陈雨佳, 宁媛, 唐坤俊,等. 基于模型剪枝与知识蒸馏的双路分支轻量级语义分割算法[J]. 智能计算机与应用,2026,16(2): 134-140. DOI:10.20169/j.issn.2095-2163.24040903

基于模型剪枝与知识蒸馏的双路分支轻量级语义分割算法

陈雨佳[1], 宁 媛[1], 唐坤俊[1], 刘聂天和[2]
(1 贵州大学 电气工程学院, 贵阳 550025; 2 贵州电网有限责任公司贵阳花溪供电局, 贵阳 550025)

摘 要:针对现有语义分割算法在自动驾驶场景下精度与实时性难以兼顾的问题,提出一种结合剪枝技术与知识蒸馏技术的轻量级语义分割算法。模型结构为双路分支结构,语义分支使用 ResNet50 作为骨干网络提取图像语义特征,加入空洞空间金字塔模块进一步提取特征,细节分支使用 3 个卷积层堆叠提取图像细节信息,将两路分支信息输入多特征融合模块 MFFM,融合上采样后获取分割结果,使用剪枝技术对模型冗余参数进行裁剪,结合解耦知识蒸馏技术对模型精度进行恢复提升。实验结果表明,在 Cityscape 数据集上,模型以 10.26 M 参数量,76.3%的 mIoU,达到了 64.16 ms 的推理时间,相较于 BiSeNetV2 模型,在精度上高 3.09%,参数量多 6.9 M,推理时间慢 18.74 ms。

关键词:轻量级语义分割; 模型剪枝; 解耦知识蒸馏
中图分类号:TP391.41 **文献标志码**:A **文章编号**:2095-2163(2026)02-0134-07

Lightweight dual-branch semantic segmentation algorithm based on model pruning and knowledge distillation

CHEN Yujia[1], NING Yuan[1], TANG Kunjun[1], LIUNIE Tianhe[2]
(1 School of Electrical Engineering, Guizhou University, Guiyang 550025, China;
2 Guiyang Huaxi Power Supply Bureau of Guizhou Power Grid Co., Ltd., Guiyang 550025, China)

Abstract: In response to the challenge of balancing accuracy and real-time performance of existing semantic segmentation algorithms in the context of autonomous driving, a lightweight semantic segmentation algorithm combining pruning and knowledge distillation techniques is proposed. The model structure consists of a dual-branch architecture. The semantic branch employs ResNet50 as the backbone network to extract semantic features from the images, incorporating Atrous Spatial Pyramid Pooling (ASPP) modules to further extract features. The detail branch uses three stacked convolutional layers to extract detailed information from the images. The information from both branches is input into a Multi-Feature Fusion Module (MFFM) for fusion and upsampling to obtain the segmentation results. Pruning techniques are applied to trim redundant parameters of the model, and knowledge distillation techniques are combined to restore and enhance the model accuracy. Experimental results demonstrate that on the Cityscape dataset, the model achieves a parameter size of 10.26 million, 76.3% mIoU, and an inference time of 64.16 milliseconds. Compared to the BiSeNetV2 model, it achieves a 3.09% higher accuracy with 6.9 million more parameters and 18.74 milliseconds slower inference time.

Key words: lightweight semantic segmentation; model pruning; decoupled knowledge distillation

0 引 言

计算机三大视觉任务包含了图像分类、目标检测和语义分割。语义分割旨在通过对图像中每一个像素点的分类从而达到对于整幅图像的语义理解,这一技术在自动驾驶、遥感图像分割、医疗图像分割等领域都有着广泛的应用。

随着深度学习技术的发展,基于各种思想相继提出了一系列的语义分割算法。通过整合上下文信息,研发出编码器-解码器架构的 U-Net[1],SegNet[2]; 以及扩大感受野和整合多尺度信息的 DeepLab[3-5] 系列模型,这类模型普遍适用于对分割精度要求较高而对

基金项目:贵州省科技计划基金(黔科合 ZK2022135)。
作者简介:陈雨佳(1994—),男,硕士研究生,主要研究方向:机器学习,图像处理; 唐坤俊(1998—),男,硕士研究生,主要研究方向:目标检测,模型部署; 刘聂天和(1998—),男,硕士,主要研究方向:电网技术,边缘计算。
通信作者:宁 媛(1968—),女,教授,主要研究方向:计算机视觉,图像处理。Email:ee.yning@gzu.edu.cn。
收稿日期:2024-04-09

实时性要求不高的场景，例如医疗图像分割、遥感图像分割等等；而在自动驾驶场景下，不仅要求模型精度要高，推理速度也要快，实际应用中，受限于边缘设备功耗、计算力、内存容量，对模型提出了更高的需求，因此设计高效的轻量级模型已然成为近年来研究热点。

目前，对于轻量级模型的设计大致可以分为 2 类方法。第一类方法是直接进行轻量级模型的设计，具体设计思路上，有基于多路分支结构的 BiSeNet，Fast-SCNN，ICNet 等等[6-8]；有基于轻量级骨干网络的 Laspp，ENet，DFANet 等等[9-14]；还有基于高效特征提取模块堆叠的 CGNet，ESPNetV1，ESPNetV2 等等，这类模型由于本身在设计时着重考虑了高效性，因此速度上普遍较快，但是精度上稍低。肖哲璇等学者[15]通过设计双路分支结构的模型实现了精度与速度的一定平衡，唐雪瑾等学者[16]，胡继涛等学者[17]基于轻量级骨干网络 MobileNetV2 设计分割模型也取得了不错的分割效果。第二类方法则是通过将一个精度高、参数量大的模型通过模型剪枝技术对冗余参数进行裁剪，再结合知识蒸馏策略对模型精度进行恢复，从而获得较为轻量的模型。这类模型的特点在于大型模型设计时并未考虑高效性，因此即使经过模型剪枝后，运行速度上稍微低于第一类设计方法的模型，不过最终会达到较高的模型精度。黄启灏等学者[18]，刘媛媛等学者[19]通过剪枝技术与知识蒸馏技术的结合分别获取到轻量级目标检测模型，图像分类模型在其应用中也获得了良好的效果。

综上所述，本文提出一种双路分支的语义分割模型。在编码阶段，通过对两路分支输入进不同分辨率的图像以此减少计算量，引入 ASPP[5] 模块来获取不同尺度的上下文信息；在解码阶段，通过设计特征融合模块（MFFM）有效融合细节分支与语义分支的信息，同时融入注意力机制对图像语义信息进行聚焦进一步强化模型的学习能力；最后，结合模型剪枝技术与知识蒸馏技术对模型冗余参数进行裁剪，对裁剪后的模型进行精度的恢复，最终获取到实时性与精度较为平衡的轻量级分割模型。

1　模型整体设计

模型整体结构如图 1 所示，主要分为特征提取网络、多尺度池化的 ASPP 模块、细节分支网络、多特征融合模块 MFFM。由图 1 可知，输入图像首先经过细节分支网络，原始图像尺寸缩小 2 倍再输入进特征提取网络，同时获取到细节信息与语义信息，再将二者输入进多特征融合模块进行信息的整合，最后输出上采样获得分割结果。

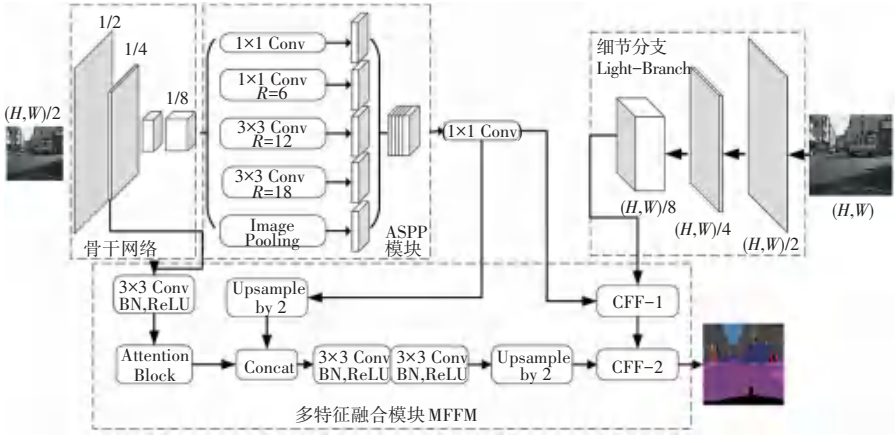

图 1　模型整体结构

Fig. 1　Overall model structure

1.1　特征提取网络

特征提取网络作为全局语义信息提取的关键，在自动驾驶这样对精度有着一定要求的场景中，特征提取网络首先应当尽可能强大，其次实时性的需求又要求设计中不能选择过于耗时的网络或者不利于轻量化的网络，综合各骨干网络在 ImageNet 数据集上的表现和易于做轻量化的程度，选取 ResNet[20] 作为骨干网络，ResNet 是 He 等学者[20]为解决深度卷积神经网络退化现象而提出的网络，其基本残差模块结构如图 2 所示。ResNet50 网络相比于 ResNet101 网络在少 16.99 M 的参数量的情况下，ImageNet 上 Top-1 准确率只低 1.054%，ResNet101 网络额外多出的 16.99 M 参数和计算量相比于其特征提取能力微弱的提升，在对实时性的要求比较高的场景中是不

利的,因此最终选择 ResNet50 作为骨干特征提取网络。其基本残差单元结构如图 2 所示。

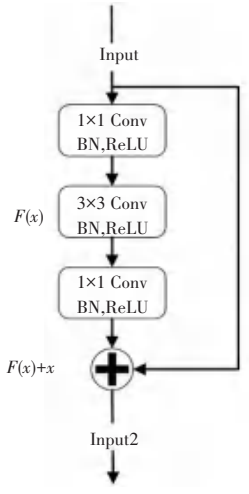

图 2 基本残差模块

Fig. 2 Basic residual module

1.2 引入空洞空间金字塔池化模块

多尺度信息的整合在 DeepLabV2[5]网络中被证明是有效的,故在本文模型中引入了空洞空间金字塔池化模块 ASPP(Atrous Spatial Pyramid Pooling)用以整合多尺度信息,这个模块包含了多个并行的膨胀卷积模块,每个都具有不同的膨胀率,用以获取不同尺度的特征信息,增强了网络对于不同尺度目标的预测能力。其结构如图 3 所示。

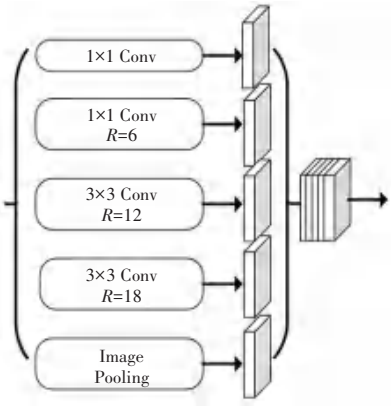

图 3 ASPP 模块结构

Fig. 3 Structure of ASPP module

输入特征图并行经过一个 1×1 卷积, 一个膨胀率 $R=6$ 的 1×1 卷积,一个 $R=12$ 的 3×3 卷积,一个 $R=18$ 的 3×3 卷积, 还有一个全局平均池化结构帮助提取全局信息,通过不同膨胀率的膨胀卷积后,得到的特征图都具有不同的尺度的信息,最后将这些得到的特征图进行拼接融合,将多尺度信息整合输出。由于该模块整合了多尺度信息,从而能够捕获图像中目标的不同尺度的语义信息。这使得模型在进行语义分割时能够更好地理解图像中不同大小和比例的物体,并有效提高分割的准确性和鲁棒性。

1.3 细节分支网络

多路分支结构的网络中,图像细节信息的提取依靠细节分支,输入进细节分支网络的图像一般都是原始分辨率的图像,因此细节分支的设计上首先不能过于复杂,否则会带来较大计算量,从而降低推理速度。

其次,该分支也需要有足够的感受野能尽可能对全局细节信息进行捕捉,本文所设计的细节分支由 3 个卷积层组成,卷积核步距为 2,膨胀系数为 2,扩大感受野的同时将原始输入图像下采样 8 倍,充分获取图像细节信息,其结构如图 4 所示。

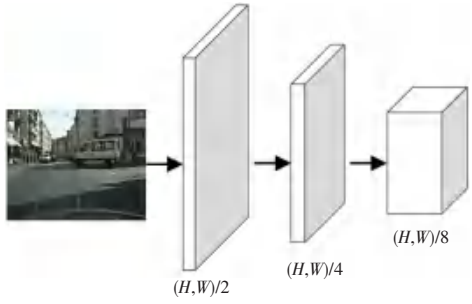

图 4 细节分支结构

Fig. 4 Structure of detail branch

1.4 多特征融合模块

一般特征图通道数越多,其包含的图像语义信息越丰富,语义层次也更加抽象。然而一般来说为了减少计算量,通道数加深的同时也会对特征图进行下采样,由此则会导致图像细节信息损失,例如损失了图像中目标边缘轮廓信息、位置信息等等。在 1.3 小节设计了一个细节分支对这部分损失的信息进行补充,还需要对语义分支与细节分支的信息进行有效的整合,本文设计了一种多特征融合模块 MFFM(Multi-Feature Fusion Module),参见图 1 下方虚线框,其主要作用在于融合不同尺度、不同语义层次的信息;将骨干网络的 4 倍下采样特征与 ASPP 模块输出进行融合,通过 CA[21]注意力机制进行信息的聚焦,并最终上采样至原图 1/2 大小,这部分输出同时具有语义信息和细节信息,其次将 ASPP 模块输出与细节分支输出输入进 CFF-1[7]模块进行融合,融合的结果输入 CFF-2 模块并与后一级输出进行融合,最后通过双线性插值上采样 2 倍,CFF 模块结构如图 5 所示。

1.5 通道剪枝

剪枝技术旨在减少神经网络参数量从而获得计算量的减少提升推理速度,一般来说,通过识别并移

除深度神经网络中的冗余参数(通常是权重较小的神经元、通道、组别)来减小模型大小。这些参数通常对模型的总体性能贡献较小,若能够正确实施剪枝,不会显著影响模型的性能,有时候甚至可以通过消除冗余和噪声改善性能。但过度的剪枝可能会损害网络的学习能力,导致精度下降。

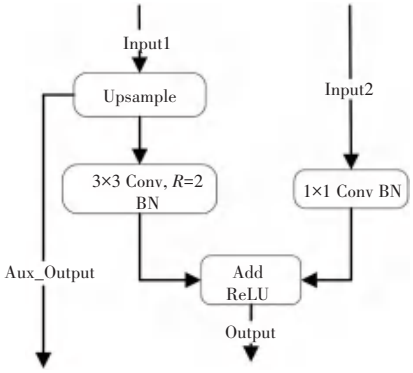

图 5　CFF 模块结构

Fig. 5　Structure of CFF module

基于稀疏因子的通道剪枝方法基本原理是利用批次归一化层(Batch Normalization,BN)中缩放因子 γ 来衡量通道的重要性。若对于一个通道特征图 X_i, 经过 BN 层归一化之后,得到新的通道特征图 X_{i+1}, 则有如下公式:

$$X_{i+1} = \gamma \times \mathrm{Norm}(X_i) + \beta \tag{1}$$

因此,若缩放因子 γ 的值趋于 0,则输出的通道特征图的值全部趋于 β, 这样该通道特征图的方差基本为 0,基本不包含任何信息,这样的通道对于神经网络来说是冗余的,因此可以进行裁剪。要实现基于稀疏因子的剪枝办法,就需要将大部分缩放因子 γ 训练为趋于 0 的值,这样才能够裁剪更多的通道。通过在优化目标中添加 L1 正则项,可以使得训练参数更好地趋近于 0 值,从而可以将缩放因子变得稀疏。具体优化目标如下:

$$L = \sum_{(X,Y)} l(f(X,W),Y) + \lambda \sum_i \sum_j \| \gamma_{ij} \|_1 \tag{2}$$

其中, X,Y 分别表示输入和输出; f 表示网络结构; l 表示损失函数; λ 表示平衡因子,用以调整正则项权重; γ_{ij} 表示第 i 层中第 j 个缩放因子。

1.6　解耦知识蒸馏

知识蒸馏主要用于模型的精度提升,旨在利用参数量大、性能强的教师模型提升参数量少、性能差的学生模型的精度。通过蒸馏温度 T 对教师模型的输出概率分布进行平滑来提取信息,此后又将损失重新表述为了 2 部分:学生模型输出与教师模型输出概率分布的 KL 散度损失,学生模型输出与标签的交叉熵损失,由此总损失可以表述为:

$$\mathrm{Loss} = \alpha \times \mathrm{KL}(y^T \| y^S) + (1-\alpha) \times \mathrm{CE}(y^S, y^L) \tag{3}$$

其中, y^T,y^S,y^L 分别表示教师模型的输出、学生模型的输出和标签值; KL(·)表示 KL 散度计算; CE(·)表示交叉熵损失计算; α 为权重系数,取值范围为 [0,1]。

文献[22]提出的解耦知识蒸馏则将损失分为了 3 部分:目标类知识蒸馏(Target Class Knowledge Distillation,TCKD)损失和非目标类知识蒸馏(Non-target Class Knowledge Distillation,NCKD)损失和交叉熵损失。由此推得网络总体损失可以表示为:

$$\mathrm{Loss} = \alpha \times \mathrm{TCKD} + \beta \times \mathrm{NCKD} + \eta \times \mathrm{CE}(y^S, y^L) \tag{4}$$

其中, α,η,β 均为大于等于 0 的实数,用于调整各个损失项的比重。

2　实验和分析

2.1　实验数据集及实验配置

本次实验使用公共数据集 Cityscape[23-24],该数据集是自动驾驶领域获得广泛使用的数据集。共包含 50 个城市、不同季节、不同时间段、不同道路环境等各类情形,共 5 000 张精细标注图片,其中 2 975 张训练集,500 张验证集,1 525 张测试集,共计 19 个类别,分辨率为 2 048×1 024。实验平台配置见表 1。

表 1　实验平台参数

Table 1　Experimental platform parameters

名称	配置
操作系统	Ubuntu20.04
CPU	i7-13700KF
内存	32 GB
显卡	RTX 4070ti
Pytorch 版本	1.10.0
Cuda 版本	11.3

2.2　评价指标

语义分割任务中,最常用的精度评价指标为平均交并比(mean Intersection over Union, mIoU)。mIoU 表示真实标注的像素区域与预测像素区域之间的重合度。计算公式为:

$$\mathrm{mIoU} = \frac{1}{n_{\mathrm{class}}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ij} - n_{ii}} \tag{5}$$

其中, n_{ii} 表示真实标签为 i, 预测结果也为 i 的像素个数; n_{ij} 表示真实标签为 i 但是预测结果为 j 的像素总个数; t_i 表示在真实标签中类别为 i 的像素总个数; n_{class} 表示类别的个数。

此外,对于模型的计算量则使用指标 FLOPs(FLoating point Operations Per second)来衡量。该指标表示每秒执行浮点运算次数,既衡量了模型的计算量,也在一定程度上衡量模型的推理时间。

2.3 与轻量级网络性能对比

本小节将在 Cityscape 数据集的基础上,将本文模型与主流轻量级网络进行对比,对比指标包含精度指标 mIoU、模型参数量、模型浮点运算量、模型推理时间。训练则在商汤科技的深度学习语义分割训练框架 mmsegmentation 下完成,共 40 000 个 iter,每 4 000 个 iter 验证一次,训练时 mIoU 和 loss 变化趋势如图 6、图 7 所示,性能对比结果见表 2。

表 2 不同轻量级分割算法性能对比

Table 2 Performance comparison of different lightweight segmentation algorithms

模型	参数量/M	mIoU/%	FLOPs/G	推理时间/ms
Lraspp	3.21	69.54	18.49	36.81
BiSeNetV2	3.36	73.21	98.86	45.42
ICNet-ResNet50	47.53	73.82	122.00	76.22
Fast-Scnn	1.40	70.96	7.48	**25.82**
CGNet	**0.50**	68.27	27.73	49.19
本文模型	40.96	**76.77**	497.00	144.11

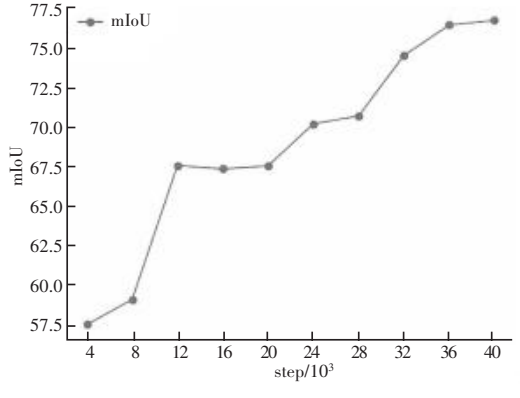

图 6 训练时 mIoU 变化

Fig. 6 Changes in mIoU during training

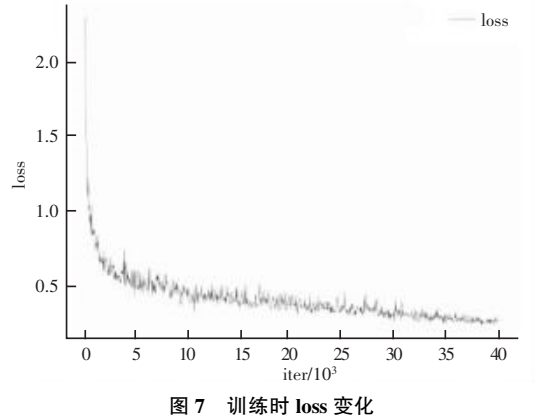

图 7 训练时 loss 变化

Fig. 7 Changes in loss during training

从表 2 中可以看到,与主流轻量级模型相比,本文所设计的模型在精度上具有优势,而在推理速度和参数量上相比于主流轻量级模型却并无优势。

2.4 结合通道剪枝与知识蒸馏

原始模型在参数量和推理时间上相比于主流轻量级模型暂时未显出优势,本小节将结合模型剪枝技术对模型进行裁剪。首先对模型进行稀疏训练,然后通过设置不同通道剪枝率,探索不同通道剪枝率对模型精度、参数量、计算量和推理时间的影响,结果见表 3。

表 3 不同剪枝率对模型性能影响

Table 3 Impact of different pruning rates on model performance

通道剪枝率	参数量/M	mIoU/%	FLOPs/G	推理时间/ms
0	40.96	75.8	497.00	144.11
0.2	26.21	75.5	316.12	107.69
0.3	20.07	**76.0**	249.16	96.23
0.4	14.74	75.5	185.93	83.75
0.5	10.26	75.3	133.49	64.16
0.6	6.54	74.6	88.08	57.02
0.7	**3.67**	67.4	**53.37**	**44.16**

在表 3 中,可以看到模型经过稀疏训练后,精度(75.8%)相比未经过稀疏训练的模型(76.77%)下降约 0.97%,经过通道剪枝、并对模型进行微调训练后,可以看到精度相比于经过稀疏训练的模型略有提升或者下降,在剪枝率为 0.2~0.6 时,此时计算量下降较多,且推理速度得到大幅度提升,精度降低幅度也在可接受范围内;而在剪枝率为 0.7 时,精度下降较多,已难以满足要求。同时注意到剪枝率在 0.4~0.6 之间的时候,精度下降较少,推理时间几乎成倍减少,取得了一定速度与精度的平衡。

接下来将结合解耦知识蒸馏技术对精度进行提升, DeepLabv3+模型在 Cityscape 数据集上精度为 79.61%,精度较高,可以作为教师模型,设置蒸馏温度 $T=4$,权重 $\alpha=1$, $\eta=1$, $\beta=8$,对剪枝率为 0.4~0.6 的剪枝后模型进行蒸馏,结果如表 4。

表 4 知识蒸馏前后对模型性能影响

Table 4 Impact on model performance before and after knowledge distillation

通道剪枝率	参数量/M	蒸馏前 mIoU/%	蒸馏后 mIoU/%	推理时间/ms
0.4	14.74	75.5	76.3	83.75
0.5	10.26	75.3	76.3	64.16
0.6	6.54	74.6	75.8	57.02

可以看到,经过知识蒸馏后,精度与速度最为均衡的是剪枝率为 0.5 的模型,精度为 76.3%,而推理

时间仅为 64.16 ms，已然达到了主流轻量级模型的推理时间水平；剪枝率为 0.6 时，推理速度上差异较小、只有约 7 ms，但是精度损失 0.5%，因此可知剪枝率为 0.5 时，精度与速度的平衡性更好，研究中则选取其作为本文最终的模型。

相较于表 2 中精度最高的轻量级模型 ICNet-ResNet50，精度上高 2.48%，参数量少 37.27 M，推理时间快 12.06 ms。相比于精度与速度上最平衡的轻量级模型 BiSeNetV2，精度上高 3.09%，推理时间上慢 18.74 ms，参数量多 6.9 M。

2.5　结果可视化展示

将分割结果进行可视化展示。图 8 中，第一行为标签图像，第二行开始依次为 BiSeNetV2，CGNet，Fast-scnn，ICNet，Lraspp，本文最终模型预测结果。

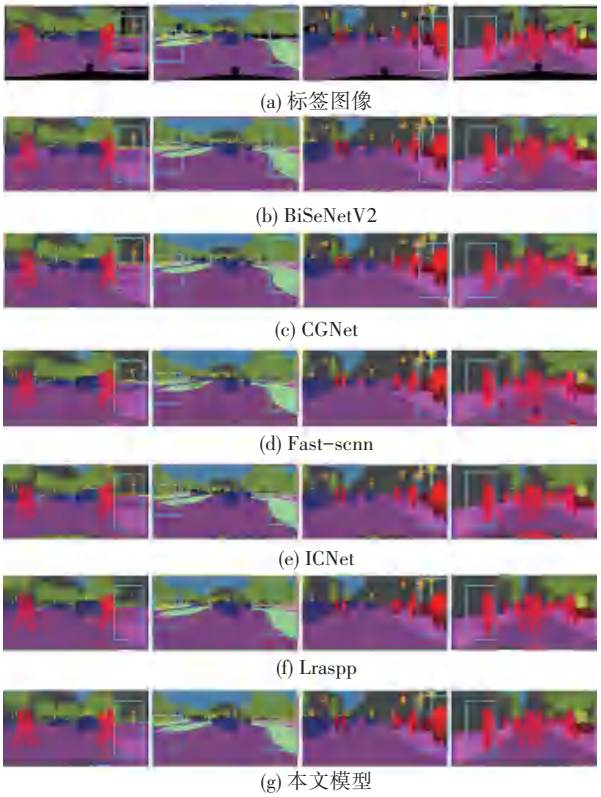

图 8　分割可视化结果

Fig. 8　Segmentation visualization results

图 8 中，第 1 幅场景明显看到右边粉色人行道区域分割结果更加精细且连续；第 2 幅场景右边人行道的分割结果也更加精准；第 3 幅场景中，对于人物轮廓的分割结果也比前几个模型精确。第 4 幅场景中左侧人行道人物与人行道的交界处本文的模型分割结果更加连续，少有空洞地方，且本文设计的模型 64.16 ms 的推理速度也达到了主流轻量级模型的速度基准，其参数量相比较于主流轻量级模型稍多、为 10.26 M。

3　结束语

针对自动驾驶这类实时性要求较高的场景，本文提出了一种双路分支结构的模型，通过设计细节分支与语义分支，获取图像不同语义层次的信息，再将二者信息输入 MFFM 模块进行信息的整合，得到精度较高、但是参数量较大、推理时间稍慢的模型，接着通过模型剪枝技术对模型参数量进行裁剪，对裁剪后模型通过知识蒸馏技术进行精度的恢复提升，最终获取到速度和精度较为平衡的模型，在 Cityscape 数据集上的实验结果表明，本文所提模型在 0.5 倍剪枝率时以 10.26 M 的参数量、76.3% 的 mIoU、64.16 ms 的推理速度，相较于 BiSeNetV2 模型精度高 3.09%，参数量多 6.9 M，推理时间慢 18.74 ms，也到达了速度与精度的平衡。

参考文献

[1] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015). Lecture Notes in Computer Science. Cham: Springer, 2015, 9351: 234-241.

[2] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[3] CHEN L C, ZHU Yukun, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 801-818.

[4] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv, 1706.05587, 2017.

[5] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[J]. arXiv preprint arXiv, 1412.7062, 2014.

[6] YU Changqian, WANG Jingbo, PENG Chao, et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 325-341.

[7] ZHAO Hengshuang, QI Xiaojuan, SHEN Xiaoyong, et al. ICNet for real-time semantic segmentation on high-resolution images[C]// Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 405-420.

[8] POUDEL R P K, LIWICKI S, CIPOLLA R. Fast-scnn: Fast semantic segmentation network[J]. arXiv preprint arXiv, 1902.04502, 2019.

[9] 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.

[10] LI Hanchao, XIONG Pengfei, FAN H, et al. DFANet: Deep

feature aggregation for real-time semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2019:9522-9531.
[11] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 1314-1324.
[12] WU Tianyi, TANG Sheng, ZHANG Rui, et al. CGNet: A light-weight context guided network for semantic segmentation[J]. IEEE Transactions on Image Processing, 2020, 30:1169-1179.
[13] WATANABE S, HORI T, KARITA S, et al. ESPNet: End-to-end speech processing toolkit[J]. arXiv preprint arXiv, 1804.00015, 2018.
[14] MEHTA S, RASTEGARI M, SHAPIRO L, et al. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019:9190-9200.
[15] 肖哲璇,陈辉,王硕. 基于双分支多尺度特征融合的道路场景语义分割[J]. 宁夏师范学院学报,2024,45(1):81-92.
[16] 唐雪瑾,杨卫华,于晋伟. 基于多尺度池化与特征融合的轻量级语义分割算法[J]. 微电子学与计算机, 2024,41 (12):1-9.
[17] 胡继涛,马晓锋,赵荣丽,等. 基于轻量级 DeepLabV3+网络的焊接熔池图像分割方法[J]. 计算机集成制造系统, 2025 ,31 (1):126-134.
[18] 黄启灏,靳国旺,熊新,等. 通道剪枝与知识蒸馏相结合的轻量化 SAR 目标检测[J]. 测绘学报, 2024 ,53 (4):712-723.
[19] 刘媛媛,王定坤,邬雷,等. 基于知识蒸馏和模型剪枝的轻量化模型植物病害识别[J]. 浙江农业学报,2023,35(9):2250-2264.
[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2016:770-778.
[21] LIU Zhuang, LI Jianguo, SHEN Zhiqiang, et al. Learning efficient convolutional networks through network slimming[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway,NJ:IEEE, 2017:2736-2744.
[22] ZHAO Borui, CUI Quan, SONG Renjie, et al. Decoupled knowledge distillation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2022:11953-11962.
[23] 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,NJ:IEEE, 2016:3213-3223.
[24] HOU Qibin, ZHOU Daquan, FENG Jiashi. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2021:13713-13722.