DOI: 10.13382/j.jemi.B2306470

全距离深度平衡立体匹配网络*

覃业宝[1,2]　孙　炜[1,3]　范诗萌[1,2]　张　星[1,3]　刘　剑[1,2]

(1. 湖南大学电气与信息工程学院　长沙　410082;2. 湖南大学汽车车身先进设计制造国家重点实验室　长沙　410082;
3. 湖南大学深圳研究院　深圳　518000)

摘　要:针对当前视差估计网络在将视差转换成深度时,存在深度精度受相机参数影响,且在远距离处产生深度精度急剧下降的问题,提出一种全距离深度平衡立体匹配网络(FRDBNet)。首先构建深度代价体,使网络学习到全距离深度的概率分布,进行深度回归直接生成深度;然后采用视差与深度损失融合的训练策略使网络同时关注远中近三分段全距离的深度估计;最后,基于初始视差右图对应点 7 邻域特征设计视差优化模块进一步提高网络的深度估计精度。在大型真实驾驶场景 DrivingStereo 数据集上的实验表明,针对全距离[1,100]m 的深度估计,FRDBNet 在[1,30]m 近距离、[30,60]m 中距离和[60,100]m 远距离处深度精度相比 CVPR2022 性能表现优越的 ACVNet 分别提高 10.38%、15.11%和 20.35%,达到了良好的深度精度平衡。

关键词: 立体匹配;深度代价体;视差与深度损失融合;7 邻域特征;视差优化;深度精度

中图分类号: TP391;TN98　　**文献标识码**: A　　**国家标准学科分类代码**: 510.40;520.20

Full range depth balanced stereo matching network

Qin Yebao[1,2]　Sun Wei[1,3]　Fan Shimeng[1,2]　Zhang Xing[1,3]　Liu Jian[1,2]

(1. School of Electrical and Information Engineering, Hunan University, Changsha 410082, China;
2. State Key Laboratory of Advanced Vehicle Design and Manufacturing, Hunan University, Changsha 410082, China;
3. Shenzhen Research Institute, Hunan University, Shenzhen 518000, China)

Abstract: In view of the problem that depth accuracy is affected by camera parameters when disparity is converted into depth in the current disparity estimation network, and depth accuracy decreases sharply at long distance, a full range depth balanced stereo matching network (FRDBNet) is proposed. Firstly, the depth cost volume is constructed to make the network learn the probability distribution of the full distance depth, and the depth is directly generated by depth regression. Then, the training strategy of disparity and depth loss fusion is used to make the network pay attention to the depth estimation of the long, middle and near three segments distance at the same time. Finally, a disparity optimization module is designed based on the seven neighborhood features corresponding to the original disparity right map to further improve the depth estimation accuracy of the network. Experiments on the DrivingStereo dataset of large real-world driving scenarios show that for the full distance[1,100]m depth estimation, the depth accuracy of FRDBNet at[1,30]m short distance,[30,60]m middle distance and[60,100]m long distance is 10.38%, 15.11% and 20.35% higher than that of ACVNet with superior performance of CVPR2022, respectively, achieving a good balance of depth accuracy.

Keywords: stereo matching; depth cost volume; disparity and depth loss fusion; seven neighborhood features; disparity optimization; depth accuracy

收稿日期: 2023-04-25　Received Date: 2023-04-25

*基金项目:国家自然科学基金(U22A2059)、深圳科技计划项目(2021Szvup035)、湖南大学汽车车身先进设计制造国家重点实验室自主研究项目、电子制造业智能机器人技术湖南省重点实验室开放课题项目资助

0　引　言

三维空间中目标物体的深度信息对于:3D 重建[1]、自动驾驶[1-2]、自主导航[3]和虚拟现实[4]等应用至关重要。不同于激光雷达等其他高成本且大体积的测距方案,基于双目立体匹配的深度估计有着极高的成本优势和易于安装的部署优点。通常,给定一对极线校正后的标准立体图像,找到同名点在左图和右图相对水平方向的位移,即视差,是立体匹配的目标。

立体匹配方法可分为传统立体匹配算法和基于学习的立体匹配算法。传统立体匹配算法计算成本低,但精度粗糙,而基于学习的立体匹配算法有着全面优于前者的精度优势,但需要更大的计算开销。

Hirschmuller[5-6]提出的基于互信息的半全局立体匹配算法 SGM、Mei 等[7]提出的结合颜色差和 Census 变换[8]信息来计算代价的立体匹配方法 AD-Census 和 Bleyer 等[9]提出的基于倾斜支持窗的立体匹配算法 PatchMatch Stereo 是目前得以广泛应用的 3 个传统立体匹配算法。这些方法都是通过计算每个像素在所有视差水平上的相似度,然后采用赢者通吃(WTA)的策略选择相似度最大的视差作为最终的估计视差。

传统的立体匹配方法有着速度快,部署简单的优势,但匹配精度较低。与此同时,神经网络在计算机视觉其他任务上的成功应用[10],让一些学者看到了神经网络在立体匹配上的应用前景,认为神经网络是拥有从大量的立体图像数据集中学习到寻找左右同名点的能力。

Mayer 等[11]提出的 DispNet 通过构建相关代价体来计算每个像素在所有视差水平上的相似度,而 Chang 等[12]提出的 PSMNet 则通过级联代价体计算相似度。相关代价体是一种十分有效的特征相似度的表示方法,但丢失了很多有用信息,而级联代价体包含了所有可用信息,但却无法有效表示特征相似度。为此,Guo 等[13]结合两者的优势构建分组相关体。Xu 等[14]提出的 ACVNet 则利用相关代价体来计算得到级联代价体的视差权重,让高置信度的视差特征得以充分表达,并抑制低置信度的视差特征,达到了十分优越的性能。

这些方法同其他网络[15-23]都是通过计算每个像素在所有视差水平上的相似度,然后通过视差回归[24]的方式生成最终估计视差。

以上立体匹配方法都有一个共同点:它们是视差估计方法,都追求着理想的视差精度。正因此,它们都存在着同样的问题:如自动驾驶[1-2]等下游任务需要的是物体深度,这需要将这些方法估计得到的视差转换成实际物体深度。在将视差转换成深度的过程中,深度精度受到相机参数的影响,同时在远距离处急剧下降。为此,本文提出一种全距离深度平衡立体匹配网络(full range depth balanced stereo matching network, FRDBNet)来解决这个问题。

本文的主要贡献有:

1)提出一种全距离深度平衡立体匹配网络,以解决当前视差估计网络存在估计视差无法直接应用于下游任务,且深度精度受相机参数影响,在远距离处急剧下降的问题,其达到了很好的深度估计性能。

2)为解决 1)中问题,提供了构建深度代价体和进行深度回归的有效思路。

3)通过视差与深度损失融合的训练策略,使模型同时关注远中近三分段全距离区域的立体匹配。

4)设计视差优化模块有效提高模型预测的深度精度。

1　FRDBNet 模型设计

FRDBNet 输入一对立体图像 I_{left} 和 I_{right},通过立体几何和语义信息对图像中的物体进行深度估计,输出左图像的稠密深度图 $I_{l,depth}$。FRDBNet 基本结构如图 1 所示。

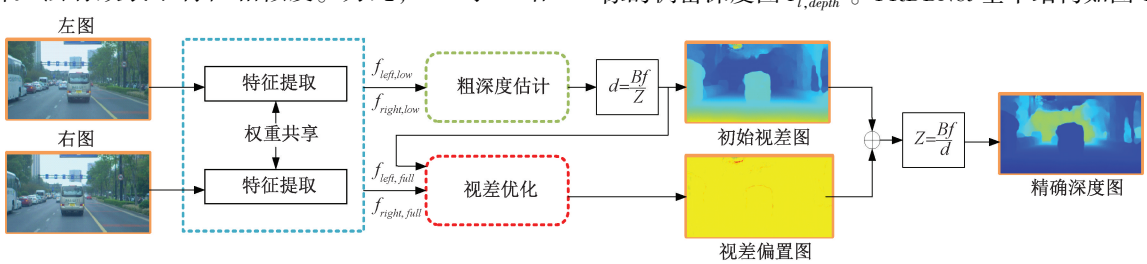


图 1　全距离深度平衡立体匹配网络结构

Fig. 1　The structure of full range depth balance stereo matching network

首先,通过特征提取模块提取到两个不同分辨率的左右图像特征,即 1/4 尺度的 $f_{left,low}$,$f_{right,low}$ 和全尺度的 $f_{left,full}$,$f_{right,full}$。然后基于 $f_{left,low}$ 和 $f_{right,low}$,通过粗深度估计模块构建深度代价体,将深度代价体聚合,深度回归得到初始深度图。最后,先将初始深度图转换成视差图,而后将 $f_{left,full}$,$f_{right,full}$ 和初始视差图一同送入视差优化模块,

学习得到视差偏置。初始视差与视差偏置相加得到精确视差,然后通过三角测量公式转换得到精确深度图。

值得注意的是,FRDBNet 基于深度采样,直接输出深度图,而不是基于视差采样[12-13],输出视差图。

1.1 特征提取

特征提取模块从原始输入图像 I_{left} 和 I_{right} 提取 1/4 尺度的特征 $f_{left,low}$、$f_{right,low}$ 用于粗深度估计阶段,提取全尺度的特征 $f_{left,full}$、$f_{right,full}$ 用于视差优化阶段。图 2 所示为 FRDBNet 中特征提取模块的结构。

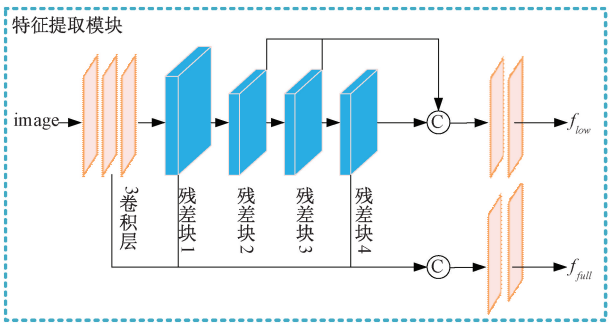


图 2 特征提取模块结构

Fig. 2 The structure of feature extraction module

输入图像首先通过 3 卷积层提取浅层信息,然后依次通过残差块 1~4,其中残差块 1~4 分别包含 3、16、3 和 3 个基本残差单元。将残差块 2~4 中同为 1/4 尺度的特征级联之后,通过包含两层 2D 卷积的 1/4 尺度卷积层,输出 1/4 尺度 32 通道的特征 f_{low}。然后将 3 卷积层输出的 1/2 尺度和残差块 1 与 4 中同为 1/4 尺度的特征反卷积到全尺度之后进行级联,而后通过包含两层 2D 卷积的全尺度卷积层作为全尺度特征 f_{full} 输出,其中 f_{full} 有 32 个通道。

1.2 粗深度估计

粗深度估计阶段主要解决当前视差估计网络普遍存在的两个问题:

1)基于视差采样构建视差代价体,存在图像远距离处深度采样分辨率低的问题;

2)估计视差需要转换成深度,才能应用于其他下游任务,且深度精度受相机参数的影响,并在远距离处发生深度精度急剧下降的问题。

图 3 是粗深度估计模块的结构图。为了解决问题 1),不同于构建视差代价体[12-13],本文基于深度采样来构建深度代价体。

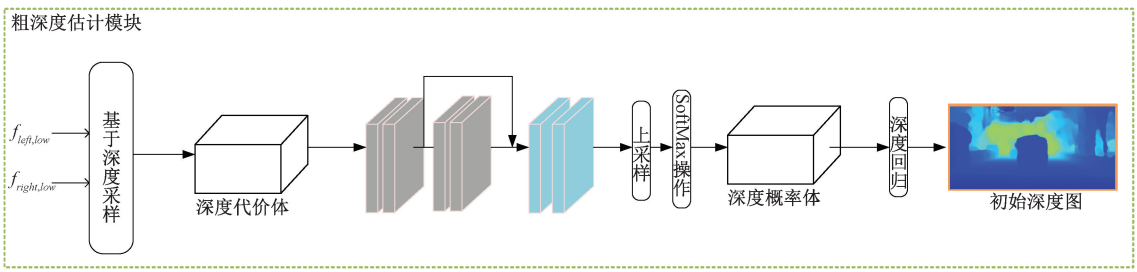


图 3 粗深度估计模块结构

Fig. 3 The structure of coarse depth estimation module

图 4 为目前众多立体匹配网络[11-23]基于视差采样构建视差代价体的示意图。通常会先设置一个最大视差 $d_{\max}$,然后对于左图中的每个像素点 (x_l, y_l) 沿着视差维度 0 ~ $d_{\max}$ 构建视差代价体。这种作法保证固定的视差采样分辨率,但如图 5 所示,视差从 6 跳变到 7 只改变了一个视差单位,对于不同的相机参数 B 和 f,真实深度跳变量不同且与视差跳变量成反比,即在将视差转换成深度时,深度精度受到了相机参数的影响,同时观察曲线 2($Z=(0.54\times1\,003)/d$),在近距离处,视差只跳变一个单位,而真实深度却跳变近 13 m,这是由于深度与视差成反比,远距离处视差跳变一个单位足以引起很大的深度跳变,即深度采样分辨率在远距离处会急剧下降,且随着距离的不断增大,这种现象越来越明显。

图 6 为本文基于深度采样来构建深度代价体的示意

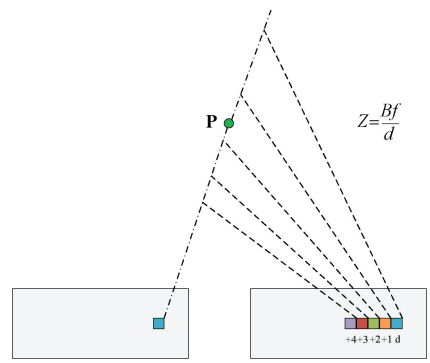


图 4 构建视差代价体示意图

Fig. 4 The schematic of disparity cost volume construction

图。Yang 等[25]为了计算距离注意指标,通过将视差图转

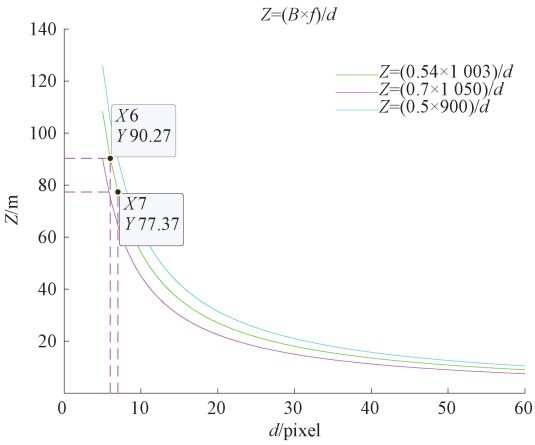


图 5　三角测量公式示意图

Fig. 5　The schematic of triangulation formula

换成深度图,然后按照一定的深度采样间隔去计算每个分段距离内的深度绝对相对误差。受其启发,不同于视差代价体的构建,本文设定最大深度为 $Z_{\max}$m,每隔固定物理深度 ΔZm 在 0 ~ $Z_{\max}$ 内进行深度采样来构建代价体。这种作法能够强制网络在全距离范围内拥有一个固定的深度采样分辨率,而不受距离变化的影响。值得注意的是:Yang 等[25]是在视差估计完成之后才进行深度采样的,网络还是基于视差代价体来直接估计视差,而本文的深度采样是在构建深度代价体时进行的。构建完成的深度代价体通过 3D 聚合之后,进行深度回归直接生成深度图。

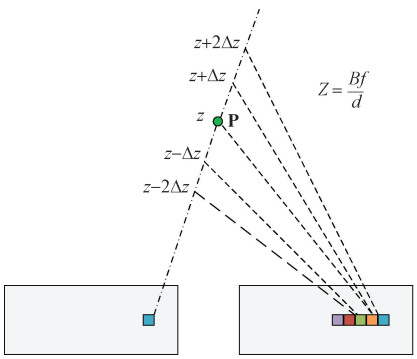


图 6　构建深度代价体示意图

Fig. 6　The schematic of depth cost volume construction

首先输入 1/4 尺度的左右图像特征 $f_{left,low}$,$f_{right,low}$。第 i 采样深度 z_i 由式(1)计算:

$$z_i = Z_{\min} + i \cdot \Delta Z \tag{1}$$

其中,$Z_{\min}$ 为最小深度,一般为 0,$i \in \{0,1,\cdots,N\}$,N 为深度采样点数,由式(2)计算得到:

$$N = \frac{Z_{\max} - Z_{\min}}{\Delta Z} \tag{2}$$

然后,由于数字图像的本质特征,只能以整像素单位对图像进行处理,即只能在视差水平上进行左图和右图特征的拼接,所以需要将深度 z_i 转换成邻近的整像素视差 $d_i = round((Bf)/z_i)$,再构建代价体。另外,这个做法能够显式地将视差与深度的三角测量公式引入模型当中,使模型学习到视差与深度的转换规律,从而很大程度上减弱了深度精度受相机参数的影响。但是,由于在远距离处,单位深度往往对应着亚像素的视差,直接将深度转换成邻近的整像素视差会引入新的深度损失误差,故下节本文设计了一种视差优化结构来对这一部分的误差进行补偿并进一步提高初始估计视差。

如果事先知道双目立体相机的基线 B 和焦距 f,左图像素点 (x_l,y_l) 的第 i 匹配候选对象在右图中的位置 $Position_i^r(x_l,y_l)$,由以式(3)计算:

$$Position_i^r(x_l,y_l) = (x_l - d_i, y_l) \tag{3}$$

最后,基于深度采样的 4 D ($2C \times N \times H \times W$) 代价体 CV_c,由式(4)计算:

$$CV_c(x_l,y_l,i) = f_{left,low}(x_l,y_l) \,\|\, f_{right,low}(Position_i^r(x_l, y_l)) \tag{4}$$

其中,‖ 表示级联操作。如图 3 所示,CV_c 经过相关 3 D 卷积之后,学习到左右图像对应点的相似关系,之后从 1/4 尺度上采样到全尺度,进行 *SoftMax* 操作之后,变成全尺度 3 D ($N \times H \times W$) 的深度概率体 P_c 进行深度回归得到初始深度图。

为了解决问题 2),本文不采用视差回归[25]的方式计算视差,而是通过深度回归直接计算深度,让网络隐式地学习到视差与深度之间的转换关系,可有效缓解将视差转换成深度应用于下游任务时,深度精度受相机参数影响,在远距离处急剧下降的问题。不同于视差回归,本文对像素点 (x,y) 进行深度回归得到预测深度,如式(5)所示:

$$D_c(x,y) = \sum_{i=0}^{N} (Z_{\min} + i \cdot \Delta Z) \cdot P_c(x,y,i) \tag{5}$$

1.3　视差优化

由粗深度估计阶段已经得到了较为粗略的深度图,但因在构建深度代价体时,直接将深度转换成邻近的整像素视差会引入新的深度损失误差,故本节设计了一种视差优化结构以补偿这种损失,并进一步细化深度估计结果。图 7 为本文的视差优化模块结构图。从视差优化模块中得到初始视差的偏置量,其表示着每个像素点的视差修正值。

首先将初始深度图转换成初始视差图,同 $f_{left,full}$,$f_{right,full}$ 送入视差优化模块。在粗深度估计之后,大部分像素点的视差绝对误差在 3 pixels 之内,所以本文采用初始视差右图对应点的 7 邻域像素特征来对初始视差进行修正。

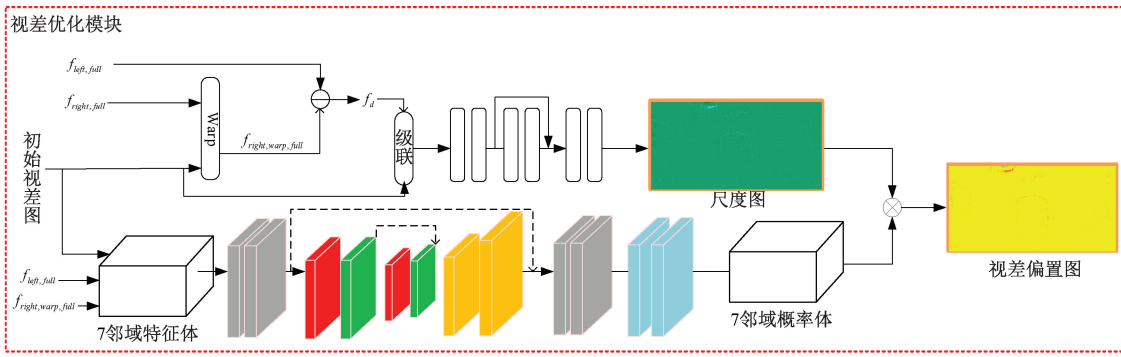


图 7 视差优化模块结构

Fig. 7 The structure of disparity refine module

根据左图像的初始视差图将右图像全尺度特征 $f_{right,full}$ 进行反变换[26]后与左图像全尺度特征 $f_{left,full}$ 取绝对误差,得到重构特征 f_d [20],其能够表征初始视差的估计准确度。

当重构特征 f_d 较大时,初始视差偏离真实视差较远,应给予对应像素较大的修正范围,反之,则给予较小的修正范围,同时初始视差较小时对应着较远的距离,不宜采用较大的修正范围,不然会引起很大的深度波动,使得估计深度更加偏离正确深度值,而需采取较小的修正范围,若初始视差较大时对应着较近的区域,则有着较大的视差容错空间,可以采用较大的修正范围。故本文将重构特征 f_d 和初始视差图级联之后,经过多层卷积得到每个像素点的视差修正范围,即视差尺度图 $d_s(\cdot)$ 。

此外,当右图对应点与左图像素点特征越相似时,这个对应点越有可能是正确的视差点,应给予较大的修正概率,反之给予较小的修正概率。故如图 8 所示,本文基于初始视差根据全尺度左右图特征 $f_{left,full}$ 和 $f_{right,full}$ 来构建 7 邻域特征体,并通过多层卷积使网络学习到初始视差相对于右图对应点 7 邻域的修正概率。对于左图全尺度像素点 (x_l,y_l) 的初始视差 d ,右图对应 7 邻域像素点 (x_r,y_r) ,由式(6)计算:

$$(x_r,y_r)=(x_l-(d+i),y_l) \tag{6}$$

其中,$i\in\{+3,+2,+1,0,-1,-2,-3\}$ 。基于全尺度左图特征 $f_{left,full}$ 和右图特征 $f_{right,full}$,构建 $(4C,7,H,W)$ 7 邻域特征体,其中前 $2C$ 通道是级联特征[12],包含初始视差右图对应点 7 邻域的丰富信息,中间 C 通道是重构特征[20],表示着初始视差的重构误差,能够衡量初始视差的准确度,最后 C 通道是相关特征[11],表示着初始视差右图对应点 7 邻域的相关特征。如图 7 所示,将 7 邻域特征体通过两层 3 D 卷积之后,通过一层堆叠沙漏模块[13],最后通过 4 层 3D 卷积,得到大小为 $(7,H,W)$ 7 邻域修正概率体 $P_{refine}(\cdot)$ 。

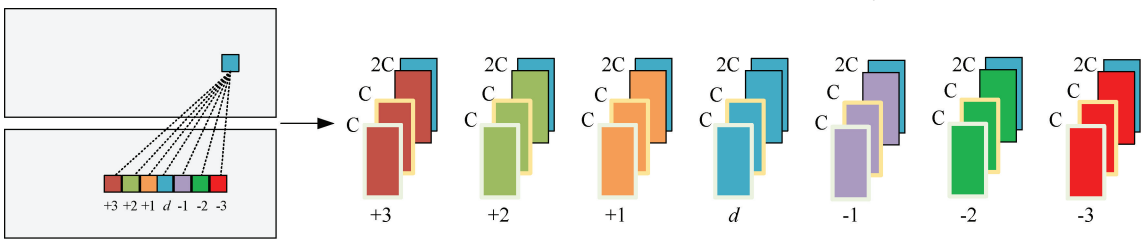


图 8 7 邻域特征体构建示意图

Fig. 8 The schematic of seven neighborhood feature volume construction

最后,通过式(7)算得到初始视差的修正偏置量:

$$offdisp(x,y)=d_s(x,y)\sum_{i=3}^{-3} i\times P_{refine}(x,y,i) \tag{7}$$

通过以式(8)、(9)得到最终的精确深度 Z :

$$D=d+\Delta d \tag{8}$$

$$Z=\frac{Bf}{D} \tag{9}$$

其中,$d,\Delta d$ 分别为初始视差和修正量,B,f 分别为双目立体系统的相机基线和焦距。

1.4 视差与深度损失融合

基于视差损失的网络往往更加关注大视差近距离的立体匹配,而基于深度损失的网络则更加关注小视差远距离的立体匹配。受 Chuah 等[27]启发,本文设计了一种视差和深度损失融合的损失函数来训练 FRDBNet,让网

络同时关注远中近三分段全距离的立体匹配。式(10)为本文采用的损失函数。

$$Loss = \lambda \cdot loss_{disp} + (1 - \lambda) \cdot loss_{depth} \tag{10}$$

$$loss_{disp} = \sum_{i=0}^{1} \alpha_i \cdot Smooth_{L1}(d_i - d_{gt}) \tag{11}$$

$$loss_{depth} = \sum_{i=0}^{1} \beta_i \cdot Smooth_{L1}(z_i - z_{gt}) \tag{12}$$

式中：λ 表示视差损失的权重，α_0,α_1 分别表示初始视差和精确视差的损失权重，β_0,β_1 分别表示初始深度和精确深度的损失权重。$Smooth_{L1}$ 表示平滑 $L1$ 损失函数[28]，由式(13)计算得到：

$$Smooth_{L1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{其他} \end{cases} \tag{13}$$

2　实验及结果分析

本文在公开数据集 Scene Flow[11] 和 DrivingStereo[25] 上进行实验分析。

不同于其他视差估计网络[12-13]在如 EPE[29]、D1 误差[30]等视差指标上追求高精度。本文更加关注深度指标，即期待网络在近距离处能够拥有足够低且安全的深度误差，而随着距离的不断增加，多数应用对深度误差的容忍程度上升，允许深度误差分段上升，故本文采用分段距离平均深度绝对误差作为评价指标。平均深度绝对误差(mean depth absolute error，MDAE)由式(14)计算：

$$MDAE = \frac{1}{N} \cdot |z^* - z_{gt}| \tag{14}$$

其中，z^*、z_{gt} 分别为预测深度和真实深度值，N 为总像素点数。

以下所有列表中的数值都为特定距离范围内的平均深度绝对误差，这些数值越小，说明模型在对应距离范围内的深度精度越高。

2.1　数据集与实验细节

FRDBNet 应用 Pytorch 框架实现，使用两个 Nvidia 3090 GPU 进行训练，使用一个 Nvidia 3090 GPU 进行测试。采用 Adam[31] 优化器进行优化且设置 $\beta_1 = 0.99$，$\beta_2 = 0.999$。在训练阶段，将图像随机裁剪成 512×256 的大小。

1) Scene Flow 数据集。该数据集是一个虚拟合成的大型数据集，包含 34 454 对训练图像和 4 370 对测试图像，并且提供精确且稠密分辨率为 960×540 的视差图。在训练前，本文先将视差图转换成深度图，分 3 阶段训练。第 1 阶段只训练粗深度估计模块，第 2 阶段只训练视差优化模块，第 3 阶段训练整个网络。3 个阶段分别训练 60、60、80 代。

2) DrivingStereo 数据集。该数据集是一个大型真实驾驶场景的数据集，包含 174 437 对训练图像和 7 751 对测试图像。训练集只提供半分辨率 881×400 的图像，测试集则提供全分辨 1 762×800 的图像和半分辨率 881×400 的图像。由于真实视差和深度难以获取，DrivingStereo 只提供了稀疏的视差图和深度图，而本文只用到深度图。同 Scene Flow 一样，分 3 阶段训练，3 个阶段分别训练 6、7、8 代。

2.2　性能比较与分析

本文将 FRDBNet 与 PSMNet[12]、GwcNet[13] 和 CVPR2022 性能最好的视差估计网络 ACVNet[14] 在 Scene Flow 和 DrivingStereo 数据集上的进行性能比较和分析。

如表 1 和 2 所示，所有模型在[1,20] m 近距离的深度精度极高，但随着距离的不断增大，深度精度不断下降，这是符合先验知识的。对于视差估计模型，因为其深度受相机参数影响，且与估计视差成反比，其深度精度必然会随着距离的增长而不断下降，且距离越远，下降幅度越大。对于 FRDBNet，因为远距离的物体在图像中占据着极小的区域，这势必会增加对远距离物体的匹配难度，导致远距离的深度精度下降，但如图 9 所示，在[1,30] m 近距离，所有模型的平均深度绝对误差 MDAE 十分接近，但随着距离的不断增大，FRDBNet 脱颖而出，在 MDAE 深度指标上，FRDBNet 与其他模型拉开的差距越来越大。并且在深度精度上，FRDBNet 全面优于 CVPR2022 的表现最好的视差估计模型 ACVNet。

如表 1 所示，在虚拟合成 Scene Flow 数据集上，FRDBNet 除了在[1,20] m 距离范围内相比 ACVNet 具有轻微的深度精度下降外，在[20,100] m 距离范围内的深度精度全面优于其他模型。在[1,20] m 距离范围导致模型精度下降的原因是因为：FRDBNet 采用了以 m 为单位的深度间隔进行采样，相比于单像素单位进行视差采样，这无疑会降低如[1,20] m 较近区域的视差采样分辨率。但在[20,100] m 取得了良好的估计精度则说明 FRDBNet 在较少牺牲近距离的精度，并且不会对真实驾驶场景造成很大的安全问题下，在[1,100] m 全距离范围取得了较好的深度精度平衡。

而如表 2 所示，在大型驾驶场景 DrivingStereo 数据集上，FRDBNet 全面优于 ACVNet，且深度精度在[1,30] m 近距离、[30,60] m 中距离和[60,100] 远距离处分别提高 10.38%、15.11% 和 20.35%，这呈现出随着距离的增大而性能表现越好的趋势，表明 FRDBNet 在较远距离，拥有着相比较近距离更加优越的性能，且其在真实驾驶场景下表现比虚拟场景更加优越，这利于真实场景下的应用。

值得注意的是，所有模型在 Scene Flow 数据集上[1,100] m 的平均深度绝对误差要远小于 DrivingStereo 数据集上的结果，这是因为 Scene Flow 数据集深度分布不均

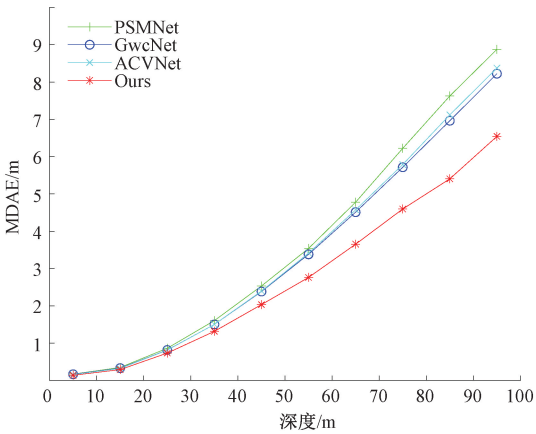


图 9 各模型在 DrivingStereo 上误差曲线

Fig. 9 The error curves of each model on DrivingStereo

衡而导致的,Scene Flow 数据集中大部分区域深度范围在[1,20]m 之间,这直接决定了[1,100]m 的深度精度,而 DrivingStereo 数据集的深度分布较为均衡,其[1,100]m 全距离范围内的平均深度绝对误差更能体现出测试模型在全距离范围内深度估计的精度。

如图 10 所示为 FRDBNet 与其他方法的示例估计深度误差图,黑色区域表示无效点或者深度误差为 0 的区域,而颜色越深的蓝色区域表示越小的深度误差,颜色越深的橙色或者红色等明亮颜色区域表示越大的深度误差。所有模型在近距离处取得了很高的深度精度,但在矩形方框内较远区域,其他模型的深度精度急剧下降,而 FRDBNet 却能保持着良好的估计性能,说明 FRDBNet 在[1,100]m 全距离范围内达到了良好的深度精度平衡,可适用于如自动驾驶等需要全距离范围深度精度平衡的深度估计场景。

表 1 所有模型在 Scene Flow 数据集上的测试结果

Table 1 Test results of all models on the Scene Flow dataset (m)

模型	[1, 10]	[11,20]	[21, 30]	[31,40]	[41, 50]	[51,60]	[61, 70]	[71,80]	[81, 90]	[91,100]	[1,100]
PSM-Net[12](2018)	0. 248	0. 749	2. 426	3. 308	2. 967	3. 925	4. 874	5. 886	6. 720	7. 133	0. 411
Gwc-Net[13](2019)	0. 222	0. 691	2. 186	3. 457	2. 722	3. 603	4. 534	5. 266	5. 877	6. 088	0. 375
ACVNet[14](2022)	**0. 174**	**0. 481**	1. 457	2. 420	1. 529	1. 905	2. 096	2. 471	2. 699	2. 874	**0. 265**
FRDBNet	0. 183	0. 607	**1. 287**	**1. 306**	**1. 092**	**1. 345**	**1. 433**	**1. 673**	**1. 655**	**2. 127**	0. 288

表 2 所有模型在 DrivingStereo 数据集上的测试结果

Table 2 Test results of all models on DrivingStereo dataset (m)

模型	[1, 10]	[11,20]	[21, 30]	[31,40]	[41, 50]	[51,60]	[61, 70]	[71,80]	[81, 90]	[91,100]	[1,100]
PSM-Net[12](2018)	0. 172	0. 350	0. 865	1. 607	2. 534	3. 542	4. 780	6. 222	7. 626	8. 874	1. 204
Gwc-Net[13](2019)	0. 168	0. 332	0. 817	1. 504	2. 384	3. 382	4. 512	5. 715	6. 963	8. 223	1. 129
ACVNet[14](2022)	0. 159	0. 323	0. 813	1. 505	2. 397	3. 416	4. 575	5. 788	7. 122	8. 374	1. 139
FRDBNet	**0. 140**	**0. 292**	**0. 735**	**1. 318**	**2. 034**	**2. 764**	**3. 655**	**4. 596**	**5. 405**	**6. 543**	**0. 912**
提高率	**11. 95%**	**9. 60%**	**9. 60%**	**12. 37%**	**14. 68%**	**18. 27%**	**18. 99%**	**19. 58%**	**22. 38%**	**20. 43%**	
分段提高率	[1,30]m 近距离 10. 38%			[30,60]m 中距离 15. 11%				[60,100]m 远距离 20. 35%			

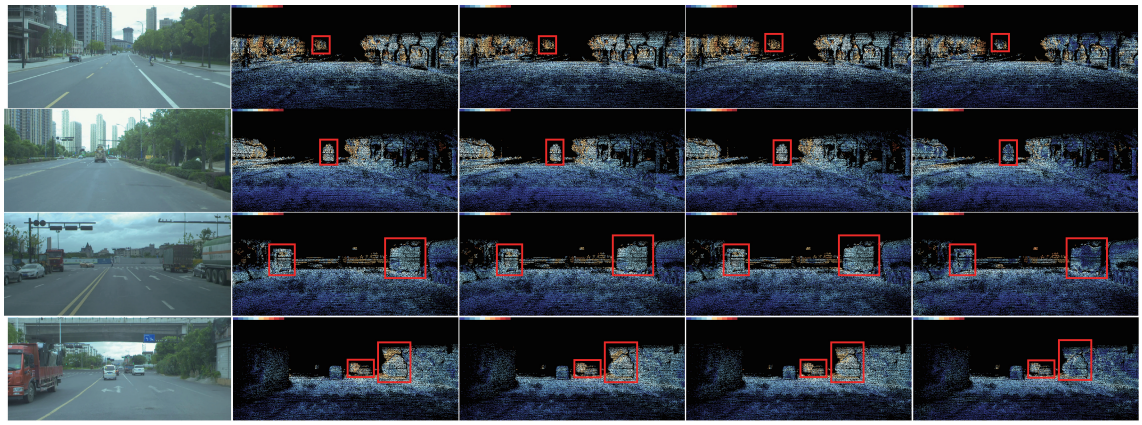
2. 3 消融实验

本文在 DrivingStereo 数据集上进行消融实验以验证深度代价体、视差和深度损失融合训练策略与视差优化模块的有效性。

实验结果如表 3 所示,表中第 1 个模型为基于视差代价体,采用视差损失的基线模型,第 2 个模型为基于深度代价体,采用深度损失的模型,第 3 个模型为基于深度代价体,采用视差与深度损失融合训练策略的模型,第 4 个模型为 FRDBNet 全模型。

模型 2 较模型 1,因为基于深度代价体,降低了较近距离的视差采样分辨率,而导致在[1,10]m 较近距离的深度精度有所下降,但这种厘米级的精度下降针对室外场景的应用是可接受的,同时换来了[10,100]m 距离范围内深度精度的有效提高,达到了一个较好的全距离深度精度平衡。模型 3 较模型 2,因为采用视差与深度损失融合的策略训练模型,使网络同时关注远中近三分段全距离范围的深度估计,而全面提高[1,100]m 全距离的深度精度。模型 4 较模型 3,因为采用了视差优化模块来优化初始深度,有效弥补了构建深度代价体时在远距离处引入的深度损失误差,并进一步提高了[1,100]m 全距离范围的深度精度。

表 3 消融实验结果证明了本文全距离深度平衡立体匹配网络 FRDBNet 能够较好地解决当前视差估计网络在实际应用时,深度精度受相机参数影响,且在远距离处深度精度急剧下降的问题。同时也说明了本文基于深度代价体、视差与深度损失融合训练策略和视差优化模块的有效性。



(a)左图　(b) PSMNet　(c) GwcNet　(d) ACVNet　(e) FRDBNet

图 10　各模型在 DrivingStereo 上示例误差图

Fig. 10　The sample error graphs for each model on DrivingStereo

表 3　本文模型在 DrivingStereo 上的消融实验

Table 3　Ablation experiments of our model on DrivingStereo　(m)

模型	[1, 10]	[11,20]	[21, 30]	[31,40]	[41, 50]	[51,60]	[61, 70]	[71,80]	[81, 90]	[91,100]	[1,100]
基线模型	0.177	0.358	0.880	1.604	2.546	3.566	4.921	6.055	7.458	8.984	1.204
模型 2	0.192	0.334	0.758	1.369	2.142	3.073	4.244	5.429	6.441	7.652	1.028
模型 3	0.155	0.311	0.755	1.342	2.100	2.864	3.790	4.918	5.798	6.912	0.955
FRDBNet	**0.140**	**0.292**	**0.735**	**1.318**	**2.034**	**2.764**	**3.655**	**4.596**	**5.405**	**6.543**	**0.912**

3　结　论

本文针对当前视差估计网络存在估计视差无法直接应用于如自动驾驶、虚拟现实等下游任务,且将视差转换成深度时,受到相机参数的影响,同时在远距离处的深度精度急剧下降的问题,提出一种全距离深度平衡立体匹配网络 FRDBNet。FRDBNet 通过构建深度代价体,使网络隐式地学习到视差与深度之间的转换关系,通过深度回归直接估计深度,使网络输出深度可直接应用于下游任务,同时缓解了众多视差估计网络在远距离处深度精度急剧下降的问题,然后通过视差与深度损失融合的训练策略使得网络对远中近三分段全距离范围的立体匹配给予同样的关注,最后设计视差优化模块来进一步优化初始估计深度,从而在一定程度上缓解了上述问题。实验表明,FRDBNet 在极少牺牲[1,20]m 近距离范围深度精度的前提下,提高了[20,100]m 距离范围的深度精度,有效平衡了[1,100]m 全距离范围的深度精度。本文提出的深度估计方法可以为自动驾驶等需要全距离范围深度估计的任务提供安全有效的深度估计思路。

参考文献

[1] MA X, WANG Z, LI H, et al. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6851-6860.

[2] WANG Y, CHAO W L, GARG D, et al. Pseudo-Lidar from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 8445-8453.

[3] ZHAO X, ZHENG R, YE W, et al. A robust stereo semi-direct SLAM system based on hybrid pyramid[C]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 5376-5382.

[4] KESELMAN L, ISELIN WOODFILL J, GRUNNET-JEPSEN A, et al. Intel real sense stereoscopic depth cameras[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017: 1-10.

[5] HIRSCHMULLER H. Accurate and efficient stereo processing by semi-global matching and mutual information[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05). IEEE, 2005, 2: 807-814.

[6] HIRSCHMULLER H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 328-341.

[7] MEI X, SUN X, ZHOU M, et al. On building an accurate

stereo matching system on graphics hardware [C]. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE, 2011: 467-474.

[8] ZABIH R, WOODFILL J. Non-parametric local transforms for computing visual correspondence [C]. European Conference on Computer Vision. Springer, Berlin, Heidelberg, 1994: 151-158.

[9] BLEYER M, RHEMANN C, ROTHER C. Patchmatch stereo-stereo matching with slanted support windows [C]. British Machine Vision Conference (BMVC), 2011, 11: 1-11.

[10] KHAN S, RAHMANI H, SHAH S A A, et al. A guide to convolutional neural networks for computer vision[J]. Synthesis Lectures on Computer Vision, 2018, 8(1): 1-207.

[11] 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, 2016: 4040-4048.

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

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

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

[15] ZHANG F, PRISACARIU V, YANG R, et al. Ga-net: Guided aggregation net for end-to-end stereo matching[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 185-194.

[16] DUGGAL S, WANG S, MA W C, et al. Deeppruner: Learning efficient stereo matching via differentiable patchmatch [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 4384-4393.

[17] BARNES C, SHECHTMAN E, FINKELSTEIN A, et al. PatchMatch: A randomized correspondence algorithm for structural image editing [J]. ACM Transactions on Graphics, 2009, 28(3): 24.

[18] XU H, ZHANG J. Aanet: Adaptive aggregation network for efficient stereo matching [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1959-1968.

[19] PANG J, SUN W, REN J S J, et al. Cascade residual learning: A two-stage convolutional neural network for stereo matching [C]. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017: 887-895.

[20] LIANG Z, FENG Y, GUO Y, et al. Learning for disparity estimation through feature constancy [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2811-2820.

[21] YANG G, ZHAO H, SHI J, et al. Segstereo: Exploiting semantic information for disparity estimation [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 636-651.

[22] 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.

[23] TONIONI A, TOSI F, POGGI M, et al. Real-time self-adaptive deep stereo[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 195-204.

[24] 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, 2017: 66-75.

[25] YANG G, SONG X, HUANG C, et al. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 899-908.

[26] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 270-279.

[27] CHUAH W, TENNAKOON R, HOSEINNEZHAD R, et al. Semantic guided long range stereo depth estimation for safer autonomous vehicle applications [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18916-18926.

[28] GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[29] BAKER S, SCHARSTEIN D, LEWIS J P, et al. A database and evaluation methodology for optical flow[J]. International Journal of Computer Vision, 2011, 92(1):

1-31.

[30] MENZE M, GEIGER A. Object scene flow for autonomous vehicles [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3061-3070.

[31] KINGMA D P, BA J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv: 1412.6980, 2014.

作者简介

覃业宝,2021 年于福州大学获得学士学位,现为湖南大学电气与信息工程学院硕士研究生,主要研究方向为机器视觉和双目立体视觉。

E-mail: qinyebao@hnu.edu.cn

Qin Yebao received his B. Sc. degree in 2021 from Fuzhou University. He is currently a M. Sc. candidate at Hunan University. His main research interests include machine vision and binocular stereo vision.

孙炜(通信作者),1996 年于湖南大学获得学士学位,1999 年于湖南大学获得硕士学位,2003 年于湖南大学获得博士学位,现为湖南大学教授,主要研究方向为机器人和人工智能。

E-mail: david-sun@126.com

Sun Wei (Corresponding author) received his B. Sc. degree in 1999 from Hunan University, received his M. Sc. degree in 1999 from Hunan University, received his Ph. D. degree in 2003 from Hunan University. He is currently a professor at Hunan University. His main research interests include robot and artificial intelligence.