文章编号：2095-2163(2021)05-0019-07 中图分类号：TP391.4 文献标志码：A

基于 CBAM 的深度序数回归方法

高永彬，王慧星，黄 勃
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：本文针对单目深度估计模型深度序数回归算法中全图像编码器易丢失较大像素值像素特征信息和位置信息的缺点，提出一种基于 CBAM 的深度序数回归方法。首先，将 CBAM 嵌入到深度序数回归算法中作为全图像编码器，依次采用通道注意力机制和空间注意力机制来捕获图像完整的特征信息和位置信息，通过获得的注意力图重新调整原始特征；其次，对像素的深度值进行离散，将深度估计重新转化为序数回归问题；最后，使用回归损失函数对网络进行训练。实验结果表明，相比于其他有监督学习、半监督学习和无监督学习的方法，该方法在 KITTI 数据集上取得更好的效果。

关键词：单目深度估计；CBAM；通道注意力；空间注意力；序数回归

CBAM-based deep ordinal regression method

GAO Yongbin, WANG Huixing, HUANG Bo
(School of Electronic and Electrical Engineering, Shanghai University of Engineering and Science, Shanghai 201620, China)

【Abstract】 Aiming at the shortcomings of the full-image encoder in the monocular depth estimation model deep oridinal regression network that it is easy to lose the feature information and position information of the truncated length value, a CBAM-based depth ordinal regression method is proposed. First embed CBAM into depth ordinal regression algorithm as a full image encoder, and use channel attention mechanism and spatial attraction mechanism to capture the complete feature information and position information of the image, and readjust the original features through the obtained attention map; then perform the corresponding depth value Discrete, transform the depth estimation into an ordinal regression problem; finally use the regression loss function to train the network. The experimental results show that, using other supervised learning, semi-supervised learning and unsupervised learning methods, this method achieves better results on the KITTI data set.

【Key words】 monocular depth estimation; CBAM; channel attention; spatial attention; ordinal regression

0 引 言

单目深度估计对三维场景理解任务具有重要意义，在三维重建、自动驾驶、视觉跟踪、三维目标检测、增强现实等领域有着广泛的应用。随着深度学习的迅速发展，利用有监督学习方法进行单目深度估计的研究大量涌现，这些方法通常将深度估计建模作为一个回归问题，使用深度卷积神经网络获取图像的层次信息和层次特征，并通过最小化均方误差来训练回归网络。然而，这些方法往往存在缺点：一方面，使用最小化均方误差来训练回归网络，往往会导致网络收敛慢和局部解不理想的问题；另一方面，为了获得高分辨率的深度图，需要使用跳跃连接或多层反卷积网络结构，这使网络训练更加复杂，计算量大大增加；最后，利用多尺度网络对图像进行特征提取，往往会丢失像素的特征信息和位置信息，对较小目标的深度估计效果较差。为此，Fu 等人提出了用于单目深度估计的深度序数回归网络（Deep Ordinal Regression Network），使用 ASPP（Atrous Spatial Pyramid Pooling）获取不同尺度的特征，并通过全图像编码器捕获全局上下文信息[1]。采用离散策略对深度值进行离散，将深度估计转化为序数回归问题，通过一个普通回归损失函数训练网络，提高网络训练效率。

本文主要对深度序数回归网络深度序数回归算法进行研究，主要贡献如下：

（1）提出了一种基于 CBAM（convolutional block attention module）的深度序数回归方法，通过 CBAM 代替深度序数回归算法中的全图像编码器，获取更完整的像素特征信息和位置信息，提高全局上下文信息的表示能力；

（2）将 CBAM 中的通道注意力机制和空间注意力机制以不同的顺序融入到网络中，以发现注意力机制的顺序与网络结构的相适应性，探索出最佳的

基金项目：国家自然科学基金（61802253）。

作者简介：高永彬（1988-），男，博士，副教授，主要研究方向：人工智能、机器学习、图像处理等；王慧星（1994-），男，硕士研究生，主要研究方向：计算机视觉；黄 勃（1985-），男，博士，讲师，主要研究方向：需求工程、软件工程、形式化方法等。

收稿日期：2020-11-27

网络模型;

(3)实验结果证明,本文提出的网络模型可以有效地提高深度估计的精度,在 KITTI 数据集上进行测试,效果比当前最佳方法提高 1%左右。

1 单目深度估计研究现状

近年来,深度学习被广泛应用于计算机视觉领域,并在单目深度估计方面取得了显著的成就。Eigen 等首次将深度学习应用于单目深度估计研究中,提出了一种多尺度神经网络用于深度估计的思想,首先使用粗尺度网络预测图像的全局深度,然后使用细尺度网络优化局部细节,最终获得像素级别的深度信息[2];在此方法的基础之上,他们又提出了一种用于多任务的多尺度网络框架,使用了更深层次的网络结构,利用 3 个细尺度的网络进一步增添细节信息,使用不同的损失函数和数据集分别对深度预测、表面法向量估计和语义分割任务进行训练,最终获得了良好的效果[3];由于多尺度网络只是使用几个串联的浅层网络对图像进行分层细化,因此最终得到的深度图分辨率是偏低的,为了提高深度图的分辨率,Li 等在多尺度网络之间加入跳跃连接,在第一个网络中使用跳跃连接,对池化后的特征图进行上采样,进而与第二个网络中的特征图进行拼接,同样地,第二个网络中的特征图与第三个网络中的特征图进行拼接,使网络同时将较深层的低空间分辨率深度图与较低层的高空间分辨率深度图融合,提高了深度图的分辨率[4];Laina 等提出了一种残差学习的全卷积网络,用于单幅图像的深度估计,网络结构更深,提高输出分辨率的同时又优化了效率[5];Liu 等提出了将条件随机场(conditional random field, CRF)与 CNN 相结合来估计单幅图像深度的方法,使用 CRF 的一阶项和二阶项综合训练 2 个 CNN,然后将这两个网络通过 CRF 能量函数统一于一个训练框架中,这种方式可以提供更多的约束[6];同样使用 CRF 方法,Xu 等提出了一种结构化注意力模型,它可以自动调节不同尺度下对应特征之间传递的信息量,并且可以无缝集成到 CRF 中,允许对整个架构进行端到端训练[7];Cao 等把深度估计问题看作像素分类问题,首先将深度值进行离散,然后使用残差网络来预测每个像素对应的类别,最终使用 CRF 模型进行优化[8];Chang 等提出了使用金字塔池化模块来捕捉更多的全局信息,使单幅图像的深度估计精度得到提高[9]。

以上方法虽然都利用了有监督学习的方法对单幅图像进行深度估计,但使用多尺度网络结构往往会丢失像素的特征信息和位置信息,对深度估计精度造成影响。通过最小化均方误差训练网络,存在收敛慢和局部解不理想的缺点。加入跳跃连接等结构,使网络训练复杂,计算量增加。

目前还有一些使用无监督学习进行深度估计的方法,Chen 等提出了一种场景网络来对物体的几何结构进行建模,通过增强立体图像对之间的语义一致性来执行区域感知深度估计[10];Lee 等提出了一种利用相对深度图进行单目深度估计的方法,使用 CNN 在不同的尺度上估计区域对之间的相对深度和普通深度,进而将普通深度图和相对深度图分解,并对分解之后的深度图进行优化重组,以重建最终的深度图[11]。虽然无监督学习方法在一定程度上克服了数据标注工作量大的问题,但是始终达不到有监督学习的方法的精度。

针对以上问题,本文对有监督学习的单目深度估计模型深度序数回归算法进行了研究,发现深度序数回归算法中使用的全图像编码器存在易丢失较大特征值像素特征信息和位置信息的缺点。本文引入 CBAM,提出了一种 CBAM 的深度序数回归方法。使用全局最大池化和全局平均池化替代局部平均池化,解决较大特征值像素特征信息易丢失的问题。使用空间注意力机制生成的注意力特征图与原始特征图相乘替代简单的复制操作,解决像素位置信息易丢失的问题。

2 网络框架

本文方法的整体网络框架如图 1 所示。主要由 3 部分组成,特征提取网络、场景理解模块和序数回归模块。

首先将单幅图像输入到特征提取网络中进行初步的特征提取,特征提取网络采用 ResNet-101,通过在 ImageNet 数据集上预训练好的模型对其进行初始化。由于前几层的特征只包含一般的低级信息,在初始化后固定 ResNet-101 前 2 个卷积层的参数,且在训练过程中为 BN(Batch Normalization)层直接进行初始化;然后将得到的特征送入场景理解模块,场景理解模块包括全图像编码器、空洞空间卷积池化金字塔模块 ASPP 和跨通道信息学习器。全图像编码器主要作用是捕获全局特征的上下文信息,在这里使用 CBAM 取代全图像编码器结构,依次使用通道注意力机制和空间注意力机制捕获像素更好的特征信息和位置信息;ASPP 模块主要使用

采样率分别为 6、12 和 18 的空洞卷积对输入的特征图进行并行采样，进而得到多尺度融合特征，来表征不同大小区域的图像特征；跨通道信息学习器主要使用 1×1 的卷积对各个通道之间的相互作用进行学习。进一步地将全图像编码器、ASPP 模块和跨通道信息学习器输出的特征图分别经过一个 1×1 的卷积，进而将 3 个模块的所有输出进行合并，再经过一个 1×1 的卷积，输入到序数回归模块。最后根据深度值的序数相关性，使用间隔递增离散化策略（spacing-increasing discretization，SID）在对数空间中对深度值进行离散，以降低深度值较大区域的训练损失。使用普通的序数回归损失来学习网络参数，获得更高的精度。

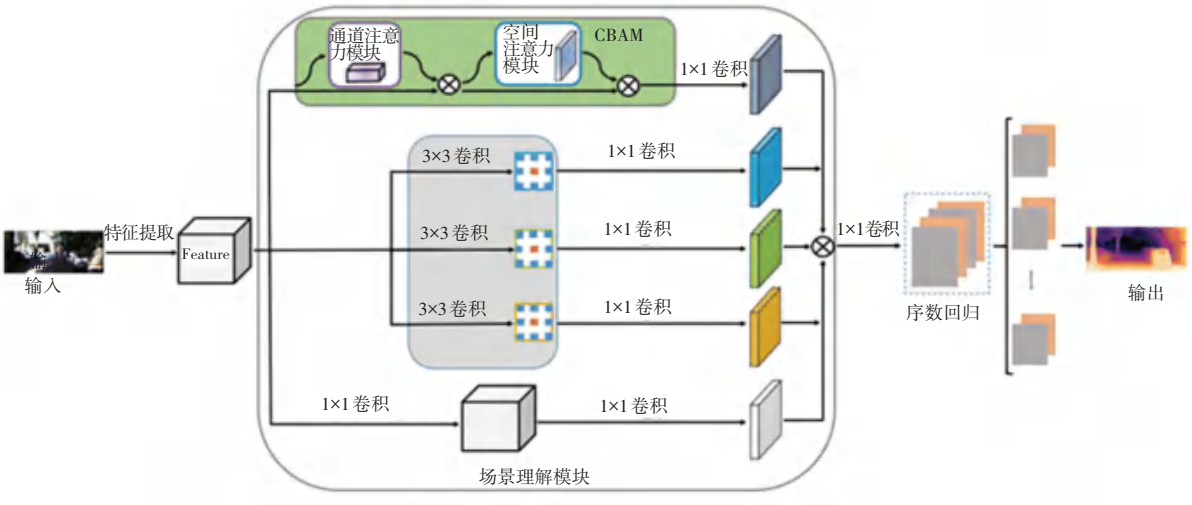

图 1　整体网络结构

Fig. 1　Overall network structure

2.1　全图像编码器

深度序数回归算法中的全图像编码器结构如图 2 所示。为了从尺寸为 $C \times h \times w$ 的 F 中获得相同尺寸的全局特征 $F^{''}$，首先要通过局部平均池化对原始特征进行降维，将降维之后的特征通过全连接层得到一个 C 维的特征向量；将特征向量视为空间维数为 1×1 特征图的 C 通道，并添加一个核尺寸为 1×1 的卷积层作为特征向量跨通道参数池化结构；最后，将特征向量复制到 $F^{''}$，使 $F^{''}$ 的每个位置对整个图像有相同的理解。

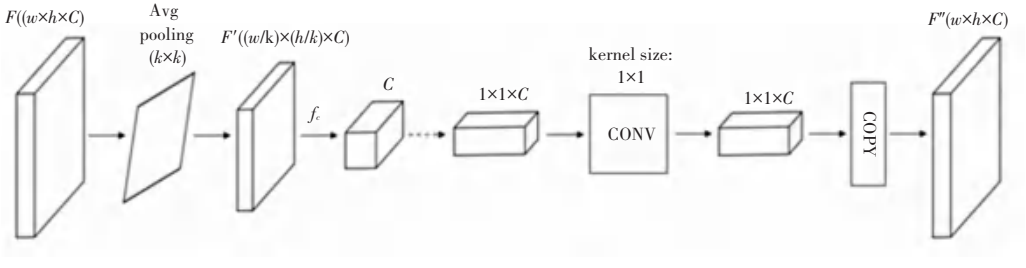

图 2　全图像编码器

Fig. 2　Full image encoder

通过研究发现全图像编码器存在以下缺点：

（1）只使用平均池化存在 2 个弊端：一方面，由于图像中感兴趣的对象往往会产生较大的像素值，因此只使用平均池化会丢失较大特征值像素的特征信息；另一方面，局部的平均池化只是使用小尺寸的卷积核在图像中进行局部卷积，难以很好地整合图像的全局信息；

（2）针对图像每个位置的信息，只将特征图简单地复制到整个图像，会丢失重要像素的位置信息。基于以上全图像编码器的缺点，本文使用 CBAM 替代全图像编码器，通过全局最大池化和全局平均池化更好地捕获较大特征值像素的特征信息。通过空间注意力机制生成的注意力图与原始特征图相乘替代简单的复制操作，保留完整的位置信息。

2.2　CBAM（Convolutional Block Attention Module）

如图 1 中的绿色部分所示，CBAM 依次通过通

道注意力机制和空间注意力机制,下面分别对通道注意力机制和空间注意力机制进行详细介绍。

通道注意力机制如图 3 所示。首先在空间维度上使用全局最大池化和全局平均池化操作对输入特征 $F \in R^{C\times H\times W}$ 进行压缩,生成 2 个不同的特征描述符;将 2 个描述符分别送入一个由多层感知机(multi-layer perceptron,MLP)构成的共享网络进行计算,进一步将共享网络输出的最大池化特征向量和平均池化特征向量以元素求和的方式进行合并;最终使用 *sigmoid* 函数将合并之后的特征向量映射到[0,1],进而得到通道注意力图。通道注意力图 $M_c \in R^{C\times 1\times 1}$ 的计算过程如式(1):

$$M_c(F) = \sigma\begin{pmatrix} MLP(MaxPool(F)) + \\ MLP(AvgPool(F)) \end{pmatrix}. \tag{1}$$

其中,σ 代表 *sigmoid* 函数。

图 3 通道注意力机制

Fig. 3 Channel attention module

空间注意力机制如图 4 所示。首先在通道维度上对经过通道注意图提炼之后的特征 $F' \in R^{C\times H\times W}$ 使用全局最大池化和全局平均池化操作,得到 2 个不同的特征描述符;使用卷积层对它们进行连接合并;最终使用 sigmoid 函数将合并之后的特征向量映射到[0,1],进而得到空间注意力图。空间注意力图 $M_s \in R^{H\times W}$ 的计算过程如式(2):

$$M_s(F) = \sigma(f^{7\times 7}([MaxPool(F');AvgPool(F')])). \tag{2}$$

其中,$f^{7\times 7}$ 代表卷积核尺寸为 7 × 7 的卷积运算。

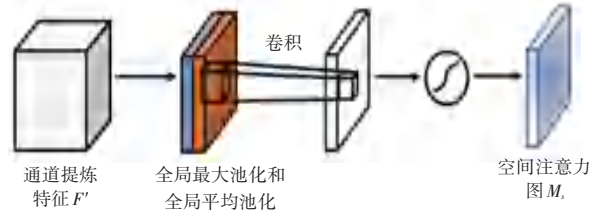

图 4 空间注意力机制

Fig. 4 Spatial attention module

得到通道注意力图和空间注意力图后,将通道注意力图与输入特征相乘得到 F',然后计算 F' 的空间注意力图,并将二者相乘得到最终的特征 F''。该过程可表示为式(3)和式(4):

$$F' = M_c(F) \otimes F. \tag{3}$$

$$F'' = M_s(F') \otimes F'. \tag{4}$$

其中,$\otimes$ 代表逐元素相乘。

将原始特征依次经过通道注意力图和空间注意力图的调整,使最终特征图中的较大特征值像素特征信息和位置信息更加完整。

2.3 损失函数和离散策略

总的序数损失被表示为每个像素的序数损失的平均值。每个像素的序数损失函数为式(5)和式(6):

$$\Psi(h,w,X,\Theta) = \sum_{k=0}^{l_{(w,h)}-1} \log(p^k_{(w,h)}) + \sum_{k=l_{(w,h)}}^{K-1} \log(1-p^k_{(w,h)}), \tag{5}$$

$$p^k_{(w,h)} = P(l'_{(w,h)} > k \mid X,\Theta). \tag{6}$$

其中,$l_{(w,h)} \in \{0,1,\cdots,K-1\}$ 代表在空间位置 (w,h) 通过使用 SID 离散策略得到的离散标签;$l'_{(w,h)}$ 代表预测的离散深度值;$p^k_{(w,h)}$ 通过 softmax 函数计算。

总的序数损失函数为式(7):

$$L(X,\Theta) = -\frac{1}{N}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}\Psi(h,w,X,\Theta). \tag{7}$$

其中,$N = W \times H$。

由于随着深度值的增大,用于深度估计的信息会逐渐减少,进而导致较大深度值的估计误差通常较大。因此使用 SID 策略进行离散化,该策略在对数空间中统一离散给定深度区间,以降低大深度值区域的训练损失,合理估计大深度值。假设深度区间$[\alpha,\beta]$ 需要离散为 M 个子段,SID 策略可表示为式(8):

$$s_i = e^{\log(\alpha)+\frac{\log\left(\frac{\beta}{\alpha}\right)\times i}{M}}, \tag{8}$$

其中,$s_i \in \{s_0,s_1,\cdots,s_M\}$ 代表离散阈值。

最终预测的深度值为式(9):

$$d_{(w,h)} = \frac{s_{l'_{(w,h)}} + s_{l'_{(w,h)}+1}}{2} - \varepsilon. \tag{9}$$

其中,ε 为偏移值,$\alpha + \varepsilon = 1$。

3 实验过程及结果

3.1 实验设置

KITTI 数据集主要包含室外场景,数据由装载在行驶汽车上的相机和深度传感器捕获,图像大小为 375×1241 像素[12]。本文算法在 KITTI 数据集上进行训练和测试,数据切分方式从 29 个场景中切分出 697 幅图像进行测试,其余的 32 个场景中的

23 488幅图像用于训练和交叉验证，其中22 600幅用于训练，剩余的图像用于验证。实验中，网络结构使用 Pytorch 框架实现，训练时将输入图像大小调整为 385×513。网络使用 SGD 优化器进行优化，动量缩减参数设置为 0.9，权重缩减参数设置为 0.0005，初始学习率设置为0.000 1，mini－batch 尺寸设置为 4。

将训练模型的实验结果与其它相关方法进行对比，采用常用的评价指标来评估结果，其中 d_i 表示真实深度；d_i^* 表示预测深度；N 表示图像的像素总数。指标表达式为：

• 绝对相对误差（absolute relative error，AbsRel），式(10)：

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i^* - d_i}{d_i}, \tag{10}$$

• 平方相对误差（squared relative error，SqRel），式(11)：

$$SqRel = \frac{1}{N}\sum_{i=1}^{N}\frac{\| d_i^* - d_i \|^2}{d_i}, \tag{11}$$

• 均方根误差（root mean squared error，RMSE），式(12)：

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\| d_i^* - d_i \|^2}, \tag{12}$$

• 准确率：满足如下条件的像素占总像素的百分比，式(13)：

$$\delta = \max\left(\frac{d_i^*}{d_i}, \frac{d_i}{d_i^*}\right) < thr. \tag{13}$$

其中，$thr = 1.25, 1.25^2, 1.25^3$。

3.2　实验结果与分析

本文方法与几个先进的单目深度估计方法的对比结果见表 1，这些方法中包括了有基于监督学习的方法（Eigen et al.[2]、Liu et al.[13] 和 Gan et al.[14]）、半监督学习的方法（Kuznietsov et al.[15]）和无监督学习的方法（Garg et al.[16] 和 Yin et al.[17]）。从实验结果可以看出，本文算法的深度估计效果明显优于无监督学习方法的效果，同时也达到甚至超过了有监督学习方法的效果，这主要得益于在训练过程中，通过使用通道注意力机制和空间注意力机制提高了全局信息的表示能力。为了证明算法改进部分的有效性，在表 1 中还提供了在 KITTI 数据集上的消融实验结果，各项指标的结果证明了改进部分的有效性。

表 1　KITTI 数据集上的实验结果对比

Tab. 1　Comparison of experimental results on the KITTI dataset

方法	有/无监督	AbsRel	SqRel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
		误差（越小越好）			精度（越大越好）		
Eigen et al.	有	0.215	1.515	7.156	0.692	0.899	0.967
Liu et al.	有	0.202	1.614	6.523	0.678	0.895	0.965
Gan et al.	有	0.098	0.666	3.933	0.890	0.964	0.985
Kuznietsov et al.	半监督	0.113	0.741	4.621	0.862	0.960	0.986
Garg et al.	无	0.169	1.080	5.104	0.740	0.904	0.962
Yin et al	无	0.155	1.296	5.857	0.793	0.931	0.973
Fu et al	有	0.072	0.307	2.727	0.932	0.984	**0.994**
Ours	有	**0.064**	**0.286**	**2.697**	**0.944**	**0.989**	0.992

3.3　消融实验

消融实验的结果见表 2。主要对网络中 CBAM 中通道注意力和空间注意力机制的使用顺序进行了分析。在只使用通道注意力机制、先空间注意力机制后通道注意力机制和先通道注意力机制后空间注意力机制 3 个方面进行实验。通过分析表 2 可知，先通道注意力机制后空间注意力机制的精度比只使用通道注意力机制和先使用空间注意力机制后使用通道注意力机制的效果都高，说明先通道注意力机制后空间注意力机制的顺序结构可以捕获像素更完整的特征信息和位置信息。

表 2　消融实验结果

Tab. 2　Results of ablation experiment

方法	AbsRel	SqRel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
	误差（越小越好）			精度（越大越好）		
只使用通道注意力	0.275	1.796	6.364	0.598	0.781	0.893
空间注意力+通道注意力	0.149	0.638	3.507	0.887	0.932	0.957
通道注意力+空间注意力	**0.064**	**0.286**	**2.697**	**0.944**	**0.989**	**0.992**

KITTI 数据集上的深度估计的效果图如图 5 所示。与其它方法相比,该模型在细节处理方面具有更强大的能力,主要表现在小物体、行人以及树木等区域保留了更为丰富的纹理信息,细节处理更加平滑,且前景和背景分离效果更好。

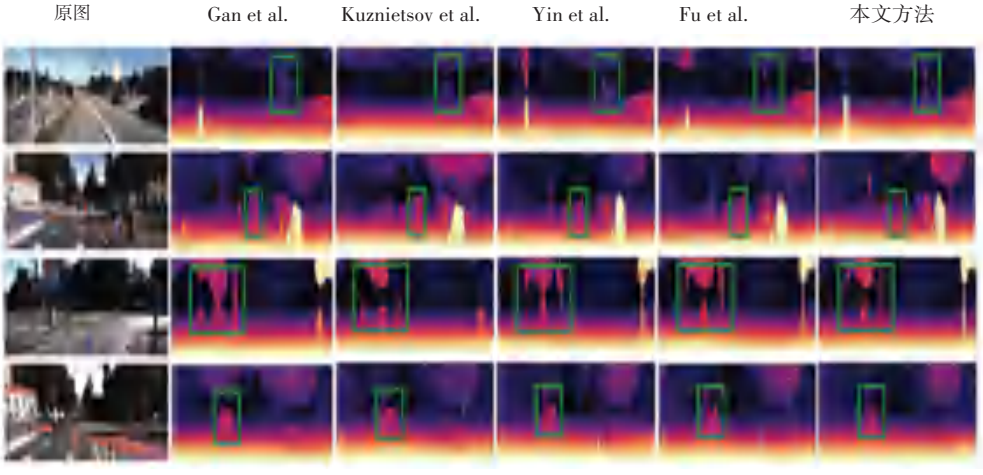

图 5 各模型深度预测结果

Fig. 5 Depth prediction results of each model

为了评估本文方法的泛化能力,本文还在 Cityscapes 数据集做了测试实验,效果如图 6 所示。该方法只使用 KITTI 数据集进行训练和评估,而没有使用 Cityscapes 数据集。虽然两个数据集的场景类型存在一定差异,但是该方法仍然可以输出效果很好的深度图像。

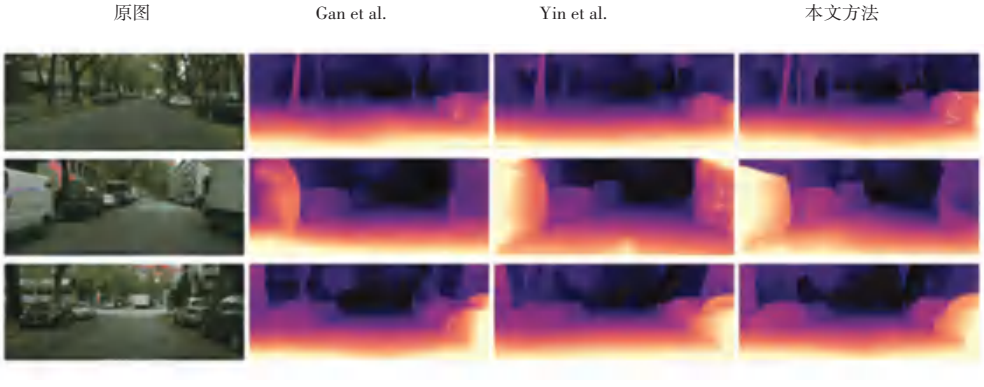

图 6 在 Cityscapes 数据集上的测试效果图

Fig. 6 Test effect diagram on Cityscapes dataset

4 结束语

针对有监督学习的单目深度估计模型深度序数回归算法中全图像编码器易丢失较大像素特征信息和位置信息的问题,本文提出一种基于 CBAM 的深度序数回归方法。通过一系列的对比试验和消融实验,展示出了该方法的优异性和合理性。对比基础网络,该方法的网络模型捕获了更多目标的特征信息和位置信息,更加完整地保留了图像中较小目标或其他细节的特征。通过利用 KITTI 数据集和 Cityscapes 数据集对该方法进行验证,表明其高于现有的大部分深度估计方法。

参考文献

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

[2] EIGEN D,PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]// InAdvances in neural information processing systems,2014:2366-2374.

[3] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scaleconvolutional architecture[C]// InProceedings of the IEEE international conference on computer vision,2015:2650-2658.

[4] Li, Jun,Reinhard Klein, Angela Yao. Learning fine-scaled depth maps from single RGB images[J].arXiv preprint,2016.

[5] LAINA I, RUPPRECHT C, BELAGIANNIS V, et al. Deeper

depth prediction with fully convolutional residual networks[C]// Proceedings of the 2016 international conference on 3D vision. Piscataway：IEEE，2016：239-248.
[6] LIU F，SHEN C，LIN G. Deep convolutional neural fields for depth estimation from a single image[C]//InProceedings of the IEEE conference on computer vision and pattern recognition，2015：5162-5170.
[7] XU D，RICCI E，OUYANG W，et al. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation[C]// InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：5354-5362.
[8] 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.
[9] CHANG J R，CHEN Y S. Pyramid stereo matching network[C]// InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：5410-5418.
[10] CHEN P Y，LIU A H，LIU Y C，et al. Towards scene understanding：Unsupervised monocular depth estimation with semantic-aware representation[C]// InProceedingsof the IEEE Conference on Computer Vision and Pattern Recognition，2019：2624-2632.
[11] LEE J H，KIM C S. Monocular depth estimation using relative depth maps[C]// InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2019：9729-9738.
[12] GEIGER A，LENZ P，STILLER C，et al. Vision meets robotics：Thekittidataset[J]. The International Journal of Robotics Research，2013，32(11)：1231-1237.
[13] LIU F，SHEN C，LIN G，et al. Learning depth from single monocularimages using deep convolutional neural fields[J]. IEEE transactionson pattern analysis and machine intelligence，2015，38(10)：2024-2039.
[14] GAN Y，XU X，SUN W，et al. Monocular depth estimation with affinity，vertical pooling，and label enhancement[C]// InProceedings of the European Conference on Computer Vision (ECCV)，2018：224-239.
[15] KUZNIETSOV Y，STUCKLER J，LEIBE B. Semi-supervised deep learning for monocular depth map prediction[C]// InProceedings of the IEEE conference on computer vision and pattern recognition，2017：6647-6655.
[16] GARG R，BG V K，CARNEIRO G，et al. Unsupervised cnn for single view depth estimation：Geometry to the rescue[C]// InEuropean conference on computer vision. Springer，Cham，2016：740-756.
[17] YIN Z，SHI J. Geonet：Unsupervised learning of dense depth，optical flow and camera pose[C]// InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：1983-1992.

(上接第 18 页)

[3] YANG C，XIE L，QIAO S，et al. Training deep neural networks in generations：A more tolerant teacher educates better students[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019，33(1)：5628-5635.
[4] AHN S，HU S X，DAMIANOU A，et al. Variational information distillation for knowledge transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：9163-9171.
[5] TUNG F，MORI G. Similarity-preserving knowledge distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：1365-1374.
[6] YIM J，JOO D，BAE J，et al. A gift from knowledge distillation：Fast optimization，network minimization and transfer learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：4133-4141.
[7] KOMODAKIS N，ZAGORUYKO S. Paying more attention to attention：improving the performance of convolutional neural networks via attention transfer[C]//ICLR. 2017，3：4.
[8] HOU Y，MA Z，LIU C，et al. Learning lightweight lane detection cnns by self attention distillation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019：1013-1021.
[9] 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. 2016：770-778.
[10] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015：1-9.
[11] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv：1409.1556，2014.
[12] LONG J，SHELHAMER E，DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015：3431-3440.
[13] RONNEBERGER O，FISCHER P，BROX T. U-net：Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer，Cham，2015：234-241.
[14] ZHAO H，SHI J，QI X，et al. Pyramid scene parsing network[C]//Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition，F，2017.
[15] CHEN L-C，PAPANDREOU G，SCHROFF F，et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv：1706.05587，2017.
[16] 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，2017，40(4)：834-848.
[17] PASZKE A，CHAURASIA A，KIM S，et al. Enet：A deep neural network architecture for real-time semantic segmentation[J]. arXiv preprint arXiv：1606.02147，2016.
[18] ROMERA E，ALVAREZ J M，BERGASA L M，et al. Erfnet：Efficient residual factorized convnet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems，2017，19(1)：263-272.
[19] PAN X，SHI J，LUO P，et al. Spatial as deep：Spatial cnn for traffic scene understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence，2018.
[20] EVERINGHAM M，GOOL L V，WILLIAMS C，et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision，2010，88(2)：303-338.