面向交通场景基于双注意力机制和自适应代价卷的自监督单目深度估计

武 港[1],刘 威[2,3*],胡 骏[2,3],程 帅[2],杨文兴[2,3],孙令岿[1]
(1. 东北大学信息科学与工程学院,辽宁沈阳 110167;2. 东软睿驰汽车技术有限公司,辽宁沈阳 110179;
3. 东北大学计算机科学与工程学院,辽宁沈阳 110167)

摘 要: 针对当前交通场景下自监督单目深度估计存在特征表达能力弱、深度图局部细节模糊、深度估计精度低的问题,提出一种基于双注意力机制和自适应代价卷的自监督单目深度估计方法. 该方法首先利用双注意力机制的特征提取网络,结合通道注意力和空间注意力,对提取的场景特征进行自适应加权,增强特征表达能力. 其次,根据提取的全局特征自适应的构建代价卷,引导网络学习精细的深度特征,提升网络模型对深度图局部细节的学习能力,解决现有方法深度估计精度低的问题. 在自动驾驶公开数据集KITTI、Cityscapes上的实验结果表明,本文方法优于目前主流方法.

关键词: 单目深度估计;自监督;注意力机制;自适应;代价卷
基金项目: 辽宁省“兴辽人才计划”项目(No.XLYC1902029);辽宁省“揭榜挂帅”科技重大专项项目(No.2022JH1/10400030);国家自然科学基金(No.U22A2043)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)05-1670-09
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220710

Self-Supervised Monocular Depth Estimation for Traffic Scenes Based on Dual Attention Mechanism and Adaptive Cost Volume

WU Gang[1], LIU Wei[2,3*], HU Jun[2,3], CHENG Shuai[2], YANG Wen-xing[2,3], SUN Ling-kui[1]
(1. *College of Information Science and Engineering*, *Northeastern University*, *Shenyang*, *Liaoning* 110167, *China*;
2. *Reachauto*, *Shenyang*, *Liaoning* 110179, *China*;
3. *College of Computer Science and Engineering*, *Northeastern University*, *Shenyang*, *Liaoning* 110167, *China*)

Abstract: Aiming at the problems of self-supervised monocular depth estimation in current traffic scenarios, such as weak feature expression ability, fuzzy local details of depth map and low accuracy of depth estimation, a self-supervised monocular depth estimation method based on dual attention mechanism and adaptive cost volume is proposed. Firstly, a dual attention mechanism combining channel attention and spatial attention is used to adaptively weight the extracted scene features to enhance the feature expression ability of the feature extraction network. Secondly, according to the adaptively constructed cost volume of extracting global features, the network is guided to learn fine depth features, which improves the learning ability of the network model for the local details of the depth map and solves the problem of low accuracy of existing depth estimation methods. Experimental results on public datasets KITTI and Cityscapes show that the proposed method is superior to the current mainstream methods.

Key words: monocular depth estimation; self-supervision; attention mechanism; adaptive; cost volume
Foundation Item(s): Liaoning Province “Xingliao Talent Plan” Project (No.XLYC1902029); Liaoning Province’s “Jie Bang Gua Shuai” Science and Technology Major Special Project (No.2022JH1/10400030); National Natural Science Foundation of China (No.U22A2043)

收稿日期:2022-06-20;修回日期:2023-02-06;责任编辑:王天慧
*通讯作者:刘威

1　引言

深度估计是计算机视觉领域的一个重要研究内容,它旨在从特定视角的图像出发生成像素级别的深度图,这种深度图有助于更好地理解场景3D结构,也有助于处理许多计算机视觉任务,如无人驾驶、增强现实、三维重建等. 如何高精度地从图像估计出场景深度是一个研究热点问题. 目前,单目深度估计方法大致可以分为基于有监督学习的方法和基于自监督学习的方法. 基于有监督学习的方法将利用激光雷达或者其他传感器获取的深度作为标签,进行有监督训练,如文献[1~4]. 该类方法能够从单个图像中高精度地估计深度信息,但是需要将带有深度真值的数据作为训练数据,而获取场景深度值数据,需要激光雷达等额外传感器,成本较高.

另一类基于自监督学习的深度估计方法是利用图像序列中相邻帧的光度一致性,构建自监督框架,不需要真值标签数据,即可完成模型训练. 文献[5]首次提出并使用单目图像序列进行自监督深度估计,该方法由深度估计网络和位姿估计网络构成,分别用来预测目标帧深度图和相邻帧到目标帧的相机位姿变化. 利用预测出的深度图、相机位姿和相邻帧来合成目标帧,将合成的目标帧与真实目标帧之间的光度误差作为损失函数,进行自监督训练. 文献[6]为了解决场景中由于动态物体干扰而导致的光度损失函数失效问题,提出了利用掩膜将动态物体标记出来,并对动态物体单独进行损失计算的方法. 文献[7]提出了三维空间损失函数来训练网络,三维空间损失是基于点云一致性设计的,并使用最近点迭代算法来对齐相邻帧的点云. 文献[8]证明了基于深度特征的重建误差比基于图像的光度误差效果要好,并且把基于深度特征的损失定义为特征空间中的光度损失. 文献[9]采用三维卷积进行上采样和下采样,得到了较好的深度细节图,但是三维卷积需要占用较大的显存,参数量大,使算法的运行速度下降. 以上方法均采用传统的特征提取网络,特征提取能力较弱,对深度图的局部细节估计精度不高.

本文在自监督深度估计框架下,提出一种双注意力机制和自适应代价卷的自监督单目深度估计网络. 以相邻帧图像作为输入,通过源图像和合成图像之间的光度差异构建损失函数,完成自监督学习. 在深度估计网络的主体结构设计上,采用双注意力机制特征提取网络,对特征图的重要维度进行自适应加权,使得提取的特征更加关注于对深度估计任务重要的区域. 进一步地,在相邻两帧提取的特征图之间通过平面扫描的方式构建自适应代价卷,自适应代价卷可以根据场景的深度分布特点自适应计算代价值,引导网络学习到细节清晰的深度图.

2　注意力机制

在单目深度估计任务中,许多研究人员采用典型的特征提取网络如Visual Geometry Group network(VGG)[10]、Residual Network(ResNet)[11]来进行场景特征提取,存在难以提取到区分性强的特征而导致估计的深度图局部细节信息模糊的问题. 文献[12]提出了一种基于通道注意力的特征提取网络,该网络通过在不同通道的特征图上施加不同的权重,强化对视觉任务作用大的通道,抑制不必要的通道,从而提升模型的估计精度. 对于每一个特征图,采用全局平均池化操作在空间维度上进行信息压缩,得到全局特征描述子,接着利用全连接层对该描述子进行重新学习赋权并采用ReLU函数激活通道,将通道信息与原有特征图进行逐通道相乘,最后得到了带有通道注意力的特征图.

通道注意力机制在分类任务上表现较好,但是在深度估计任务中,不同物体间的相对位置关系具有重要信息,该做法则损失了特征图中的空间关系,不太适用.

3　代价卷

在立体匹配任务中,文献[13~15]使用代价卷,提升深度估计精度. 文献[16]将代价卷引入到自监督单目深度估计中,通过使用线性深度平面构建代价卷,以便在特征层面对网络进行额外的约束,进一步提高深度估计精度.

代价卷构建方法如图1所示,F_{t-1}、F_t为通过特征提取网络得到的源帧和目标帧的特征图,$\boldsymbol{T}_{t\to t-1}$为位姿估计网络输出的两帧之间位姿变化矩阵,在深度范围内按线性方式划分假设深度d_i,构建合成的特征图$\hat{F}_t$,合成的特征图$\hat{F}_t$和特征图F_t在每个假设深度平面d_i上计算L1距离,最后在每个通道上合并得到代价卷. 线性

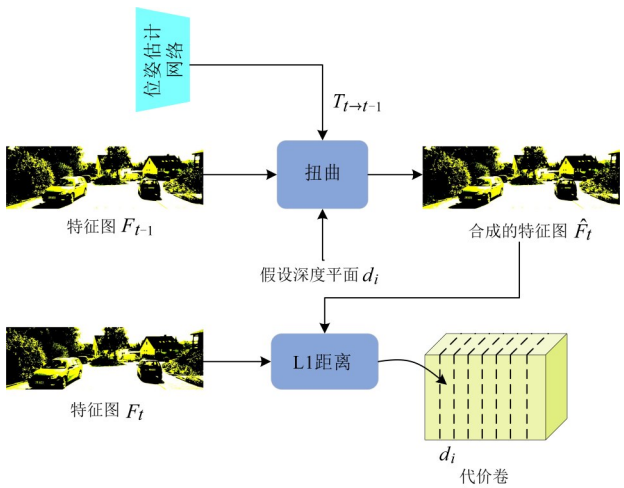


图1　代价卷的构建方法示意图

划分深度平面公式为

$$d_i = d_{\min} + (d_{\max} - d_{\min}) \times i/k \tag{1}$$

其中，$d_{\min}$ 和 $d_{\max}$ 分别为场景的最小深度和最大深度，k 表示划分的深度数量，d_i 表示划分的第 i 个深度．由假设深度平面 d_i 合成的特征图计算公式为

$$\hat{F}_t(\boldsymbol{p}_t) = F_{t-1}\left\langle \boldsymbol{K}\boldsymbol{T}_{t\to t-1} d_i \boldsymbol{K}^{-1}\boldsymbol{p}_t \right\rangle \tag{2}$$

其中，$\boldsymbol{K}$ 为相机的内参矩阵，$\boldsymbol{p}_t$ 表示目标帧特征图的像素坐标，$\langle\ \rangle$ 为双线性插值操作．

使用线性深度平面来构建代价卷，没有考虑输入场景的深度分布，会导致算法对场景中深度信息稀疏的区域给予过高的关注，从而难以收敛到较为准确的深度上．

4 基于双注意力机制和自适应代价卷的自监督单目深度估计方法

4.1 自监督深度估计方法整体结构

本文的自监督网络构建方式参考了自监督单目深度估计框架[17]．对于单目图像序列，由深度估计网络和位姿估计网络分别预测目标帧的深度图、源帧到目标帧之间的相机位姿变化．随后利用目标帧深度图、相机位姿变化和源帧来合成目标帧．最后将合成的目标帧与真实的目标帧之间的光度误差作为网络约束进行自监督训练．本文训练阶段的自监督网络结构如图2所示．

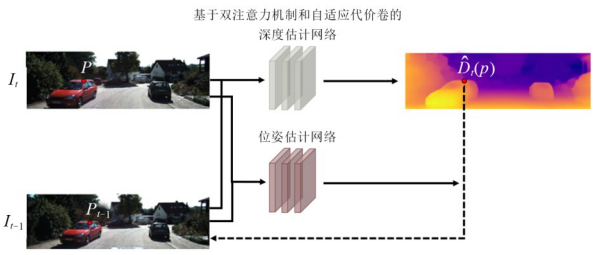


图2 基于自监督的深度估计网络结构

在训练完成后，预测阶段将仅使用其中的基于双注意力机制和自适应代价卷的深度估计网络进行预测，本文的工作主要在于对深度估计网络的设计，其主要结构如图3所示．目标帧 I_t 和源帧 I_{t-1} 分别经过双注意力特征提取网络得到尺寸缩小的特征图 F_t 和 F_{t-1}，利用位姿估计网络输出的相邻帧之间的相机位姿，构建衡量两特征图相似性的自适应代价卷，最后经过代价聚合生成最终的深度图．该方法主要创新点包括双注意机制和自适应代价卷两部分．

4.2 双注意力特征提取网络

通道注意力机制在分类任务上表现良好[12]．但是在深度估计任务中，通道注意力机制将每个通道的特征图都经过全局平均池化得到一个值，该做法会损失特征图中空间位置关系信息，对于像素级的深度估计任务来说，特征图的空间信息包含了物体间的相对位置关系，该信息对深度估计任务有很大的帮助．为了有效解决这个问题，本文提出一种双注意力特征提取网络，特征经过通道维度和空间维度的赋权重新学习，在通道层面，增强有用通道，抑制冗余通道，在空间层面，通过非全局池化来保留原本特征中的空间信息，使网络可以提取到相对位置信息．网络结构如图4所示，特征图为 X_{in}，主通路上经过两个卷积操作得到特征图 X_{res}，副通路部分经过平均池化操作得到包含全局信息的特征图，然后该特征图经过两个卷积层来重新学习注意力权重得到 $\hat{X}_{\text{att}}$，$\hat{X}_{\text{att}}$ 经过上采样操作恢复分辨率后得到 X_{att}，X_{att} 与 X_{in} 逐像素相乘后与 X_{res} 相加得到最终的特征，该特征包含通道和空间双重注意力．

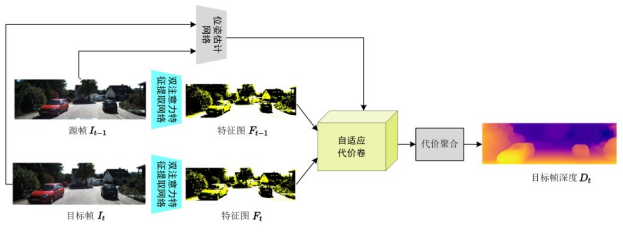


图3 基于双注意力机制和自适应代价卷的深度估计网络结构图

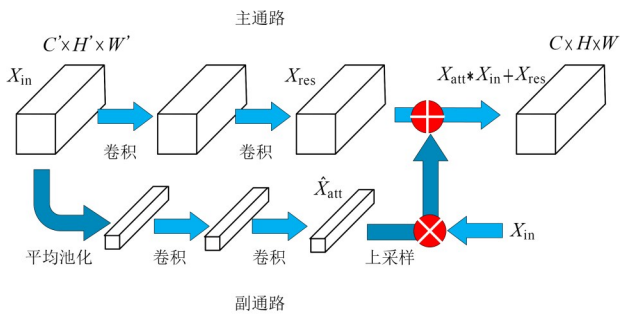


图4 双注意力特征提取网络结构

4.3 自适应代价卷

代价卷的引入可以在特征层面对网络作额外的约束，但是以往使用线性深度平面来构建代价卷的方式会导致算法对深度信息稀疏的区域关注度过高而造成该区域深度估计误差较大．一般来说，对于深度值大的区域，场景中能提供给深度估计的信息相比于深度值较小的区域会少很多，只使用线性划分方式难以应对场景中复杂多变的情况，应该根据实际分布将假设深度值合理地离散化，在深度分布较为密集的区域多离散化一些深度值，在深度分布稀疏的区域（通常是较大深度值的区域）少离散一些深度值，通过较为精确的深度值假设构建代价卷，引导网络收敛到合适深度值，进而输出较为精准的深度图．

本文提出自适应代价卷来解决这一问题,设计一种根据输入帧的特征来划分假设深度构建代价卷方法,通过假设较为精确的深度平面,引导网络输出更加准确的深度值.

图5为自适应假设深度划分设计,对于输入帧I_t经过特征提取网络得到该帧的特征图F_t,经过回归网络得到学习到的场景深度间隔值b_i',b_i'经过Sigmoid激活函数输出范围在0到1间的浮点数,经过归一化操作得到b_i,最后经过深度值还原操作得到最后的离散深度值d_i,网络最终输出k个离散化后的深度值.

本方法没有让网络直接输出离散后的k个深度值,而是让回归网络先学习场景深度间隔然后再经过归一化和还原操作输出结果,相当于让网络直接输出离散化后的深度值,这样间接的方式可以让网络学习到更为有用的信息,降低网络的学习难度. 自适应输出离散化后深度值的整个过程如图5所示,其中归一化操作采用式(3),ϵ的引入是为了防止分母为0,还原操作如式(4)所示,其中b_i'为回归网络输出的深度间隔值,b_i为归一化后的深度间隔值,$d_{\min}$和$d_{\max}$则分别为场景的最小、最大深度值,d_i则为离散后深度值.

$$b_i=\frac{b_i'+\epsilon}{\sum_{j=1}^{N}\left(b_j'+\epsilon\right)} \tag{3}$$

$$d_i=d_{\min}+\left(d_{\max}-d_{\min}\right)\left(b_i/2+\sum_{j=1}^{i-1}b_j\right) \tag{4}$$

针对回归网络的实现,采用全连接层会使网络的参数量暴增,严重影响整个算法的运行速度,因此本文提出用全卷积实现回归网络,该方式的参数量仅取决于卷积核的大小和个数,相比于全连接,参数量锐减,同时也加快了推理速度. 鉴于回归网络的输出需要生成k个离散化后的深度值,这里采取4个全卷积操作逐渐下采样,最终生成维度为$1\times1\times k$的特征图.

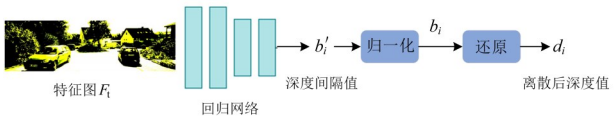


图5　自适应假设深度划分设计

4.4　损失函数

代价卷中包含的信息只对场景中静态物体和纹理明显的地方有效,对移动物体和无纹理区域,代价卷将是一个不可靠的深度信息来源,在不可靠区域会导致网络过度依赖代价卷而产生错误的深度估计结果. 针对这个问题,引入额外的单目深度估计网络Monodepth2[17],该网络对于无纹理区域和场景中移动物体的深度估计表现良好并且没有构建代价卷,因此可以利用Monodepth2教会本文所述网络忽略不可靠区域的代价卷信息.

对于一张输入的RGB图像,Monodepth2网络输出深度图为$\hat{D}_t$并且训练后被丢弃且阻断梯度回传,确保信息只从Monodepth2网络传递到本文网络. Monoodepth2网络和本文网络共享位姿估计网络,确保输出深度图尺度一致. 这里使用掩膜过滤掉不可靠区域的重投影损失,采用式(5)约束本文网络的输出D_t与Monodepth2网络的输出一致.

$$L_{\text{kd}}=\sum \boldsymbol{M}\left|D_t-\hat{D}_t\right| \tag{5}$$

其中,掩膜$\boldsymbol{M}$为二值表示,值为1表示该区域的像素不可靠,0表示可靠. 在代价卷可靠的区域,Monodepth2网络输出的深度值$\hat{D}_t$应与本文网络中代价卷沿通道的argmin值D_{cv}是一致的,因此掩膜$\boldsymbol{M}$计算公式由式(6)给出:

$$\boldsymbol{M}=\max\left(\frac{D_{\text{cv}}-\hat{D}_t}{\hat{D}_t},\frac{\hat{D}_t-D_{\text{cv}}}{D_{\text{cv}}}\right)>1 \tag{6}$$

本文方法总损失函数由式(7)给出,其中最小重投影损失L_{p}和平滑损失L_{smooth}采用文献[17]中的方法.

$$L=(1-\boldsymbol{M})L_{\text{p}}+L_{\text{kd}}+L_{\text{smooth}} \tag{7}$$

5　实验与分析

为了验证本文方法的有效性,选取目前交通场景广泛使用的KITTI和Cityscapes数据集进行实验. KITTI数据集[18]包括93 000张图片和对应的深度图,其中RGB图片的分辨率为1 242×375,深度图的分辨率为1 224×368. 实验采用深度估计常用的Eigen Split方式[19]划分训练集和测试集,其中39 810张用于训练集,4 424张用于验证集,为减少计算量,本文将输入图像分辨率调整为640×192,深度图也调整为相同尺寸. Cityscapes数据集[20]包含3 250个训练数据序列和1 250个测试数据序列共计15万张图片,图片的分辨率为2 048×1 024,深度图由半全局匹配算法得到,本文实验将69 731张图片用于训练集,7 621张图片用于验证集,文献[3]的裁剪策略,训练时采用图片分辨率为512×192,评估时将输出深度图裁剪为416×128.

5.1　实验环境与参数设置

本文在Ubuntu18.04上进行实验,CPU型号为Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz,内存为32 GB,硬盘为1 TB固态硬盘+ 2 TB机械硬盘,显卡为两张显存为11 GB的GeForce RTX 2080 Ti,训练框架采用Pytorch1.7. 对输入图片按50%几率进行数据增强,包括水平翻转、图像亮度、对比度、饱和度、色调随机加减0.2. 训练时每个批次送入8张图片,共训练30个轮次,梯度优化算法采用Adam算法,初始学习率为0.000 1,权重衰减系数为0.1,学习率每过15个轮次衰减为原来

的0.1,本文提出的双注意力特征提取网络在KITTI数据集上的参数如表1所示.

表1 网络参数设计

网络层	卷积操作	输出特征图分辨率
卷积层	7×7,64 步长=2	320×96
最大池化层	3×3 max pooling 步长=2	160×48
注意力块1	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	160×48
注意力块2	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	160×48
注意力块3	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	160×48
注意力块4	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	160×48

5.2 实验结果与分析

5.2.1 定量对比

为了验证方法有效性,将本文提出方法与Monodepth2[17]、Packnet[9]、Manydepth[16]等12种方法在KITTI数据集上进行对比,与Struct2depth[21]等6种方法在Cityscapes数据集上进行对比.对比结果如表2、表3所示,评价指标采用平均相对误差(Abs Rel,Absolute Relative error)、平方根相对误差(Sq Rel,Square root Relative error)、均方根误差(RMSE,Root Mean Square Error)、对数均方根误差(log RMSE,log Root Mean Square Error)、给定阈值下的精度δ_1、δ_2、δ_3共7个指标,其中前4个指标越低、后3个指标越高,表示深度估计结果越精确.分割列表示方法是否用到额外的分割结果,辅助深度估计.由表2可见,本文方法相比于目前最好的自监督单目深度估计方法Manydepth,在平均相对误差上提升了0.003,在平方根相对误差上提升了0.031,在均方根误差上提升了0.062,对数均方根误差上提升了0.001,δ_1和δ_2则分别提升了0.003和0.001,有5个指标都超出其他自监督单目深度估计方法.由表3在Cityscapes上的实验结果可见,本文方法在所有指标上均超过其他方法.

深度估计方法中,有学者使用分割结果作为额外的约束对移动物体进行处理,提升深度估计精度.尽管如此,本文方法深度估计的结果已经比借助分割算法的Struct2Depth[21]和Video in the wild[22]方法效果更好.

表2 本文方法与其他自监督方法在KITTI数据集上定量对比

方法	分割	Abs Rel↓	Sq Rel↓	RMSE↓	log RMSE↓	δ_1↑	δ_2↑	δ_3↑
Ranjan et al[23]		0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++[24]		0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2Depth[21]	·	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Video in the wild[22]	·	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Packnet[9]	·	0.102	0.698	4.381	0.178	0.896	0.964	**0.984**
Johnston et al[25]		0.106	0.861	4.699	0.185	0.889	0.962	0.982
Monodepth2[17]		0.115	0.903	4.863	0.193	0.877	0.959	0.981
Packnet[9]		0.111	0.785	4.601	0.189	0.878	0.960	0.982
Li et al[26]		0.130	0.950	5.138	0.209	0.843	0.948	0.978
Patil et al[27]		0.111	0.821	**4.365**	0.187	0.883	0.961	0.982
Wang et al[28]		0.106	0.799	4.662	0.187	0.889	0.961	0.982
Manydepth[16]		0.098	0.770	4.459	0.176	0.900	0.965	0.983
本文方法		**0.095**	**0.739**	4.397	**0.175**	**0.903**	**0.966**	0.983

注:黑色点代表使用了额外的分割数据.

表3 本文方法与其他自监督方法在CityScape数据集上定量对比

方法	分割	Abs Rel↓	Sq Rel↓	RMSE↓	log RMSE↓	δ_1↑	δ_2↑	δ_3↑
Struct2Depth[21]	·	0.145	1.737	7.280	0.205	0.813	0.942	0.976
Pilzer et al[29]		0.240	4.264	8.049	0.334	0.710	0.871	0.937
Monodepth2[17]		0.129	1.569	6.876	0.187	0.849	0.957	0.983
Video in the Wild[22]	·	0.127	1.330	6.960	0.195	0.830	0.947	0.981
Li et al[26]		0.119	1.290	6.980	0.190	0.846	0.952	0.982
Manydepth[16]		0.114	1.193	6.223	0.170	0.875	**0.967**	**0.989**
本文方法		**0.112**	**1.164**	**6.221**	**0.168**	**0.877**	**0.967**	**0.989**

注:黑色点代表使用了额外的分割数据.

5.2.2　定性分析

图6为本文方法和Monodepth2、PackNet方法在KITTI数据集上的定性对比. 由图6可见,本文方法在路标、杆等体积较小的物体上表现较好,输出的深度图细节特征丰富,较为平滑. 在第一行中,Monodepth2方法没有将路标与背景区分开,PackNet方法将路标与背景区分开了一些,但不是很明显,而本文方法将路标与背景很好地区分开了. 第二行中,对于细小的路标杆,本文方法的深度估计效果也比其他两种方法好,且相邻很近的杆也可以区分开. 第三行中左下角,对于汽车车窗的深度估计,Monodepth2和PackNet方法没有考虑到周边像素深度估计值的一致性,均估计出错误的深度,而本文方法估计出了较为平滑的深度值,与车身整体深度一致,这得益于本文提出的局部空间注意力机制. 第四行中,左下角位置的物体较多且聚集在一起,对于该区域本文方法可估计出较为明显的边界,而其他两种方法则将该区域估计为一整片的区域,对于旁边的小柱子,本文方法可以清晰地将其与周围物体区分开,效果较好.

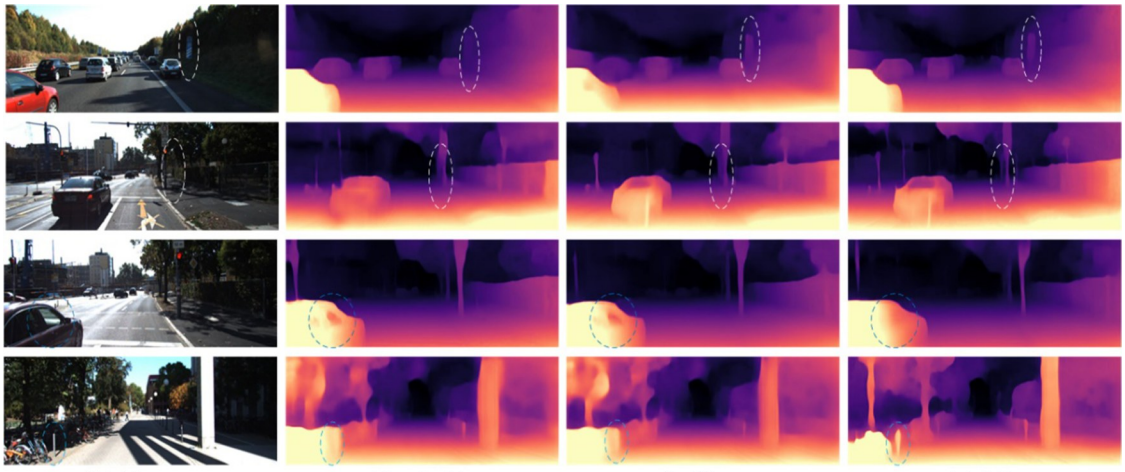


图6　本文方法和其他方法在KITTI数据集上定性对比

图7为本文方法和Manydepth方法在Cityscapes数据集上的定性对比. 第一行中,车辆为移动物体,Manydepth方法将车辆区域估计为空洞,而本文方法则较好的估计出了车辆的深度图. 在第二行中,对于前方较远处车辆,Manydepth方法也将其估计为一个洞,即将其估计为深度值无限远的区域,而本文方法则要好些,洞的面积小于Manydepth方法,这得益于本文使用额外的算法来引导网络忽略代价卷中不可靠区域,使其能够学习到移动物体区域的深度. 第三行中,对于较近的交通指示牌区域,Manydepth方法深度估计结果较为模糊,而本文方法则可以明显地将交通指示牌与其周围背景区分开. 第四行中对于较远处的树,对比方法失效,本文方法表现较好.

5.2.3　消融实验

为了进一步探究本文提出双注意力特征提取网络和自适应代价卷的有效性,在KITTI数据集上进行消融实验. 消融实验结果在图8中给出. 图中ResNet、通道注意力(SENet)、双通道注意力(LS-SENet)表示单目深度估计采用的特征提取方式,代价卷和自适应代价卷分别表示是否使用代价卷和本文提出的自适应代价卷,预训练表示特征提取网络是否在ImageNet上做过预训练. 由图8可见,在其他条件不变的情况下,特征

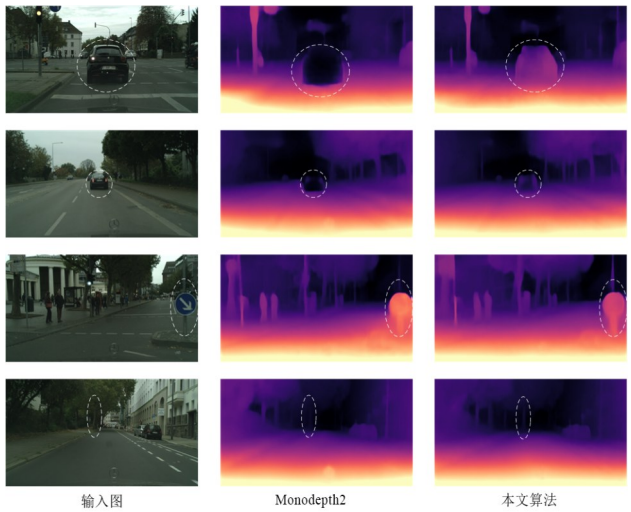
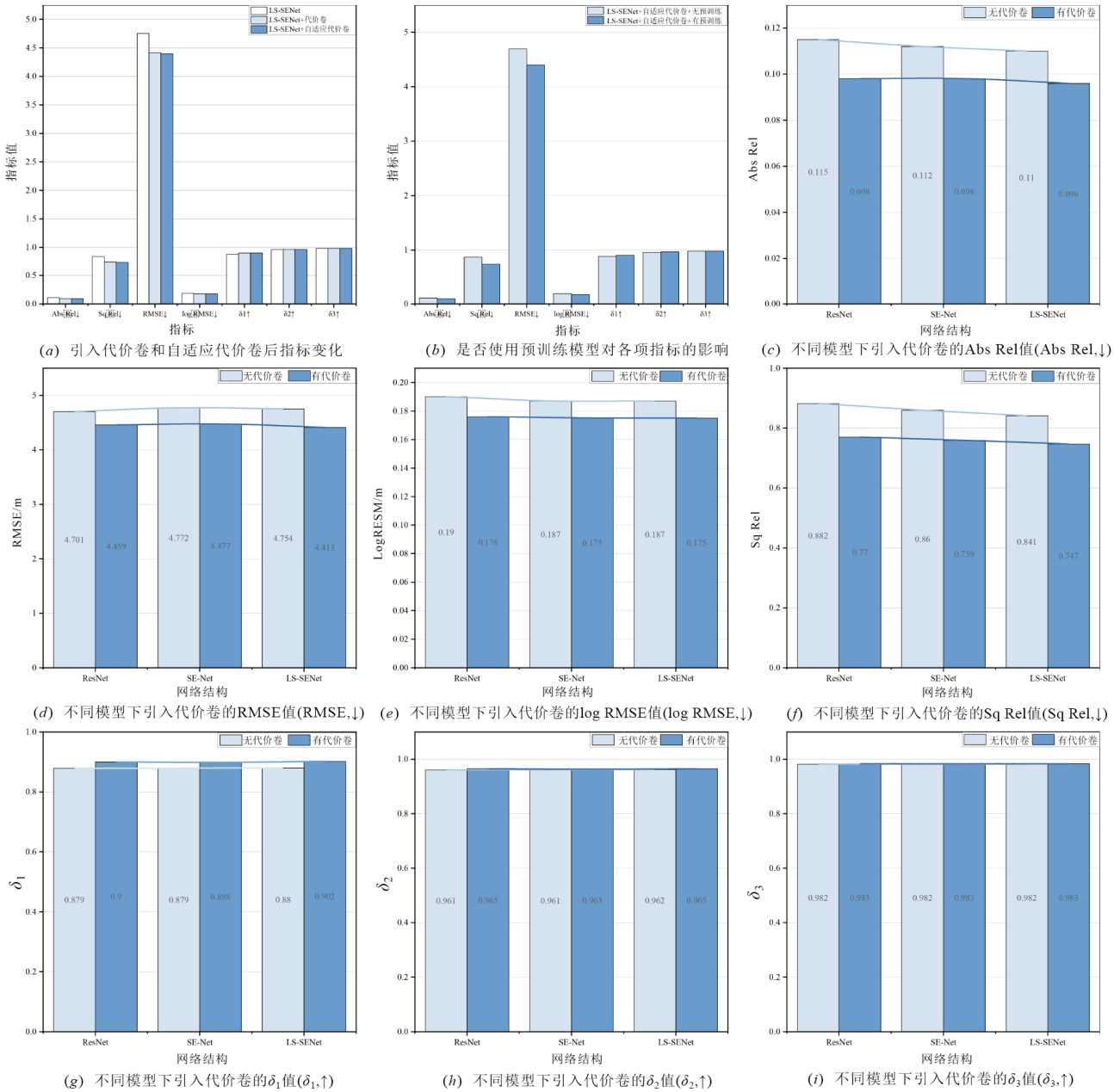


图7　本文方法和其他方法在CityScape数据集上定性对比

提取网络分别选取ResNet、通道注意力、双注意力,平均相对误差指标分别为0.115、0.112、0.110,其他指标也稳步改善. 双注意力特征提取网络不仅继承了通道注意力机制的优点,还能在空间层面对特征进行自适应加权,使得网络可以捕获物体空间相对位置关系,使提取的特征鲁棒性更强. 在其他实验条件固定的情况下,

通过增加代价卷和自适应代价卷，实验指标也逐步提升．通过采用本文提出的双注意力特征提取网络和自适应代价卷，最终平均相对误差达到0.095，超出目前大部分方法，证明了本文提出改进点的有效性．



(*a*) 引入代价卷和自适应代价卷后指标变化 (*b*) 是否使用预训练模型对各项指标的影响 (*c*) 不同模型下引入代价卷的Abs Rel值(Abs Rel,↓)

(*d*) 不同模型下引入代价卷的RMSE值(RMSE,↓) (*e*) 不同模型下引入代价卷的log RMSE值(log RMSE,↓) (*f*) 不同模型下引入代价卷的Sq Rel值(Sq Rel,↓)

(*g*) 不同模型下引入代价卷的δ_1值(δ_1,↑) (*h*) 不同模型下引入代价卷的δ_2值(δ_2,↑) (*i*) 不同模型下引入代价卷的δ_3值(δ_3,↑)

图8 消融实验结果

6 结论

本文提出面向交通场景基于双注意力机制和自适应代价卷的自监督单目深度估计方法，首先通过双注意力特征提取网络，对场景特征在通道、空间层面进行自适应加权，得到带有双注意力的特征，使得所提取的特征更具鲁棒性，其次根据双注意力特征提取网络提取的特征自适应的构建代价卷，引导网络学习到更精细的深度，提升算法在局部细节深度图上的表现．在KITTI和Cityscapes数据集上的实验结果表明，本文方法相比于其他方法有一定的提升，消融实验也表明了本文方法的有效性和合理性．在未来的研究中，将进一步分析不同特征提取网络对自监督单目深度估计任务的有效性．此外，还将研究更有针对性的策略来解决局部深度图模糊问题．

参考文献

[1] LI B, SHEN C, DAI Y, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs[C]//Computer Vision & Pattern Recognition. Piscataway: IEEE, 2015: 1119-1127.

[2] WANG P, SHEN X, LIN Z, et al. Towards unified depth and semantic prediction from a single image[C]//Computer Vision & Pattern Recognition. Piscataway: IEEE, 2015: 2800-2809.

[3] CAO Y, WU Z, SHEN C. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(11): 3174-3182.

[4] FU H, GONG M, WANG C, et al. Deep ordinal regression network for monocular depth estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2018: 2002-2011.

[5] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1851-1858.

[6] VIJAYANARASIMHAN S, RICCO S, SCHMID C, et al. Sfm-net: Learning of structure and motion from video[EB/OL].(2017)[2022]. https://arxiv.org/abs/1704.07804.

[7] MAHJOURIAN R, WICKE M, ANGELOVA A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5667-5675.

[8] ZHAN H, GARG R, WEERASEKERA C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Rrecognition. Piscataway: IEEE, 2018: 340-349.

[9] GUIZILINI V, AMBRUS R, PILLAI S, et al. 3d packing for self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2485-2494.

[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014)[2022]. https://arxiv.org/abs/1409.1556.

[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Ccomputer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[13] DAI Y, ZHU Z, RAO Z, et al. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry[C]//2019 International Conference on 3D Vision (3DV). Piscataway: IEEE, 2019: 1-8.

[14] HUANG P H, MATZEN K, KOPF J, et al. Deepmvs: Learning multi-view stereopsis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2821-2830.

[15] GU X, FAN Z, ZHU S, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2495-2504.

[16] WATSON J, AODHA O MAC, PRISACARIU V, et al. The temporal opportunist: Self-supervised multi-frame monocular depth[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 1164-1174.

[17] GODARD C, AODHA O MAC, FIRMAN M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 3828-3838.

[18] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 3354-3361.

[19] EIGEN D, FERGUS R. Predicting Depth, Surface normals and semantic labels with a common multi-scale convolutional architecture[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2014: 1-12.

[20] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 3213-3223.

[21] CASSER V, PIRK S, MAHJOURIAN R, et al. Unsupervised monocular depth and ego-motion learning with structure and semantics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2019: 1-10.

[22] GORDON A, LI H, JONSCHKOWSKI R, et al. Depth

from videos in the wild: Unsupervised monocular depth learning from unknown cameras[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 8977-8986.

[23] RANJAN A, JAMPANI V, BALLES L, et al. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 12240-12249.

[24] LUO C, YANG Z, WANG P, et al. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding[J]. IEEE transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2624-2641.

[25] JOHNSTON A, CARNEIRO G. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 4756-4765.

[26] LI H, GORDON A, ZHAO H, et al. Unsupervised monocular depth learning in dynamic scenes[EB/OL]. (2020)[2022]. https://arxiv.org/abs/2010.16404.

[27] PATIL V, VAN GANSBEKE W, DAI D, et al. Don't forget the past: Recurrent depth estimation from monocular video[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6813-6820.

[28] WANG J, ZHANG G, WU Z, et al. Self-supervised joint learning framework of depth estimation via implicit cues [EB/OL]. (2020)[2022]. https://arxiv.org/abs/2006.09876.

[29] PILZER A, XU D, PUSCAS M, et al. Unsupervised adversarial depth estimation using cycled generative networks[C]//2018 International Conference on 3D Vision (3DV). Piscataway: IEEE, 2018: 587-595.

作者简介

武 港 男,1997 年 2 月出生于山西省临汾市. 就读于东北大学信息科学与工程学院. 主要研究方向为深度估计、目标检测等.

E-mail: 914766938@qq.com

刘 威 男,1975 年 6 月出生于辽宁省沈阳市. 就职于东北大学计算机科学与工程学院. 主要研究方向为计算机视觉、深度学习、多传感器融合及路径规划控制.

E-mail: lwei@neusoft.com

胡 骏 男,1985 年 12 月出生于安徽省滁州市. 就读于东北大学计算机科学与工程学院. 主要研究方向为计算机视觉、自动驾驶.

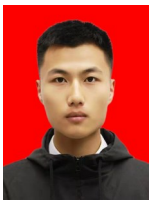
E-mail: hu.jun@reachauto.com

程 帅 男,1987 年 8 月出生于内蒙古呼伦贝尔市. 就职于东软睿驰汽车技术有限公司自动驾驶业务线. 主要研究方向为计算机视觉、深度学习.

E-mail: cheng.shuai@reachauto.com

杨文兴 男,1992 年 11 月出生于内蒙古呼伦贝尔市. 现为东北大学计算机科学与工程学院在职博士,就职于东软睿驰汽车技术有限公司自动驾驶业务线. 研究方向为自动驾驶相关的视觉感知及预测.

E-mail: yang.wx@reachauto.com



孙令岿 男,1998 年 12 月出生于山东省济宁市. 就读于东北大学信息科学与工程学院. 主要研究方向为自动驾驶感知算法.

E-mail: sunlingkui@163.com