文章编号：2095-2163(2023)03-0025-09 中图分类号：TP391.41 文献标志码：A

一种面向城市场景的轻量级实时语义分割网络

顾嘉城，龙英文
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：在自动驾驶系统、无人机、机器人和视频监控等移动终端设备中，语义分割深度卷积神经网络的复杂程度也随着网络层数的增加而增加。虽然这能更好地提高网络的精度，但是在日常生活真实场景下，需要考虑网络的参数和运算速度，并同时确保较好的网络精度。基于上述需求，开发一种对移动设备内存和计算能力固定且相对较小的实时语义分割场景理解系统是有必要的。根据先前的实时语义分割模型的启发思想，设计了一种轻量化的基于注意力机制的实时语义分割模型 ALRNet（Attention-Light-Realistic Net）。通过在城市场景的 2 个数据集 Cityscapes 和 Camvid 上进行实验并与其他模型进行比较，结果表明，所设计的网络模型能够在提升分割速度的同时，也同样保证了分割的精度。做到了分割精度与速度的平衡。

关键词：实时语义分割；城市场景；轻量化网络；注意力机制

A lightweight real-time semantic segmentation network for urban scene

GU Jiacheng, LONG Yingwen
（School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China）

【Abstract】In mobile terminal devices such as autonomous driving systems, drones, robots and video surveillance, the complexity of semantic segmentation deep convolutional neural networks also increases with the increase of network layers. Although this can better improve the accuracy of the network, in the real scene of daily life, the parameters and operation speed of the network should be considered first, and at the same time to ensure better network accuracy. Based on the above requirements, it is necessary to develop a real-time semantic segmentation scenario understanding system with fixed and relatively small memory and computing power of mobile devices. Based on the previous real-time semantic segmentation model, a lightweight real-time semantic segmentation model based on attention mechanism ALRNet is designed. The results show that the effectiveness of the network model of this design is verified by experimenting on two datasets Cityscapes and Camvid and comparing them with other existing models. The designed network model can improve the speed of segmentation while also ensuring the accuracy of the segmentation. A balance between segmentation accuracy and the speed is achieved.

【Key words】real-time semantic segmentation; urban scene; lightweight network; attention mechanism

0 引 言

计算机视觉[1]是人工智能的一个重要部分，能为计算机提供解释和理解现实世界的能力。可以让机器识别和分类物体的图像，并对机器所看到的东西做出反应。计算机视觉三大基本任务有：图像分割、目标检测和图像分类。其中，图像分割包括了语义分割、实例分割、全景分割。研究可知，语义分割[2]作为图像分割的常用任务之一，在 2015 年提出的全卷积神经网络（Fully Convolution Network，FCN）即是其开山之作，代表着深度学习技术首次被应用于图像分割之中。语义指的是图像中每一个物体的含义，比如车、建筑、墙、道路，语义分割可理解成按像素对图像进行分类。但是不区分属于相同类别的不同实例（即同一物体的不同实例无需单独分割出来，且需要分割出背景信息）[3]。目前，语义分割技术常应用于自动驾驶[4]、人机交互[5]、手术中的医疗设备检测[6]等。

在自动驾驶和无人机技术等核心技术中，与语义分割技术密切相关的是对环境信息的处理，需要

基金项目：国家自然科学基金（61603241）。

作者简介：顾嘉城（1994-），男，硕士研究生，计算机学会 CCF 会员（K3336G），主要研究方向：图像处理、模式识别；龙英文（1974-），男，博士，副教授，主要研究方向：人工智能、图像处理。

通讯作者：龙英文 Email：1825405229@qq.com

收稿日期：2022-07-15

高质量高水平的语义分割技术作为保障,能够给车辆的安全驾驶提供周围环境信息的重要分析。使得车辆内部系统能够对周围环境做出正确的判断,以此保证车辆的安全行驶。由此可见,对城市场景语义分割任务的研究具有极其重要的现实意义[7],已经成为当下的一个热门方向。

以往的语义分割方法,如 PSPNet[8] 和 SegNet[9],都是为了获得更高的 *MIoU* 等其他评价指标,并使用 GPU 硬件的算力获取更高的精度,但运算速度比较低。另一方面,ENet[10] 和 BiSeNet 都设计了较小的编码器和解码器模型,更倾向于提高速度,但却造成了精度的下降。同时总结了其他研究者为了减少网络模型参数的研究成果。有些工作是使用裁剪来减少输入图像的大小,但很容易失去边界周围的空间细节和小对象;或是减少网络通道的数量;或是使用更少的卷积计算操作,而不是平方卷积来减少模型参数,如利用深度可分离卷积,可以提高模型的运算速度[11]。还有的研究者[12]使用多分支框架结合上下文信息,来提高网络模型的运算准确性,以及通过添加注意机制,使得处理后分割的边缘更加平滑。

在 2018 年,提出了 ICNet[13],ICNet 采用多尺度输入,在大分辨率采用较少的卷积核与层,在小分辨率使用较深网络,最后进行融合,并且在 3 个尺度提取出来的特征图进行预测分类来辅助整个损失函数,上采样部分采用空洞卷积和双线性采样。其优势在于:

(1)该网络是新颖、且独特的图像级联网络用于实时语义分割,利用了低分辨率语义信息和高分辨率图像的细节。

(2)提出的级联特征融合单元和级联标签引导能够以较低的计算成本逐步恢复和细化分割预测。

(3)ICNet 速度快,内存占用小。

在 2018 年还同时提出了 BiSeNet。BiSeNet 采用单尺度原图输入,2 个分支,路径使用 3 层卷积来避免破坏边缘信息且降低计算量,上下文模块使用深层网络获得更好的上下文信息,使得感受野更大,并在上下文模块增加了注意力机制和预测分类来辅助损失函数,最后进行融合。

2 种算法都没有采用常见的 U 型结构,而是使用了多路分支,既要提取分辨率大时的信息,又要提取分辨率小时的信息。且在分辨率大的特征图采用浅层网络,在分辨率小的特征图使用深层网络,BiSeNet 相较于 ICNet 的提升较大。受到先前研究的 ICNet 和 BiSeNet 的启发,本文设计了一种轻量化基于注意力机制的实时语义分割网络模型。该模型使用的是以轻量级非对称的编码器—解码器结构型网络进行实时的城市道路场景分割,以追求速度和准确性之间的平衡。使用 2 个方向来实现网络可以实时分割城市道路场景。并且使用非对称卷积思想来减少卷积操作参数操作。与非对称卷积导致的特征图精度降低相比,以残差块作为主干,弥补了通过跳跃连接来提高准确性,同时使用扩张卷积[14]来增加感受野。在残差块中,使用分组卷积进一步减少参数,并且网络通道可以通过编码器网络中的通道分段和打乱操作相互通信。在解码器 decoder 部分,结合特征金字塔网络(Feature Pyramid Networks,FPN)结构思想,利用自注意力机制和通道注意机制来提升网络的性能,同时利用显示通道内嵌空间信息模块进行上采样并恢复图像。在 Cityscapes 数据集和 Camvid 数据集上测试了该网络验证其有效性。

1 网络结构

在本节中,描述了非对称卷积、群卷积和所设计的 ALRNet 网络结构,包括了其内部编码器网络中的 ALR 模块和解码器部分的 ARPN 模块,ARPN 模块即结合了注意机制和 ECRE 块的特征金字塔网络结构,在解码器中实现网络复杂性和分割性能之间的平衡。

1.1 ALR 模块

在 ALRNet 网络中结合了 ResNet 模块的网络结构,提出了编码器网络中的 ALR 模块,其卷积过程如图 1 所示。

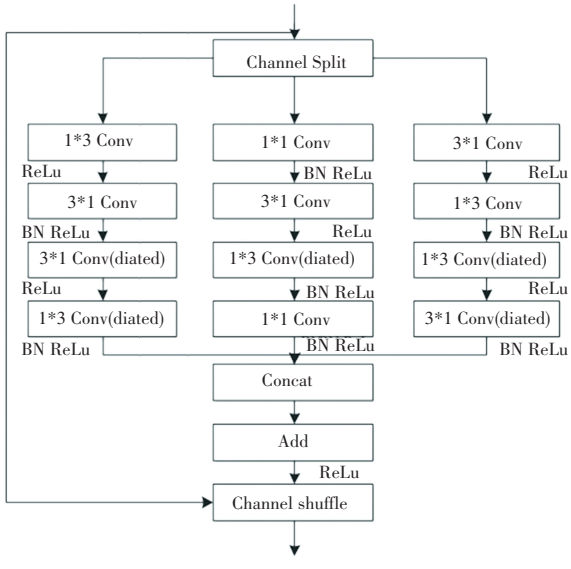


图 1 ALR 模块示意图

Fig. 1 ALR module diagram

ALR 模块的输入通道分为了 3 组，第一组的通道采用了 1 * 3、3 * 1 和 1 * 3、3 * 1 的卷积核进行卷积。第二组的通道采用了 1 * 1、3 * 1、1 * 3、1 * 1 的卷积核进行卷积。第三组则采用 3 * 1、1 * 3 和 1 * 3、3 * 1 的卷积核来卷积（diated 代表空洞卷积）。其中，空洞卷积的好处在于不增加参数量为前提，还能够增加感受野的大小。为了清晰对比 ALR 模块运算的过程，图 2 和图 3 分别列出了 ENet 的常规卷积流程和非对称卷积的流程。

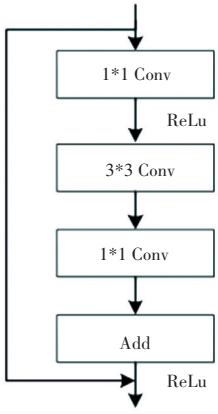


图 2　ENet 中卷积示意图

Fig. 2　Schematic diagram of convolution in ENet

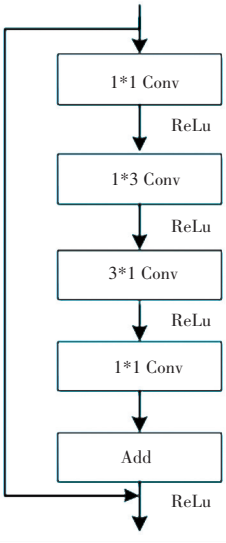


图 3　非对称卷积示意图

Fig. 3　Schematic diagram of asymmetric convolution

1.2　非对称卷积和群卷积运算

非对称卷积能够降低运算量，非常接近于平方核卷积运算。其特点总结为：

（1）先进行 $n * 1$ 卷积，再进行 $1 * n$ 卷积。这与直接进行 $n * n$ 卷积的结果是等价的。

（2）该卷积的目的是降低运算量，假设原为 $n * n$ 次的卷积，变更之后为 $2 * n$ 次卷积，如果 n 越大，那么减少的运算量会越多。

在实时语义分割任务中，减少参数是首要目标，这可以提高速度和效率，但还需要进一步的研究来确保网络的准确性。残差块用于道路场景分割、目标分类等各种应用场景。不仅可以跳转来连接卷积操作前的特征图和卷积操作后的特征图，还可以很好地提高网络的精度。不对称卷积经常被添加到网络模型中减少降采样过程中的计算量。然而，非对称卷积会导致精度的损失。本文在 ALR 块中加入了空洞卷积，以增加网络的感受野。

群卷积的应用最早始于 AlexNet。因为在那时硬件条件有限。当机器训练 AlexNet 模型网络时，无法在一个 GPU 中同时处理全部卷积操作。所以当时把特征图分配于多个 GPU 中分别进行处理运算，运算后再把多个 GPU 的结果进行融合。这样就可以减少训练参数，且不容易过拟合。

图 4 是一个常规的且没有分组的卷积层 CNN 结构。图 4 中展示了 CNN 的结构，一个卷积核对应一个输出通道。研究发现当网络的层数不断变大时，通道数会随之增加，空间维度也随之减少，因为卷积层的卷积核越来越多，以及卷积池化的操作，则使得特征图会越来越小。因此在深层网络中，通道会显得越来越重要。

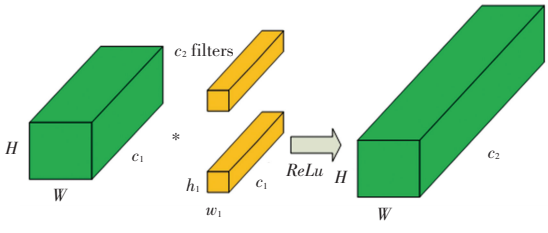


图 4　卷积层 CNN 结构

Fig. 4　Convolutional layer CNN structure

图 5 是一个群卷积的 CNN 结构。卷积核被分成了 2 个组。每个组都只有原来一半的大小。

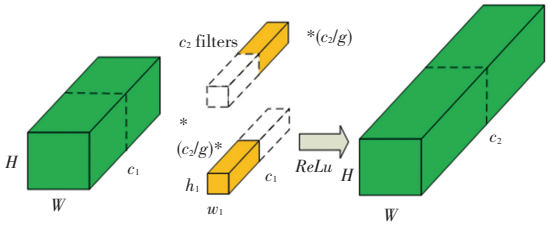


图 5　群卷积的 CNN 结构

Fig. 5　CNN structure of group convolution

为了进一步减少网络参数,在下采样过程中增加了群卷积操作,该设计的目的是为了减少卷积操作的运算量以及运算参数。采用了普通卷积后,比采用了非对称卷积在参数上多了约 1/5。

1.3 ARPN 模块

所提出的 ARPN 模块用于特征融合和上采样。该解码器采用了特征金字塔网络(Feature Pyramid Networks,FPN)结构,结合了通道注意机制和显示通道内嵌空间信息的方法,从而实现了网络复杂度和分割性能之间的平衡。

1.3.1 显示通道内嵌空间信息模块

特征金字塔网络被许多的网络模型所使用。特征金字塔网络目的是为了进行多尺度增强来提高网络的性能,但却会增加计算量。由于硬件计算能力的限制,于是实时语义分割网络从多尺度增强的思想中学习,并采用注意力机制,从而加强注意力的集中程度。受 ExfuseNet[15] 的启发,超分辨率上采样可以提高网络的精度和处理数据不平衡问题,因此在上采样过程中添加了显示通道内嵌空间信息方法(Explicit Channel Eesolution Embedding,ECRE)。该方法是采用子像素上采样,即重建空间与通道维度,把 4 个像素拼接在一起放大特征图。并且无需调参的方式来替代反卷积层。显示通道内嵌空间信息模块的运行结构如图 6 所示。

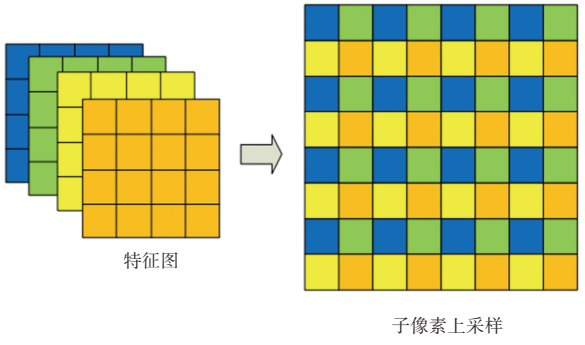


图 6 显示通道内嵌空间信息模块

Fig. 6 Displays the channel embedded spatial information module

1.3.2 注意力机制模块

注意力机制来源于人类大脑,甫一面世就被引入到自然语言处理技术中,随后才将其运用到计算机视觉的应用范畴内。

基于通道注意力机制的特征处理模块可以对特征通道之间的相互依赖关系进行精确建模,以提高网络产生的表示质量,使网络运用全局信息来有选择地强调信息特征[16]。

图 7 为通道注意模块,设输入特征映射 $X = [x_1, x_2, \cdots, x_C] \in R^{C\times H\times W}$,文中应用全局平均池化,输出的 $Z \in R^{C\times 1\times 1}$ 可以表述如下:

$$z_c = \frac{1}{H * W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i, j) \tag{1}$$

其中,z_c 表示与第 c 个通道相关联的输出,H、W 分别表示特征图的高度和宽度。该操作可以使网络能够收集全局信息。以下操作可以表示为:

$$U = X \otimes \sigma(g) \tag{2}$$

其中,"$\otimes$"表示信道乘法;σ 为 *Sigmoid* 函数;U 表示最终输出结果;$g \in R^{C\times 1\times 1}$ 为转换操作生成的最终注意向量结果,可用下式进行计算:

$$g = T_2(ReLU(T_1(z))) \tag{3}$$

这里,T_1 和 T_2 是 2 个不同的 $1 * 1$ 卷积层,可以捕获通道之间的相关性。通过第一次卷积,可以得到一个中间注意张量 $g_1 \in R^{C/r\times 1\times 1}$。$r$ 是控制块大小的减小比,r 对模型的性能有重要影响。这里,把 r 设置为 8,并讨论不同的降低比对性能的影响。接下来,通过第二次卷积,可以得到最终的注意张量 g。通道注意力机制可以聚合全局信息来捕获更重要的信息。

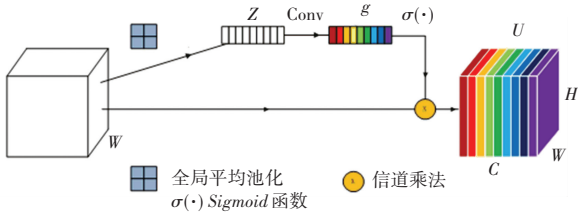


图 7 通道注意模块

Fig. 7 Squeeze-and-Excitation block

1.4 网络结构 ALRNet

图 8 是 ALRNet 网络模型架构,模型结构见表 1。使用了非对称的编码器-解码器的结构。其中,编码部分使用卷积运算和多重 ALR 模块以进行特征提取;解码部分使用 ARPN 模块进行上采样。编码部分先是进行了下采样,可以去除冗余信息,使特征图的信息更加紧凑。在 ALR 模块之后,执行特征提取。受 DeepLab 模型的启发,在编码部分增加了空洞卷积以增加网络的感受野,可以提高网络的准确率,同时避免使用大卷积带来的计算量增加的问题。在解码器中,把特征图输入进去,以此进行不同尺寸的卷积核的下采样的操作。

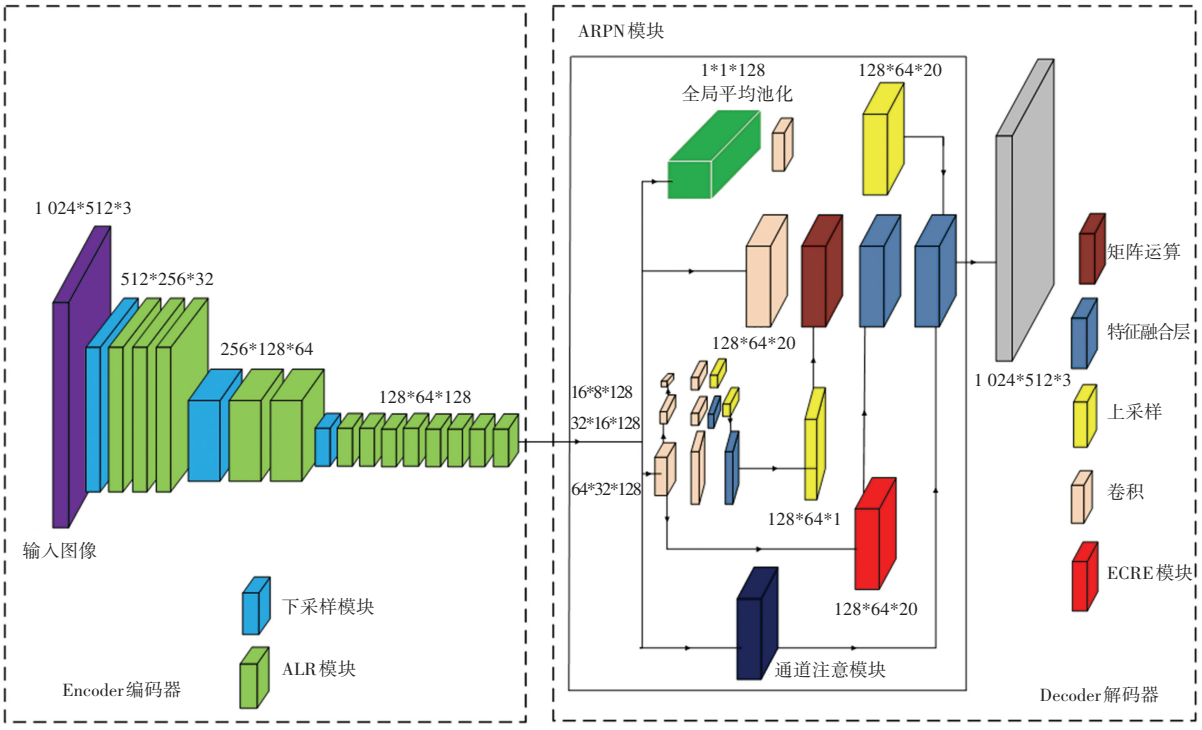


图 8　ALRNet 网络模型

Fig. 8　ALRNet network model

表 1　ALRNet 网络模型结构表

Tab. 1　ALRNet network model structure table

结构	类型	尺寸
编码器	下采样模块	512 * 256 * 32
Encoder	ALR 模块 * 3	512 * 256 * 32
	下采样模块	256 * 128 * 64
	ALR 模块 * 2	256 * 128 * 64
	ALR 模块(dilted n = 1)	128 * 64 * 128
	ALR 模块(dilted n = 2)	128 * 64 * 128
	ALR 模块(dilted n = 5)	128 * 64 * 128
	ALR 模块(dilted n = 9)	128 * 64 * 128
	ALR 模块(dilted n = 2)	128 * 64 * 128
	ALR 模块(dilted n = 5)	128 * 64 * 128
	ALR 模块(dilted n = 9)	128 * 64 * 128
	ALR 模块(dilted n = 17)	128 * 64 * 128
解码器	ARPN 模块	512 * 256 * N
Decoder	上采样单元 * 2	1 024 * 512 * N

解码部分对输入的特征图进行了全局平均池化，3 * 3、5 * 5、7 * 7 和步长为 2 的卷积操作，得到不同尺度的特征图进行上采样。使用 7 * 7 卷积核、步长为 2 的卷积得到的特征图使用显示通道内嵌空间信息的方法进行上采样，N 表示卷积 Concat 操作后的结果。由于参数的考虑，ECRE 只执行上采样。通过矩阵点加法合并特征图，再通过双线性插值恢复特征图，实现端到端训练。本文设计的网络模型没有后处理操作，也没有特征图级联方法增加计算压力，所以也可以有效地进行城市道路场景分割。

1.5　评价指标

对于图像分割中的语义分割，针对算法网络性能的评价指标有 3 个，详述为：设共有 n 个类别的物体和 1 个背景类，P_{ii} 是第 i 类被正确分为 i 类的像素数量，P_{ij} 表示属于 i 类但是被分为 j 类的像素数量，P_{ji} 表示属于 j 类但是被错分为 i 类的像素数量。

(1)像素精度(Pixel Accuracy，PA)：分类正确的像素数量与所有像素的比值，公式为：

$$PA = \frac{\sum_{i=0}^{n} P_{ii}}{\sum_{i=0}^{n}\sum_{j=0}^{n} P_{ij}} \tag{4}$$

(2)平均像素精度(Mean Pixel Accuracy，MPA)：所有类别的像素精度平均值，公式为：

$$MPA = \frac{1}{n+1}\frac{\sum_{i=0}^{n} P_{ii}}{\sum_{i=0}^{n}\sum_{j=0}^{n} P_{ij}} \tag{5}$$

(3)交并比(Intersection over Union，IoU)：模型检测出的目标区域与目标实际区域的重合部分占两者共同组成区域的比值，公式为：

$$IoU = \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} (P_{ij} - P_{ii})} \tag{6}$$

(4)平均交并比(Mean Intersection over Union, *MIoU*):所有类别的真实标签与预测结果的交集和并集的比值,公式为:

$$MIoU = \frac{1}{n+1} \sum_{i=0}^{n} \frac{P_{ii}}{\sum_{j=0}^{n} P_{ij} + \sum_{j=0}^{n} P_{ji} - P_{ii}} \tag{7}$$

(5)每秒传输帧数(Frames per Second, *fps*):在实时语义分割场景中往往需要速度和时间等衡量指标。*fps* 是衡量速度的指标,即图像的刷新频率。目标网络每秒可以处理或检测多少帧,为时间的倒数。这里假设目标检测网络处理 1 帧要 0.02 s,此时 *fps* 为1/0.02=50。公式为:

$$fps = \frac{1}{T} \tag{8}$$

2 实验

2.1 实验环境配置

为了验证提出的模型有效性,本文实验中硬件、软件系统配置环境见表 2。

表 2 实验配置环境表

Tab. 2 Experimental configuration environment table

内容	参数
操作系统	Ubuntu16.04
CPU	Core 6600X
GPU	GeForce GTX3090
GPU 显存	32 G
CUDA	Cuda 11.1 with cudnn
深度学习框架	Pytorch

2.2 数据集介绍

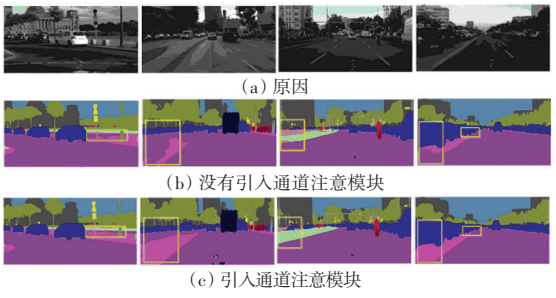
Cityscapes 数据集是一个从 50 座不同城市的街景中收集到的大型像素级注释数据集。该数据集中的训练集、验证集和测试集的数量分别为 2 975 张、500 张和 1 525 张。此外,为了评估基于弱监督学习的分类网络的性能,还提供了 20 000 张粗分割的图像。

Camvid 数据集包含 5 个不同的视频序列,由标注软件手动标注 700 帧,每幅图像的分辨率大小为 960 * 720。Camvid 共包括有 32 个语义类别 710 张图片。大部分视频都是用固定位置的相机拍摄的,一定程度上解决了对实验数据的需求。32 类建筑,如建筑物、墙壁、树木、人行道和交通灯等。

2.3 实验结果与对比

2.3.1 不同方案下的性能对比

为了验证在本文所提出的模型注意力机制有效性,在 Cityscapes 数据集上做了测试。注意力模块能保留空间细节的信息,对语义边界的信息有着较好的识别分割效果。例如图 9 中的 4 个例子,分别为原图、没有引入通道注意模块的方案、引入通道注意模块的方案。由图 9 可以看出,没有引入通道注意模块的方案没能完整地分割出道路,以及后排车辆的轮胎边缘分割也不够完整。而使用了通道注意模块的方案很好地解决了这问题。



(a) 原因

(b) 没有引入通道注意模块

(c) 引入通道注意模块

图 9 对比分割效果

Fig. 9 The effect comparison of different modules

2.3.2 Camvid 数据集下实验结果分析

由于轻量化网络模型对实际应用中硬件移动端的存储量有局限性,把 Camvid 数据集中输入图片的分辨率从 960 * 720 变化至 360 * 480,初始学习率设置为 1e-4。为了进一步验证本文 ALRNet 网络模型的有效性,以经典网络模型 SegNet 和 ENet 在 Camvid 数据集上的测试结果作为评判基准,表 3 即为各个模型在 Camvid 数据集上各个样本类别的分割像素精度结果对比。这里,以%为单位,且结果范围在±0.05变化范围之间。

表 3 各模型在 Camvid 数据集上像素精度对比

Tab. 3 Comparison of pixel accuracy of each model on the Camvid dataset

Label_name	SegNet	ENet	ALRNet
Sidewalk	96.13	94.62	95.45
Car	81.32	82.61	80.15
Building	86.59	75.86	89.51
Pedestrian	57.44	65.31	70.07
Fence	48.31	52.73	57.38
Road	83.23	85.92	88.36
Bicyclist	31.42	35.91	41.37
Tree	83.93	74.38	84.95
Sky	91.54	96.12	97.23
SignSymbol	20.63	57.64	58.19
Pole	29.56	34.23	26.87
MPA	**64.56**	**68.69**	**71.70**

从表 3 中可以看出,ALRNet 在 Camvid 上有 8 个类别(Building、Pedestrian、Sky、Fence、Road、Bicyclist、Tree、SignSymbol)的分割像素精度优于

SegNet 和 ENet，且平均像素精度也大于 2 个基准模型精度。表 4 为各个模型的分割准确度和处理一幅预测图像时间的对比结果。

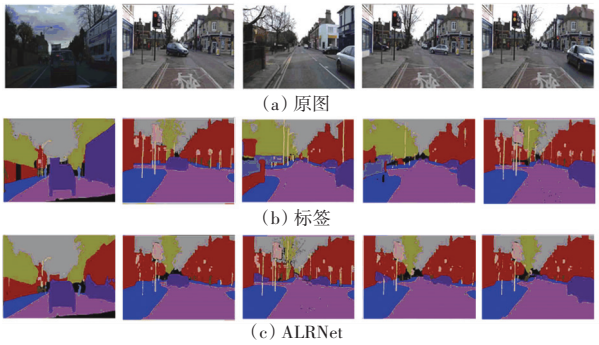
表 4　ALRNet 模型与其他模型在 Camvid 上对比

Tab. 4　Comparison of ALRNet with other methods on Camvid

方法	*MIoU* /%	*fps*
SegNet	57.12	19.5
ENet	54.22	59.6
ALRNet	**58.41**	**37.5**

通过对比实验结果可以看到，ALRNet 在分割精度上均优于 SegNet 和 ENet，在不失准确度情况下，本文提出的基于轻量注意力机制的语义分割网络有着较高的运算速度。因此，ALRNet 能够平衡速度和精度。能够满足自动驾驶、无人机飞行、智能机器人等实时应用需求。

由表 4 可以看出，相比于 SegNet 和 ENet，ALRNet 的 *MIoU* 值分别比 SegNet 高出了 1.29%、4.19%。在 *fps* 上比 SegNet 高出了 18.0，比 ENet 低了 22.1。图 10 则为 ALRNet 网络模型在 Camvid 数据集上的可视化结果。图 10(a)～图 10(c)中，从左到右依次为测试图 1～测试图 5。



(a) 原图

(b) 标签

(c) ALRNet

图 10　ALRNet 网络模型在 Camvid 数据集上的可视化结果

Fig. 10　Visualization results of ALRNet network model on Camvid dataset

2.3.3　Cityscapes 数据集下实验结果分析

为了充分验证 ALRNet 的有效性，将模型在 Cityscapes 数据集上进行实验。当硬件处理高分辨率的图片时，往往需要高配置的硬件，且训练时间较长。因此将 Cityscapes 数据集输入图像分辨率从 1024 * 2048 调整为 1024 * 512(分辨率降为一半)，初始学习率设置为 $1e^{-4}$。表 5 为 ALRNet 网络模型和 SegNet、ENet 模型在 Cityscapes 数据集上各个类别的像素精度为指标的测试结果。以%为单位，且结果范围在±0.05 变化范围之间。

从表 5 中可以看出，ALRNet 在 Cityscapes 内含的 9 个类别(Wall、Fence、Pole、Traffic light、Traffic sign、Rider、Truck、Train、Motercycle)的分割像素精度均优于 SegNet 和 ENet，且 ALRNet 网络模型的平均像素精度上也超过其他 2 个模型。以此判断本文提出的 ALRNet 的网络模型，可以在速度上和精度上做到了较好的平衡。

表 5　各模型在 Cityscapes 数据集上像素精度对比

Tab. 5　Comparison of pixel accuracy of each model on the Cityscapes dataset

Label_name	SegNet	ENet	ALRNet
Road	92.76	89.37	89.38
Sidewalk	83.66	88.91	89.68
Building	91.88	96.56	94.99
Wall	49.22	54.53	83.48
Fence	72.40	62.45	76.50
Pole	78.88	67.15	80.75
Traffic light	68.77	77.88	85.47
Traffic sign	72.89	79.32	89.99
Vegetation	92.91	96.62	96.03
Terrain	54.65	71.55	69.82
Sky	95.25	98.16	96.98
Person	80.42	90.99	87.38
Rider	61.16	74.58	77.28
Car	92.52	96.80	95.44
Truck	67.23	76.01	85.35
Bus	83.95	88.09	85.89
Train	78.92	52.54	94.31
Motercycle	59.86	70.42	79.96
Bicycle	71.89	84.46	82.85

为了增加实验的可靠性，在表 6 实验数据中，增加了 ICNet 以及 Deeplabv3+作为对照(其中硬件以及实验配置环境都相同)来进行实验。从结果上看，Deeplabv3+的 *MIoU* 最高，但是其模型较大，因此实时性运算速度较差。相较于 SegNet、ENet 和 ICNet，ALRNet 的 *MIoU* 值分别高出了 1.17%、4.09%、1.50%。在 *fps* 上比 SegNet 高出了 26.2，比 ENet 低了 11.7，比 ICNet 低了 10.4。

表 6　ALRNet 与其他模型在 Cityscapes 上对比

Tab. 6　Comparison of ALRNet with other models on Cityscapes

方法	*MIoU*/ %	*fps*
SegNet	58.24	13.5
ENet	55.32	51.4
ICNet	57.91	50.1
Deeplabv3+	67.14	–
ALRNet	**59.41**	**39.7**

在参数方面，Deeplabv3+的参数量最大，其次分别是 SegNet、ICNet、ALRNet、ENet。ENet 的参数量最小，因此分割效果较差，但是分割速度最高。ALRNet 的参数量与 ENet 相比多了 2 倍，但是平均交并比以及平均分割精度却有着较大提升。

由表 4~表 6 证明,本文所设计的 ALRNet 模型可以实现分割的速度以及准确度上的平衡,因为缩短了处理时间,以此能够为自动驾驶、无人机飞行等方面的使用提供了可能。同时也满足了网络模型对城市道路场景在分割精度上的要求。图 11 是 SegNet 与 ALRNet 模型的对比分割效果可视化结果。图 11(a)~图 11(c)中,从左至右依次为测试图 1~测试图 5。

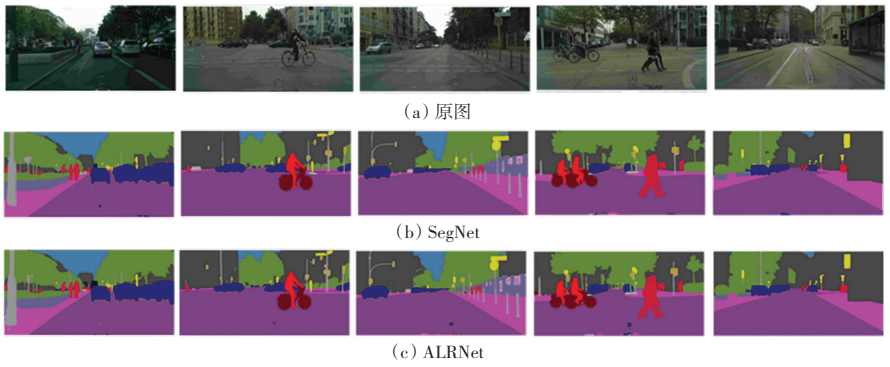


图 11 ALRNet 和 SegNet 网络模型在 Cityscapes 数据集上的可视化结果

Fig. 11 Visualization results of ALRNet and SegNet network model on Cityscapes dataset

3 结束语

本文基于通道注意力机制提出了用于实时道路场景分割的模型。该模型以端到端的方式进行训练。在编码器部分,采用非对称卷积、群卷积和扩展卷积的组合进行特征提取;解码器部分采用显示通道内嵌空间信息方法、并利用了通道注意机制的思想进行上采样。对城市场景数据集 Cityscapes 和 Camvid 进行实验,在权衡速度和分割精度两个方面,本文显示了较好的结果。体现在以下 2 个方面:

(1)速度和精度:由于 ALR 模块和 ARPN 模块的设计,网络参数大大降低。做到了网络模型的运算速度和分割精度的平衡,并且具有很好的可视化分割效果。

(2)简洁性:ALRNet 网络由编码器和解码器组成,其中 ALR 模块和 ARPN 模块可以很容易地移植到其他网络中,以此方便后续的研究。

参考文献

[1] WILEY V, LUCAS T. Computer vision and image processing review[J]. International Journal of Artifcial Intelligence Research, 2018,2(1):22-31.

[2] CAO Fude, BAO Qinghai. A survey on image semantic segmentation methods with convolutional neural network[C]// IEEE International Conference on Communications, Information System and Computer Engineering. Kuala Lumpur, Malaysia: CISCE,2020:458-462.

[3] HU Yaosi, CHEN Zhenzhong, LIN Weiyao. Rgb-D semantic segmentation[C]//International Conference on Multimedia & Expo Workshops. San Diego, CA, USA:IEEE,2018.

[4] GEIGAR A, LENZ P, URTASUN R. Are we ready for autonomous driving The KITTI vision benchmark suite[C]// IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA:IEEE,2012: 3354-3361.

[5] CORDTSM, OMRAN M, RAMOS S, et al. Hands deep in deep learning for hand pose estimation[C]//The Cityscapes Vision and Pattern Recognition. Seggau, Austria: 20th Computer Vision Workshop ,2016: 213-223.

[6] OBERWEGER M, WOHLHART P, LEPETIT V. Deep learning for hand pose estimation [C]//2016 IEEE Conference on Computer Version and Pattern Recognition. Las Vegas: IEEE, 2016:3140-3200.

[7] 沈启金,龙观潮,陈羽中. 基于图像分类的弱监督 RGBD 图像显著性检测方法[J]. 小型微型计算机系统,2022,43(01):61-68.

[8] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 2881-2890.

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

[10] PASZKE A, CHAURASIA A, Kim S, &Culurciello, E.ENet: A deep neural network architecture for real - time semantic segmentation[J]. arXiv preprint arXiv:1606.02147,2016.

[11] 王龙飞,严春满. 道路场景语义分割综述[J]. 激光与光电子学进展, 2021,58(12):44-66.

[12] 苗晨,李艳梅,陶卫国,等. 全卷积神经网络在道路场景语义分割中的应用研究[J]. 太原师范学院学报(自然科学版),2020, 19(02):58-63.

[13] ZHAO Hengshuang, QI Xiaojuan, SHEN Xiaoyong, et al. ICnet for real-time semantic segmentation on high-resolution images [M]//FERRARI V, HEBERT M, SMINCHISESCU C, et al. Computer Vision - ECCV 2018. ECCV 2018. Lecture Notes in Computer Science( ). Cham:Springer, 2018,11207: 405-420.

[14] MEHTA S, RASTEGARI M, SHAPIRO C A. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich ,Germany:dblp,2018:522-556.

(下转第 38 页)