文章编号：2095-2163(2022)11-0204-06 中图分类号：TP181 文献标志码：A

基于 Transformer 的道路场景分割算法研究

魏鹏磊，雷菊阳
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要： 图像语义分割技术作为计算机视觉领域的关键技术之一，可以识别并理解图像中每一个像素的内容，并已应用在自动驾驶、医疗诊断、地理信息系统以及图像搜索等很多场景。相对于深度卷积神经网络，Transformer 模型基于纯注意力机制，没有任何卷积层或循环神经网络层。本文在 Swin Transformer 的基础上进行了改进，提出了一种新的网络结构 SwinLab。实验结果表明改进后的 SwinLab 模型相比于深度卷积神经网络的模型算法以及原 Swin Transformer 模型的分割精度不相上下，*mIoU* 可达 80.1，同时在 CityScapes 数据集上也进行了对比实验，从而进一步证明了该结构的有效性和泛化性。综上，本文在以 Swin Transformer 为骨干网络的基础上做了相关工作，从而使模型结构更简单，训练和推理速度更快，且准确率也相当可观。

关键词： 语义分割；卷积神经网络；Transformer；注意力机制

Research on road scene segmentation algorithm based on Transformer

WEI Penglei, LEI Juyang
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 As one of the key technologies in the field of computer vision, images semantic segmentation technology can identify and understand the content of each pixel in an image, and is used in many scenarios such as autonomous driving, medical diagnosis, geographic information systems, and images search. Compared with deep convolutional neural networks, the Transformer model is based on a pure attention mechanism without any convolutional or recurrent neural network layers. Improvements have been implemented on the basis of Swin Transformer, and a new network structure SwinLab is proposed in this paper. The experimental results show that the segmentation accuracy of the improved SwinLab model is comparable to that of the deep convolutional neural network model algorithm and the original Swin Transformer model, and the *mIoU* can reach 80.1. At the same time, a comparative experiment is also carried out on the CityScapes dataset, so the effectiveness and generalizability of this structure is furtherly demonstrated. In summary, this paper has performed related work on the basis of Swin Transformer as the backbone network, so that the model structure is simpler, the training and inference speed is faster, and the accuracy rate is also considerable.

【Key words】 semantic segmentation; Convolutional Neural Network; Transformer; attention mechanism

0 引 言

常见的基于深度卷积神经网络的模型，如 FCN[1]、Deeplab v3[2]、SegNet[3] 等在传统语义分割任务上有着很好的效果，但是对于城市道路场景的分割仍然难以达到理想的状态。随着 Transformer 在 NLP 领域表现优异性能之后，越来越多的人尝试将其应用在 CV 领域，并取得了可观进步。继 ViT[4] 之后，出现了很多 Transformer 运用在 CV 各个任务上的工作，而 Swin Transformer[5] 是第一个备受青睐的可以在下游任务中使用的纯 Transformer 结构的方式，但却有着如下缺点：参数量过大、显存占用高、训练时间长。究其原因，下游任务，如语义分割是高密度预测任务，对于分割精度要求很高，从而使得训练参数量巨大，增加了训练成本。故本文在 SwinT 的基础上改进了网络结构，可以明显加快训练速度，也可以很好地定位分割边界。其次，针对特征信息学习不充分问题，传统做法为通过设置不同参数的卷积层或池化层，先提取到不同尺度的特征图，再将这些特征图送入网络做融合。但是由于图像金字塔的多尺度输入，在计算时需要保存大量梯度，故对硬件的要求很高。而本次研究是将网络进行多尺度训练，在测试阶段进行多尺度融合，这样可减少参数和内存占用，且由于引入多尺度信息，可以更好地定位分割边界，提高了网络性能。

1 相关工作

1.1 数据集

本文使用 Pascal VOC 2012 扩增数据集做基础研究，Cityscapes 数据集做进一步验证。

作者简介：魏鹏磊(1997-)，男，硕士研究生，主要研究方向：计算机视觉；雷菊阳(1966-)，男，博士，副教授，主要研究方向：深度学习。

收稿日期：2022-02-21

Pascal VOC 挑战赛是一个世界级的计算机视觉挑战赛。Pascal VOC 挑战赛整体上可分为如下几类：图像分类、目标检测、目标分割、行为识别等。在 Pascal VOC 数据集中主要包含 20 个目标类别和 1 个背景类别。

对于图像语义分割，Pascal VOC 2012 中共有训练集图像 1 464 张、验证集图像 1 449 张、测试集图像 1 456 张，但是对于语义分割，特别是基于 Transformer 骨干网络而言，拥有大量的数据是很有必要的，所以本文使用了 Pascal VOC 的扩增数据集，共有训练集图像10 582张。

另外，在语义分割中对应的标注图像（.png）用 PIL 的 *Image.open*( ) 函数读取时，默认是 P 模式（调色板模式），即一个单通道的图像。在背景处的像素值为 0，目标边缘处用的像素值为 255，目标区域根据目标类别的类别索引信息进行填充，如图 1 所示，人对应的目标索引是 15，所以目标区域的像素值用 15 填充。具体调色板信息见表 1。

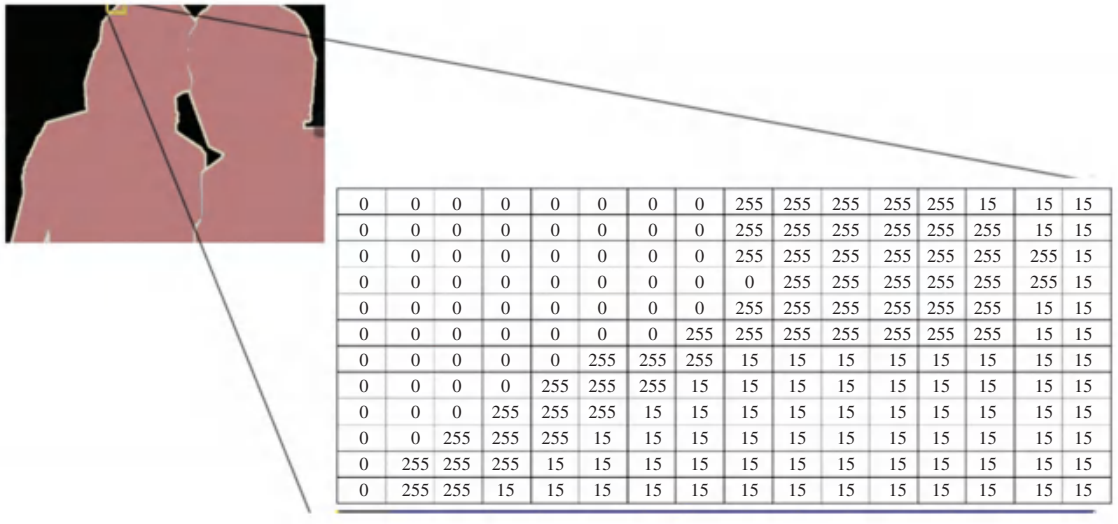


图 1　P 模式下的标签图

Fig. 1　Label map in P mode

表 1　不同类别的索引值

Tab. 1　Index values for different categories

类别名称	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
类别索引	0	1	2	3	4	5	6	7	8	9	10
类别名称	diningtable	dog	horse	motorbike	person	pottedplan	sheep	sofa	train	tvmonitor	
类别索引	11	12	13	14	15	16	17	18	19	20	

Cityscapes 数据集于 2016 年发布，在自动驾驶领域是权威且热门的语义分割数据集之一，该数据集含有国外 50 个道路场景的高分辨率图像，其中精细标注图像共有 5 000 张，粗略标记图像共 19 998 张，为保证能够最大限度地获取充足的数据信息。本文使用含粗略标注和精细标注数据集 24 998 张，共分为建筑、行人、天空等 19 个类别。

1.2　数据预处理

基于 Transformer 网络架构相对于深度卷积神经网络更容易出现过拟合现象，除需对网络中的模型结构进行优化外，拥有大量的数据也能够减少过拟合的发生，故考虑对图像进行预处理。本文的数据扩增操作是在 OpenCV 上完成的，包括对图像进行 -10° ~ 10° 的旋转、随机裁剪 *crop_size* 的 0.5 ~ 2 倍、随机水平翻转以及模糊图像等操作。

2　方法

本文算法是由 2 个路径组成的，分别是编码器提取路径与解码器提取路径。其中，编码器块是在 Swin Transformer 的基础上改进后得到的，不仅加快了训练速度，而且也缓解了过拟合。解码器块中的 Prediction Head 则是基于 ASPP+模块，考虑通过利用跳跃连接以及捷径分支优化模块结构，使其可以更好地解决目标多尺度的问题。具体来说，是以改进后的 Swin Transformer 模型 SwinLab 为骨干网络，再对 ASPP 模块进行优化，并构建模块 ASPP+，使

ASPP+可以多尺度理解上下文信息的能力。整体模型通过 3 个阶段构建不同大小的特征图,且又在 SwinT 的基础上剔除掉 Patch Partition 和 Linear Embeding 模块,并添加 1 个和后 2 个阶段同样的 Patch Merging 层进行下采样。网络总体模型结构如图 2 所示。

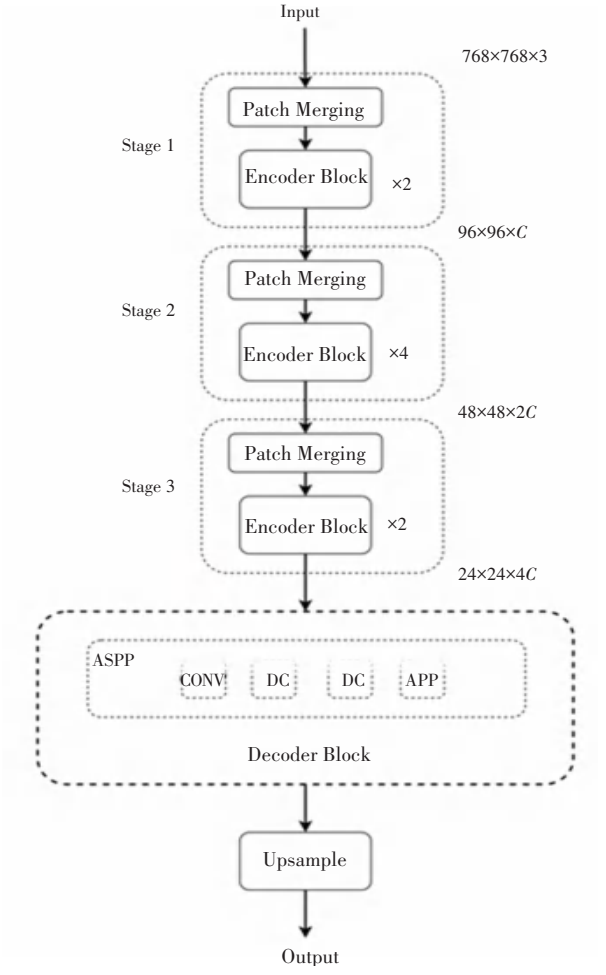


图 2 网络总体模型结构

Fig. 2 Overall model structure of the network

编码块是有 2 个结构,一个使用了 W-MSA 结构,另一个使用 SW-MSA 结构。一般情况下,这 2 个结构是成对使用的,先使用 W-MSA 结构,而后使用 SW-MSA 结构,具体编码器模型如图 3、图 4 所示。

解码块包括 ASPP+模块和 Prediction Head 模块。ASPP+在 ASPP 的基础上摒弃了膨胀系数为 36 的空洞卷积层,并采用自适应平均池化层,即共有 4 个并行分支,分别为 1 个 1×1 卷积层、3 个 3×3 空洞卷积层,以及 1 个自适应全局平均池化层,该层目的是可以增加 1 个全局上下文信息。其中,使用 *concat* 方法对 4 个并行分支进行拼接之前,先利用自注意力机制对不同分支获得的信息进行注意力处理,这样有利于不同特征信息的融合,而虚线部分的捷径分支则使用 1×1 卷积核进行维度处理。对于 Prediction Head 模块来说,得到 ASPP+模块的输出后,添加一个跳跃连接残差模块[6],其后续接一个 Layer Norm 层,再通过一个 1×1 卷积层来融合信息。Prediction Head 通过双线性插值的方法还原输入图像的尺寸大小[7],网络模型细节如图 5 所示。

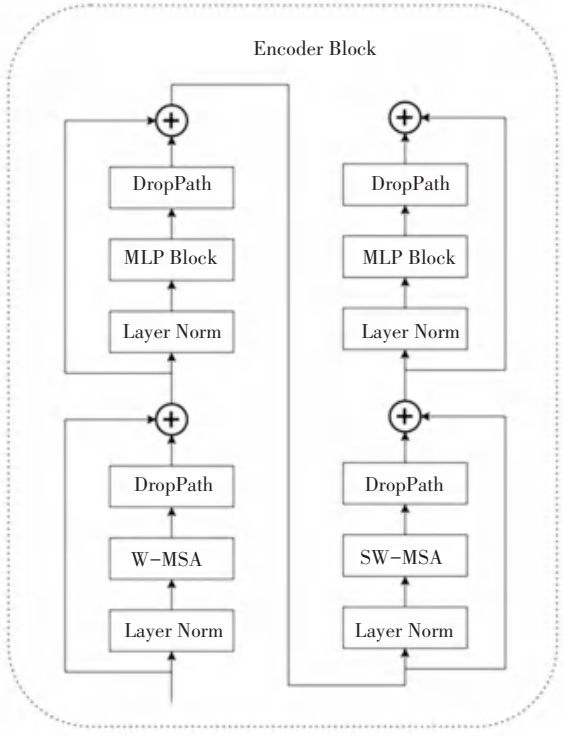


图 3 编码器结构图

Fig. 3 Encoder structure diagram

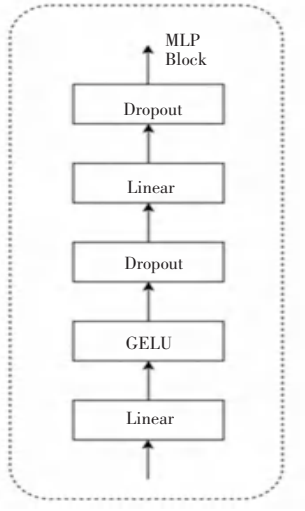


图 4 MLP 结构图

Fig. 4 MLP structure diagram

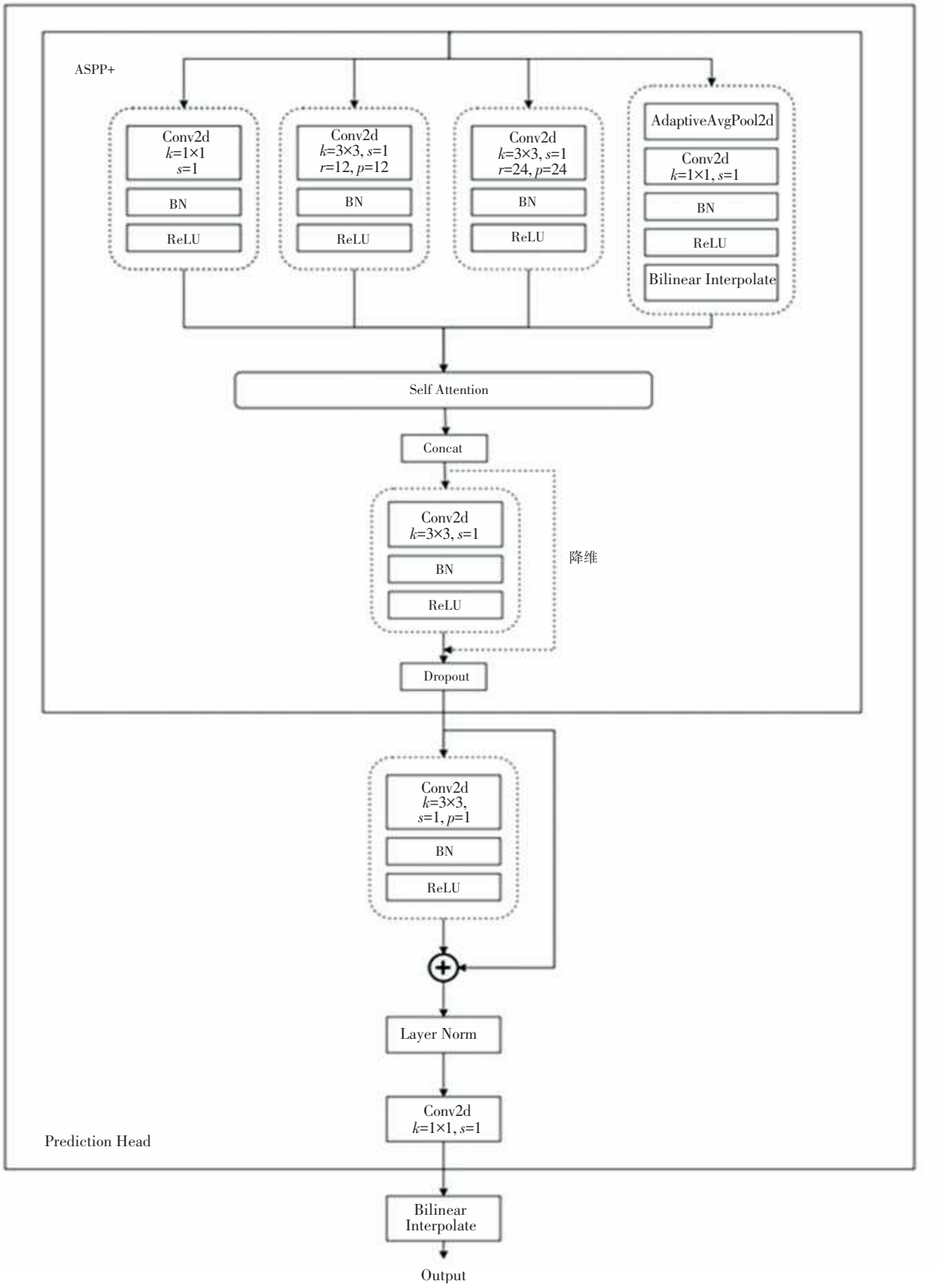
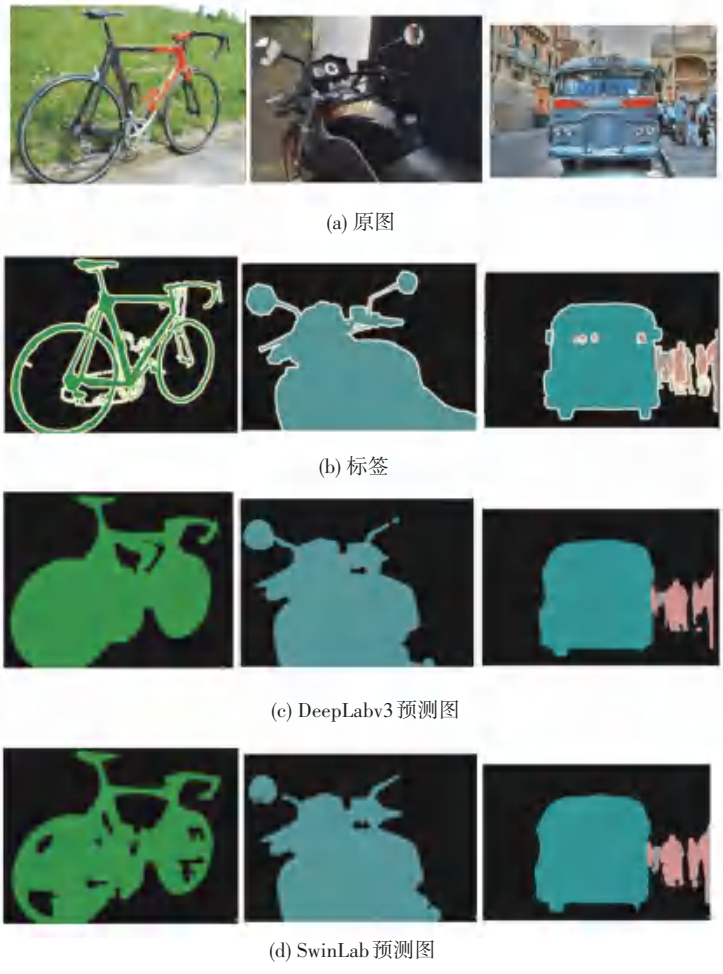


图 5　解码器结构图

Fig. 5　Decoder structure diagram

3 实验结果分析

在 Pascal VOC2012 数据集和 Cityscapes 数据集上的分割效果如图 6、图 7 所示。图 6、图 7 中，从(a)到(d)分别是原图、标签、DeepLabv3 预测图以及 SwinLab 预测图。



(a) 原图

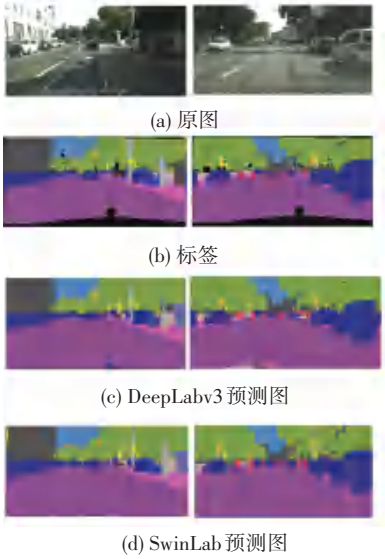
(b) 标签

(c) DeepLabv3 预测图

(d) SwinLab 预测图

图 6 Pascal VOC2012 数据集

Fig. 6 Pascal VOC2012 dataset



(a) 原图

(b) 标签

(c) DeepLabv3 预测图

(d) SwinLab 预测图

图 7 Cityscapes 数据集

Fig. 7 Cityscapes dataset

4 结束语

针对道路场景识别任务，本文提出了一种基于 Transformer 的 SwinLab 模型架构。该网络架构增强了网络在多尺度下多类别分割时的鲁棒性，同时使用不同的采样比例与感受野提取特征，使其可以在多个尺度上捕获上下文信息。实验结果表明，基于 Transformer 构建的 SwinLab 模型网络相比于传统基于深度卷积神经网络的语义分割模型，效果，及性能均获提升，虽不及 SOTA，但在 Pascal VOC2012 数据集上 *mIoU* 可达 80.1，在 Cityscapes 数据集上也有不错的效果。除此之外，本文重点关注的训练速度也得到了显著改善，对于后续的研究有着实际参考意义。另外，本文使用的显卡为单张英伟达最新 3090

(下转第 215 页)