge, occlusion, parallax discontinuity and weak texture, a stereo matching algorithm based on parallax attention mechanism and introducing edge and semantic information is proposed. In the cost calculation and cost aggregation using parallax attention mechanism, the edge detail information is introduced to improve the problem of large matching errors in edge and occluded regions. The timing of the fusion of the edge information and the different scale feature maps obtained in the process of feature extraction is discussed, and it is determined that the introduction of edge information in the shallow large-scale feature map can improve the matching accuracy. The semantic information is introduced to improve the parallax discontinuous and weak texture regions in parallax optimization, and the influence of semantic information extracted from feature maps of different scales on the matching accuracy is discussed, and semantic information extracted from deep small-scale feature map can improve the matching accuracy. The proposed method is evaluated on the SceneFlow dataset and compared with other algorithms, and the error of the benchmark network PASMNet is reduced by 49.05%. Experiments show that the introduction of edge and semantic information improves the disparity solution of existing algorithms in edge, occlusion and weak texture, so as to improve the overall prediction accuracy.

Key words: stereo matching; binocular vision; edge information; semantic information; parallax-attention mechanism

0 引 言

深度信息是自动驾驶、机器人、物体检测等计算机视觉应用中非常重要的信息。通过对双目立体相机拍摄得到的两个图像进行立体匹配,计算参考图像的稠密视差图,是获取深度信息的重要途径之一。

Zbontar 和 LeCun 首次将 CNN 应用到立体匹配中,利用卷积神经网络(MC-CNN)[1]计算匹配代价。FlowNetC[2]、DispNetC[3]等算法将立体匹配通过端到端的有监督深度学习方式实现,采用编码-解码的结构回归视差。文献[4-5]利用从边缘检测任务中获得的边缘信息补充视差图中丢失的细节信息。SegStere[6]将语义特征融入特征图中,并设计语义损失项来改善视差学习效果。

近年来,注意力机制在许多计算机视觉任务中也

收稿日期:2021-12-13 **修回日期**:2022-04-13
基金项目:辽宁省教育项目(LJGD2020006)
作者简介:魏 东(1968-),男,硕士,副教授,研究方向为计算机图形学、虚拟现实;通讯作者:何 雪(1997-),女,硕士研究生,研究方向为计算机图形学。

得到了广泛的应用[7-11]，将注意力机制应用在立体匹配任务上，可以有效捕捉图像中有用的区域，从而提升立体匹配算法性能。PASMNet[12] 中的视差注意力机制（Parallax-Attention Mechanism，PAM）针对固定的最大视差阻碍了具有较大视差变化的图像对在进行匹配代价计算时，由于差异回归的模糊性导致不合理的代价分配问题，采用将极线约束与注意力机制相结合，计算沿极线的特征相似性的方法来解决。

虽然基于 PAM 的立体匹配方法可以避免以上问题，仍然难以克服边缘处误差较大、视差不连续以及一些遮挡和弱纹理区域误匹配率较高的问题。

针对以上问题，提出在视差注意力机制上引入语义和边缘引导信息的立体匹配算法以提高匹配精度。

1 算法设计

为了解决立体匹配算法在物体边缘处和遮挡区域匹配精度较低的问题，在边缘处匹配误差较大的主要原因是卷积操作引起的细节信息丢失，可以引入边缘特征信息来弥补。在弱纹理和视差不连续区域匹配误差较大的主要原因是在这些区域上进行视差估计的匹配特征不足，可以通过语义分割获得更多的特征，如个体的语义一致性等特征，使得在弱纹理和视差不连续区域更好地实现特征匹配。算法整体网络结构如图 1 所示，包括边缘提取（Edge extraction）、特征提取（Feature extraction）、视差注意力机制（PAM）、语义信息提取（Semantic information extraction）、视差估计（Disparity prediction）、视差优化（Refinement）六部分。

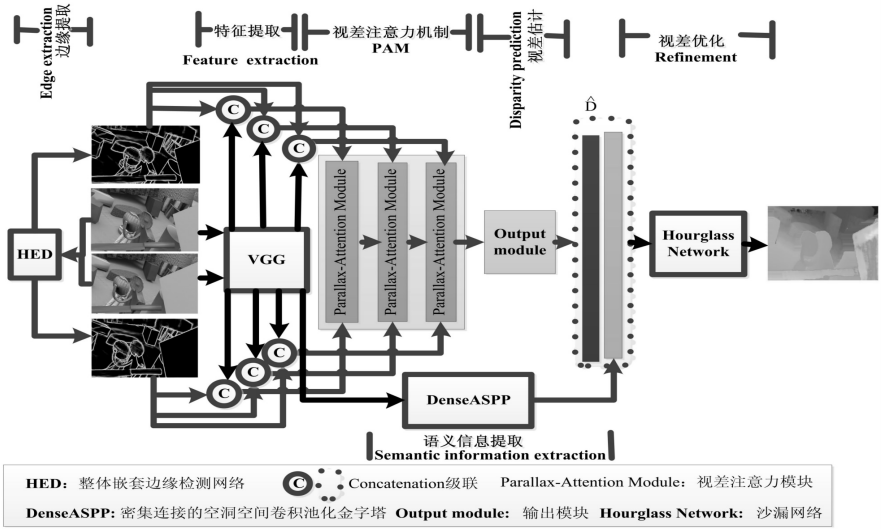


图 1 网络结构

该网络将左右图像送入 VGG 网络进行特征提取，获取不同尺度的图像特征，同时把左右图像送入 HED[13]（Holistically-nested Edge Detection）网络提取边缘细节信息。将经过特征提取得到的特征与边缘信息进行融合，送入 PAM 回归匹配代价，在特征中引入边缘细节改善由于遮挡使回归的视差图中物体边缘处误差较大的问题。PAM 通过矩阵乘法计算匹配代价，得到视差注意力图，再利用输出模块回归视差注意力图得到初始视差图。把经过特征提取网络得到的最后 1 个尺度特征送入 DenseASPP[14] 网络得到语义特征图。将初始视差图和语义特征图级联送入沙漏型网络进行视差优化，融合语义信息改善弱纹理和视差不连续区域，得到最终的视差图。

1.1 基准网络

基准网络是 PASMNet，网络结构如图 2 所示。左右图像经过沙漏网络特征提取，将提取的特征送入级

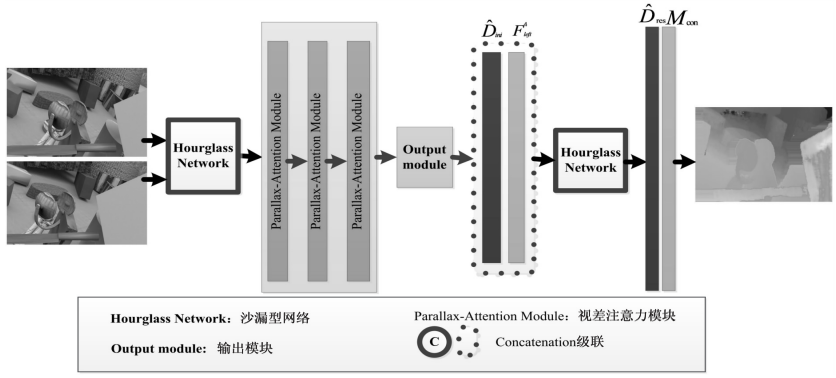


图 2 PASMNet 网络结构

联的视差注意力模块回归匹配成本,然后使用输出模块从匹配成本中获得一个初始视差。最后,利用沙漏网络进一步细化初始视差,以产生最后的视差图。

PASMNet 使用视差注意力计算匹配代价,与 3D cost volume 计算匹配代价不同的是:视差注意力是将极线约束与注意力机制相结合,沿极线计算特征相似性,如图 3 所示。具体来说,对于左图像中的每个像素 P ,在右图像沿极线上的所有像素中找到特征最相似的像素。

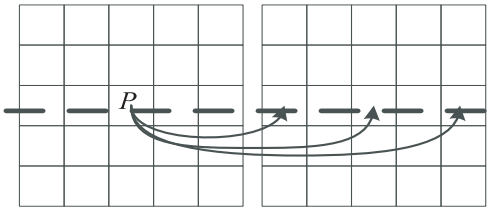


图 3　视差注意力

PAM 由 3 个视差注意力模块级联组成,每个视差注意力模块由 4 个相同的结构块构成,如图 4 所示。C 为匹配代价表示两个像素之间的特征相关性,沙漏型网络提取的特征经过卷积层得到特征映射 A 、B $\in R^{H\times W\times C}$,将 A 、B 送到 1×1 卷积以进行特征匹配。具体来说, A 送入 1×1 卷积以生成查询特征映射 Q $\in R^{H\times W\times C}$,同时 B 送入另一个 1×1 卷积中,生成一个关键特征映射 $K \in R^{H\times W\times C}$,再将其重塑为特征空间是 $R^{H\times C\times W}$ 的特征映射。然后,在 Q 和 K 之间执行矩阵乘法,通过矩阵乘法,可以有效地将沿极线的任意两个位置之间的特征相关性编码到视差注意力图中。

将级联的视差注意模块得到的匹配代价送到输出模块。输出模块如图 5 所示,其中匹配代价 C 首先被送到 Softmax 层,以分别产生通道数为 1 的视差注意力图 M ,对得到的视差注意力图回归计算得到初始视差,计算公式如公式(1):

$$\hat{D} = \sum_{k=0}^{\frac{w}{4}-1} k \times M^{3}_{\text{right}\to\text{left}}(:,:,k) \tag{1}$$

其中, $\hat{D}$ 表示初始视差, $M^{3}_{\text{right}\to\text{left}}$ 是第三个视差注意力模块计算得到的目标图像图对参考图像的视差注意力, w 和 k 是特征维度。

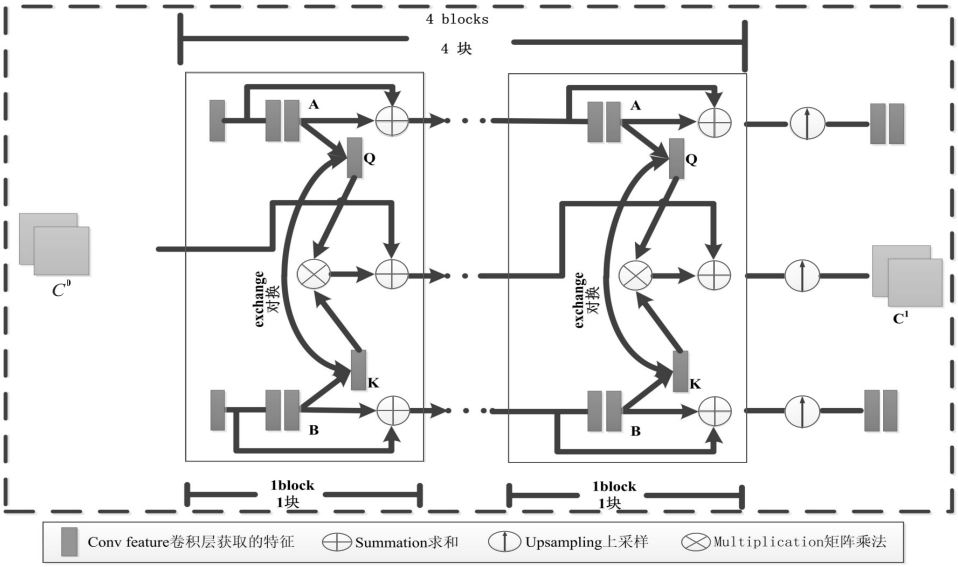


图 4　视差注意力模块

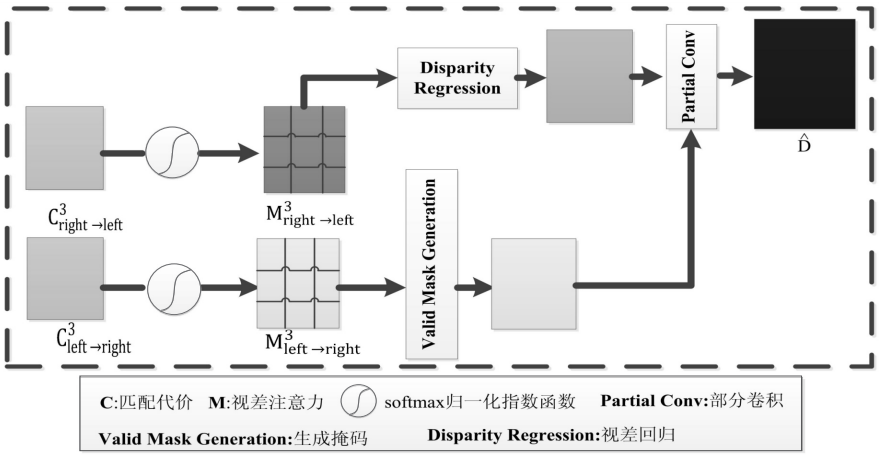


图 5　输出模块

1.2 特征提取

特征提取部分使用 VGG 网络,VGG 网络可以产生丰富的多尺度几何信息特征,如图 6 所示。VGG 网络可以获取多个尺度的特征,第 1、2、3 尺度特征用于视差注意力模块,第 5 尺度特征用于语义信息提取模块。

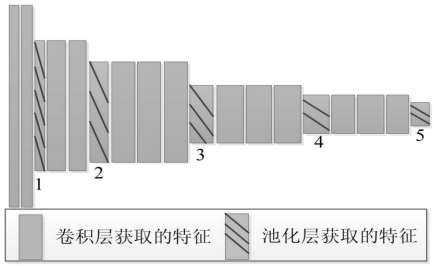


图 6 VGG 网络结构

1.3 边缘提取

边缘是由于图像中像素值发生较大变化而导致不连续的结果,它存在于目标与背景、目标与目标、区域与区域之间。针对立体匹配中边缘误差较大和遮挡问题,引入边缘细节可以得到改善。在深度学习出现之前,边缘检测有几种常用的方法,如 Sobel、Canny 等。传统的边缘检测算法到现在还在使用,但是过于依赖人工设定阈值,不能在通用场景下工作。

该文使用 HED 算法提取边缘,如图 7 所示。HED 是以 VGGNet 与 FCN 作为基础网络进行改进,将 VGG 网络的多个特征层的输出,利用 FCN 全卷积网络,通过权重融合实现各个层相连接,得到边缘特征。

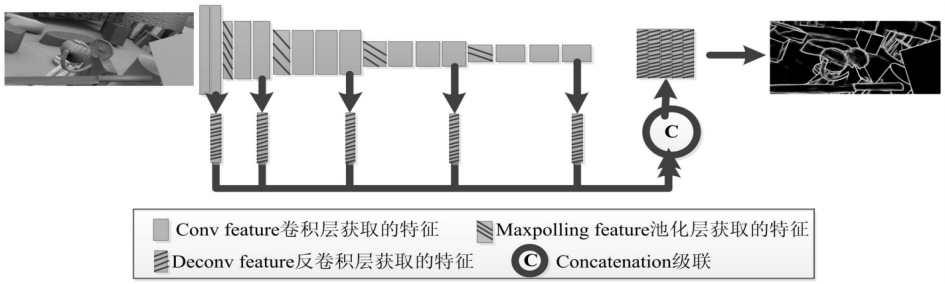


图 7 HED 网络结构

1.4 语义信息提取

针对立体匹配中视差不连续、弱纹理区域,引入语义分割获取的语义信息可以得到改善。使用深度学习方法之前,语义分割方法比较常用的是 TextonForest 和基于随机森林分类器等方法。现在很多使用深度学习进行语义分割的模型基本上都是由 FCN 改进的,但是 FCN 模型需要池化操作,池化操作可以通过扩大感受野进而能够很好地整合上下文信息,但通过池化进行下采样操作也使分辨率降低,削弱了位置信息,而语义分割中对齐操作需要丰富的位置信息。

该文使用文献[13]提出的 DenseASPP 算法提取语义信息,网络结构如图 8 所示,DenseASPP 将 ASPP 和 DenseNet 中的密集连接相结合,具有更大的感受野和更密集的采样点。扩张卷积用于解决特征图分辨率和感受野之间的矛盾;用密集连接获得更好的性能,将每个扩张卷积使用密集连接的方式输出结合到一起。

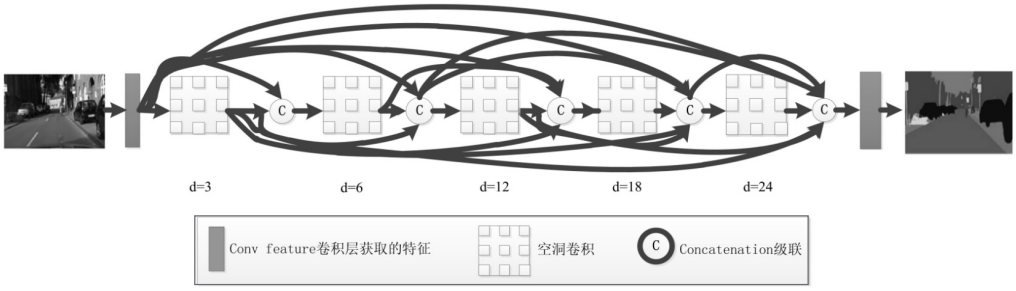


图 8 DenseASPP 网络结构

1.5 损失函数

使用 Groundtruth 视差数据在有监督学习的模式下对模型进行端到端的训练,为了更好地监督生成边缘清晰、物体表面平滑和语义明确的视差图,利用常规的视差回归损失。

对于视差回归,采用 smooth_{L_1} 损失函数来训练视差分支,与 L_2 损失相比,smooth_{L_1} 损失具有很好的鲁棒性和对异常值的低敏感性[15-16]。

损失函数定义如下:

$$L_s = \frac{1}{N}\sum_{i}^{N} \text{smooth}_{L_1}(D_i - \hat{D}_i) \tag{2}$$

smooth_{L_1} 定义如下:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5 \times x^2 & |x| < 1 \\ |x| - 0.5 & |x| \geqslant 1 \end{cases} \tag{3}$$

其中,N 表示所有带 Groundtruth 的像素数目,D_i 和 $\hat{D}_i$ 分别表示实际视差和估计视差。

2　实验结果

2.1　数据集及评价指标介绍

将所提方法在 SceneFlow[2] 上进行评估。SceneFlow 是一个大规模的合成数据集,包括分辨率为 540×960 像素的 35 454 组训练图像以及 4 370 组测试图像,且提供了稠密的视差图作为 Groundtruth。使用端点误差(End-Point-Error,EPE),即预测视差图与实际视差图的平均绝对误差,和 t 像素(t-pixel,tpx)误差作为评价指标。

2.2　实验环境、参数设置及效果图

在 SceneFlow 数据集上训练网络。在训练阶段,将左右图像随机裁剪为 256×512 像素的图像作为输入。所有模型均采用 Adam 方法进行优化,批次大小为 2。初始学习率设定为前 5 个 epoch 的 1×10^{-3},后 15 个 epoch 的学习率降低到 1×10^{-4}。所有实验都是在 NvidiaGTX1080 GPU 的 PC 上进行的。图 9 分别为 SceneFlow 数据集中部分数据的左图、预测视差图和视差真值图。

(从上到下依次为输入左图、预测视差图、视差真值图)

图 9　SceneFlow 测试结果

2.3　消融实验、实验环境及效果图

特征提取:VGG 网络可以获取多个尺度的特征,用于视差注意力模块和语义信息提取模块。VGG 网络可以产生丰富的多尺度几何信息特征,可以实现更好的性能。

引入边缘信息:将边缘信息与特征提取的特征融合,引入到视差注意力模块中,对边缘信息融合特征位置进行实验。从表 1 中可以看出,将边缘信息与特征提取中池化后第 1、2、3 尺度特征融合与池化后第 3、4、5 尺度特征融合对比, EPE 从 2.614 增加到 3.809, 1px/3px 错误率从 19.112/12.195 增加到 28.825/18.273。并在 PAM 网络上 EPE 减少了 31.32%。因此,将边缘信息与池化后第 1、2、3 尺度特征融合引入 PAM 模块中可以取得更好的效果。引入语义信息:从特征提取中获取的不同尺度特征中选出一个特征送入语义信息提取模块提取语义信息,利用语义信息进行视差优化。对送入语义信息提取模块的不同尺度特征选取进行实验。从表 1 中可以看出,传入特征提取中池化后第 5 个尺度特征比池化后第 1 个尺度相比, EPE 从 2.685 增加到 2.711, 1px/3px 错误率从 18.551/12.121 增加到 19.415/12.754。并在 PAM 网络上 EPE 减少了 29.45%,因此,利用特征提取模块中第 5 个尺度的特征进行语义信息的提取,可以使得最终的视差估计取得更好的效果。

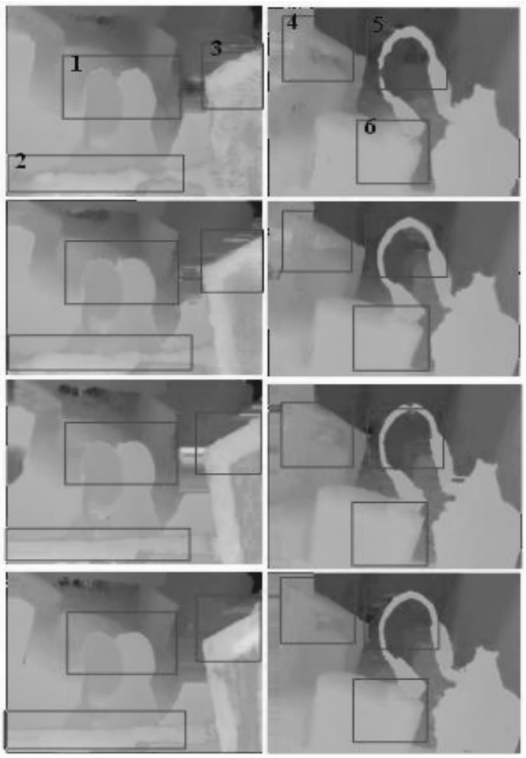
表 1　SceneFlow 消融实验

Model	Feaextract	PAM	Edge		Semantic		EPE	1px	3px
			123	345	1	5			
PAM	√	√					3.806	23.534	15.149
PAM+Edge	√	√	√				2.614	19.112	12.195
PAM+Edge	√	√		√			3.809	28.825	18.273
PAM+Semantic	√	√			√		2.795	19.415	12.754
PAM+Semantic	√	√				√	2.685	18.355	12.121
PAM+Edge+Semantic	√	√	√			√	2.313	16.324	11.953

从表 1 中可以看到,在 PAM 网络上分别引入边缘和语义信息都可以降低错误率,提升网络性能。当边缘和语义信息同时引入时,EPE 在 PAM 网络上提升了 39.23%,实验证明引入边缘和语义信息可以有效提升网络回归视差图性能。

图 10 是不同网络结构得到的视差图对比,矩形框 1 处物体边缘、3 处圆柱体边缘和 5 处耳机边缘在 PAM+Edge 网络得到的视差图对应位置与 PAM 网络得到的视差图对应位置对比边缘轮廓更圆滑且误差较小;矩形框 2 处长矩形与矩形框 1 处物体遮挡处在 PAM+Edge 网络得到的视差图对应位置与 PAM 网络得到的视差图对应位置对比边缘轮廓更清晰准确;矩形框 2 处长矩形和矩形框 5 处耳机在 PAM+ Semantic 网络得到的视差图对应位置与 PAM 网络得到的视差图对应位置对比内部视差相对统一、视差连续;矩形框 4 处的长方体和矩形框 6 处的处圆柱体由于弱纹理导

致特征匹配有误的问题在 PAM+Semantic 网络得到的视差图也得到改善。而且在同时引入边缘和语义信息的 PAM+Edge+Semantic 网络得到的结果整体效果好于分别引入边缘和语义信息。通过可视化的结果可以更直接看到引入边缘信息得到的视差图边缘更平滑，降低边缘处误差，引入语义信息后视差图中视差不连续和弱纹理处效果有提升，这得益于边缘和语义信息的引入能指导生成更加精准的视差图。



（从上到下依次为 PAM 网络、PAM + Edge、PAM + Semantic、PAM+Edge+Semantic）

图 10 不同网络结构得到的视差图

2.4 与其他算法比较

在 SceneFlow 数据集中的量化误差指标比较如表 2 所示。从表 2 中可以看出，文中方法与基准网络 PASMNet 进行对比，将 PASMNet 的误差降低了 49.05%，相对于其他经典的匹配算法在计算端点误差时，误差率较小。

表 2 SceneFlow 测试结果

Method	EPE	1px	3px
PASMNet	4.54	18.99	15.91
GC-Net[17]	2.51	16.9	–
DenseMapNet[18]	5.36	–	–
Base+PAM[19]	3.99	–	–
文中方法	2.313	16.324	11.953

3 结束语

为改善立体匹配中边缘处的误差较大和遮挡、视差不连续、弱纹理等区域匹配精度不高的问题，提出一种在利用视差注意力机制进行立体匹配时引入边缘和语义信息的立体匹配算法。在利用视差注意力机制进行匹配代价计算过程中引入边缘信息，改善边缘处的误差较大、遮挡问题，以提高整体匹配精度；在视差优化过程中引入语义信息，改善视差不连续、弱纹理问题，以提高整体匹配精度。通过实验表明，该算法在 SceneFlow 数据集上将基准网络 PASMNet[12] 的误差降低了 49.05%。该算法相较于基准方法虽然精度有所提升，但是需要大量的数据集进行训练，因此，在小数据集上的效果表现不好。在未来的研究中需要对网络结构进一步优化，可以考虑如何减小网络大小，这可能进一步改善视差的估计效果。

参考文献：

[1] ŽBONTAR J，LECUN Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of Machine Learning Research，2016，17(1)：2287-2318.

[2] DOSOVITSKIY A，FISCHER P，ILG E，et al. Flownet：learning optical flow with convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. Washington：IEEE，2015：2758-2766.

[3] MAYER N，ILG E，HAUSSER P，et al. A large dataset to train convolutional networks for disparity，optical flow，and scene flow estimation[C]//IEEE conference on computer vision and pattern recognition. Long Beach：IEEE，2019：4040-4048.

[4] SONG X，ZHAO X，HU H，et al. EdgeStereo：a context integrated residual pyramid network for stereo matching[C]//Asian conference on computer vision. Berlin：Springer，2018：20-35.

[5] SONG X，ZHAO X，FANG L，et al. EdgeStereo：an effective multi-task learning network for stereo matching and edge detection[J]. International Journal of Computer Vision，2020，128(4)：910-930.

[6] WU Z，WU X，ZHANG X，et al. Semantic stereo matching with pyramid cost volumes[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul：IEEE，2019：7484-7493.

[7] 王昊飞，李俊峰. 基于注意力机制的改进残差网络的人体行为识别方法[J]. 软件工程，2021，24(11)：51-54.

[8] 秦庭威，赵鹏程，秦品乐，等. 基于残差注意力机制的点云配准算法[J/OL]. 计算机应用：1-11. http://kns.cnki.net/kcms/detail/51.1307.tp.20211029.1616.008.html.

[9] 张 纠，刘晓芳，杨 兵. 基于双通道级联注意力网络的医学图像配准[J]. 计算机工程与设计，2021，42(10)：2894-2901.

（下转第 172 页）