文章编号：2095-2163(2021)05-0013-07 中图分类号：TP391.41 文献标志码：A

面向图像语义分割任务的多级注意力蒸馏学习

刘佳琦[1,2]，杨 璐[1,2]，王龙志[3]

(1 天津理工大学 天津市先进机电系统设计与智能控制重点实验室，天津 300384；2 天津理工大学 机电工程国家级实验教学示范中心，天津 300384；3 奥特贝睿(天津)科技有限公司，天津 300300)

摘 要：传统的蒸馏学习仅通过大网络对轻量网络进行单向蒸馏，不但难以从轻量网络的学习状态中得到反馈信息，对训练过程进行优化调整，同时还会限制轻量网络的特征表达能力。本文提出结合自身多级注意力上下文信息进行自我学习优化的方法(MAD，Multi Attention Distillation)，以自监督的方式使自身成熟的部分约束不成熟部分，即浅层可以从深层中提取有用的上下文信息，让浅层特征学习高层特征的表达，从而提升网络的整体表达能力。使用轻量级网络 ERFNet、DeepLab_V3 在两个不同任务的数据集 CULane、VOC 上进行验证。实验结果表明，MAD 可以在不增加推理时间的前提下，提升网络的特征提取能力，使 ERFNet 在 CULane 任务的 F_1 - *measure* 指标提升 2.13，DeepLab_V3 在 VOC 任务的 mIoU 指标提升 1.5。

关键词：蒸馏学习；语义分割；注意力；卷积神经网络

MAD：Multi attention distillation learning for semantic segmentation

LIU Jiaqi[1,2]，YANG Lu[1,2]，WANG Longzhi[3]

(1 Tianjin Key Laboratory for Advanced Mechatronic System Design and Intelligent Control，School of Mechanical Engineering，Tianjin University of Technology，Tianjin 300384，China；2 National Demonstration Center for Experimental Mechanical and Electrical Engineering Education (Tianjin University of Technology)，Tianjin 300384，China；3 Autobrain(Tianjin) Technology，LTD，Tianjin 300300，China)

【Abstract】 Traditional distillation learning only uses one-way distillation of large network to light-weight network，which not only is difficult to get feedback information from the learning state of light-weight network and optimize the training process，but also limits the feature expression ability of light-weight network. This paper proposes a self-learning optimization method (MAD，Multi Attention Distillation) based on multi-level attention context information，which makes the mature part restrain the immature part by self-monitoring，that is，the shallow layer can extract useful context information from the deep layer，and let the shallow layer learn the expression of high-level features，so as to improve the overall expression ability of the network. Using lightweight network ERFnet and DeepLab_V3 to verify on two different task datasets，CULane and VOC，mad can improve the network performance without increasing the reasoning time. Improved the F_1 - *measure* index of ERFNet in CULane task by 2.13，and improved the mIoU index of DeepLab_V3 in VOC task by 1.5.

【Key words】 distillation learning；semantic segmentation；attention；convolutional neural network

0 引 言

知识蒸馏是深度学习领域一项重要的模型压缩技术。传统的蒸馏学习思想是通过提前训练好的大网络对轻量网络进行知识传递，从而使轻量网络能达到大网络的表达能力，实现知识迁移。基于传统蒸馏学习的模型训练主要分为两个步骤：首先充分训练一个结构复杂、学习能力强的教师网络，使其具有优秀的表达能力；其次在教师网络的基础上设计一个结构简单、参数量小的学生网络，使用教师网络的特征约束作为软标签进行监督，使学生网络通过软标签对真实标签辅助训练，逐渐逼近教师网络的表达水平。从模型推理方面分析，教师网络只在训练阶段对学生网络起到约束作用，不参与学生网络的独立推理过程的计算与部署，因此知识蒸馏在神经网络模型轻量化领域有着重要的意义。

由于传统蒸馏学习中的教师网络对学生网络的知识传递是单向的，难以从学生网络的学习状态中得到反馈信息，来对训练过程进行优化调整，从而对学生网络的训练产生负影响；其次，采取教师网络产生软标签结合真实标签进行监督的形式，当软标签权重过高时，学生网络会过于模仿教师网络，从而限

基金项目：天津市自然科学基金(16JCQNJC04100)。

作者简介：刘佳琦(1996-)，男，硕士研究生，主要研究方向：机器视觉、模式识别；杨 璐(1982-)，女，博士，副教授，主要研究方向：机器视觉、模式识别；王龙志(1983-)，男，博士，工程师，主要研究方向：无人车。

通讯作者：杨 璐 Email：yanglu8206@163.com

收稿日期：2021-01-31

制学生网络的特征表达能力;由于针对不同困难程度的数据集任务,所需要的教师网络的软标签监督权重也有所不同,因此增加了训练过程的难度。

近年来,人们尝试将网络自身的特征作为软标签,不需要训练教师网络,实现了轻量网络模型自身的监督优化。SAD 尝试使用两个相邻层级间的上下文信息,使得浅层特征学习高层特征,从而实现网络性能的整体提升。但是,以相邻层级信息进行约束不能对上下文信息进行充分的利用。本文借鉴了 Densenet 对 Resnet 的优化改进思想,使得上下文特征信息可以在整个网络中被充分利用,进而保证每一层级学习的特征约束,都能作用到网络之后的所有层级。

综上所述,针对已有研究工作的不足本文提出了一种自适应多级注意力蒸馏学习方法(MAD)。该方法使用网络自身的高层特征对浅层逐级进行约束,以自身的深层单元来约束浅层单元,以此实现模型知识由深层向浅层的传递。在充分利用上下文信息的基础上,对各层级间的表达能力进行提升。

1 相关工作

1.1 知识蒸馏

知识蒸馏的提出,是为了实现知识迁移。Hinton[1]系统的诠释了知识蒸馏的概念,并以教师网络的输出作为软标签来监督学生网络训练,从而验证了知识迁移的可行性。在后续的大量研究中,探究了提高知识迁移效率的方法。FITNETS[2]提出添加教师网络中间层的特征,作为学生网络学习的软标签,使得学生网络在关注教师模型输出的同时,实现了中间层的特征约束。文献[3]认为硬标签会导致模型在训练过程中发生过拟合,而使用软标签更能提高模型的泛化能力,在训练任务中选择几个具有最高置信度分数的类,可以作为软标签来计算损失。为了更好的表征神经网络中间层的特征,文献[4-5]认为,可以优化中间层特征的表征编码方式;文献[6]认为,神经网络层与层之间的特征关系更能作为特征提取能力的指标,因此让学生网络学习到教师网络的层级之间的特征关系更为重要;文献[7]认为,以特征图作为软标签进行传递效果不佳,提出使用注意力图代替特征图,最小化教师网络与学生网络之间注意力图的欧氏距离,可以获得更好的效果;文献[8]中认为,基于激活的注意力蒸馏会产生明显的性能提高,而基于梯度的注意力蒸馏提升相对较小。

1.2 语义分割

语义分割属于像素级别的分类任务,通过对每个像素进行密集的预测来实现细粒度的推理。语义分割体系结构被广泛认为是编解码结构(Encoder-Decoder)。其中编码器通常为特征提取网络,典型的编码特征提取网络有:Resnet[9]、GoogleNet[10]、VGGNet[11]。解码器将编码器学习提取到的语义特征投影到像素空间上得到密集的分类。

FCN[12]作为经典的编解码语义分割结构,取得了瞩目的成就,使用反卷积对卷积特征进行上采样,对每个像素类别进行预测,实现图像语义分割。U-Net[13]在 FCN 的基础上,通过加入更多的底层特征,实现了对小物体细节分割质量的提升,PSPNet[14]通过利用空间金字塔池化模块进行编码,实现了对多尺度信息进行特征融合。同时,DeepLab_V3[15]在 DeepLab_V2[16]的基础上,对空间金字塔模块进行改进,取得了分割精确性的提升。ENet[17]使用下卷积层并行池化层来进行下采样,解码阶段使用空洞卷积,在获得大感受野的同时获得丰富的上下文信息。ERFNet[18]在 ENet 的基础上进行改进,使用非对称卷积,实现了网络准确性与实时性的提升。

2 多级注意力蒸馏学习

2.1 多级蒸馏

将卷积神经网络划分为几个单元,使得前一单元可以从后续的各个单元中提取有用的上下文信息。本文以 ERFNet 网络结构为例,将网络的 Encoder 部分拆分成 6 个单元,如图 1 所示,以自身的深层单元来约束浅层单元,以实现模型知识由深层向浅层的传递,提高浅层的表达能力,从而提升模型整体的表达能力。

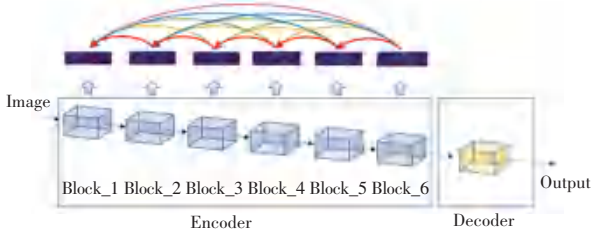


图 1 MAD 算法结构图

Fig. 1 Structure of MAD algorithm

对于一个共有 L 个单元的网络,如图 2 所示,共包含 N 级蒸馏。将各级蒸馏中每个分支进行加和,各级蒸馏损失函数如下:

$$N = \frac{L \times (L-1)}{2}, \tag{1}$$

$$Loss\_n = \sum_{i=1}^{L-n} Loss\_n\_i. \tag{2}$$

式中，n 为蒸馏级别，i 代表了当前级别下的序数。

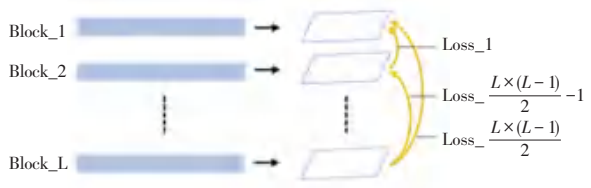


图 2　多级注意力蒸馏示意图

Fig. 2　Multi-Res Distillation

将总蒸馏损失函数设为各级蒸馏加权损失之和，计算公式如下：

$$Loss\_d = \sum_{n=1}^{N} \alpha_n \times Loss\_n. \tag{3}$$

式中，a_n 为一个自适应的权重系数，与初设的单元数量有关。设定相邻层级蒸馏可获得高权重，蒸馏层级跨度越大获得权重越低，因此自适应权重系数 α_n 为：

$$\alpha_n = \frac{1}{n}. \tag{4}$$

如公式(5)所示，通过计算加权的分割损失与蒸馏损失，可得到总的损失函数。本文使用一个权重系数 β 来平衡分割损失与蒸馏损失对最终任务的影响。(权重系数 β 将在 3.3 节进行讨论)

$$Loss\_Tol = Loss\_seg + \beta \cdot Loss\_d. \tag{5}$$

所有蒸馏部分只在训练期间进行计算，因此在测试推理过程中不增加计算量。

2.2　激活注意力

注意力图主要可以分为两类：基于激活的注意力图与基于梯度的注意力图。文献[7]的研究发现，基于激活的注意力蒸馏可显著提高性能；基于梯度的注意力蒸馏，提升效果不明显。因此，本文使用基于激活的注意力蒸馏方法。

为了定义一个空间注意力映射，可以将隐藏神经元激活的绝对值，作为该神经元相对于输入的注意力。因此，通过计算这些值在通道维度上的统计，来构建空间注意力图。生成注意力图时，在特征经过 softmax 操作之前，使用双线性上采样，使不同层输出的特征图尺寸保持统一。注意力图生成过程如图 3 所示。

针对不同的任务，注意力图的计算方式会有所不同，这主要与任务中特征的种类、数量、复杂程度有关。对于车道线检测任务，计算注意力图使用 $A \in R^C \times R^H \times R^W$ 表示网络卷积层的激活输出，其中 C、H 和 W 分别表示通道、高度和宽度。通过计算通道维度上的统计来表征注意力图，有以下 2 种有效操作[7]可以作为映射函数：

$$F_{sum}(A) = \sum_{c=1}^{C} |A_c|, \tag{6}$$

$$F^p_{sum}(A) = \sum_{c=1}^{C} |A_c|^p. \tag{7}$$

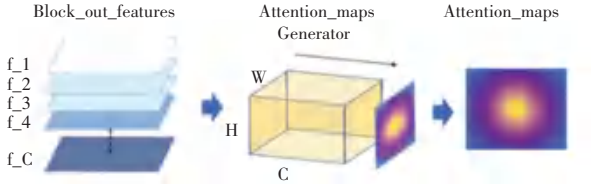


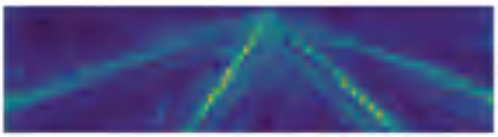
图 3　注意力图生成过程

Fig. 3　The generating process of attention maps

针对车道线检测任务，希望网络注意力集中在车道线附近的区域，本文可视化了 2 种映射函数提取得到的注意力图，如图 4 所示。通过对比两种映射函数发现：$F^p_{sum}(A)$ 作为映射函数，可以使注意力图更加集中在特征区域。其中 p 越大，特征区域的激活程度越高。根据对比分析，当 p 取 2 时，会使注意力图的偏差更小。

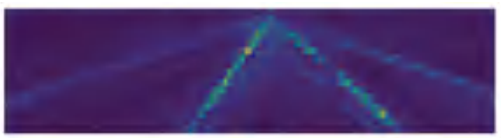
(a) 输入图片

(a) Input image



(b) $F_{sum}(A)$ 映射输出

(b) Mapping output of Fsun $F_{sum}(A)$



(c) $F^p_{sum}(A)$ 映射输出

(c) Mapping output of $F^p_{sum}(A)$

图 4　不同映射方法结果

Fig. 4　Results of different mapping methods

因此，本文注意力图计算公式为：

$$A_i = \sum_{c=1}^{C} A\_c\_i \times A\_c\_i. \tag{8}$$

式中，C 为卷积层的通道数，即特征图的数量。注意力图即为当前单元各通道特征图的平方和，从而可以计算出每个层级单元间注意力蒸馏损失函数：

$$Loss\_n\_i = \| A_i^{previous} - A_i^{back} \|_2^2. \quad (9)$$

式中，$A_i^{previous}$、A_i^{back} 分别代表当前蒸馏级别序数下，2 个单元中前一单元与后一单元的注意力输出。

3 实 验

使用轻量级网络 ERFNet、DeepLab_V3 在 2 个不同难度任务的分割数据集 CULane、VOC2012 上进行验证。

3.1 数据集

3.1.1 CULane

CULane 数据集[19]是一个大规模车道检测数据集，包含了许多具有挑战性的驾驶场景。如，拥挤的道路条件或照明不足的道路。其是由安装在北京不同司机驾驶的 6 辆不同车辆上的摄像头采集而得。收集了超过 55 h 的视频，提取了 133 235 帧。将数据集分成 88 880 个训练集、9 675 个验证集、34 680 个测试集。测试集分为正常和 8 个挑战性类别，对应于表 1 中的 9 个示例。

3.1.2 PASCAL VOC

PASCAL VOC 2012 数据集[20]是常用语义分割的数据集。该数据集拥有 1 464 张训练图片、1 449 张验证图片和 1 456 张测试图片。其中包括 20 个前景类别和一个背景类别共 21 个语义分类，该数据集中的大部分图像的分辨率接近 500×500。

3.2 评价指标

实验中采用 3 个评价指标来衡量车道线检测算法的性能，分别是精确度（*precision*）、召回率（*recall*）、F_1 度量（$F_1 - measure$）。其中，精确度表示正确预测为真的正样本占全部预测为真的样本的比例，即正确检测为道路的像素占全部检测为道路像素的百分比；召回率表示正确预测为真的正样本占全部正样本的比例，即正确检测为道路的像素占全部道路像素的百分比；$F_1 - measure$ 作为综合指标，是精确率和召回率的加权调和平均，受平衡准确率和召回率的影响。其计算公式如下：

$$F_1 - measure = \frac{2 \times precision \times recall}{(precision + recall)} \times 100\%. \quad (10)$$

同时，针对 VOC 数据集的语义分割性能评估，以 mIoU 的值作为评价指标，来说明 MAD 方法对语义分割任务性能的提升。分别计算每个类别的 IoU（Intersection over Union）值再求平均来计算。评估过程还包括整体准确度（*Acc*）、分类准确度（*Acc_class*）和带权重交并比（*fwavacc*）。

3.3 ERFNet

为验证 MAD 方法的有效性，进行了对比试验。以 ERFNet 网络为 baseline，使用 MAD 多级注意力蒸馏进行优化，设置仅使用第一级蒸馏损失优化作为单级注意力蒸馏（SAD）方法作为对比试验。使用 SGD 优化器，学习率为 5e-2；在训练阶段使用了预训练权重，对模型训练了 12 个 epoch；设置 batch_size 为 12，共优化 88K 次迭代数。通过在数据集 CULane 上进行训练，并使用 CULane 的验证集对 9 个场景任务进行测试，分别计算出 3 个指标值，并结合训练过程的验证信息进行综合评价。

针对不同类别的测试结果，进行了详细的实验结果分析。在表 1、表 2 中，对比了在不使用优化方法下模型训练后的测试结果（baselines），与使用单级（SAD）、多级（MAD）注意力不同蒸馏权重 $\beta = \{100, 200\}$ 的结果。

见表 1，针对 9 种类别场景中的车道线分割任务，无论是基于单级（SAD）还是多级（MAD）注意力蒸馏的结果，均较原始网络有了明显的提升。根据表 2 所展示综合评价结果，在不同的权重因子下，通过比较两种方法可以发现，多级（MAD）蒸馏优化都要高于单级（SAD）蒸馏优化。同时，本实验可视化了单级（SAD）、多级（MAD）注意力蒸馏优化方法的验证结果与不同蒸馏权重的对比结果（$\beta = \{10, 50, 100, 200, 500\}$），如图 5 所示。

表 1 基于单级（SAD）、多级（MAD）注意力蒸馏的 9 种场景结果

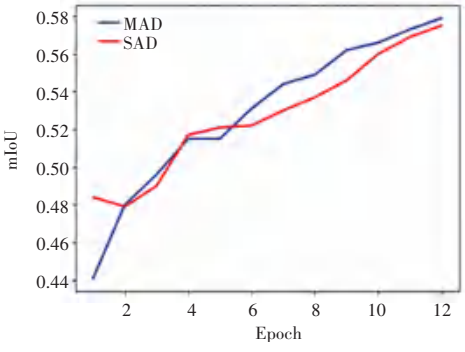
Tab. 1 Results of 9 scenarios based on single level（SAD）and multi-level（MAD）attention distillation

Accuracy / model	*Normal*	*Crowded*	*Night*	*no line*	*shadow*	*arrow*	*Dazzle light*	*curve*	*crossroad*
Org（baselines）	0.883 9	0.673 1	0.481 0	0.409 2	0.567 6	0.825 0	0.576 0	0.620 5	1 977
SAD200	0.884 4	0.675 5	0.560 7	0.401 8	0.591 6	0.811 5	0.582 7	0.620 4	1 646
SAD100	0.883 7	0.668 8	0.526 7	0.406 3	0.560 2	0.815 7	0.593 1	0.645 4	2 145
MAD200	0.888 8	0.681 4	0.539 0	0.408 3	0.616 3	0.820 7	0.598 9	0.636 0	1 755
MAD100	**0.889 0**	**0.685 7**	**0.547 0**	**0.431 125**	**0.626 5**	**0.833 2**	**0.600 3**	**0.622 0**	**1 783**

表 2　基于单级(SAD)、多级(MAD)注意力蒸馏的综合评价结果

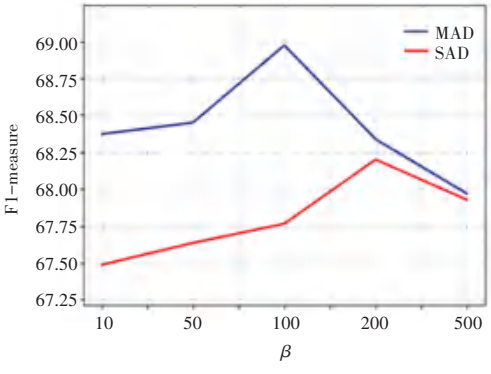
Tab. 2　Comprehensive evaluation results based on single level (SAD) and multi-level (MAD) attention distillation

Accuracy / model	*precision*	*recall*	F_1 – *measure*
Org(baselines)	0.706 3	0.634 4	66.845 5
SAD200	0.713 3	0.653 3	68.199 8
SAD100	0.700 9	0.646 6	67.265 0
MAD200	0.715 4	0.654 1	68.337 4
MAD100	**0.722 4**	**0.659 9**	**68.973 4**



(a) SAD 与 MAD 验证结果

(a) Comparison of validation results of SAD and MAD



(b) 不同蒸馏权重的对比

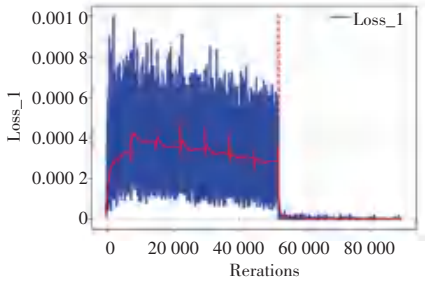
(b) Comparison of different distillation weights

图 5　对比结果

Fig. 5　Comparison results

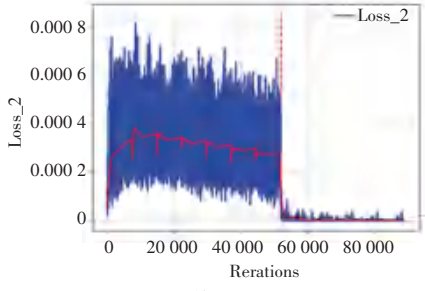
图 5(a)为 $\beta = 200$ 的训练过程评估。在随机初始化后，以同样的参数训练 5 个 epoch，在第 6 个 epoch 时引入优化方法。图中可以证明，在第 6 个循环周期(epoch)后，多级(蓝线)方法始终优于单级(红线)。图 5(b)所示为本文探索不同权重因子的两种蒸馏优化结果对比。从图中可以看出，不同权重因子下多级(蓝色)方法较单级(红色)方法仍有明显的优势。

图 6 所示为 $\beta = 200$ 的训练过程损失曲线。在实验中，设置第 6 个循环周期(epoch)后加入优化。MAD 优化方法在不同级别的蒸馏分支中，加入之前模型收敛速度慢，加入之后实现快速的收敛。相对整体的模型收敛状态，并不会产生震荡的后果。具体测试结果如图 7 所示。



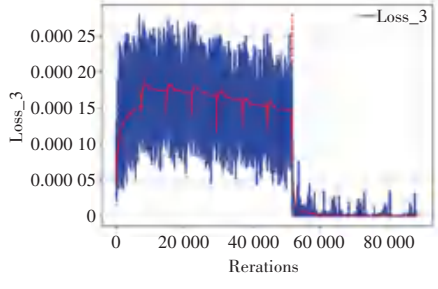
(a) 第一级蒸馏损失

(a) The first-stage distillation loss



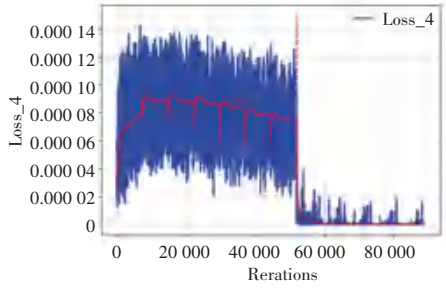
(b) 第二级蒸馏损失

(b) The second-stage distillation loss



(c) 第三级蒸馏损失

(c) The third-stage distillation loss



(d) 第四级蒸馏损失

(d) The fourth-stage distillation loss

图 6　各级蒸馏损失函数

Fig. 6　Distillation loss function of each stage

3.4　DeepLab_V3

为探索多级自适应蒸馏学习(MAD)的普适性，即是否可以应用于多种类别的分割任务，本文将实

验扩展为基于 DeepLab_V3 网络在 VOC2012 数据集上训练 20 类别的语义分割实验。使用 DeepLab_V3 为基础网络,将网络特征提取部分分为 4 个单元,在 PASCL VOC 上训练了 200 个循环,在训练 80 次循环后加入 MAD 模块。batch_size = 16,学习率为 0.01。实验训练结果见表 3,验证结果见表 4。

表 3 DeepLab_V3 在 VOC2012 数据集上的训练结果

Tab. 3 Training results of Deeplab_V3 on VOC2012 datasets

LOSS/Evaluation	*Acc*	*Acc_class*	*mIoU*	*fwavacc*
org	94.32	86.84	79.93	89.41
MAD	94.55	87.16	80.33	89.82

表 4 DeepLab_V3 在 VOC2012 数据集上的测试结果

Tab. 4 Testing results of Deeplab_V3 on VOC2012 datasets

LOSS/Evaluation	*Acc*	*Acc_class*	*mIoU*	*fwavacc*
org	81.72	46.74	37.15	69.81
MAD	83.10	48.38	38.65	72.45

根据表 3、表 4 中结果可以看出,MAD 方法针对 20 类别的语义分割具有明显的提升效果。其中 Acc 为整体准确度,Acc_class 为分类准确度,fwavacc 为带权重交并比。在训练集 mIoU 指标中 MAD 方法在原网络的基础上提升 0.4,在验证集 mIoU 指标中本文方法提升 1.5。

(a) 弯道场景 (b) 匝道场景 (c) 常规场景

(a) Curve scene (b) Ramp scene (c) Conventional scene

图 7 ERFNet 的检测结果

Fig. 7 ERFNet predict results

4 结束语

本文提出了一种多级注意力蒸馏学习方法(MAD),以自身网络的浅层特征学习高层特征的表达。实验证明,该方法可普遍提高网络中不同层次的视觉注意力,使 ERFNet 在 CULane 任务的 F_1 – *measure* 指标提升 2.13,DeepLab_V3 在 VOC2012 任务的 mIoU 指标提升 1.5,在提升网络特征提取能力方面具有重要意义。

选择合适的不同层级之间的权重,对训练时间与收敛具有一定影响,在后续的工作中,可以考虑探究每个蒸馏级别中不同层之间的权重,从而实现模型收敛性能的提高。

参考文献

[1] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J]. Computer ence, 2015, 14(7):38-39.

[2] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.

(下转第 25 页)