·图形图像处理·　　文章编号：1000-3428(2022)03-0220-09　　文献标志码：A　　中图分类号：TP391

基于视差优化的立体匹配网络

刘建国[1,2,3,4]，纪　郭[1,2,3,4]，颜伏伍[1,2,3,4]，沈建宏[5]，孙云飞[5]

（1.先进能源科学与技术广东省实验室佛山分中心（佛山仙湖实验室），广东 佛山 528200；
2.武汉理工大学 现代汽车零部件技术湖北省重点实验室，武汉 430070；3.汽车零部件技术湖北省协同创新中心，武汉 430070；
4.湖北省新能源与智能网联车工程技术研究中心，武汉 430070；5.宁波华德汽车零部件有限公司，浙江 宁波 315000）

摘　要：现有的立体匹配算法通常采用深层卷积神经网络提取特征，对前景物体的检测更加精细，但对背景中的小物体及边缘区域匹配效果较差。为提高视差估计质量，构建一个基于视差优化的立体匹配网络CTFNet。分别提取浅层与深层特征，并基于深层特征构建全局稀疏代价卷，从而预测初始视差图。在预测的初始视差图和浅层特征的基础上构建局部稠密代价卷并进行视差优化，以细化预测视差值邻域的概率分布，提高特征不明显区域的匹配精度。此外，引入新的概率分布损失函数，监督softmax函数计算的视差值概率分布在真实视差值附近成单峰分布，提高算法的鲁棒性。实验结果表明，该网络在SceneFlow和KITTI数据集上的误匹配率分别为0.768%和1.485%，在KITTI测评网站上的误差率仅为2.20%，与PSMNet网络相比，精度和速度均得到一定提升。

关键词：立体匹配；视差优化；浅层特征；匹配代价卷；损失函数

开放科学（资源服务）标志码（OSID）：

中文引用格式：刘建国，纪郭，颜伏伍，等. 基于视差优化的立体匹配网络[J]. 计算机工程，2022，48(3)：220-228.

英文引用格式：LIU J G，JI G，YAN F W，et al. Stereo matching network based on disparity optimization[J]. Computer Engineering，2022，48(3)：220-228.

Stereo Matching Network Based on Disparity Optimization

LIU Jianguo[1,2,3,4]，JI Guo[1,2,3,4]，YAN Fuwu[1,2,3,4]，SHEN Jianhong[5]，SUN Yunfei[5]

（1.Foshan Xianhu Laboratory of the Advanced Energy Science and Technology Guangdong Laboratory，Foshan，Guangdong 528200，China；
2.Hubei Key Laboratory of Advanced Technology for Automotive Components，Wuhan University of Technology，Wuhan 430070，China；
3.Hubei Collaborative Innovation Center for Automotive Components Technology，Wuhan 430070，China；
4.Hubei Research Center for New Energy & Intelligent Connected Vehicle，Wuhan 430070，China；
5.Ningbo Huade Automobile Parts Co.，Ltd.，Ningbo，Zhejiang 315000，China）

【Abstract】Existing stereo matching algorithms usually use deep convolutional networks to extract features，and can improve the accuracy of foreground object detection，but display poor matching results for small objects and boundary areas in the background.In order to improve the quality of disparity estimation in these areas，a stereo matching network named Coarse To Fine Net（CTFNet ）is proposed based on disparity optimization.The network extracts shallow and deep features separately and a global sparse cost volume is constructed based on the deep features to predict the initial disparity map.Then a local dense cost volume is constructed based on the predicted initial disparity map and shallow features，which optimizes the disparity and refines the probability distribution of the neighborhood of predicted disparity value to improve the matching accuracy of areas with less obvious features.At the same time，a new loss function for probability distribution is introduced to supervise the probability distribution calculated by the softmax function in a unimodal distribution near the true disparity value and improve the robustness of the algorithm.The experimental results show that the mismatching rate of the proposed network is 0.768% on the SceneFlow dataset and 1.485% on the KITTI data set，and its error rate on the KITTI evaluation website is only 2.20%.Compared with the PSMNet network，the proposed algorithm displays an improvement in both accuracy and speed.

【Key words】stereo matching；disparity optimization；shallow feature；matching cost volume；loss function

DOI：10. 19678/j. issn. 1000-3428. 0060806

基金项目：国家自然科学基金（51975434）；先进能源科学与技术广东省实验室佛山分中心（佛山仙湖实验室）开放基金（XHD2020-003）。

作者简介：刘建国（1971—），男，副教授、博士，主研方向为机器视觉、智能驾驶；纪　郭，硕士；颜伏伍，教授、博士；沈建宏、孙云飞，学士。

收稿日期：2021-02-03　　**修回日期：**2021-03-13　　**E-mail：**ljg424@163.com

0　概述

随着图像处理技术的发展，基于视觉的深度估计逐渐发展成为无人驾驶、机器人等领域的重要测距方法之一，其中基于立体匹配的双目测距方法以兼顾精度、速度和成本的优势受到广泛关注，该方法通过匹配同一场景的左右视点两幅图像中的相应像素点来计算视差，并根据相似三角形原理计算深度距离。传统的立体匹配算法将匹配过程划分为匹配代价计算、代价聚合、视差计算和视差优化4个部分[1]，并基于代价函数的约束范围及搜索策略分为局部、全局和半全局立体匹配算法[2]。但传统算法采用手工设计的特征描述符，缺乏全局上下文信息，且受经验参数的影响，算法鲁棒性较差，不适合在复杂环境下应用[3]。

近年来，随着深度学习在计算机视觉领域中的发展，研究人员开始基于深度学习方法解决立体匹配问题。LECUN等[4]引入卷积神经网络提取左右图特征，通过学习两者的相似性计算匹配代价，有效提高了算法鲁棒性，但该方案仍然需要配合传统算法中的十字交叉代价聚合[5]、半全局优化[6-7]及滤波操作等完成立体匹配。LUO等[8]在此基础上将匹配代价计算转化为多分类问题，训练网络直接输出所有潜在视差值下的匹配代价，大大提高算法效率。

上述方法利用卷积神经网络计算匹配代价减少了传统算法中手工设计特征的误差，但仍需结合传统算法中的其他步骤求解视差图，运行速度较低。因此基于卷积神经网络的端到端立体匹配算法应运而生。MAYER等[9]提出以左右图像为输入，以视差图为输出的端到端立体匹配网络DispNet，并发布一个带有真实视差图的大型合成数据集用于训练网络。在此基础上，KENDALL等[10]提出GCNet，首次通过级联不同视差值下的特征图构建匹配代价卷，并通过3D卷积进行代价聚合，最终通过视差回归的方式计算视差图，为后续算法发展提供重要思路。PANG等[11]提出一种两阶段网络结构，第1阶段学习初始视差，第2阶段学习修正初始视差的残差，最终将两阶段的和作为结果输出，有效提高匹配精度。CHANG等[12]提出PSMNet网络，利用空间金字塔池化（Spatial Pyramid Pooling，SPP）模块[13]融合不同尺度特征，同时采用堆叠的编码解码结构进行代价聚合，有效提高了视差预测精度。ZHANG等[14]基于传统的半全局匹配算法提出GANet，设计了半全局，引导聚合层从不同方向对代价卷进行聚合取代3D卷积，并结合局部引导聚合层，有效提升立体匹配的性能。MA等[15]结合光流、视差估计及实例分割3种任务，将各个实例的光流、视差及语义线索编码成能量函数进行最小化求解，实现多任务间互相融合，但运行时间过长。XU等[16]采用3D代价卷并设计尺度内及尺度间代价聚合模块代替3D卷积，有效提高算法实时性，但匹配精度相对较低。ZHU等[17]基于多尺度特征，设计十字形空间金字塔模块以不同的比例和位置聚合上下文信息构建代价卷，并设计多尺度3D特征匹配和融合模块聚合代价卷，有效提高算法在不适定区域的匹配精度。

随着高性能计算平台的发展，立体匹配的网络结构更加复杂，特征提取及代价聚合网络不断加深。深层网络有助于提取更加抽象的特征，对于目标检测、语义分割等对语义信息要求较高的视觉任务具有重要意义。但立体匹配作为低层级视觉任务，除了依赖深层特征完成前景物体的基本匹配，还需要浅层特征和局部上下文信息细化小物体、边缘等区域的匹配。而大多数立体匹配网络采用深层特征提取网络和堆叠的编码解码结构，在反复上下采样过程中造成浅层特征中的细节信息丢失。同时，传统的代价卷构造方式对每个像素的完整视差范围都构建匹配代价进行计算，虽然通过稀疏化视差的方式可以降低计算量，但是仍然造成代价卷在非真实视差处的计算冗余。

针对上述问题，本文引入视差优化思想，基于PSMNet构建一种改进网络CTFNet。采用特征提取网络分别提取下采样程度不同的浅层和深层特征，基于深层特征构建所有潜在视差值范围内的全局稀疏代价卷，并通过代价聚合和视差计算预测初始视差图。此外，基于局部上下文信息丰富的浅层特征和初始视差图，对每个像素点构建初始预测视差邻域范围内的局部稠密代价卷，并通过简单的代价聚合和视差计算进行视差优化。在损失函数部分，本文基于文献[18]引入softmax操作后的概率分布损失函数，在预测初始视差图过程中通过限制每个像素点处视差值概率，使其分布在真实视差值附近，形成高斯分布，提高初始视差图精度，保证视差优化阶段利用初始视差图构造局部稠密代价卷的可靠性，从而优化视差图精度。

1　网络结构

本文以PSMNet作为骨干网络进行改进，其网络结构如图1(a)所示（彩色效果见《计算机工程》官网HTML版）。PSMNet采用残差网络和空间金字塔池化（SPP）模块提取特征，其中残差网络由3个3×3卷积层和4个残差块共53层卷积层构成，配合SPP模块可以得到多尺度深层特征，基于左右特征图构建的匹配代价卷通过3个相同的编码-解码结构进行聚合并实现多级监督，最终视差回归得到预测视差图。PSMNet的特征提取网络相对复杂，计算成本高，且三次编码-解码结构始终对完整视差范围的代价卷进行聚合，造成计算冗余。基于上述问题，本文提出一种改进后的立体匹配网络CTFNet，其网络结构如图1(b)所示，由特征提取、初始视差图预测和视差优化3部分构成。

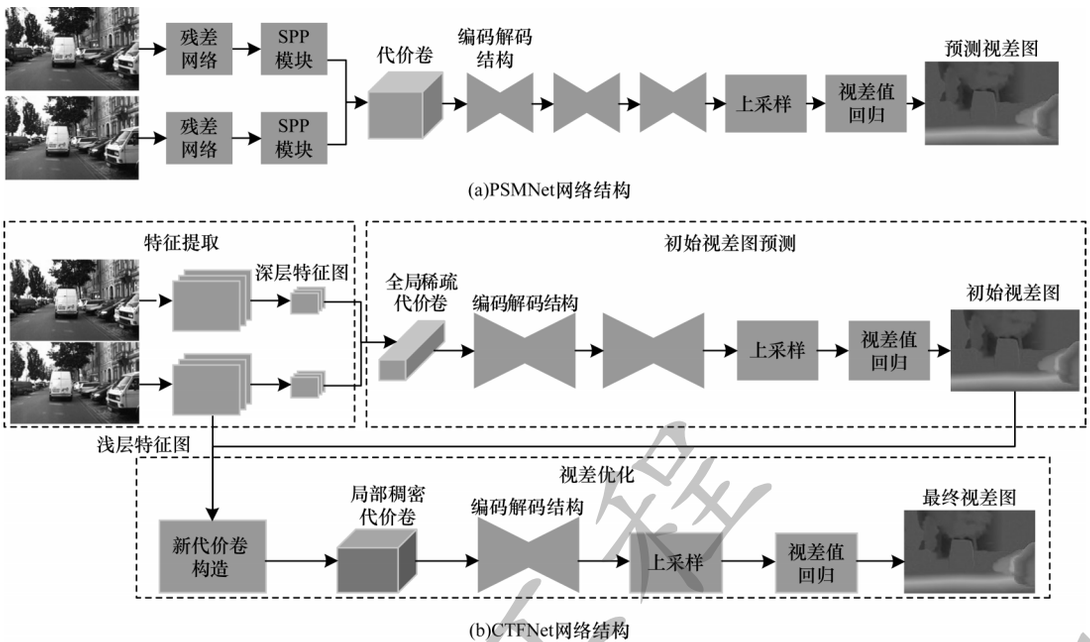


图1 PSMNet与CTFNet网络结构对比

Fig.1 Comparison of PSMNet CTFNet network structures

1.1 特征提取

相对于目标检测、语义分割等视觉任务，立体匹配对于特征的抽象程度要求相对较低，而且更加注重全局信息与局部细节信息的结合，全局信息有利于保证前景物体匹配精度和视差的连续性，局部细节信息对于提高不适应区域如小物体、边缘等区域的匹配精度具有重要作用。因此与PSMNet网络所采用的复杂特征提取结构不同，本文采用深浅层特征两阶段输出的特征提取网络。具体来说，由浅层特征提取结构和深层特征提取结构组成，如图2所示。

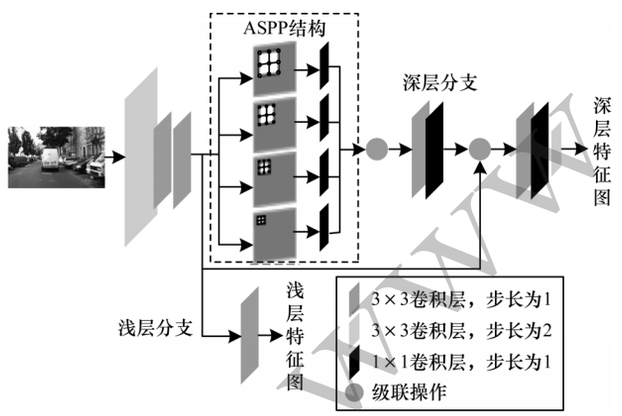


图2 特征提取网络结构

Fig.2 Structure of feature extraction network

浅层特征提取结构由3个卷积核尺寸为3×3的卷积层构成，每个卷积层后都跟随着批标准化层和ReLU激活函数层，其中第1个卷积层步长为2，将图像下采样尺寸设置为原图尺寸的1/2，其他卷积层步长均为1，以保留更多的空间细节信息。将浅层特征提取模块输出的初始特征图用来继续提取深层多尺度特征，同时也经过卷积核尺寸为3×3卷积层对通道进行调整，并将得到的浅层特征图输出到视差优化阶段，从而构建局部稠密代价卷。

将深层特征提取结构引入多孔空间金字塔池化(ASPP)结构[19]，以提取多尺度空间信息，并分别通过卷积核尺寸为1×1的卷积操作实现跨通道信息整合。每个卷积操作后面都同样跟随批标准化层和激活函数层，最终采用级联的方式将包含不同尺度信息的特征图级联起来。ASPP结构的使用可以保证使用较少的卷积层实现较大的感受野，有利于匹配对全局信息要求较高的前景物体。级联后的多尺度特征经过卷积核尺寸为3×3及1×1的卷积层后与初始特征图级联，再通过卷积核尺寸为3×3及1×1的卷积层以及一个卷积核大小为3、步长为2的卷积层下采样得到最终输出的1/4原图大小的深层特征图，并用作初始视差图预测。

1.2 初始视差图预测

经过共享权重的特征提取网络得到左右特征图后，本文将每个潜在视差值下的左图特征和对应右图下的特征级联起来，封装成一个4维的匹配代价卷。针对4维代价卷，本文采用3D卷积来聚合上下文信息并通过编码-解码结构聚合匹配代价卷。如图3所示，通过4个3D卷积层对匹配代价卷进行初步的代价聚合，为了补充浅层特征信息，将第2次卷积的结果与第4次卷积的结果进行跳跃连接。接着，采用基于3D卷积的编码-解码结构对代价卷进行聚合。编码-解码结构如图3中虚线框所示，编码与解码阶段分别使用2个步长为2的3D卷积与3D反卷积进行下/上采样，提高对全局信息的利用程度并降低计算量。为弥补上下采样引起的局部上下文

信息的损失，在反卷积时将编码阶段对应尺寸大小的代价卷通过跳跃连接与解码阶段的代价卷进行连接。本文采用2个编码-解码结构串联使用进行代价聚合和多级监督，将每个编码-解码结构输出的匹配代价卷，通过线性插值的方式上采样到原图尺寸，用作视差回归。

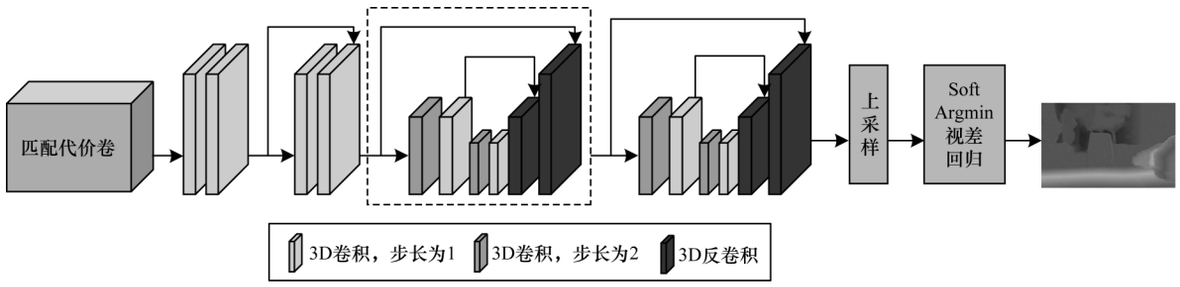


图3　初始视差图预测网络结构

Fig.3　Structure of initial disparity map prediction network

本文采用完全可微的Soft Argmin[10]操作进行视差回归，将预测的匹配代价 C_d 取负，把匹配代价转化为匹配可能性（匹配代价越高，可能性越低），然后使用softmax操作进行归一化，输出每个像素在不同视差值 d 下的概率，最终基于概率对视差进行加权求和得到预测视差值 $\hat{d}$，如式（1）所示：

$$\hat{d}=\sum_{d=0}^{D_{\max}} d\times \mathrm{softmax}\left(-C_d\right) \tag{1}$$

其中：d 表示预测视差值；$D_{\max}$ 表示最大视差；C_d 表示在视差 d 下的匹配代价；softmax(·)表示softmax操作，其数学表达式如下：

$$\mathrm{softmax}\left(z_i\right)=\frac{\mathrm{e}^{z_i}}{\sum_{c=1}^{C}\mathrm{e}^{z_c}},\ i=1,2,\cdots,C \tag{2}$$

1.3　视差优化

网络采用左右特征图构造全局稀疏视差值（0，4，…，$D_{\max}$，$D_{\max}$=192）下的代价卷，经过2个沙漏结构聚合后通过上采样和视差回归得到与原图分辨率相同的初始视差图。代价聚合过程中，用于视差回归的代价卷需要通过插值的方式恢复到原图尺寸和完整稠密视差值范围（0，1，…，$D_{\max}$，$D_{\max}$=192），这就使最终的视差结果在目标边缘、小物体等细节区域引入大量误差。同时，由于特征图经过多次编码-解码结构，其特征随着网络的加深不断抽象，最终的结果在前景物体的匹配精度指标上表现良好，但是针对背景物体或小物体，由于其对上下文细节信息要求较高，因此匹配误差率会显著增大。基于上述问题，本文提出基于浅层特征和局部稠密代价卷的视差优化模块。

为降低反复上下采样对局部细节信息造成的损失，本文在视差优化阶段采用特征提取阶段的1/2原图尺寸的浅层特征图构造新代价卷。为减少冗余计算，提高视差回归精度，本文利用初始视差图预测阶段输出的原图尺寸的视差图构造局部稠密代价卷，仅针对预测视差值附近的视差范围求解详细的概率分布。基于初始预测的视差图，本文将每个像素的预测视差值线性扩展为其邻域内的 $2n$ 个视差从而构造预测视差卷，其中 n 为超参数。假设某像素点初始预测视差为 d'，则以 $[d'-n,d'+n]$ 作为该点的局部视差范围，并限制其不超出 $[0,D_{\max}]$。然后将该视差范围平均划分为 $2n$ 个视差值作为候选视差。由于初始视差值是亚像素级的，而传统构造方式只能对视差值为整数的情况进行代价卷构造，因此本文采用一种新的代价卷构造方式，如图4所示（彩色效果见《计算机工程》官网HTML版）。

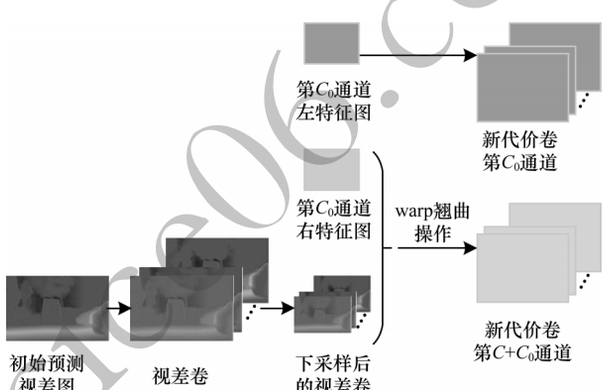


图4　新代价卷构造方式示意图

Fig.4　Schematic diagram of construction mode of new price volume

如图4所示，假设左右特征图组通道数均为 C，以同为 C_0 通道的左右特征图为例，由于特征图尺寸为原图的1/2，因此首先需要将视差卷下采样至原图的1/2尺寸，同时所有视差值相应除以2。然后将 C_0 通道的左特征图复制 $2n$ 次作为第 C_0 通道的代价卷，将右特征图基于视差卷进行warp翘曲操作[20]得到第 $C+C_0$ 通道的代价卷。其中warp翘曲操作如图5所示。首先，根据视差图计算得到1个与左特征图尺寸相同的坐标网格，网格中每一点 (x,y) 的值为左特征图中 (x,y) 处的像素点在右特征图中的对应匹配点的坐标 $(x-d,y)$，其中 d 代表该点候选视差值。然后，利用坐标网格将右特征图中匹配点 $(x-d,y)$ 处的像素值全部填充到左特征图的 (x,y) 处，从而产生1张新的特征图。由于视差值 d 为亚像素级，则计算得到的坐标 $(x-d,y)$ 不一定是整数值，因此要用插值的方式从 $(x-d,y)$ 邻域的像素值得到 $(x-d,y)$ 处的像素值。将第 C_0 通道的右特征图基于所有候选视差产

生的$2n$个特征图作为第$C+C_0$通道的代价卷。最终对所有特征通道的特征图构造代价卷即可得到1个$\left(2C, 2n, \frac{H}{2}, \frac{W}{2}\right)$的4维代价卷，其中$H$和$W$分别代表原图尺寸的高和宽。

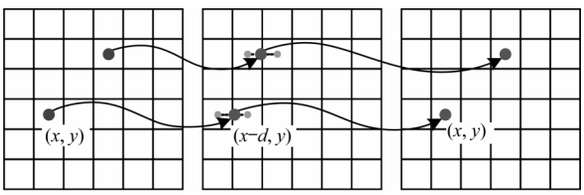


图5 warp翘曲操作示意图

Fig.5 Sketch mapnew of warp operation

为保存细节信息，本文仅使用1次编码-解码结构对新代价卷进行聚合，聚合后的代价卷仅需通过1次上采样即可恢复成原图尺寸。对于聚合后的代价卷，本文同样采用Soft Argmin操作进行视差回归，输出每个像素在不同视差值d下的概率，但由于此时代价卷代表的不再是全局范围的视差，而是基于初始视差图构造的邻域局部范围内的视差。因此，本文利用事先构造的视差卷，针对每个像素，仅在预测视差值d'邻域范围$[d'-n, d'+n]$内进行视差回归，达到视差优化的目的。

1.4 损失函数

本文的损失函数由2部分构成，如式(3)所示，一部分是基于多级监督的视差损失$L_{\text{Loss}_{\text{disp}}}$，另一部分是初始视差图预测阶段的softmax后概率分布损失$L_{\text{Loss}_{\text{distrib}}}$。

$$L_{\text{Loss}} = L_{\text{Loss}_{\text{disp}}} + L_{\text{Loss}_{\text{distrib}}} \tag{3}$$

1.4.1 视差损失

本文采用兼具鲁棒性和稳定性的Smooth L_1函数作为网络的基础视差损失函数，如式(4)所示：

$$L\left(d, \hat{d}\right) = \frac{1}{N}\sum_{n=1}^{N} \text{Smooth L}_1\left(d_n - \hat{d}_n\right) \tag{4}$$

其中：N表示有效像素点个数；d_n表示真实视差值；$\hat{d}_n$表示预测视差值；Smooth $L_1(\cdot)$表示平滑的L1损失，其表达式如下：

$$\text{Smooth L}_1(x) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, \text{其他} \end{cases} \tag{5}$$

本文采用多级监督的训练方式，对网络每个编码-解码结构输出的代价卷进行视差回归并计算损失，最终通过加权求和的方式计算总的视差损失，从而实现视差逐级细化，如式(6)所示：

$$L_{\text{Loss}_{\text{disp}}} = \sum_{i}^{M} w_i L_i\left(d, \hat{d}\right) \tag{6}$$

其中：w_i表示不同阶段输出的视差损失的权重；M表示视差结果受监督的层级数；和参考文献[12]相同，i取3；各视差结果对应权重参数分别为$w_1=0.5, w_2=0.7, w_3=1$。

1.4.2 softmax后概率分布损失

由于新代价卷的构造对初始预测的视差图精度提出较高的要求，即需要初始预测视差图中每个像素的预测视差值能够在真实视差值邻域范围内。为约束初始视差图，本文引入softmax操作后的概率分布损失。匹配代价卷被用以反映候选匹配像素对之间的相似度，代价卷经过softmax操作后输出每个像素在不同视差值d下的概率$\hat{P}(d)$，其中真实视差值具有最高概率，且概率值应随与真实视差值的距离增大而迅速下降。根据该属性，本文基于真实视差值，采用高斯分布构建真实视差概率分布，对代价卷softmax后的概率分布进行监督，约束预测视差值概率在真实视差值附近成单峰分布。基于真实视差值构建的真实视差概率分布如下：

$$P(d) = \text{softmax}\left(-\frac{\left(d-d_{gt}\right)^2}{\sigma}\right) \tag{7}$$

其中：d表示候选视差值，$d \in [0, D_{\max}]$；d_{gt}表示真实视差值；σ表示方差，用来控制视差概率分布的离散程度，σ越小则视差概率分布越集中于真实视差值附近，$\sigma>0$，基于参考文献[18]，本文σ取1.2。

根据真实视差值构建真实视差概率分布$P(d)$，同时在视差预测阶段，计算softmax后的概率分布$\hat{P}(d)$，通过交叉熵定义分布损失$L_{\text{Loss}_{\text{distrib}}}$，如式(8)所示：

$$L_{\text{Loss}_{\text{distrib}}} = \frac{1}{N}\sum_{n=1}^{N} H\left(P(d), \hat{P}(d)\right) \tag{8}$$

其中：N表示有效像素点个数；d表示候选视差值；$H(\cdot)$表示交叉熵损失函数。$H(\cdot)$的表达式如式(9)所示：

$$H\left(P(d), \hat{P}(d)\right) = -\sum_{d=0}^{D_{\max}} P(d) \times \ln(\hat{P}(d)) \tag{9}$$

为防止过拟合，本文中softmax后概率分布损失仅针对初始视差预测阶段的初始视差图进行计算。

2 实验与结果分析

为测试算法的性能，本文基于PyTorch深度学习架构实现提出的CTFNet模型，使用NVIDIA 1080Ti GPU训练及测试网络，研究网络各组成部分不同参数配置对视差图预测的影响，并将其与参考算法进行比较。

2.1 数据集

采用SceneFlow数据集[9]和KITTI 2015数据集[21]对网络进行训练和测试，其中SceneFlow数据集为合成数据集，包含图像尺寸为960×540像素分辨率的立体图像对，其中35 454张用于训练，4 370张用于测试，所有图片提供稠密视差图作为真实值。KITTI 2015数据集为真实道路场景下采集的数据集，包含200张训练集图片和200张验证集图片，图像尺寸为1 240×376像素，其中训练集提供稀疏视差图作为真实值，验证集仅提供左右图像对，预测视差图精度需将图片上传至KITTI网站进行评估。对于KITTI数据集，本文随机选取训练集中的160个图像对进行训练，剩余40个图像对用于测试。

2.2　实施细节

CTFNet的训练过程包含2个步骤，首先在SceneFlow数据集上预训练模型，在输入网络之前，对每个原始图像对进行归一化处理，将图像RGB值归一化到[−1，1]区间内，并随机裁剪成512×256分辨率的图像补丁输入到网络。网络使用Adam优化器，优化参数β_1、β_2的值分别为0.90、0.99，批尺寸和最大视差($D_{\max}$)分别设置为3和192，学习率固定为0.001，训练10个周期。在得到SceneFlow数据集上的预训练模型后，利用KITTI2015数据集对模型进行优化微调，模型训练300个周期，其中前200个周期的学习率为0.001，之后学习率调整为0.000 1。

2.3　测试及评价指标

为评估网络性能，本文基于真实视差值，分别计算SceneFlow数据集的每个训练周期的终点误差及KITTI2015训练集的三像素误差。完成训练后，使用误差最低的训练参数预测KITTI2015验证集的视差图，并将结果提交至KITTI网站进行评估。

对于SceneFlow数据集，本文计算所有像素点的预测视差值与真实视差值之间的欧氏距离并求取平均值作为终点误差(End-Point Error，EPE)，误差越小则匹配精度越高。终点误差的定义如下：

$$E_{\mathrm{EPE}}=\frac{1}{N}\sum_{i\in N}\sqrt{\left(d_i-\hat{d}_i\right)^2} \tag{10}$$

其中：N表示总像素点个数；d_i表示第i个像素点处的真实视差值；$\hat{d}_i$表示第i个像素点处的预测视差值。

对于KITTI2015数据集，本文采用三像素误差(3px Error)表征匹配的准确率，三像素误差是指预测视差值与真实视差值之间差值的绝对值超过3的像素点的数量占整幅图像的比例，比例越高说明误匹配点的数量越多，匹配准确率越低。三像素误差的定义如式(11)所示：

$$\text{3px Error}=\frac{1}{N}\sum_{i\in N}\Phi\left(\left|d_i-\hat{d}_i\right|,3\right) \tag{11}$$

其中：

$$\Phi\left(p,q\right)=\begin{cases}1, & p>q\\ 0, & p\leqslant q\end{cases} \tag{12}$$

其中：N表示总像素点个数；d_i表示第i个像素点处的真实视差值；$\hat{d}_i$表示第i个像素点处的预测视差值。

2.4　实验对比

针对CTFNet各组成部分对视差图预测的影响进行研究，并测试不同网络结构及参数配置对于视差精度及运行速度的影响。本文在SceneFlow和KITTI2015数据集上评估CTFNet网络，并在最终实验中，与本文网络相似的PSMNet进行对比。分别针对特征提取结构、局部稠密代价卷、视差优化结构、softmax操作后的概率分布损失函数等进行实验，分析其对视差结果的影响。

2.4.1　特征提取结构实验

针对构造局部代价卷时所用特征图的输出位置进行实验，结果如表1所示。在表1中，特征图输出位置代表构建局部代价卷所用的浅层特征图的输出位置，其中浅层表示图2中浅层分支的最后1层卷积层，深层表示图2中特征提取网络的倒数第2层卷积层。

表1　特征图输出位置的实验结果

Table 1　Experimental results of the output location of the feature map

序号	局部代价卷所用特征图输出位置	在SceneFlow数据集上的EPE	在KITTI2015数据集上的3px Error
1	浅层	0.768	1.485
2	深层	0.873	1.646

由表1可知，使用浅层特征构造局部代价卷使视差图的误匹配率在SceneFlow数据集上降低了12.0%，在KITTI数据集上降低了9.7%，说明浅层特征能够保留更多细节信息，在视差优化过程中能够有效地改善局部细节区域的匹配结果。

本文还针对特征提取结构输出的2组特征图的尺寸大小对视差图的影响进行实验，通过添加步长为2、卷积核尺寸为3×3的卷积层配合批标准化层和ReLU激活函数层实现特征图尺寸的调节，结果如表2所示。表2中深层和浅层特征图的尺寸分别代表用来构造稀疏代价卷和稠密代价卷的特征图尺寸与原图尺寸的比例。通过对比表2中实验结果可知，不论对于构造稀疏代价卷还是稠密代价卷，用作构造代价卷的特征图尺寸越大，局部细节信息越丰富，视差估计的误差越小。但一味增大特征图尺寸会造成代价聚合阶段的计算量过大，训练无法正常进行。通过结合视差优化的方法，采用1/4原图尺寸的特征图构造初始代价卷，同时采用1/2原图尺寸的特征图构造稠密代价卷既能保证网络正常训练，也能有效提高预测视差图精度。

表2　不同特征图尺寸的实验结果

Table 2　Experimental results of different feature map sizes

序号	深层特征图的尺寸	浅层特征图的尺寸	在SceneFlow数据集上的EPE	在KITTI2015数据集上的3px Error
1	1/4	1/2	0.768	1.485
2	1/4	1/4	0.896	1.718
3	1/8	1/4	0.992	2.089

2.4.2　对局部稠密代价卷的实验

为减少冗余计算及细化视差概率计算，在视差优化阶段针对每个像素，以其初始视差邻域内的$2n$个视差值构造局部稠密代价卷，其中$2n$是需要人为确定的超参数。为实现最优化，针对此局部视差范围参数进行实验，结果如表3所示。

表3　对局部视差范围的实验结果

Table 3　Experimental results of the local disparity range

序号	视差范围($2n$)	在SSceneFlow数据集上的EPE	在KITTI2015数据集上的3px Error	运行时间/s
1	12	0.912	1.846	0.39
2	24	0.768	1.485	0.43
3	48	0.735	1.492	0.55

由表3可知,初始视差邻域范围2n对视差估计和网络运行速度有一定影响。如果用于构造局部稠密代价卷的视差邻域范围过小,则经过多次上下采样,会引入过大的误差。如果视差邻域范围过大,最终沙漏结构进行代价聚合所需的计算时间就会增加,且对于误差率的改善收效甚微。由实验结果可知,本文选择基于初始视差构造邻域范围为24的局部稠密代价卷。

2.4.3 视差优化结构实验

为验证视差优化方案的可行性,本文通过采用相同的特征提取结构,对比视差优化方案与传统的通过3个编码-解码结构直接预测视差图方案间的匹配误差率,结果如表4所示。由表4可知,使用视差优化结构相对传统方案,视差图的误匹配率在SceneFlow数据集上降低了10.3%,在KITTI数据集上降低了11.9%。由此可知,本文提出的视差优化方案对视差图预测具有一定的提升作用。

表4 对视差优化结构的实验结果

Table 4 Experimental results of the disparity optimization structure

方案	在SceneFlow数据集上的EPE	在KITTI2015数据集上的3px Error
视差优化方案	0.768	1.485
传统方案	0.857	1.687

2.4.4 softmax后概率分布损失函数实验

为探究本文损失函数的改进对视差预测结果的影响,对CTFNet网络进行测试,验证添加和去除softmax后的概率分布损失函数对网络预测精度的影响。由表5可知,添加概率分布损失函数后,网络的视差预测精度得到了一定程度的提升。

表5 损失函数的实验结果

Table 5 Experimental results of loss function

是否添加softmax后概率分布损失函数	在SceneFlow数据集上的EPE	在KITTI2015数据集上的3px Error
是	0.768	1.485
否	0.841	1.669

通过softmax操作,代价卷被计算成每个潜在视差值的概率,所有概率和为1。本文对特征不明显区域的不同像素点在视差回归过程中计算的视差概率分布进行可视化,结果如图6所示。图中横坐标表示所有潜在视差值,纵坐标表示对应预测概率,虚线表示真实视差值。由图6可知,添加softmax后的概率分布损失函数,其视差预测的概率分布会倾向于在真实视差值附近呈现单峰分布,有效降低了其他视差值的干扰,这对于部分特征不明显区域的视差预测具有良好的改善作用。

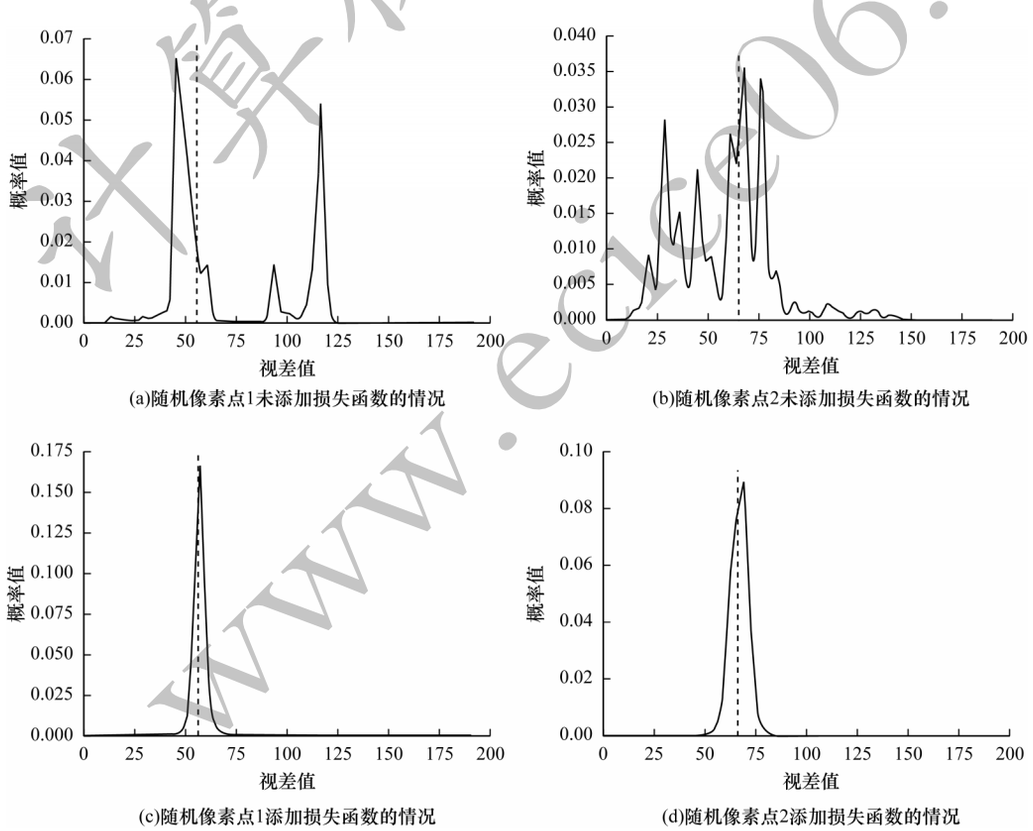


图6 视差值概率分布图

Fig.6 Probability distribution of disparity

2.5 KITTI2015排名结果

将CTFNet网络对KITTI验证集生成的视差图上传至KITTI评测网站,表6展示了KITTI2015数据集上本文网络及其他主流网络的实验结果对比。其中,D1表示视差图中误匹配点所占的比例,bg表示背景区域,fg表示前景区域,all表示整个区域。由表6可知,本文所提网络与之前的网络[10-12,15-17]相比在精度上有一定提高。与同样基于两阶段进行视差优化的CRL算法相

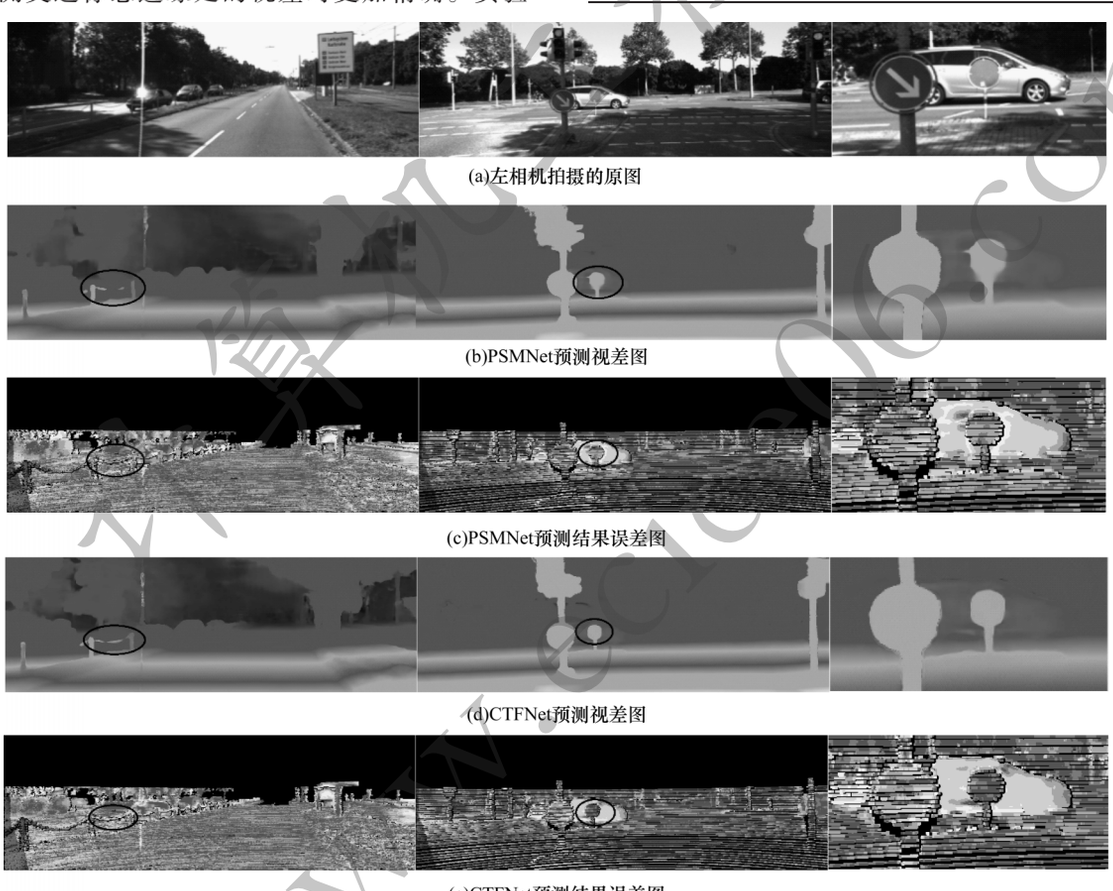
比，误匹配率降低了17.6%。与算力要求相近的PSMNet网络相比，整体的误匹配率由2.32%下降至2.20%。在运行时间方面，为保证数据的准确性，本文在Nvidia 1080Ti GPU上测试PSMNet和CTFNet网络的运行时间，PSMNet的运行时间为0.52 s，所提网络的计算时间为0.43 s，降低了约17%。

图7中第1列和第2列分别展示了本文所提CTFNet与PSMNet针对同一组图片预测视差图的对比，第3列为第4列的局部位置放大图。误差图中蓝色点表示正确匹配点，黄色点表示错误匹配点，黑色点表示忽略的点（彩色效果见《计算机工程》官网HTML版）。通过观察2种算法在图中黑色椭圆标记位置处的视差结果可以发现，与PSMNet网络相比，CTFNet网络能够准确预测图中细铁索处的视差，同时在预测交通标志边缘处的视差时更加精确。实验结果表明，通过浅层特征和视差优化的方式能够有效改善特征不明显区域的匹配结果，提高小物体及边缘等病态区域的匹配精度。

表6　KITTI2015立体匹配排名

Table 6　KITTI2015 stereo matching ranking

立体匹配网络名称	三像素误差			运行时间/s
	D1-bg /%	D1-fg /%	D1-all /%	
GCNet[10]网络	2.21	6.16	2.87	0.90
CRL[11]网络	2.48	3.59	2.67	0.47
UberATG-DRISF[15]网络	2.16	4.49	2.55	0.75
AANet[16]网络	1.99	5.39	2.55	0.062
PSMNet[12]网络	1.86	4.62	2.32	0.45
CFP-Net[17]网络	1.90	4.39	2.31	0.90
CTFNet网络	1.80	4.46	2.20	0.43



(a)左相机拍摄的原图

(b)PSMNet预测视差图

(c)PSMNet预测结果误差图

(d)CTFNet预测视差图

(e)CTFNet预测结果误差图

图7　视差图结果对比

Fig.7　Comparison of disparity map results

3　结束语

本文设计一个基于浅层特征的立体匹配网络CTFNet，通过构建稠密代价卷进行视差优化。由于深层特征网络的感受野较大，能够获取更多全局信息，从而构建全局稀疏代价卷以获取前景物体的初始视差图。浅层结构的特征提取网络减少了图像的上下采样，保留了更加完整的局部上下文信息，配合基于预测视差值构建的稠密代价卷，能够进一步细化视差。此外，softmax操作后概率分布损失函数的引入，能够对视差概率分布进行监督，有效提高了算法的鲁棒性。实验结果表明，与PSMNet网络相比，本文网络在部分病态区域如边缘及小物体处匹配效果更好，匹配精度得到一定提升。下一步将通过采用多任务网络及引入边缘检测或语义分割任务，提高算法对边缘及弱纹理区域的匹配能力，同时，还将进一步优化网络结构，降低网络参数和计算量，以实现算法在TX2等嵌入式设备上的稳定运行。

参考文献

[1] 王金鹤,车志龙,张楠,等. 基于多尺度和多层级特征融合的立体匹配算法[J]. 计算机工程,2021,47(3):243-248.
WANG J H,CHE Z L,ZHANG N,et al. Stereo matching based on multi-scale and multi-feature integration[J]. Computer Engineering,2021,47(3):243-248. (in Chinese)

[2] 赵晨园,李文新,张庆熙. 一种改进的实时半全局立体匹配算法及硬件实现[J]. 计算机工程,2021,47(9):162-170.
ZHAO C Y,LI W X,ZHANG Q X. An improved real-time semi-global stereo matching algorithm and its hardware implementation[J]. Computer Engineering,2021,47(9):162-170. (in Chinese)

[3] 陈炎,杨丽丽,王振鹏. 双目视觉的匹配算法综述[J]. 图学学报,2020,41(5):702-708.
CHEN Y,YANG L L,WANG Z P. Literature survey on stereo vision matching algorithms[J]. Journal of Graphics,2020,41(5):702-708. (in Chinese)

[4] ZBONTAR J,LECUN Y. Computing the stereo matching cost with a convolutional neural network[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2015:1592-1599.

[5] ZHANG K,LU J B,LAFRUIT G. Cross-based local stereo matching using orthogonal integral images[J]. IEEE Transactions on Circuits and Systems for Video Technology,2009,19(7):1073-1079.

[6] HIRSCHMULLER H. Accurate and efficient stereo processing by semi-global matching and mutual information[C]//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2005:807-814.

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

[8] LUO W J,SCHWING A G,URTASUN R. Efficient deep learning for stereo matching[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2016:5695-5703.

[9] 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 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2016:4040-4048.

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

[11] PANG J H,SUN W X,REN J,et al. Cascade residual learning:a two-stage convolutional neural network for stereo matching[C]//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops Venice. Washington D. C.,USA:IEEE Press,2017:878-886.

[12] CHANG J R,CHEN Y S. Pyramid stereo matching network[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2018:5410-5418.

[13] HE K M,ZHANG X Y,REN S Q,et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,37(9):1904-1916.

[14] ZHANG F H,PRISACARIU V,YANG R G,et al. GA-Net:guided aggregation net for end-to-end stereo matching[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2019:185-194.

[15] MA W C,WANG S L,HU R,et al. Deep rigid instance scene flow[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2019:3609-3617.

[16] XU H F,ZHANG J Y. AANet:adaptive aggregation network for efficient stereo matching[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2020:1956-1965.

[17] ZHU Z D,HE M Y,DAI Y C,et al. Multi-scale cross-form pyramid network for stereo matching[C]//Proceedings of the 14th IEEE Conference on Industrial Electronics and Applications. Washington D. C.,USA:IEEE Press,2019:1789-1794.

[18] ZHANG Y M,CHEN Y M,BAI X,et al. Adaptive unimodal cost volume filtering for deep stereo matching[EB/OL]. [2021-01-02]. https://www.researchgate.net/publication/335713171_Adaptive_Unimodal_Cost_Volume_Filtering_for_Deep_Stereo_Matching.

[19] CHEN L C,PAPANDREOU G,KOKKINOS I,et al. DeepLab:semantic image segmentation with deep convolutional nets,atrous convolution,and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(4):834-848.

[20] LI X T,YOU A S,ZHU Z,et al. Semantic flow for fast and accurate scene parsing[EB/OL]. [2021-01-02]. https://www.researchgate.net/publication/339471607_Semantic_Flow_for_Fast_and_Accurate_Scene_Parsing.

[21] MENZE M,GEIGER A. Object scene flow for autonomous vehicles[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2015:3061-3070.

编辑 赖玉玲