分类号：TP391　　单位代码：10363

密　级：公 开　　学　号:2210320104

题 目基于轻量化神经网络的道路场景图像分割方法研究

题 目 Research On Road Scene Image Segmentation Method Based On Lightweight Neural Network

学　生　姓　名：李雯婷

校　内　导　师：杨会成（教授）

专　　业：控制科学与工程

研　究　方　向：智能信息处理及应用

论文答辩日期：2024 年 5 月 23 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：

日期：　　年　　月

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 □。

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

基于轻量化神经网络的道路场景图像分割方法研究

摘 要

道路场景语义分割可以对图像进行像素级别的分类，实现对道路场景的精细化理解和分析，是支撑自动驾驶、交通监控以及智能交通系统实现的关键技术。目前主流的道路场景语义分割方法使用大型主干网络构建复杂模型，网络性能得到显著提高，但是存在严重内存消耗和延迟问题，难以适应实际应用中的实时性需求。通过减少模型参数量并降低计算复杂度的轻量化研究为这一问题提供了解决方案，但是会导致语义分割精度的下降。针对上述问题，轻量级网络如何在提高推理速度的前提下达到较高的分割精度成为研究的重点。本文提出一种轻量级道路场景语义分割网络，降低模型的复杂度与计算量，研究基于知识蒸馏的模型优化方法，通过知识迁移对轻量级网络进行知识蒸馏优化，进一步改善分割精度。本文主要研究内容和结论如下：

（1）提出基于轻量化网络的道路场景实时语义分割模型（Lightweight Road Real-time Segmentation Network，L2RS-Net）。设计了轻量化的特征提取网络，降低模型的复杂度；构建了一种轻量化金字塔池化模块，通过引入非对称空洞卷积降低模型参数量使模型更轻量化；引入坐标注意力机制，以增强网络捕获信息的能力；优化交叉熵损失函数，平衡不同类别样本的数量差异，提高样本分割准确率。实验结果表明，L2RS-Net 在 Cityscapes 和 Camvid 道路数据集的分割精度分别为 72.2%和 53.8%，分割速度分别为每秒 46.66 和 24.58 帧，

性能相较于 DeepLabV3+有较为显著的提升。同时将模型部署到嵌入式设备 Jeston TX2 上，完成了验证和测试。

（2）提出基于知识蒸馏的道路场景实时语义分割优化方法。针对提出的轻量化模型分割精度损失的问题，提出了一种结构化知识蒸馏。该方法采用三种知识蒸馏项实现，引入像素级蒸馏，关注像素点之间的类别差异；引入成对蒸馏，增强高阶知识迁移；引入对抗损失，轻量化网络的输出作为对抗性网络的生成器，使用生成器损失作为整体蒸馏，有效传递全局知识。实验结果表明，使用结构化知识蒸馏后的分割方法在 Cityscapes 和 Camvid 数据集上的精度为 74.3%和 57.2%，相较于 L2RS-Net 提升了 2.1%和 3.4%，同时分割速度不变。

综上所述，本文以计算量和参数量的降低为优化目标，构建了一种基于轻量化网络的道路场景实时语义分割模型，有效提升了图像的分割速度；进一步提出了轻量化道路场景实时语义分割的的优化方法，使用结构化知识蒸馏训练轻量化模型，在不增加计算能耗的前提下提升了分割精度，为道路场景下实时语义分割提供了技术支撑。

关键词：实时语义分割，道路场景，知识蒸馏，轻量级网络，嵌入式

RESEARCH ON ROAD SCENE IMAGE SEGMENTATION METHOD BASED ON LIGHTWEIGHT NEURAL NETWORK

ABSTRACT

Road scene semantic segmentation can classify images at the pixel level to achieve a refined understanding and analysis of the road scene, which is a key technology to support the realization of automatic driving, traffic monitoring and intelligent transportation systems. Currently, the mainstream road scene semantic segmentation method uses a large backbone network to construct complex models, which significantly improves the network performance, but suffers from serious memory consumption and latency problems, and is difficult to adapt to the real-time demand in practical applications. Lightweight research that reduces the number of model parameters and decreases the computational complexity provides a solution to this problem, but it leads to a decrease in semantic segmentation accuracy. Aiming at the above problems, how lightweight networks can achieve high segmentation accuracy while improving inference speed has become the focus of research. In this paper, we propose a lightweight road scene semantic segmentation

network to reduce the complexity and computation of the model, study the model optimization method based on knowledge distillation, and optimize the lightweight network with knowledge distillation through knowledge migration to further improve the segmentation accuracy. The main research content and conclusions of this paper are as follows:

(1) Lightweight Road Real-time Segmentation Network (L2RS-Net) is proposed as a real-time semantic segmentation model for road scenes based on lightweight network. A lightweight feature extraction network is designed to reduce the complexity of the model; a lightweight pyramid pooling module is constructed, which reduces the number of model parameters by introducing asymmetric null convolution to make the model more lightweight; a coordinate attention mechanism is introduced to enhance the network's ability to capture information; and the cross-entropy loss function is optimized to balance the difference in the number of samples of different categories and improve the sample segmentation accuracy.The experimental results show that the segmentation accuracy of L2RS-Net in Cityscapes and Camvid road dataset is 72.2% and 53.8% respectively, and the segmentation speed is 46.66 and 24.58 frames per second respectively, which is a more significant performance improvement compared with DeepLabV3+. Meanwhile, the model is deployed to the embedded device Jeston TX2 to complete the verification and testing.

(2) Propose a real-time semantic segmentation optimization method for road scenes based on knowledge distillation. A structured knowledge distillation is proposed to address the loss of segmentation accuracy of the proposed lightweight model. The method is implemented using three knowledge distillations, introducing pixel-wise distillation, which focuses on the category differences between pixel points; pair-wise distillation, which enhances higher-order knowledge migration; and adversarial loss, where the output of the lightweight network is used as a generator for the adversarial network, and the generator loss is used as a holistic distillation to efficiently transfer global knowledge. The experimental results show that the accuracy of the segmentation method after using structured knowledge distillation is 74.3% and 57.2% on Cityscapes and Camvid datasets, which is improved by 2.1% and 3.4% compared to L2RS-Net, while the segmentation speed remains unchanged.

In summary, this paper constructs a real-time semantic segmentation model for road scenes based on lightweight network with the reduction of computational and parameter counts as the optimization goal, which effectively improves the segmentation speed of images; and further proposes an optimization method for real-time semantic segmentation of lightweight road scenes, which uses structured knowledge distillation to train the lightweight model, which improves the segmentation accuracy without increasing the computational energy consumption, and provides a

new approach to real-time semantic segmentation of road scenes, which provides a new approach to real-time semantic segmentation of road scenes, which provides a new approach to real-time semantic segmentation of road scenes. The optimization method for real-time semantic segmentation of road scenes is further proposed.

KEY WORDS:Real-time Semantion Segmentation， Road Scene，Knowledge Distillation，Lightweight Networks，Embedded

目 录

第 1 章　绪论

1.1 研究意义与背景

随着人类对智能汽车的需求不断增长，自动驾驶技术迅速发展，并且驾驶自动化水平不断提升。根据我国汽车标准化技术委员会于 2022 年发布的《汽车驾驶自动化分级》公示，自动驾驶被划分为 L0~L5 六个等级[1]。目前国内外车企逐步生产出配备 L2 和 L3 级别的自动驾驶功能车辆，车辆可以实现车道线保持，自适应巡航等功能，但仍需要驾驶员执行决策，L4 级别以上的系统仍处在研发阶段。

自动驾驶技术系统主要分为三大模块：环境感知，决策规划和执行控制[2]。在这些模块中，环境感知是最为关键的部分。环境感知任务借助各类车载传感器获取道路信息，使智能汽车具备对外界环境语义理解的能力。

语义分割是智能汽车道路场景语义理解的核心技术之一，相比于目标识别方法[3]和目标检测方法[4]，语义分割技术能够更精确的理解道路场景信息，为智能汽车的决策和规划提供准确和详细的道路场景边界信息。

图 1-1 道路场景语义分割

Figure 1-1 Semantic segmentation of road scene

图像语义分割是指利用图像在不同层次上的特征，将其划分为若干个相互独立且无重叠的区域。在同一区域内的特征具有最大的相似性，在不同区域之间的特征具有最大的差异性[5]。在自动驾驶领域，图像语义分割主要任务是对图像中的每个像素进行分类，分割出图像中不同的目标类别进而理解道路场景信息。基于传统特征的图像语义分割主要关注图像的可见特征（边缘、纹理等表面特征）来完成图像分割，这种方法主要关注像素之间的局部关系，忽略了全局信息，分割精度较低，无法准确分割出语义类别。基于深度学习的图像语义分割主要采用了卷积神经网络进行叠加。深度卷积神经网络通过强大的特征提取能力，在图像

中提取空间特征和语义信息，从而显著提升了语义分割结果的准确性。

近年来，各类电子产品收集到海量的数据，GPU 等通用计算机硬件的计算能力大幅提升，深度学习迅速发展[6]。然而，基于深度学习的语义分割算法侧重于提升分割的准确性，却较少关注实时性能和资源消耗问题，导致算法功耗较高，这限制了图像语义分割在自动驾驶等需要实时和资源敏感的场景中的有效使用。因此，轻量级语义分割网络应运而生，通过简化网络结构和提高网络计算速度，优化整体网络结构，实现图像的实时语义分割。相较于注重高性能的深层卷积网络，轻量级语义分割网络具有效率高、硬件要求低等优点，但却在准确性上表现不佳。因此，在道路场景下，语义分割仍面临许多挑战。在自动驾驶领域，语义分割技术需要至少能够有效识别车道线，汽车，行人，道路等类别，对准确性和实时性要求极高，如何平衡准确性与实时性，正是目前轻量级语义分割技术面临的主要挑战[7]。

综上所述，深度学习为基础的图像语义分割技术在实际应用中具有巨大的价值，而面向道路环境的轻量级图像语义分割技术对于推动自动驾驶技术的进步具有不可忽视的重要性。解决在计算能力有限的设备上计算成本、准确性以及实时性等关键性能指标的平衡问题，是语义分割技术发展面临的重大挑战。

1.2 国内外研究现状

图像语义分割可以分为基于传统特征和基于深度学习两种分割方法，基于传统特征的语义分割方法依赖手工特征提取，主观因素较强，分割性能较低，每次分割时间较长[8]。基于深度学习的语义分割技术采用了深度学习方法，这种方法不仅提升了分割的精度，还增强了分割的性能。

1.2.1 基于传统特征的语义分割方法

基于传统特征的语义分割方法依赖于灰度、颜色、空间纹理和其他特征，大致可分为以下几类：

（1）基于阈值的分割方法

基于阈值的方法认为灰度值是图像的重要特征，选取一个或者多个阈值，使

用该阈值与局部或全局像素值的大小对比来识别不同的区域，完成整张图像的分割[9]。该方法仅适用于简单二值图像，对于目标与背景灰度值相近，特征差异性小的图像，实用性不高。

（2）基于边缘检测的分割方法

基于边缘检测的方法通过检测图像中不同区域的边缘来实现分割[10]。常见的边缘轮廓类型包括斜坡型、脉冲型、阶跃型和屋顶型，如图 1-2 所示：

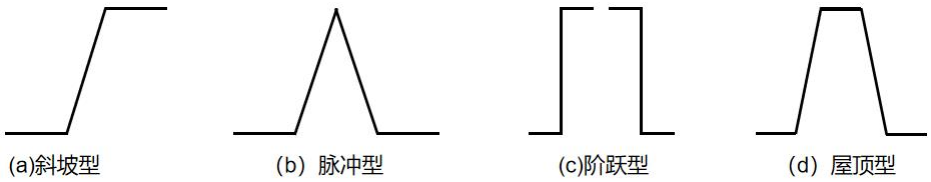


图 1-2 典型的边缘轮廓类型

Figure 1-2 Typical edge profile type

在实际图像处理中，噪声和边缘都会产生相似的间断高频变化，故该方法会混淆边缘和噪声。图像的边缘灰度值突变明显且噪声较弱时，分割效果理想。但图像边缘模糊且噪声强时，分割结果将出现不理想的情况，故该方法不适用于复杂场景。

（3）基于区域的分割方法

基于区域的分割方法首先随机选择一组代表每个区域的原始像素，然后将这些原始像素放置在相应的原始像素区域中，以判断区域中的像素是否满足特定条件，接着合并满足条件的像素，并将这些像素作为新的原始像素添加进去，最后重复这一过程，直到原始像素不满足条件时停止[11][12]。图 1-3 展示了区域生长方法的过程，该方法虽然实现比较容易，收敛较快。但是该方法对噪声也很敏感，只适合对小结构进行分割，例如伤疤，肿瘤等。

$$\begin{bmatrix} 0&1&1&3&4&0 \\ 5&0&4&4&3&1 \\ 0&0&5&4&7&7 \\ 1&1&4&6&4&4 \\ 1&0&4&4&4&4 \end{bmatrix} \qquad \begin{bmatrix} 4&4&4&4&4&4 \\ 4&4&4&4&4&4 \\ 4&4&4&4&4&4 \\ 4&4&4&4&4&4 \\ 4&4&4&4&4&4 \end{bmatrix}$$

图 1-3 区域生长过程

Figure 1-3 Regional growth process

（4）基于聚类的分割方法

基于聚类的分割方法旨在将图像中相似性的像素点聚集到一个特定区域内，反复迭代聚类直到类内像素点的相似性到最高，最终将图像中的所有像素点聚集到不同的类别中，类间的相似性到最低[13]。类聚分割方法可以用于彩色图像，其分割效果依赖于设定的初始值，流程简单。但是图像中的色彩相似但不相同时，分割效果则不理想。

1.2.2 基于深度学习的语义分割方法

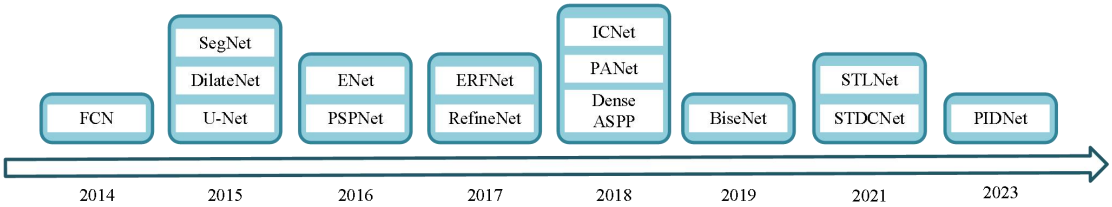


图 1-4 基于深度学习的代表性语义分割模型的发展过程

Figure 1-4 Development process of representative semantic segmentation model based on deep learning

基于深度学习的语义分割方法利用深度卷积网络自动学习和提取特征，从而显著提高了图像分割的适应性和效率。

Long 等人提出了全卷积神经网络（Fully Convolutinal Networks，FCN），该网络将全连接层全部替换为全卷积层，允许任意尺寸的图像输入，能够对图像进行有效训练和推理[14][15]；Ronneberger 等人提出了 U-Net，该网络使用了对称的编码器-解码器结构，编码器对输入图像进行特征提取和压缩，解码器与编码器中对应的特征图进行跳跃连接，有助于保留空间细节[16]；斯坦福大学研究团队提出了 DilateNet（Dilated Convolutional Network），其主要特点是引入空洞卷积，通过可调节的扩张率扩大卷积核的感受野，在不增加网络参数的前提下提升了网络分割性能[17][18]；香港中文大学研究团队提出了 PSPNet（Pyramid Scene Parsing Network），该网络引入金字塔池化模块和全局信息融合机制来提高分割得准确性[19]；悉尼科技大学研究团队提出了 RefineNet，该网络引入多层级特征融合和细化模块，能够更准确得提取语义信息[20]；中国科学院研究团队提出了 DenseASPP（Densely connected Atrous Spatial Pyramid Pooling），该网络是基于密集连接和 ASPP 模块的神经网络结构，能够提高图像语义分割性能[21]；香港大学联合腾讯实验室提出的 PANet（Path Aggregation Network），该网络引入金字塔注意力网络，采用多尺度特征融合机制，有助于捕获目标边界信息[22][23]；北

京航空航天大学研究团队提出了 STLNet（Learning Statistical Texture Network），该网络关注低层次特征提取，引入直方图均衡化，增强纹理特征的表征能力[24]。

近年来，语义分割研究在精确度上取得了较多进展，但实时语义分割研究较少。实时语义分割技术如下：设计轻量级网络结构，使用硬件加速，以及使用深度压缩和量化技术[25]。综合以上技术，可以实现在资源有限的环境下进行实时图像语义分割，但是精确度显著下降，如何在提升推理速度的前提下保持一定的模型精度是实时语义分割的一个关键问题。

英国剑桥大学研究团队提出了 SegNet（Semantic Segmentation Network），采用了编码器-解码器结构，使用最大池化索引上采样，网络结构简单参数量较小，实现了实时语义分割[26]；意大利研究团队提出了 ENet（Efficient Neural Network），该网络引入 1×1 卷积，参数共享，扩张卷积等结构设计与技巧，使用多分辨率处理思想，取得了性能和效率的平衡[27]；华中科技大学提出了 ICNet（Image Cascade Network），该网络采用快速上采样模块，分阶段训练策略，级联融合策略，适用于不同分辨率下的场景[28]；西班牙研究团队提出了 ERFNet（Efficient Residual Factorized ConvNet），该网络使用残差连接的思想，引入因果卷积实现实时语义分割的高效性能[29]；香港中文大学研究团队提出了 BiSeNet（Bilateral Segmentation Network），该网络使用了双边结构，能够综合全局与局部信息，提出了双边传播机制，为实时语义分割提供了一种有效方法[30]；美团研究团队提出了 STDCNet（Short-Term Dense Concatenate Network）,该网络通过降低特征图维数，聚合特征图进行图像表征，消除结构冗余以降低推理时间[31]；德克萨斯农工大学提出了 PIDNet（Proportional Integral Differential Networks），该网络提出三分支网络架构，分别专注于上下文，细节和边界信息的解析，并设计了边界注意力融合三分支特征，在轻量级网络上提高分割的准确性[32]。

以上方法虽然在实时语义分割任务中取得了速度方面的进展，但仍存在改进的空间；随着模型复杂度的降低，模型性能也随之下降，因此如何在降低模型复杂度的前提下保证分割精度成为了重要问题；实时语义分割任务需要大量的标定数据集，但是优良标注的数据集成本较高，这都是目前实时语义分割研究需要克服的问题[33]。

1.3 论文的主要研究内容

本文主要研究内容是基于轻量化网络的道路场景实时语义分割方法及优化，面向道路场景，实时高效的语义分割方法具备在移动端、嵌入式等计算资源有限的设备上部署的能力，具备在不同的复杂场景下对不同类别准确分割的能力，有效平衡计算速度和分割准确度。本文的主要工作概述如下：

首先介绍了语义分割的理论基础及典型工作，然后详细介绍了实时语义分割技术，本文所用数据集以及任务评价指标。

针对现有语义分割算法存在模型参数量大，很难在移动设备上使用的问题，构建一种基于轻量化网络的道路场景实时语义分割仿法。该方法对 DeepLabv3+进行改进，使用轻量级 MobileNetv2 作为主干网络来简化模型；采用轻量化金字塔池化模块（Lightweight Pyramid Pooling Module，LPPM）减少模型参数量，节约计算成本，减少推理运算过程；引入坐标注意力（Coordinate Attention，CA）加强特征学习；通过优化交叉熵损失函数，能够平衡简单困难样本学习难度，提高模型对困难样本的分割精度；解码器输出特征与编码器中相同尺度特征图进行跳跃连接，以便促进信息融合，增强特征表达。本文设计多个消融实验以验证各个模块的有效性，并在城市场景数据集 Cityscapes 和 CamVid 上与多个实时语义分割网络进行性能实验对比。

针对本文提出的轻量级网络模型实现实时分割的同时推理精度降低的问题，构建了结构化知识蒸馏优化，弥补数据集类别不平衡，在不影响推理速度的前提下提升网络训练精度。该知识蒸馏采用三种不同维度的蒸馏项，不同的蒸馏方案对训练学生网络提高精确度有着互补的贡献。并且融合了 WGAN-GP 对抗性网络，该网络减少训练过程中需要的标签数量，有利于整体蒸馏。本文在城市场景数据集 Cityscapes 和 CamVid 上设计了消融实验，验证了模型不同蒸馏项的有效性，并设计轻量级网络优化前后性能对比，验证了结构化知识蒸馏在不损失计算速度的前提下显著提升分割精度。

1.4 论文的结构安排

论文结构组织共包含五章，图 1-5 为技术路线图，细分如下：

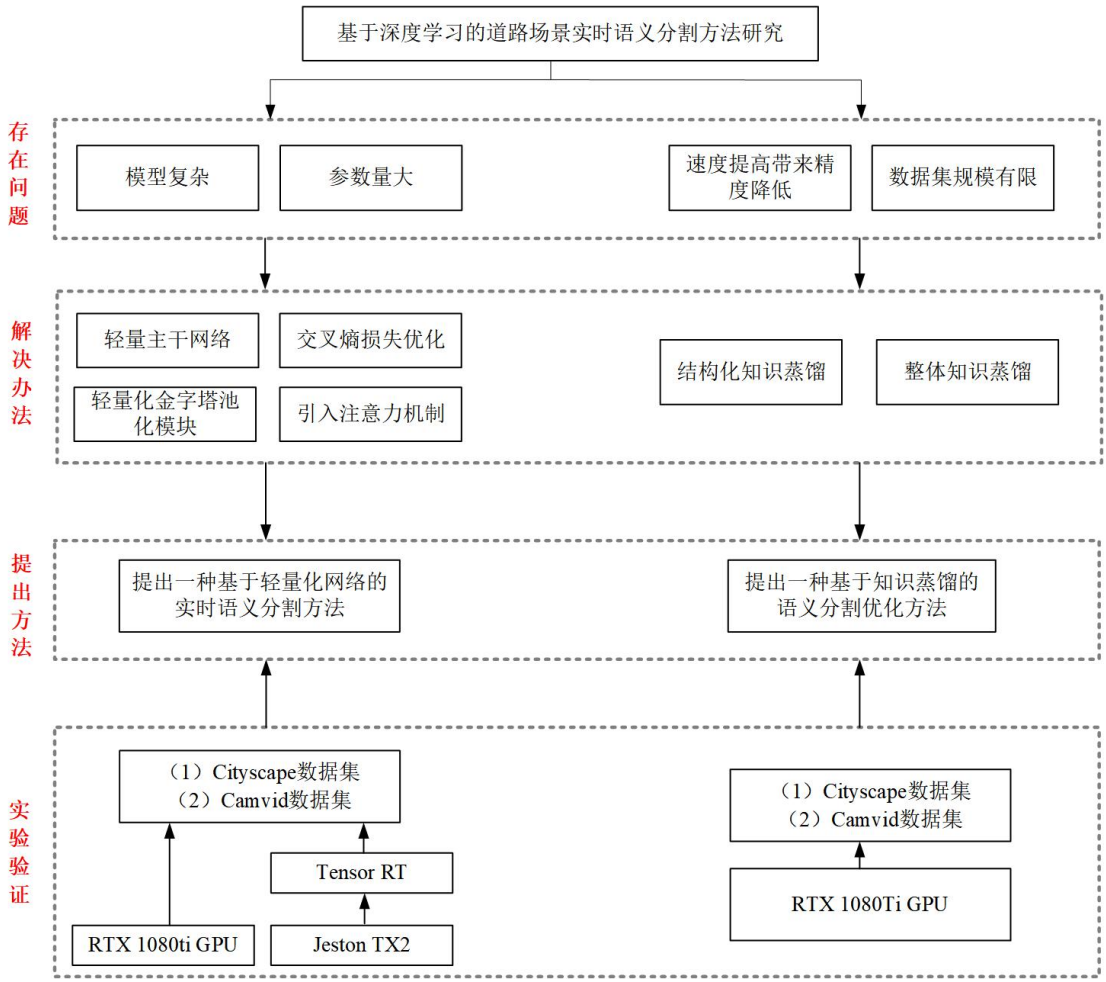


图 1-5 技术路线图

Figure 1-5 Technology road map

第一章：绪论。本章节介绍了图像语义分割的研究背景与意义，探讨了该技术所面临的困难与挑战。然后介绍了国内外语义分割的典型网络，最后对本文研究内容和技术路线图进行介绍。

第二章：图像语义分割相关理论基础。本章详细阐述了论文相关网络，对全卷积神经网络和实时语义分割的相关技术进行介绍，最后介绍了道路场景语义分割的相关数据集和图像语义分割的评价指标。

第三章：基于轻量化网络的道路场景实时语义分割方法研究。首先提出道路场景轻量化网络的设计原因和网络结构。然后对特征提取网络、特征增强模块、注意力模块、交叉熵损失优化等模块详细介绍，分析相关实验结果与道路语义分割结果的可视化分析。最后利用深度学习推理引擎 TensorRT 在嵌入式 Jeston TX2 上部署网络模型，实现图像语义分割。

第四章：基于知识蒸馏的道路场景实时语义分割优化方法研究。首先介绍了

知识蒸馏研究动机和整体蒸馏框架设计，然后详细地描述了蒸馏项的损失函数，实验参数的设置，通过对第三章轻量化网络知识蒸馏，进行优化前后网络性能实验，验证优化方法的有效性。

第五章：总结与展望。总结本文提出的语义分割网络的实验结果，对研究中的不足进行分析，并且提出了未来的研究方向。

第 2 章　图像语义分割相关理论基础

本章首先介绍了全卷积神经网络的出现及其工作原理，以及本文基础模型DeepLabv3+的网络结构及其特点，然后介绍了嵌入式设备 Jeston TX2，最后介绍了本文使用的数据集和评价模型选用的实验评估指标。

2.1 相关网络介绍

2.1.1 全卷积神经网络（FCN）

全卷积神经网络（Fully Convolutional Networks，FCN）的出现是深度学习领域的一个重要里程碑，该网络对卷积神经网络（Convolutional Neural Networks，CNN）进行了改进，带来了多项优势。传统卷积神经网络模型如图 2-1 所示，通常对输入图像尺寸有固定的要求，在卷积和下采样过程中，可能会损失图像中的细节信息。全连接层要求输入具有固定的尺寸，这使得网络在处理不同尺寸图像时需要进行额外的预处理[34]。

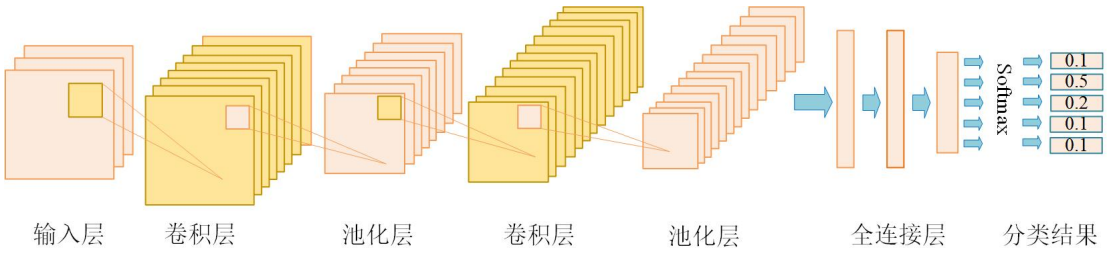


图 2-1 卷积神经网络

Figure 2-1 Conventional neural network

FCN 如图 2-2 所示，该网络通过创新性架构的设计成功克服了 CNN 的限制，其具备处理任意尺寸输入图像并输出相应尺寸密集预测结果的能力，这使得 FCN 更加灵活适用于各种场景。其次，全卷积层允许网络能够进行端到端学习，避免了传统手工设计特征提取器的复杂过程，从而简化了图像语义分割任务的流程。此外，FCN 通过跳跃连接，使低层特征与高层特征相融合，保留细节信息，提高分割精度[35]。

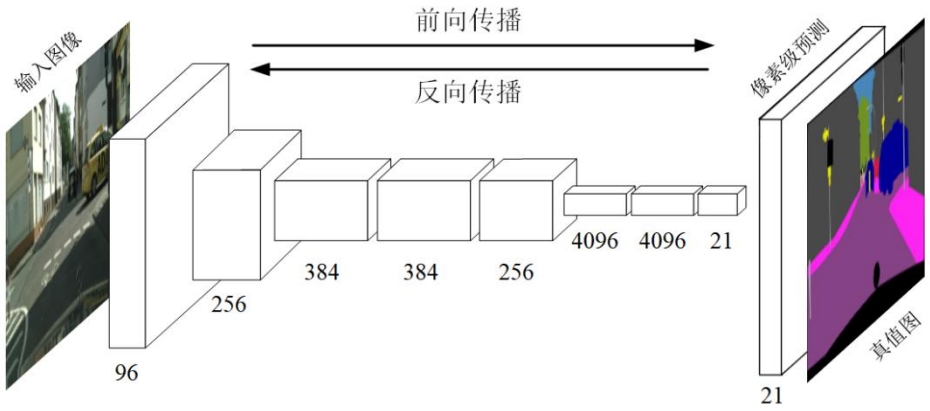


图 2-2 全卷积神经网络

Figure 2-2 Fully convolutional network

FCN 的主要结构特点包括编码器、解码器部分和跳跃连接。编码器通过卷积层和池化层提取图像特征。解码器通过反卷积或上采样操作将特征映射回原始尺寸，生成密集的像素级预测。跳跃连接是将相同层级特征图进行连接，提高分割精度。FCN 模型包括 FCN-32s，FCN-16s 和 FCN-8s，其中 FCN-8s 使用跳跃连接的方法将最后三次下采样得到的特征图进行融合，最后一次反卷积的步长为 8[36]。FCN-8s 是 FCN 系列模型里效果最好的，这表示浅层预测信息包含了更多的细节信息，FCN-8s 网络结构图如图 2-3 所示。

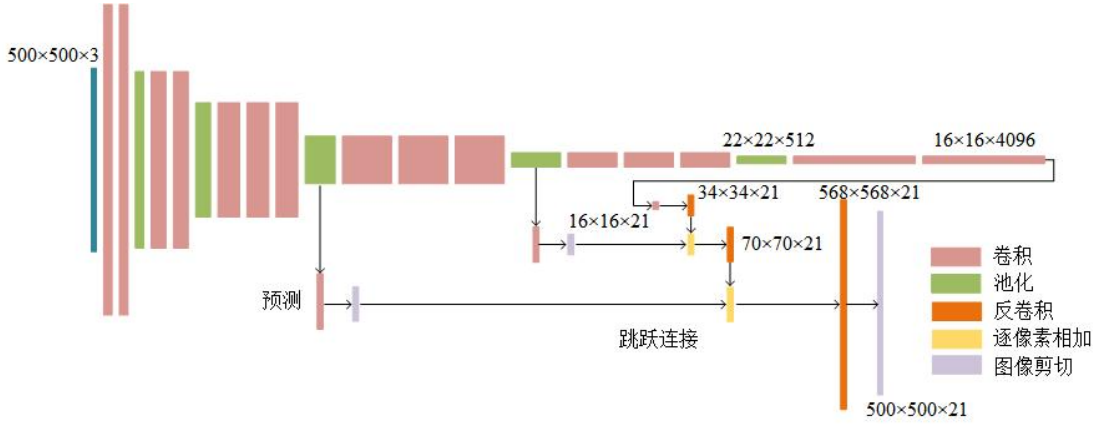


图 2-3 FCN-8s 网络结构

Figure 2-3 FCN-8s network structure

下面对全卷积神经网络的理论基础进行介绍：

（1）卷积层

卷积层主要由卷积计算构成，主要用于输入图像的初阶特征提取，以及对图像实现滤波或边缘提取。使用卷积核以一定的步长在特征图上滑动，进行乘加运算，多层卷积叠加可以实现低层到高层的特征提取。

$$F_{out}=\sum_{i=1}^{N}K_i\times F_{in}+b \tag{2-1}$$

其中，K_i表示第i个卷积核的大小，F_{in}为输入特征，F_{out}为输出特征，N为输出通道数，b表示偏置。

通常，卷积运算中$C_{in}\times H_{in}\times W_{in}$为输入特征图尺寸，$C_{out}\times H_{out}\times W_{out}$为输出特征图尺寸，$C_{in}$为输入通道数，$C_{out}$为输出通道数，$H_{in}$为输入特征高度，$H_{out}$为输出的特征高度，$W_{in}$为输入特征宽度，$W_{out}$为输出特征宽度，$H_k$为卷积核的高，$W_k$为卷积核的宽。

输入特征、输出特征与卷积核之间的关系用公式表示为：

$$H_{out}=(H_{in}+2\times p-H_k)/s+1 \tag{2-2}$$

$$W_{out}=(W_{in}+2\times p-W_k)/s+1 \tag{2-3}$$

p是边界补充，s为滑动步长。卷积过程如图2-4表示。

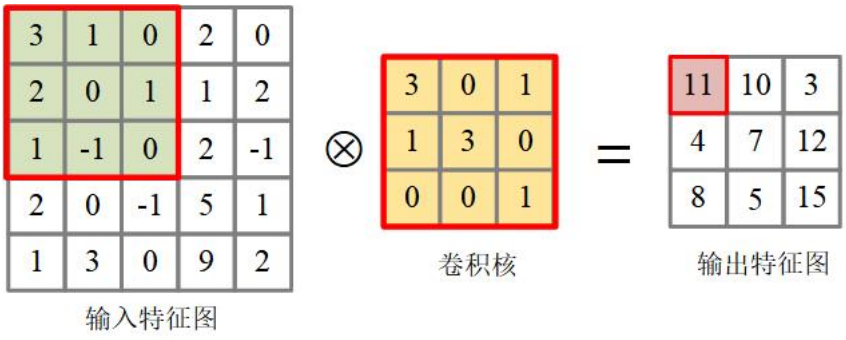


图2-4 卷积过程

Figure 2-4 Convolution process

（2）激活函数

卷积操作的本质是一种线性运算，卷积层的堆叠只能完成一些简单的任务。在实际运算中，简单的线性组合并不能解决所有问题，所以需要引入非线性函数来模拟实际情况。在神经网络中加入激活函数，相当于对模型进行去线性化操作，有更好的解决问题能力[37]。一般常见的激活函数包括 Sigmoid、tanh 和 ReLU 等。接下来具体介绍这些激活函数的定义：

Sigmoid 函数是一种常用于二分类的激活函数，其特点是输出结果在(0，1)

之间，输出稳定能够作为输出层，且函数图像连续曲线光滑易于求导。Sigmoid 函数如图 2-5（a）所示，表达式如式 2-4。

$$\sigma（x）=\frac{1}{1+e^{-x}} \tag{2-4}$$

Sigmoid 的导数取值范围较小，会导致梯度取值范围小。当多个神经网络层的梯度接近于 0 时，会出现梯度消失现象。Sigmoid 导数的函数曲线如图 2-5（b）所示，其取值较小，取值范围是(0，0.25)，Sigmoid 导数的函数表达式如式 2-5。

$$\sigma'(x)=\sigma(x)[1-\sigma(x)] \tag{2-5}$$

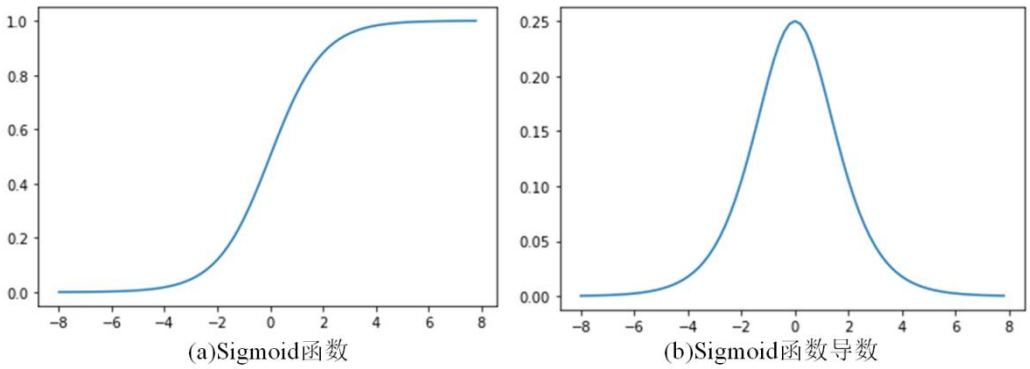


图 2-5 Sigmoid 函数及其导数图

Figure 2-5 Sigmoid function and its derivative

Tach 函数为双曲正切激活函数，输出结果在（-1，1）之间，以 0 为中心。Tach 函数图像如图 2-6（a）所示，Tach 的函数表达式如式 2-6。

$$tach(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2-6}$$

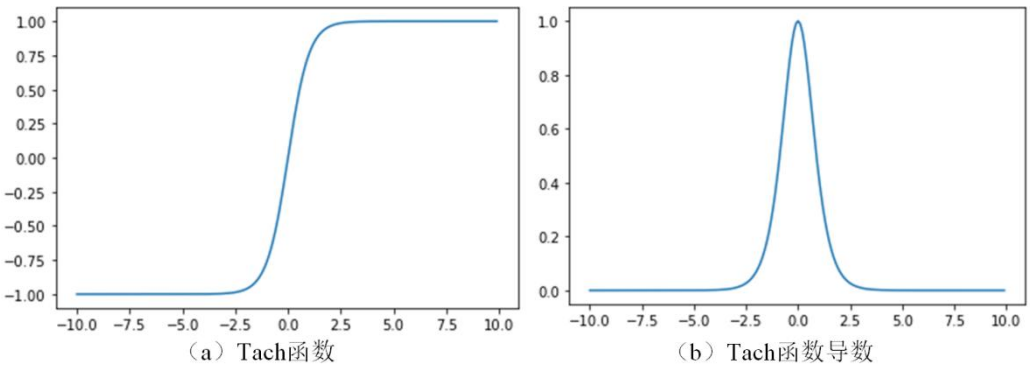


图 2-6 Tach 函数及其导数图

Figure 2-6 Tach function and its derivative

Tach 的导数范围在（0，1）之间，可以减轻梯度消失问题，容错性好。Tach

导数函数图像如图 2-6（b）所示，Tach 导数的函数表达式如式 2-7。

$$tach'(x)=\frac{4e^{2x}}{(e^{2x}+1)^2} \tag{2-7}$$

ReLU 函数通常结合卷积一起使用，其特点是输出只有 0 或正数。ReLU 是目前应用最多、最流行的激活函数之一。ReLU 的优点主要有两个，第一，简单计算量小。第二，有效改善梯度消失现象，因为当 x=0 时，ReLU 的导数恒为 1。ReLU 函数如图 2-7（a）图所示，表达式如式 2-8。

$$\max(0,x)=\begin{cases}0,(x\le 0)\\ x,(x>0)\end{cases} \tag{2-8}$$

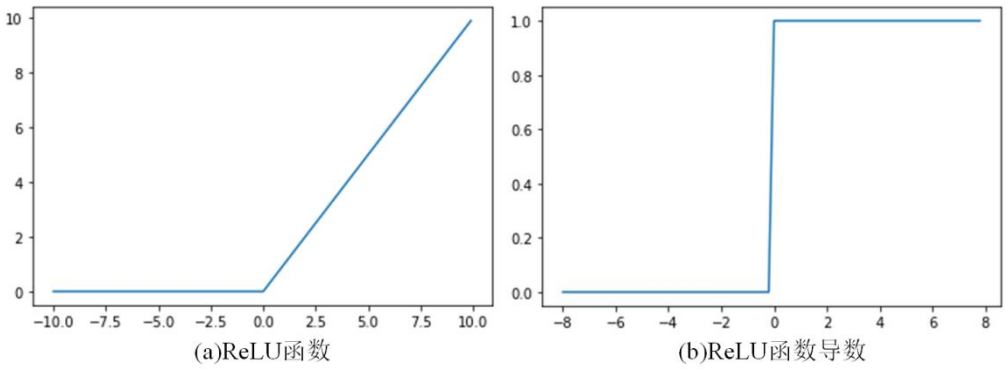


图 2-7 ReLU 函数及其导数图

Figure 2-7 ReLU function and its derivative

（3）池化层

为了避免特征维度过高，通常会使用池化层降低维度。池化操作可以减少特征图的尺寸与维度，以此降低模型复杂度。其中常见池化方式有两种，最大池化操作和平均池化操作。最大池化操作通过选取局部区域内的最大值来实现特征图尺寸的减小和维度的降低，平均池化是选取局部区域内的平均值来实现。池化的计算过程如图 2-8 所示[38]。

图 2-8 池化操作

Figure 2-8 Pooling operation

2.1.2 DeepLabV3＋网络

DeepLab 系列在语义图像分割领域中备受欢迎，由谷歌团队于 2015 年陆续提出，通过多尺度特征提取来提高网络整体性能[39][40][41]。

DeepLabv3+网络结构如图 2-9 所示。编码器阶段使用深度卷积神经网络进行特征提取，然后使用空洞空间卷积池化金字塔模块（Atrous Spatial Pyramid Pooling ，ASPP）[42][43]进行特征增强。在解码器阶段，将 ASPP 输出的特征图进行上采样操作，并与编码器中的浅层特征图进行拼接操作，再经过上采样操作至原始图像尺寸，使得上下文信息融合，得到最终的预测结果。

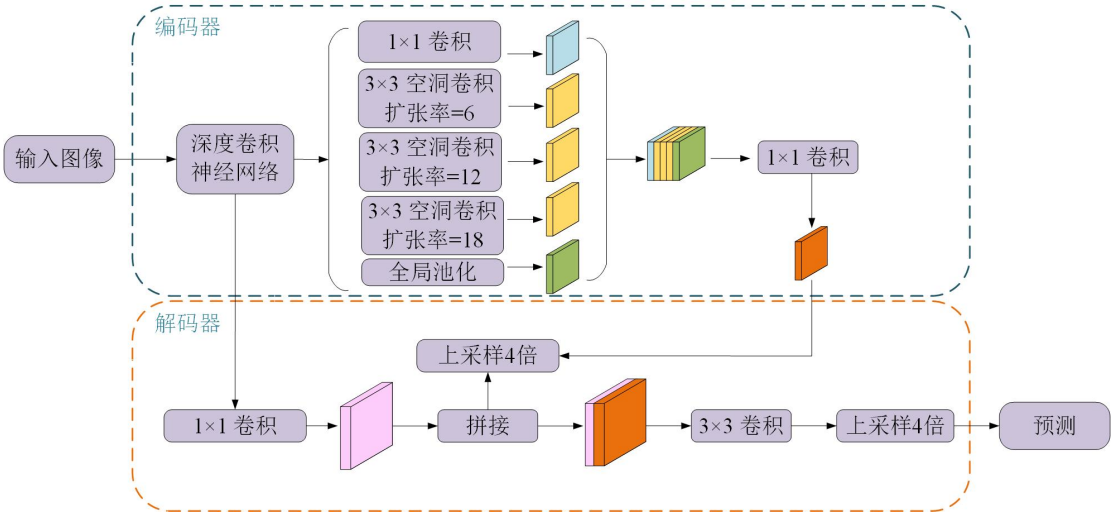


图 2-9 DeepLabv3+网络

Figure 2-9 DeepLabv3+ network

空洞卷积（Atrous Convolution）结合多尺度上下文信息，有效地扩大了感受野，从而更好地理解图像上下文，提高了分割的准确性；ASPP 模块能够在不同尺度上捕获图像的全局信息，有效地应对模糊边界问题；其使用深度卷积网络等轻量级结构作为特征提取网络，提高了模型的计算效率，使其更适用于实际应用。

（1）空洞卷积（Dilated Convolution）由 Yu 等人提出，通过扩张率（Dilation Rate），可以增加卷积感受野。空洞卷积的感受野计算公式为 2-9，其中 R 为卷积感受野，k 为卷积核尺寸，r 为扩张率。但是多次叠加相同尺寸的空洞卷积，会导致参与计算的像素不均衡，从而损失信息连续性。特别对于小物体目标，容易发生信息丢失，从而导致信息无法重建的问题。

$$R = k + (k-1)(r-1) \tag{2-9}$$

（2）ASPP 主要进行并行操作不同扩张率的空洞卷积，形成类似金字塔结构。ASPP 结构如图 2-10 所示。ASPP 结构包括五个分支，一个全局平均池化后进行 1×1 的卷积操作和上采样，一个 1×1 的卷积操作，三个扩张率分别为 6、12、18 的 3×3 的空洞卷积操作，五个分支输出进行拼接操作，再经过一个 1×1 的卷积操作输出。

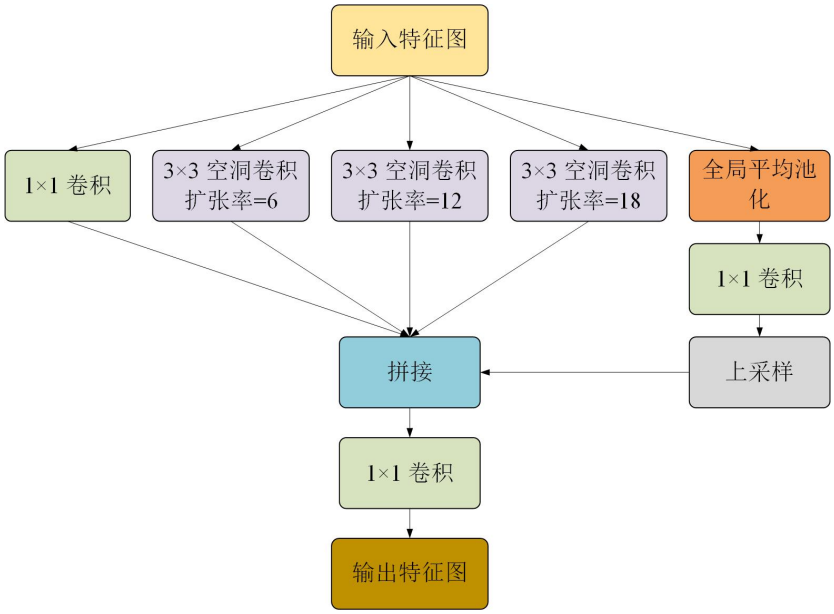


图 2-10 ASPP 结构

Figure 2-10 ASPP structure

（3）解码器能够有效地恢复边界信息，加速模型收敛和计算速度。DeepLabV3+的解码器将编码器输出的浅层信息进行降维操作，将编码器通道数降为 48，降低模型训练难度。编码器将 ASPP 输出的深层信息进行双线性 4 倍上采样，使得图像尺寸与浅层特征的尺寸相同，进而实现两者之间的拼接，提高模型表征能力，找回浅层信息，提高分割精度。对拼接后的特征，经过双线性上采样，输出预测结果，完成解码操作。解码器结构如图 2-11 所示。

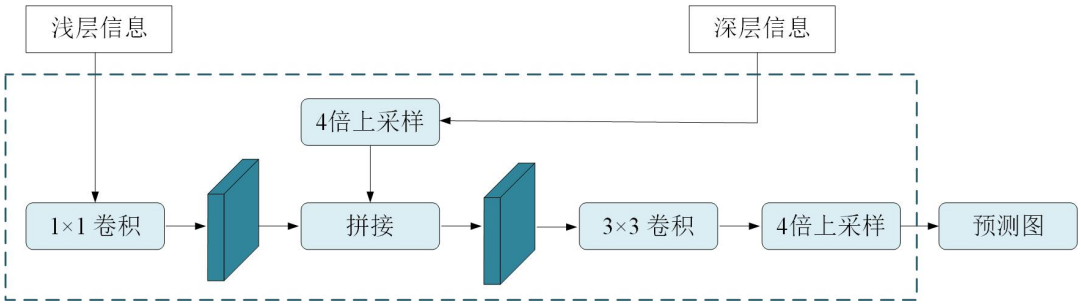


图 2-11 解码器结构

Figure 2-11 Decoder structure

（4）DeepLabV3+中常用 Xception 网络作为特征提取网络。Xception 是在 Inceptionv3 的基础上发展而来的特征提取网络，Inception 在对图像进行处理时，提出可分离卷积，即将特征图的通道信息和空间信息进行分离，进而加快训练速度，提高分类精度。Inceptionv3 采用卷积核大小和步幅不同的卷积层对其进行卷积操作，并将其卷积形成的并行特征图进行连接[44]。Inceptionv3 的网络结构如图 2-12 所示，该结构是先通过1×1的卷积对特征图在通道上进行处理，然后对输出的特征图进行3×3卷积合并。

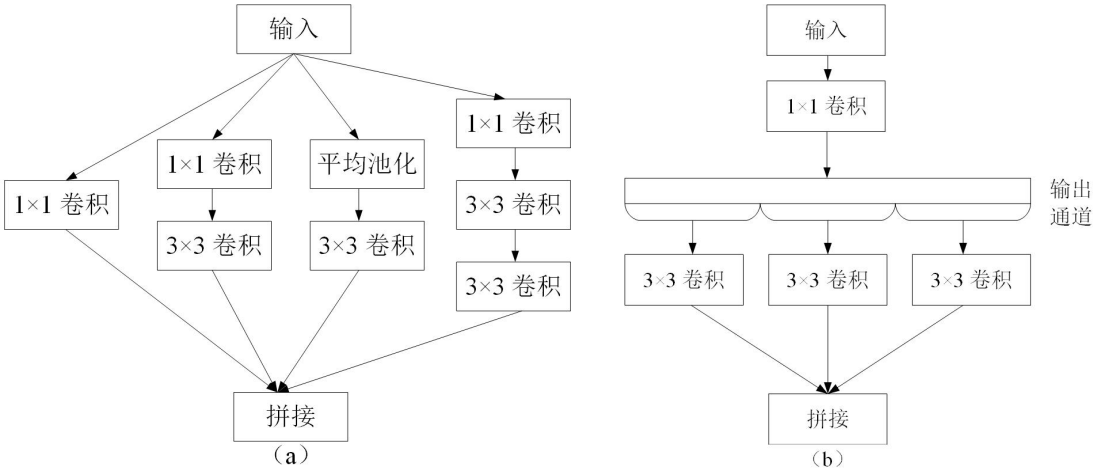


图 2-12 Inceptionv3 网络结构

Figure 2-12 Inceptionv3 network structure

Xception 是由谷歌团队提出的兼顾轻量化和高精度的模型。Xception 模型主要将 Inceptionv3 的普通卷积替换成深度可分离卷积，经过1×1的卷积拆分通道，再用3×3的深度卷积进行通道方向的拼接。Xception 网络结构如图 2-13 所示。

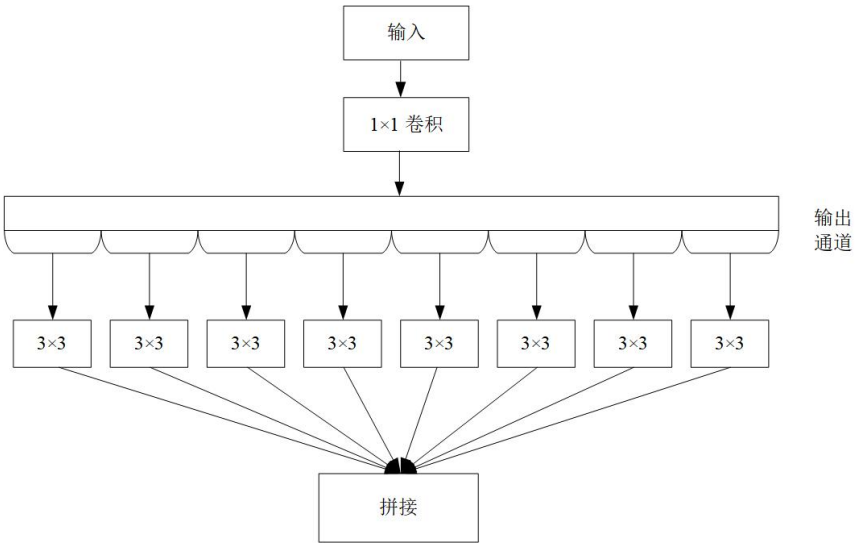


图 2-13 Xception 网络结构

Figure 2-13 Xception network structure

2.2 实时图像语义分割相关技术

2.2.1 模型轻量化技术

实时语义分割是在语义分割的基础上，采用一系列轻量化技术提高模型运算速度，使模型达到实时性的同时保证一定的精度，如使用轻量级网络、卷积分解操作、网络剪枝等。

（1）轻量级网络模型是为了在资源受限的环境中实现高性能而设计的一类深度学习模型。这类模型通过降低参数量和计算复杂度来实现模型的轻量化。SqueezeNet 在 2016 年被提出，SqueezeNet 通过使用 1x1 卷积核，极大地减少通道数，实现了模型的压缩[45]；使用全局池化代替全连接层，进一步减小了模型的复杂度；Fire 模块则通过组合不同大小的卷积核，提高了特征提取的多样性，增强了模型的表达能力。ShuffleNet 由斯坦福大学研究团队于 2017 年提出的，ShuffleNet 的通过将特征图分组、重组，以实现高效的参数共享和计算复用。模型首先将输入特征图分为多个组，然后在每个组内进行卷积操作。接下来，模型将每个组的特征图进行重组，以实现特征的交叉传递。最后，模型通过适当的连接方式将各个组的特征图进行整合，得到最终的输出特征图[46][47]。Xception 在 2017 年被提出，Xception 是一种基于 InceptionV3 的改进模型，其采用深度可分离卷积替换原来 InceptionV3 中的普通卷积，降低模型参数量，提高计算速度，但是相较其他模型网络结构较为复杂。MobileNet 系列在 2017 年被 Google 团队提出[48][49][50]，MobileNetv1 采用了深度可分离卷积，将标准卷积拆分为深度卷积和逐点卷积，减少了参数量和计算量；MobileNetv2 使用了瓶颈残差模块，通过跳跃连接来促进信息流动，并提高模型的收敛速度和泛化能力；MobileNetv3 使用了 NAS 和 NetAdapt 算法搜索最优模型结构，在瓶颈残差模块中引入压缩和激励模块（Squeeze-and-Excitation）。

（2）卷积操作分解：深度可分离卷积将卷积操作分解为深度卷积和逐点卷积。深度卷积对通道滤波，逐点卷积对通道进行转换。如图 2-14 所示，若输入特征图为 $H \times W \times C_1$，输出特征图为 $H \times W \times C_2$，标准卷积核大小为 $K \times K \times C_1$，标准卷积的过程所需要的参数量为 $(K \times K \times C_1) \times C_2$，深度卷积的过程所需要的参

数量为$(K\times K\times 1)\times C_1$，逐点卷积的过程所需要的参数量为$(1\times 1\times C_1)\times C_2$。由以上可以得出，深度可分离卷积的参数量是标准卷积的$\frac{K\times K\times C_1+C_1\times C_2}{K\times K\times C_1\times C_2}=\frac{1}{C_2}+\frac{1}{K^2}$。

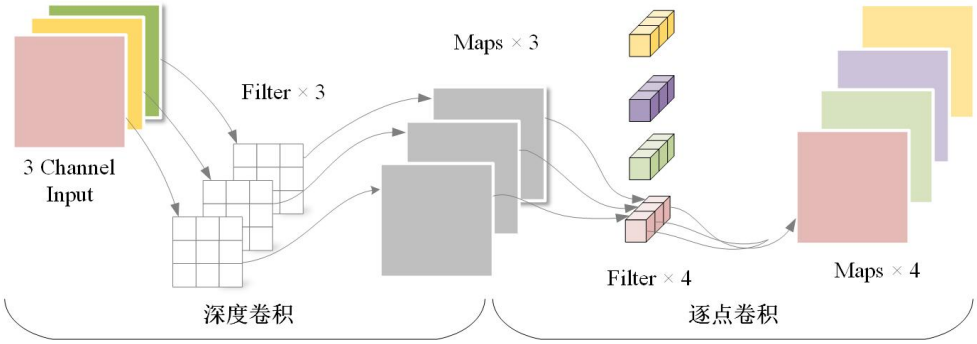


图 2-14 深度可分离卷积

Figure 2-14 Deep separable convolution

非对称卷积在水平和垂直方向采用不同大小的卷积核，将一个$n\times n$的卷积核拆解为$n\times 1$和$1\times n$两个卷积核，两种卷积操作结果相同。如图 2-15 所示，标准卷积核为$n\times n$时，一次卷积操作需要进行$n\times n$次乘法操作；若使用非对称卷积，一次卷积操作仅需要$2\times n$次乘法操作。由以上可以得出，非对称卷积的参数量是标准卷积的$\frac{2\times n}{n\times n}=\frac{2}{n}$。

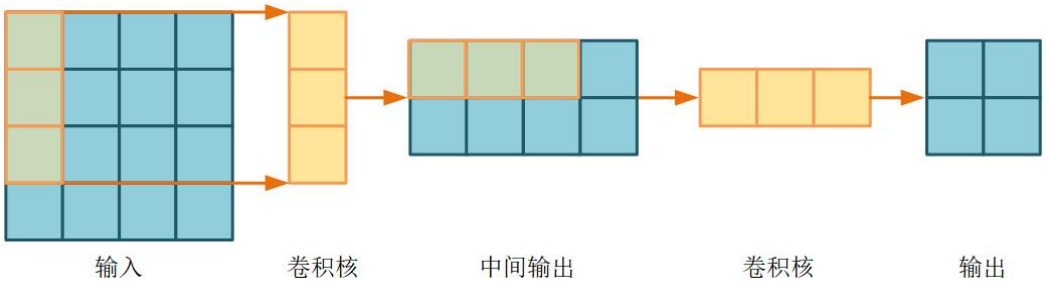


图 2-15 非对称卷积

Figure 2-15 Asymmetric convolution

（3）剪枝是将神经网络中的冗余参数移除的方法，模型剪枝如图 2-16 所示。非结构化剪枝技术基于启发式方法来排除非重要参数，例如梯度[51]、权重绝对值[52][53]、Hessian 统计[54]。该方法能够增强图像分割性能，但由于不规则的稀疏性，加速的过程更加困难。Yang 等人[55]建立了延迟表，利用贪心算法来确定裁剪层，并且还使用L_1和L_2正则化进行比较实验，在剪枝后不训练的情况下L_1比L_2的精度更好，然而在修剪之后重新训练的情况下L_2比L_1的精度更好。Neill 等人[56][57]

提出了两个权重正则化策略，一个是对齐最大化修剪，另一个是未修剪网络单元。非结构化剪枝虽然能够显著减少网络参数量，但是非结构化剪枝仅仅是将冗余神经元设置为零，而不是直接从网络结构中删除[58]。

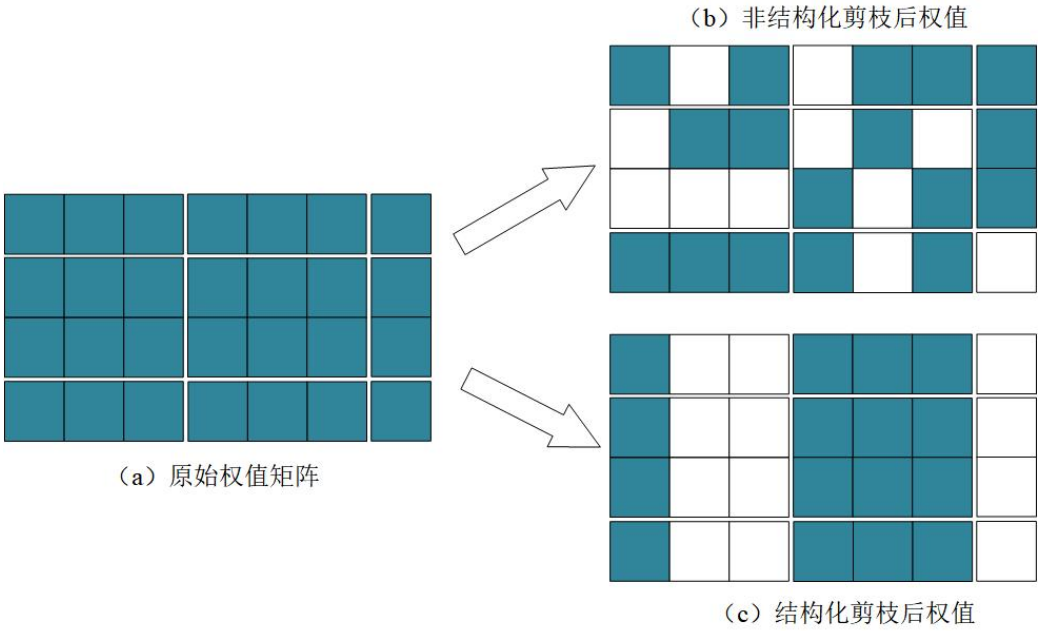


图 2-16 非结构化剪枝与结构化剪枝

Figure 2-16 Unstructured pruning and structured pruning

基于通道的结构化剪枝通过评估通道的重要性进行修剪。Li 等人[59]提出了一种融合最大平均池化和改进的通道注意力机制方法 ，通过使用两个池化函数来增强 DNN 中的特征表示。Kuang 等人[60]通过每个通道对任务相关损失函数的影响，通道重要性越低，损失函数值越低。Li 等人[61][62]在提出了一种高效的 DNN 分层精细剪枝高效方法，加速了硬件层面的推理过程[63]。目前网络剪枝方法仍存在一些问题，大多数剪枝都需要人工设置剪枝率，需要对部分参数进行微调，难以应用。

（4）通常情况下，网络模型参数量的储存以及模型计算的相关参数基本都是单精度方式，可以减少存储需求,并通过对网络模型参数存储和计算策略的调整来加快计算速度。量化技术是一种新型存储技术，在大型数据集和复杂网络拓扑结构中效果显著。量化技术实际上是一种低精度的网络模型参数存储和计算方式，合适的量化操作并不会带来严重的精度损失，并且能够有效地防止模型过拟合[64]。

2.2.2 知识蒸馏技术

神经网络模型通常呈现为单一的，宽而深的结构，或者是多个基础模型的集合，这种模型具有较强的收敛能力和任务处理性能[65][66]。相对而言，简单模型通常往往只拥有一个简单的框架，以及较浅的网络结构，因此其描述能力相对有限。知识蒸馏技术巧妙的结合了复杂模型强大处理能力和简单模型参数量小的特点，通过将复杂模型的知识迁移给简单模型，达到模型高效压缩处理的目的。知识蒸馏技术在处理相似任务时能够提升模型的精度、降低模型的时延。

知识蒸馏的实现步骤类似于老师教学的过程，如图 2-17 所示。首先，需要有一个经过充分训练的“老师”，也就是一个预先训练的性能良好的复杂模型。这个模型就像是一位经验丰富、知识渊博的老师，能够深刻理解任务并有效地处理数据。接着，从这个“老师”那里获取软标签和预测标签，这相当于学生从老师那里获取知识和指导。这些标签包含了复杂模型对任务的理解和处理方式。接下来，有了一个“学生”，也就是一个简单模型。但是，由于简单模型的结构相对简单，独立训练时很难达到理想的性能水平。因此，让“学生”模型从“老师”那里学习，通过复杂模型提供的标签来辅助训练和收敛。同时，使用真实值同时训练学生模型，相当于给出标准答案让学生学习。这样，“学生”模型就能够在“老师”的指导下更有效地学习任务的知识，从而提高学习效率。整个过程就像是在一位老师的耐心指导下，学生能够更快地掌握知识和技能一样[67]。

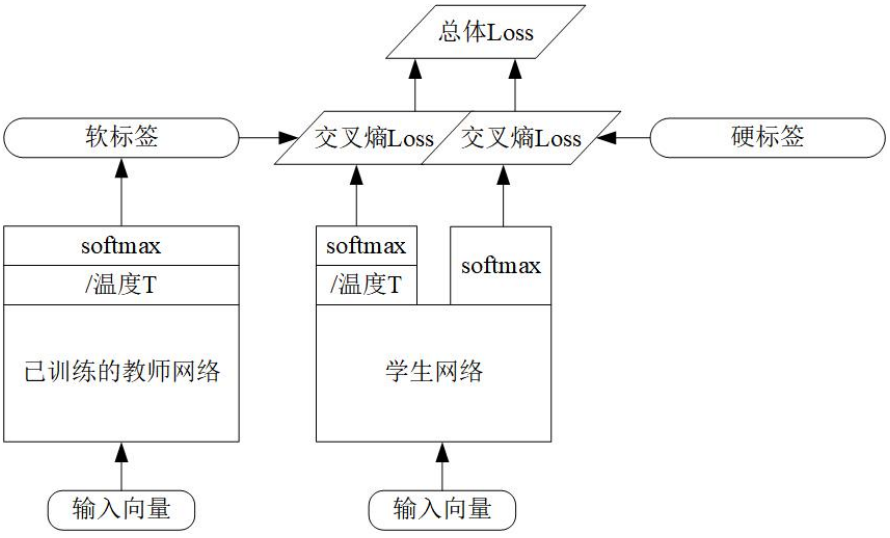


图 2-17 知识蒸馏步骤

Figure 2-17 Knowledge distillation steps

2.3 嵌入式平台与推理引擎

深度学习作为当前人工智能领域的重要技术手段，在文本，语音，图像处理等领域展现了巨大的优势。针对问题的难度不断提升，深度学习网络架构朝着复杂的方向深入，对计算资源和能力的需求不断增加。然后是在实际应用场景中，硬件资源有限，因此在嵌入式设备上的使用是深度学习实现广泛的前提。目前常见嵌入式平台有 Coral Edge TPU、FPGA、Raspberry Pi、Jeston 系列等，本文在上述嵌入式平台中选取了性能优异的 Jeston TX2。

Jeston TX2 是一种 CPU+GPU 的嵌入式架构，CPU 包含 Denver 处理器和 ARM Cortex-A57 处理器，内核最高工作频率为 2.0GHZ。GPU 包含 256 个基于 Pascal 架构的 CUDA 核心,每个内核的最高工作频率为 1.3GHz。此外，Jeston TX2 包含 8GB 128bit LPDDR4 的运行内存，板载 32GB 的 eMMC 存储空间，58.4GB/s 的带宽。Jetson TX2 配备了数据传输模块，可以实现 2.5Gbps 的通道带宽和 1.4GPi/s 的计算能力。单精度浮点运算（FP32）的计算能力可以达到 665.6 GFLOPS，功耗最低可以达到 7.5W，其实物图如图 2-18 所示。

图 2-18 Jetson TX2 实物图

Figure 2-18 Jetson TX2

尽管 Jeston TX2 拥有强大的计算能力，但是部署在嵌入式平台的深度学习模型需要花费较长时间预测，无法满足实时需求。针对以上问题，常使用深度学习加速推理引擎 TensorRT 实现对模型的推理加速。

TensorRT 是一款加速推理引擎，具有深度学习能力。一般采用深度学习框架训练模型，将其部署到嵌入式设备 Jeston TX2 中，利用 TensorRT 进行加速。

经过加速后的网络模型，能够显著减少模型的运算时间，提高计算速度，特别在对实时性要求高的场景下，具有着现实意义。此外，TensorRT 还支持低精度的数据校准。通常在网络中张量精度为 32 位浮点数(FP32)时，深度学习框架会对神经网络进行训练。但在模型部署与推理阶段，只需向前传播即可，因此为了提高性能，数据精确度可以适当降低。

深度学习加速推理引擎 TensorRT 对卷积神经网络进行合并，将卷积层，偏置层和激活层合并为一层，合并层为 CBR 层，层间纵向合并不仅有效减少了网络层数，同时对 GPU 资源进行了合理利用。合并的操作如图 2-19 所示。

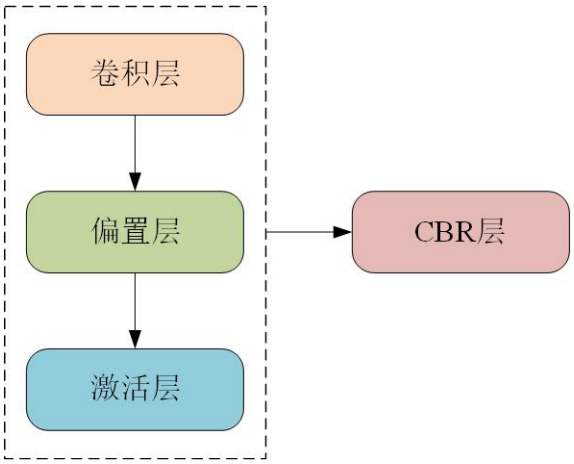


图 2-19 TensorRT 层间纵向合并

Figure 2-19 Vertical merge between TensorRT layers

图 2-19 是对网络层间进行合并，该合并减少了运算时间，但是卷积神经网络往往含有多层卷积，所以会合成多个 CBR 层。当含有多个 CBR 层时，TensorRT 会对网络进行层间横向合并，将三层 CBR 层合并为一个 CBR 层，该操作不仅精简了网络结构，还使网络更加轻量化。其中横向合并操作如下图 2-20 所示。

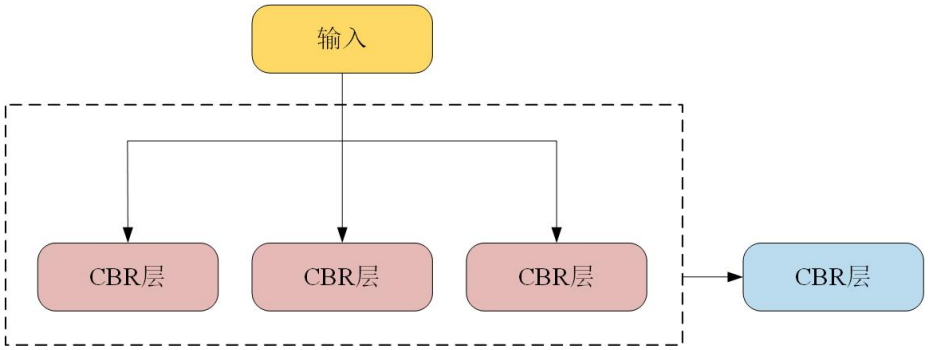


图 2-20 TensorRT 层间横向合并

Figure 2-20 Horizontal merging between TensorRT layers

TensorRT 推理首先要经过模型构建，如图 2-21 左边构建部分所示，首先要

将神经网络导入 TensorRT 优化器中进行格式转换，再对其进行序列化操作，将其转换为 TensorRT 模型。然后再进入模型部署阶段，如图 2-21 右边部署部分所示，TensorRT 模型先经过反序列操作生成 Runtime Engine。在此过程中，完成动态张量存储和 CUDA 内核的自动调整，以确保模型在 GPU 上的高效运行。完成部署后，即可进行推理运算，得到模型的预测结果。

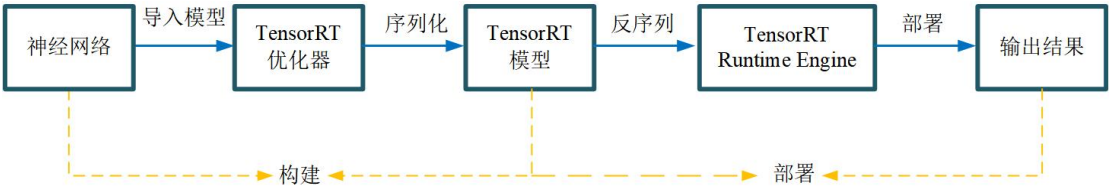


图 2-21 TensorRT 推理过程

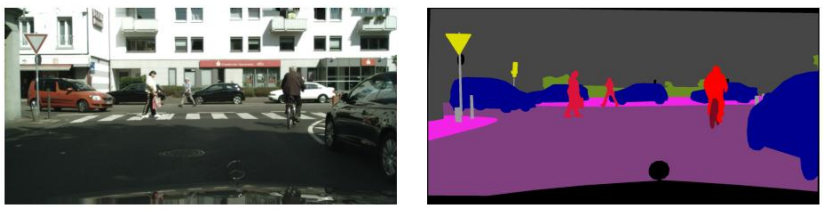
Figure 2-21 TensorRT inference process

2.4 数据集与评价指标

2.4.1 道路场景常用数据集

本文是针对道路场景提出的语义分割方法，为了验证本文研究方法，本文采用了两个不同的道路场景数据集进行实验验证，分别是 Cityscapes 城市街景数据集和 CamVid 道路场景数据集。

Cityscapes 是一个城市街景数据集，如图 2-22，该数据集从行驶车辆视角采集图像，主要在天气良好时采集。图片示例如图所示，整个数据集包含 24998 张图片。本文实验中采用 2975 张精细注释图像以及 500 张验证图进行训练和验证，1525 张测试集图像，数据集中包含 19 个类别和 1 个背景，图像尺寸为 1024×2048 像素。



（a）原图　　（b）标注图像

图 2-22 Cityscapes 数据集

Figure 2-22 Cityscapes dataset

在语义分割中，使用不同的颜色将不同类别对象分割成不同的区域，每个颜

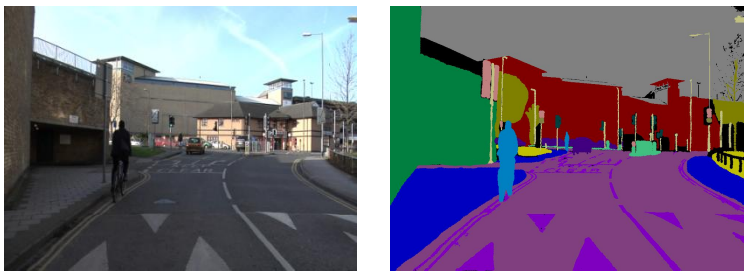
色和类别标签一一对应。在 Cityscapes 数据集中，采用的 19 个类别标签和对应的标签颜色对应关系如表 2-1 所示：

表 2-1 Cityscapes 数据集标签颜色对应

Table 2-1 Cityscapes dataset label color correspondence

编号	名称	颜色标签	编号	名称	颜色标签
0	道路	[128,64,128]	10	天空	[70,130,180]
1	人行道	[244,35,232]	11	人类	[220,20,60]
2	建筑物	[70,70,70]	12	骑行者	[255,0,0]
3	围墙	[102,102,156]	13	汽车	[0,0,142]
4	围栏	[190,153,153]	14	卡车	[0,0,70]
5	柱子	[153,153,153]	15	公交车	[0,60,100]
6	交通灯	[250,170,30]	16	火车	[0,80,100]
7	交通标志	[220,220,0]	17	摩托车	[0,0,230]
8	植被	[107,142,35]	18	自行车	[119,10,32]
9	地形	[152,251,152]			

CamVid（The Cambridge-driving Labeled Video Database）是一个道路场景数据集。数据集以汽车行驶角度采集城市道路场景图片，包括驾驶在白天和夜晚等不同时间段的情景。数据集中包含超过 700 张图像，其中含有 367 张训练集图像，101 张验证集图像，233 张测试集图像。数据集包含 11 个类别，图像尺寸大小为 960×720 像素，部分示例如图 2-23 所示：



（a）原图像　　（b）标注图像

图 2-23 Cimvid 数据集

Figure 2-23 Cimvid dataset

在 Camvid 数据集中，采用的 11 个类别标签和对应的标签颜色对应关系如表 2-2 所示：

表 2-2 Cimvid 数据集标签颜色对应

Table 2-2 Camvid dataset label color correspondence

编号	名称	颜色标签	编号	名称	颜色标签
0	天空	[128,128,128]	6	交通标志	[192,128,128]
1	建筑物	[128,0,0]	7	围墙	[64,64,128]
2	柱子	[192,192,128]	8	汽车	[64,0,128]
3	道路	[128,64,128]	9	行人	[64,64,0]
4	人行道	[0,0,192]	10	骑行者	[0,128,192]
5	树	[128,128,0]			

2.4.2 语义分割任务评价指标

为了全面评估模型的性能并了解其泛化能力，研究人员通常会采用多种评价指标。常见的评价指标包括交并比（Intersection over Union，IoU）、每秒帧数（Frames Per Second，FPS）、浮点运算数（Floating-point Operations，FLOPs）以及参数量（Parameters，Params）等。

（1）平均交并比

交并比（IoU）是一种分割精度指标，为预测结果和真实标签的交集在总并集中的比例。每张图片中含有不同类别的目标，所以使用平均交并比（mean intersection over union，mIoU）来评估图片的分割精度。平均交并比即计算每个类别的交并比平均值，其计算公式为：

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{2-10}$$

（2）每秒帧数

每秒帧数（FPS）是一种分割速度指标，计算模型每秒可处理的图像数量，FPS 越大，模型分割速度越快，当 FPS 大于 24 时被认为是实时语义分割。

（3）浮点运算数

浮点运算数（FLOPs）衡量模型复杂度的指标，反应模型运行计算力。在实际应用中，模型需要部署在资源受限的环境中，如手机和汽车等设备上，对模型的计算量也需要加以关注。

（4）参数量

参数量是衡量模型大小的一种重要指标，参数量越小，则模型越小，容易部署在移动设备等资源受限环境中。卷积神经网络中含有卷积层和全连接层，卷积层的参数量计算公式如 2-11 所示，全连接层的计算公式如 2-12 所示。

$$Para_{(conv)} = K \times K \times C_{in} \times C_{out} \tag{2-11}$$

$$Para_{(connect)} = C_{in} \times C_{out} \tag{2-12}$$

其中，K 为卷积核尺寸，C_{in} 为输入特征图的通道数，C_{out} 为输出特征图的通道数。

（5）像素准确度

像素准确度（Pixel Accuracy ，PA）是一种预测像素准确度高低的指标，像素准确度计算正确分类的像素个数和总像素个数之间的比值。像素准确度高则语义分割类别分割准确率高。

2.5 本章小结

本章首先介绍了全卷积神经网络的基本结构和工作方式。其次，介绍了本文基础网络 DeepLabV3+网络。然后介绍了实时图像语义分割相关技术。然后介绍了道路场景数据集 Cityscapes 和 Camvid 以及评价指标 MIoU、FPS、FLOPs 和参数量。

第 3 章　基于轻量化网络的道路场景实时语义分割方法研究

3.1 引言

主流的图像语义分割技术往往集中精力提高分割精度，却在推理速度上付出了代价，通常采用深且宽的网络结构。实时语义分割网络速度较快，但层数较浅，导致感受野不足，难以满足对高精度和快速决策的双重需求。主流网络设计在高分辨率和高层语义之间存在矛盾，制约了精度和速度的平衡。

DeepLabV3+网络是目前主流的优秀模型之一，在 Cityscapes 等数据集上都取得了优异成绩。但也存在不足，编码器中特征提取操作逐步降低图像维度，造成边缘信息丢失，跳跃连接也无法实现细节恢复；空洞空间卷积池化金字塔模块（Atrous Spatial Pyramid Pooling，ASPP）在模型中对分割精度的提高有着巨大贡献，提高模型对于目标边界的提取，但是主要注重局部特征信息，并未关注到局部特征之间的关系；特征提取网络为网络层数较多、参数量较大的 ResNet101[68]，虽提高了分割精度，但模型复杂度高，参数量大，网络计算负担大。ASPP 模块中使用标准空洞卷积，空洞卷积扩大了卷积感受野，但进一步增加了参数量，模型深度加深，对硬件要求高，网络训练速度慢。

针对以上问题，本章提出了一种基于轻量化网络的道路场景实时语义分割方法（Lightweight Road Real-time Segmentation Network，L2RS-Net）。本文在 DeepLabV3+的基础上，将特征提取网络替换为轻量化网络，并对该轻量化网络进行卷积操作，设计浅层网络，进一步降低参数量[69]；引入非对称空洞卷积替代标准空洞卷积，进一步降低模型参数量，提高计算速度；引入跳跃连接，对编码器-解码器进行相同尺寸特征图拼接操作，找回浅层信息，提高分割精度。

3.2 基于轻量化网络的道路场景实时语义分割方法

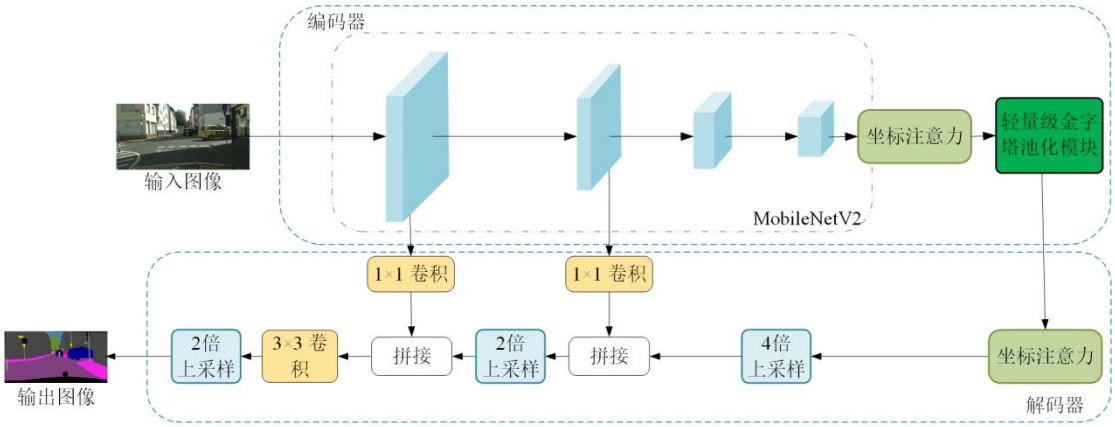


图 3-1 轻量化道路实时语义分割网络

Figure 3-1 Lightweight road real-time network

本章提出了一种基于轻量化网络的道路场景实时语义分割网络 L2RS-Net，如图 3-1 所示。本章选用传统的 DeeplabV3+网络结构为基础网络，对其进行轻量化处理。

采用轻量级网络 MobileNetv2 替代原始模型中的特征提取网络 Xception，为防止多倍下采样导致特征图尺寸过小，将 MobileNetv2 网络的下采样倍率从 32 降低为 16，为了避免标准卷积计算负担重，引入深度可分离卷积；构建了一种轻量化金字塔池化模块（Lightweight Pyramid Pooling Module，LPPM），将空洞卷积替换为非对称空洞卷积，并且使用压缩因子 a 压缩特征通道数，减少计算量；在特征提取网络和 LPPM 后同时引入坐标注意力模块，弥补分割精度损失；解码器中将编码器输出的浅层特征图与 LPPM 输出的深层特征图相融合，在浅层特征图尺寸大小为原始图像 $\frac{1}{2}$ 和 $\frac{1}{4}$ 时进行两次拼接操作，可以有效的结合上下文信息，找回浅层信息，提高小物体分割精度；引入焦点交叉熵损失函数，增加权重因子，平衡数据集中简单困难样本的学习能力。

3.2.1 特征提取网络构建

特征提取网络使用 MobileNetv2 网络，网络结构如图 3-2 所示，在该网络中首先使用一个 3×3 的普通卷积来提取图像特征，使 512×1024×3 的输入变为 256×512×3，然后采用了批量归一化和 ReLU 激活函数进一步处理这些特征。

输出特征图通过 17 个专为 MobileNetv2 设计的线性瓶颈卷积模块（从 Block0 至 Block16）继续提炼和压缩特征。在这一系列操作中，Block0 和 Block2 的输出被用作与 DeepLabV3+解码器的中间特征层拼接，而 Block16 的输出则被送往编码器里的 ASPP（Atrous Spatial Pyramid Pooling）结构以提取更深层次的特征信息。Block 的内部结构如图 3-3 所示。

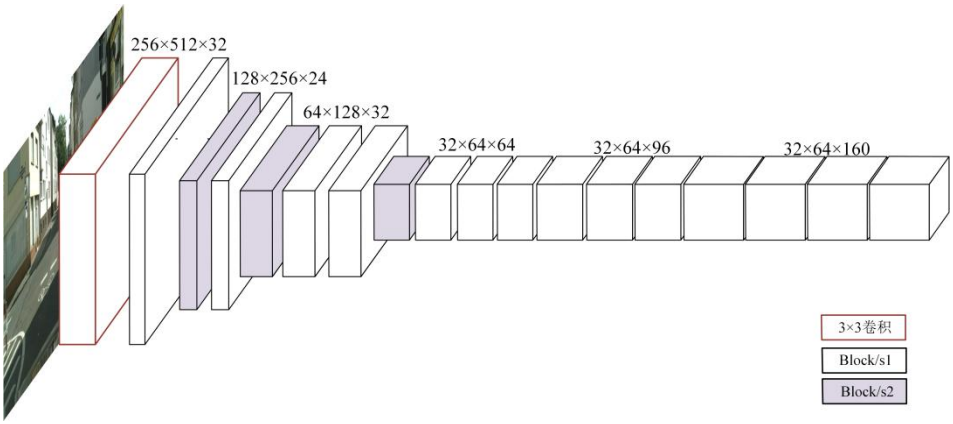


图 3-2 MobileNetv2 网络结构

Figure 3-2 MobileNetv2 network structure

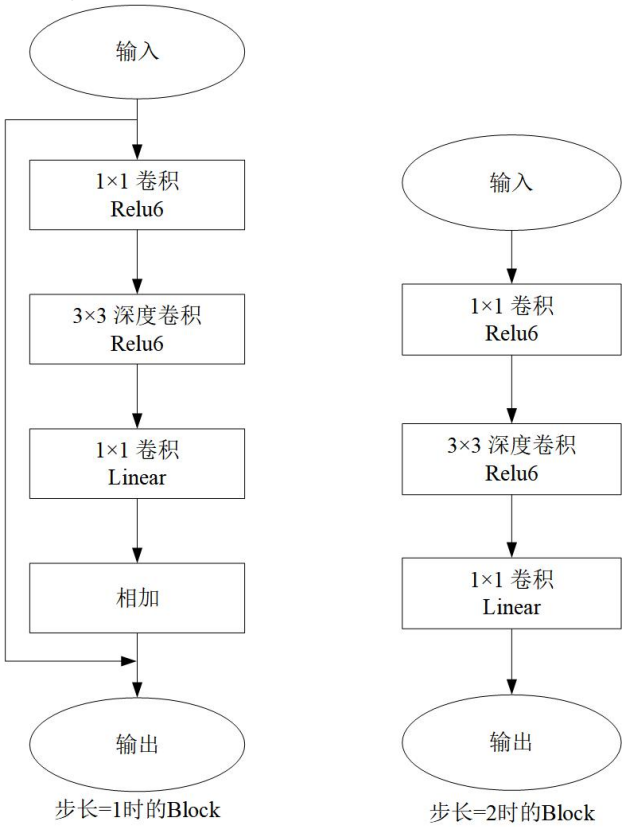


图 3-3 Block 的内部结构

Figure 3-3 Internal structure of block

具体到 Block 的工作机制，当步长设置为 1（即 stride=1）时，该结构采用

倒置残差结构。处理流程是先进行 1x1 卷积，用于扩展通道数，紧随其后的是批量归一化和 ReLU6 激活，ReLU6 的引入可以限制激活值的范围，提高模型稳定性。接下来是一步关键的3×3深度可分离卷积，区别于传统深度卷积，采用了空洞卷积策略以增强模型的感受野，再次经过批量归一化和 ReLU6 激活后，通过1×1卷积将通道数降低，并进行最终的批量归一化，此处不再应用激活函数。最后，输入和输出通过相加完成残差连接，这是提高模型性能的关键策略。

当步长增至 2（即 stride=2）时，Block 的机制则是使用步长实现下采样，以便更高效地提取特征，其结构与步长为 1 时相似，但不将输入与输出相加，以适应下采样的需求。

3.2.2 特征融合模块构建

空洞空间卷积池化金字塔严重依赖于标准的空洞卷积，并且五个金字塔分支都会产生固定数量的特征图，导致特征融合模块参数量较大。针对以上问题，对 ASPP 结构进行调整，如图 3-4 所示，建构了一种轻量化金字塔池化模块 LPPM。LPPM 由五个分支组成，一个分支使用 1×1 的卷积核对特征图进行降维操作，保留原始图像特征，降低计算复杂度；其中三个分支使用 3×3 的卷积核进行非对称空洞卷积操作，该卷积即将 3×3 的空洞卷积核因式分解为两个空洞卷积核，3×1 和 1×3 卷积核，减少模型参数量；一个分支使用自适应均值池化，该池化不用固定卷积核大小和步长，只需要指定最后的输出尺寸，在 LPPM 中输出尺寸设定为 1×1，然后使用 1×1 进行降维再进行上采样至原始尺寸。最后将五个分支的输出进行拼接操作，再进行降维操作，使输出图通道数与输入通道数相同。

五个分支生成的特征图通道数并不固定，通道数由 K/a 定义，K 是输入特征图通道数，a 是压缩因子，压缩因子 a 可以根据可用的计算能力来设置，本文将压缩因子 a 设置为 4。对于一组输入特征图 $X \in R^{C\times H\times W}$，LPPM 模块被定义为：

$$Y = F_{1,1}(F_{1,1}(X) \| A_{3,6}(X) \| A_{3,12}(X) \| A_{3,18}(X) \| F_{1,1}P_{1,1}(X)) \tag{3-1}$$

$$A_{k,r} = F_{(1\times k),r}(F_{(k\times 1),r}(X)) \tag{3-2}$$

其中 Y 表示输出特征图，$F_{1,1}$ 表示 1×1 卷积，$A_{k,r}$ 表示非对称扩张卷积分支，

$F_{(1\times k),r}$ 和 $F_{(k\times 1),r}$ 分别表示非对称卷积，r 表示扩张率， $P_{1,1}(X)$ 为自适应均值池化。

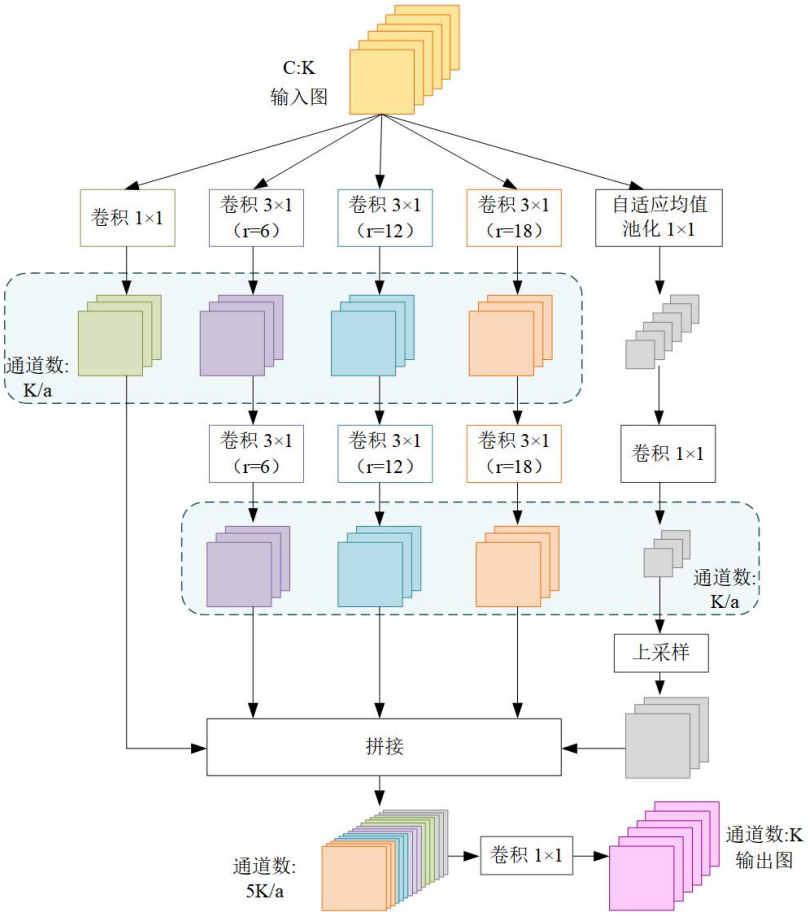


图 3-4 轻量化金字塔池化模块

Figure 3-4 Lightweight pyramid pooling module

LPPM 的主要贡献在于对计算量的减少，一方面通过非对称卷积实现，例如一个 n×n 的空洞卷积需要进行 K×n×n×F 次操作，K 为输入特征通道数，F 为输出特征通道数。而非对称卷积实现相同感受野的前提下，只需进行 2×K×n×F 次操作；另一方面通过压缩因子 a 实现，非对称卷积使用压缩因子后，只需进行 $\frac{3}{a}K\times n\times\frac{1}{a}K$ 次操作，当 a=4，n=3 时，可以减少大约 70%的参数量。

3.2.3 特征增强模块构建

特征提取网络的替换和轻量化特征融合模块的构建，降低了模型参数量，提高了计算速度，但是导致了分割精度的降低。针对这个问题，在特征提取网络以及特征融合网络后同时引入注意力机制。

注意力机制是根据图像中像素点的重要程度，将计算出的权重赋值给原始图

像，根据注意力的不同偏重，计算出不同的权重。空间注意力机制是对不同空间信息进行加权，提取关键信息，使模型关注重点区域。平均池化操作压缩空间维度，使用最大池化收集图像重要信息，该注意力对通道维度同时使用平均池化和最大池化，整合了空间信息和局部特征信息，但是空间注意力机制没有注意到通道之间的关系。压缩激励注意力机制（Squeeze and Excitation，SE）包含压缩和激励操作，输入图 F 先进行压缩操作，输入图的维度压缩后进行激励操作，先对输入图进行两次全连接操作，计算通道权重，进行归一化处理。但 SE 只考虑通道维度上的注意力，无法捕捉到空间上的注意力。所以本文选用坐标注意力机制（Coordinate Attention，CA），CA 通过考虑像素的位置信息和远程依赖来调整注意力，使模型更加关注图像中不同位置的信息，适用于处理具有空间结构的任务，能够更好捕捉到空间上的相关性。CA 分为两个步骤，坐标信息嵌入和坐标注意力生成。其基本原理如图 3-5 所示：

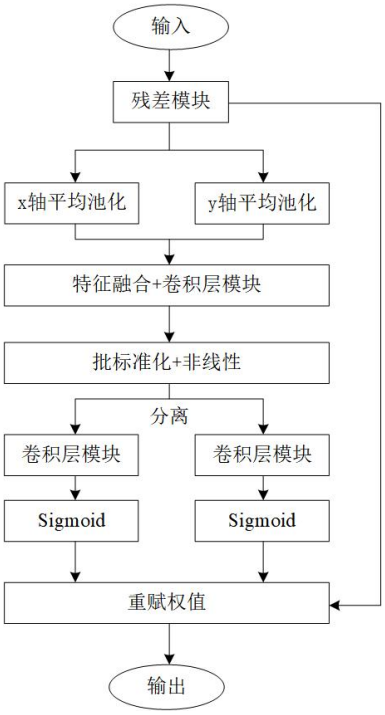


图 3-5 坐标注意力模块

Figure 3-5 Coordinate attention

（1）首先进行坐标注意力嵌入，输入图像进入残差模块后，对输出特征图在进行以为特征编码操作，分别在 X 轴和 Y 轴上进行最大池化操作，得到 X 轴和 Y 轴的方向感知特征图。

（2） 将 X 轴和 Y 轴所得出的方向感知特征图进行特征融合操作，并通过卷积层建立远程依赖关系。

（3）经过特征融合后的卷积特征图进行批归一化操作和非线性操作，其中激活函数使用 h_wish。

（4）然后进行坐标注意力生成，经过批归一化和非线性操作后的特征图进行分离操作，该操作可以通过分离函数得到 X 轴和 Y 轴的注意力值。

（5）将基于 X 轴的注意力值和基于 Y 轴的注意力值经过 Sigmoid 激活函数后，重赋权值并合并输出。

3.2.4 损失函数优化

在深度学习中，常用损失函数优化网络参数，以此计算预测值和真实值之间的差距，并使用梯度下降来降低预测值和真实值的插值。语义分割中常使用交叉熵损失函数学习难以平衡对数目较少的样本，通过衡量预测值与真实值之间的差异情况，交叉熵的值越小，则表示预测值和真实值之间差异越小，预测越准确。交叉熵损失采用类内竞争机制，所以类间的信息学习效果较好，适用于样本均匀分布的数据集。但由于本文使用的数据集类别数量有差异，比如不同类别的样本之间有数量差异，交叉熵损失主要注重学习数量较多的汽车特征，从而忽视了小样本火车的特征，导致学习能力不足；并且数据集之间的类别有学习难度之间的差异，例如数据集含有大量简单的背景，以及少量较难的前景，导致类别学习不平衡。

本文引入焦点损失（Focal loss，FL）替代交叉熵损失，通过样本难度调节因子解决简单、困难样本不平衡问题。具体的表达式为：

$$FL（p_t）=-(1-p_t)^{\gamma}\log(p_t) \tag{3-3}$$

其中，p_t 为真实概率分布，$-\log(p_t)$ 为初始交叉熵损失，$(1-p_t)^{\gamma}$ 是样本难度调节因子，γ 的取值范围大于 0。当预测越准确时，$(1-p_t)^{\gamma}$ 趋于 0；$(1-p_t)^{\gamma}$ 趋于 1，则预测不准确。焦点损失对于分类不准确的样本，并未改变其损失，但对于分类准确的样本，降低其损失，则相当于增大了分类不准确样本的样本权重。

针对数据集类别样本分布不均匀的问题，引入类别权重因子α，提升模型对

困难样本的分割能力和精度。该方法提升小样本类别的权重，降低大样本的类别权重，$\alpha \in [0,1]$，使用类别的数量反比作为对应类别的权重。经过调整之后，会加强模型对于小类别的特征提取能力，从而对数据集中各个类别的特征进行均衡学习。

$$FL（p_t）=-\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{3-4}$$

在公式 3-4 中，p_t 表示样本属于目标类别的概率，$-\log(p_t)$ 为初始交叉熵损失函数，α_t 为类别间的权重参数，$(1-p_t)^{\gamma}$ 为简单困难样本调节因子，γ 是聚焦参数。当某一样本预测正确概率较高时，样本调节因子越低。即 p_t 接近于 0 时，$(1-p_t)^{\gamma}$ 的值接近于 1；本文设置 γ=2，α_t=0.25。

3.3 实验结果与分析

3.3.1 实验环境及参数配置

本章实验基于深度学习框架 PyTorch 实现，在 NVIDIA RTX 1080Ti GPU 上运行，CPU 为 Gold 5218，内存为 32G，显存为 11G，操作系统使用 Ubuntu，环境平台的详细信息如表 3-1 所示。

表 3-1 环境平台的详细信息

Table 3-1 Environment platform details

参数	配置
CPU	Intel(R) Xeon(R) Gold 5218 CPU
GPU	NVIDIA RTX1080Ti
内存	32G
显存	11GB
Python 版本	3.7
cuDNN 版本	8.3.0
CUDA 版本	10.2.89
操作系统	Ubuntu 18.04
深度学习框架	PyTorch 1.6.0

本文方法针对道路场景下的语义分割，所以选用了两个道路场景数据集 Cityscapes 和 CamVid。针对不同的数据集，设置了不同的训练超参数，超参数设置如表 3-2 所示。

表 3-2 超参数设置

Table 3-2 Hyper parameter settings

参数	Cityscapes	CamVid
损失函数	优化交叉熵损失函数	
优化器动量	0.9	
优化器	ADAM	
学习率调整	“poly”	
初始学习率	0.001	
权重衰减	1e-5	5e-5
批量大小	6	10
迭代次数	300K	20K

在训练中损失函数使用优化交叉熵损失函数，优化器使用自适应优化算法 ADAM，该优化器在语义分割领域的多个网络中表现良好。优化器动量选择 0.9，学习率的选择对网络训练中至关重要，并且学习率衰减策略可以在训练初期加速收敛速度，在训练末期减少震荡。在实验中，采用了"Poly"学习率策略，其数学表示为公式 3-5。

$$l_c = l_i \times (1 - \frac{i_c}{i_{\max}})^{pow} \tag{3-5}$$

其中，l_i 为初始学习率，l_c 为当前学习率，$i_{\max}$ 为最大迭代次数，i_c 为当前迭代次数，pow 设置为 0.9。

在训练中，本文对训练集进行了数据增强、随机水平翻转、随机剪裁和随机缩放，这些操作可以扩充数据集，并且缓解了数据集中的类别不平衡问题。其中随机缩放的比例为(0.75, 1, 1.25, 1.5, 1.75, 2)。在 Cityscapes 数据集的训练中，将输入图片随机裁剪为 512×1024 像素。在 CamVid 数据集中则随机裁剪为 360×480 像素。

3.3.2 实验结果分析

本章进行了一系列实验，旨在验证 MobileNetv2 和 LPPM 对网络轻量化的有效性，以及注意力机制的作用。这些实验在 Cityscapes 数据集上进行训练测试，测试的图片分辨率为剪裁后的512×1024像素。消融实验结果如表 3-3 所示，评价指标包括 mIoU、网络的运行速度和网络的参数数量。

表 3-3 不同主干网络在 Cityscapes 数据集上的性能实验

Table 3-3 Performance experiments with different backbone networks on Cityscapes dataset

网络模型	MIoU（%）	FPS	参数量(M)
DeepLabV3+（ResNet101）	**78.82**	10.81	55.62
DeepLabV3+（Xception）	77.65	18.34	46.71
DeepLabV3+（MobileNetv2）	71.26	**42.67**	**4.58**

根据表 3-3 可知，本文选用 ResNet101、Xception 和 MobileNetv2 作为 DeepLabV3+网络的特征提取网络，网络模型所对应的 MIoU 分别是 78.82%，77.65%和 71.26%，所对应的 FPS 分别是 10.81，18.34 和 42.67，所对应的参数量分别是 55.62M，46.71M 和 4.58M。表格直观地展示了，采用 MobileNetv2 作为特征提取网络时，虽然 MIoU 略低于其他两种网络模型，但是 FPS 达到了 42.67，比起 ResNet101 和 Xception 分别高出 31.86 和 24.33，其参数量也明显降低，大约为 ResNet101 和 Xception 参数量的十分之一，指标明显领先于其他三种特征提取网络，实现了实时图像语义分割。

表 3-4 展示了使用不同特征提取网络下的 DeepLabv3+语义分割模型在 Cityscapes 数据集上的类别分割精度。其中 19 个类别的的分割精度在不同特征提取网络下表现各异。当使用 MobileNetv2 作为 DeepLabv3+分割网络的特征提取网络时，各类别的分割精度都略低于其他特征提取网络。在 MobileNetv2 作为特征提取网络时，火车相较于其他类别的分割精度较低，马路、天空和植被相较于其他类别的分割精度较高，在同数据集中，不同类别样本的学习难易程度不同。通过综合观察表 3-3 和表 3-4 的实验数据，选用 MobileNetv2 作为特征提取网络可以达到网络轻量化的目的，保证实时性道路场景语义分割的实现。

表 3-4 在 Cityscapes 数据集上类别分割精度

Table 3-4 Category segmentation accuracy on Cityscapes dataset

类别\模型	ResNet101	Xception	MobileNetv2
人	0.7552	0.6491	0.6081
自行车	0.6324	0.6273	0.6145
公交车	0.7035	0.6883	0.6097
马路	0.9247	0.9287	0.9301
建筑物	0.8452	0.8437	0.8023
人行道	0.7231	0.7326	0.6631
天空	0.8703	0.6845	0.8192
火车	0.5829	0.5693	0.4776
植被	0.8457	0.8573	0.8101
墙壁	0.7365	0.7365	0.6137
栅栏	0.8299	0.8325	0.7642
电杆	0.8163	0.8173	0.7896
交通灯	0.7974	0.8061	0.7362
交通标志符	0.8445	0.8388	0.7693
草地	0.8428	0.8467	0.8052
骑行者	0.7253	0.7339	0.6719
摩托车	0.7869	0.7944	0.7157
卡车	0.7163	0.7134	0.6923

为验证本文构建的 LPPM 模块的有效性，将 LPPM 模块与 ASPP 模块的性能作比较。表 3-5 展示了在以 MobileNetv2 为特征提取网络的 DeepLabv3+上，使用不同的特征融合模块 ASPP 和 LPPM 的分割精度。

表 3-5 ASPP 优化前后模型性能对比表

Table 3-5 Model performance ASPP optimization

算法模型	参数量（M）	MIoU(%)	FPS
MobileNetv2+ASPP	4.58	**71.26**	42.67
MobileNetv2+LPPM	**2.87**	70.96	**49.69**

由表3-5可知，本文提出的LPPM在模型参数量和FPS两个指标上优于ASPP，模型参数量降低了1.71，FPS达到了49.69，但是推理精度相比降低了0.3%。说明LPPM实现了轻量化，节省资源降低功耗，提高推理速度。

并且评估了LPPM模块对参数a的敏感性，从表3-6可以看出，参数a的增加会带来运行速度的加快，但是精确度会降低。综合评估后，本文设置参数a=4。表3-6为LPPM 模块的参数a的研究结果。

表3-6 参数a对LPPM 模块的性能影响

Table 3-6 Effect of parameter a on the performance of the LPPM module

参数a	MIoU(%)	Params(M)	FPS
1	**71.74**	4.16	48.25
2	71.28	4.07	49.38
4	70.96	2.87	49.69
6	70.62	**2.81**	**49.91**

消融实验原始模型是以Resnet101为主干网络的DeepLabV3+；方案一将DeepLabV3+的特征提取网络替换为轻量级网络MobileNetv2；方案二是在DeepLabV3+主干网络为MobileNetv2的基础上，在特征提取网络和特征融合模块后同时加入坐标注意力模块；方案三是在方案二的基础上，将ASPP模块替换为LPPM模块；方案四是在方案三的基础上引入优化焦点交叉熵。实验数据如表3-7所示。

表3-7 消融实验

Table 3-7 Ablation experiments

方案	PA（%）	MIoU（%）	参数量（M）	FPS	FLOPs（G）
原始模型	**93.4**	**78.8**	45.62	10.81	126.13
方案一	91.8	71.2	4.58	42.67	9.57
方案二	92.5	71.9	5.87	39.89	6.71
方案三	92.9	71.6	**4.16**	**46.91**	13.47
方案四	93.1	72.2	4.65	46.66	11.82

由表3-7消融实验结果可知，方案一模型的参数量为4.58M，FLOPS为26422.84M，相比原始模型明显减小，说明将原模型的主干网络Xception替换为MobileNetv2有效的降低了模型的计算量，明显提高了模型计算速度，但PA和MIoU均有不同程度的下降，轻量化的网络替换会导致分割精度的降低。方案二

的 PA 达到 92.5%，平均交互比达到 71.9%，均比方案一有不同程度的提升，坐标注意力的引入充分利用了像素间的位置关系，提高了语义分割精度。方案三的参数量为 4.16M，FPS 为 46.91，FLOPs 为 11.47G，相比方案二，计算速度有一定的提升，说明 LPPM 有效减少网络参数量，可以有效的提高模型推理速度。方案四像素准确率达到 93.1%，MIoU 为 72.2%，对比方案三表明了优化焦点损失函数加速网络模型收敛，有效的解决了网络训练中简单困难样本不平衡问题。

通过对模型参数量、FPS 和 FLOPs 进行分析，发现这四种方案的网络模型大小，训练时间都明显小于原始的 DeepLabv3+网络。实验结果表明，本文提出的轻量化网络替换特征提取网络和 LPPM 模块对计算速度的提升付出了贡献，优化交叉熵损失和注意力机制的引入提升了分割精度。

为了进一步验证 L2RS-Net 的分割性能，在 Cityscapes 和 Camvid 数据集上进行训练测试，将本文模型与 DeepLabv3+、ICNet、SegNet、SwiftNet 四种模型分割性能进行比较，测试结果如表 3-8 和表 3-9 所示，并从分割精度和计算速度角度分析五种网络模型。

表 3-8 主流网络在 Cityscapes 数据集上的性能对比

Table 3-8 Performance comparison of mainstream networks on Cityscapes dataset

方法	MIoU(%)	参数量(M)	FLOPs（G）	FPS
DeepLabv3+	**78.8**	55.62	126.13	10.81
ICNet	69.1	26.58	32.56	30.3
SegNet	57.0	43.84	95.72	16.7
SwiftNet	74.3	31.19	29.64	39.9
本章方法	72.2	**4.65**	**11.82**	**46.66**

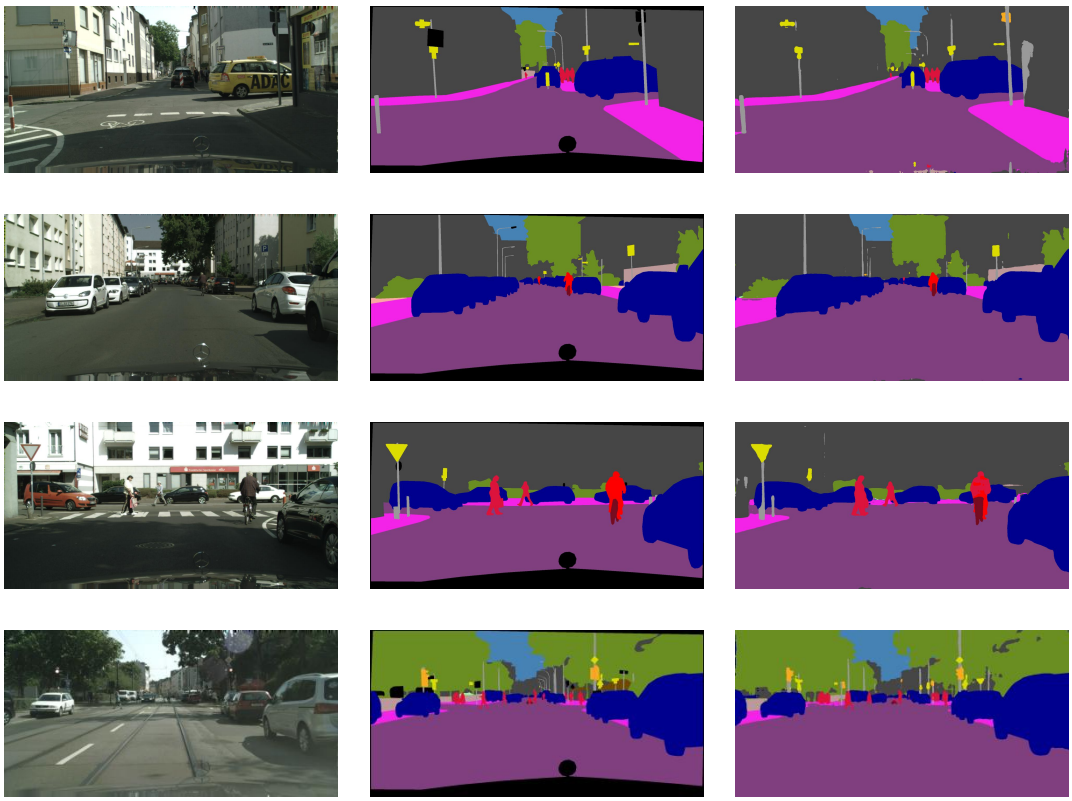
表 3-9 主流网络在 Camvid 数据集上的性能对比

Table 3-9 Performance comparison of mainstream networks on Camvid dataset

方法	MIoU(%)	参数量(M)	FLOPs（G）	FPS
DeepLabv3+	**68.5**	55.62	126.13	7.35
ICNet	51.4	26.53	32.56	13.94
SegNet	48.7	43.84	95.72	9.42
SwiftNet	63.5	31.19	29.64	19.86
本章方法	53.8	4.65	11.82	24.58

表格中呈现了原始的 DeepLabv3+、ICNet、SegNet、SwiftNet 和本文方法在性能对比实验中的结果。对模型大小的详细分析显示，本章提出的基于轻量化网络的语义分割模型相较与大多数的轻量级语义分割算法具有更快的计算速度。在 Cityscapes 数据集上，本文提出的模型不仅精度比 ICNet 高出 3.1%。且推理速度接近 ICNet 的 1.5 倍。在 Camvid 数据集上，本章提出的网络模型虽然在精度上有一定的损失，但是计算速度超过了其他的轻量化网络分割速度。实验结果验证本章提出的轻量化模型达到了实时分割，并且与其他轻量级语义分割相比计算速度也取得了较高的结果。

3.3.3 可视化结果分析



(a) 输入图像　　(b)标签图像　　(c)本节方法

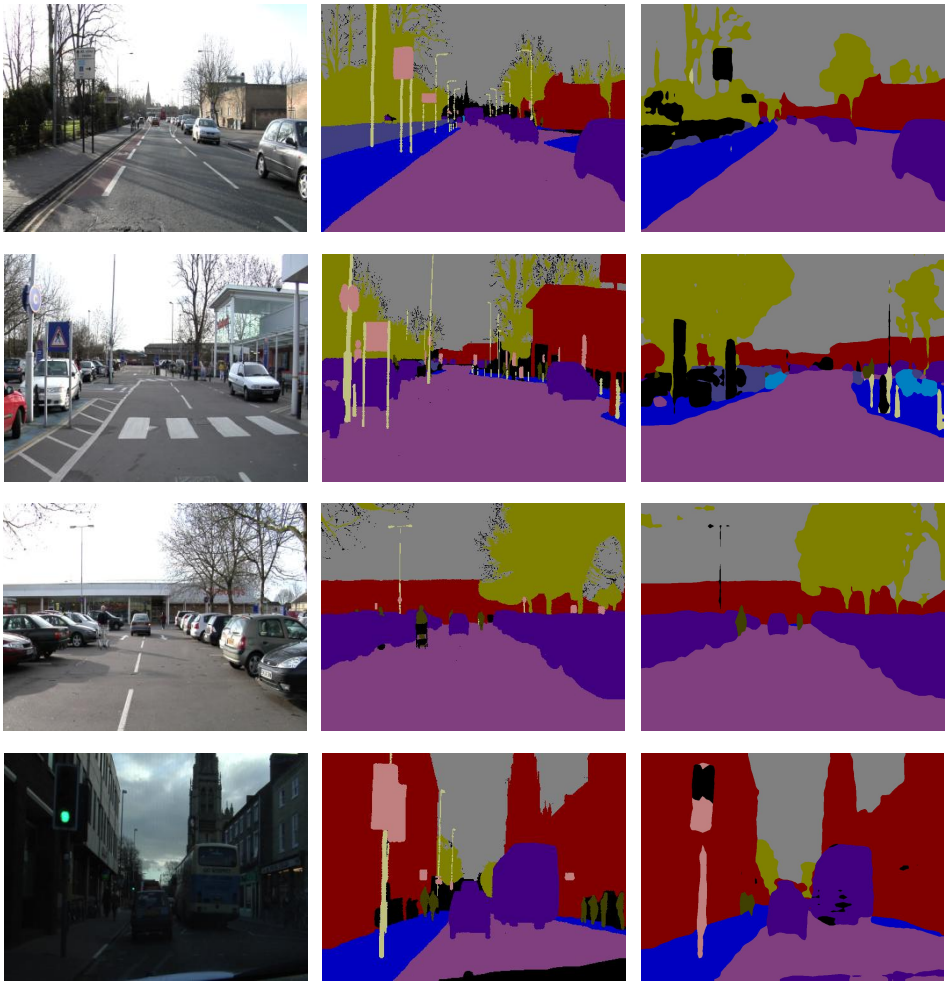
图 3-6 在 Cityscapes 数据集上的分割结果

Figure 3-6 Segmentation effect on Cityscapes dataset

图 3-6 展示了 L2RS-Net 在 Cityscapes 数据集上分割结果，其中第一列是输入图像，第二列是真实标签，第三列展示的是本节方法产生的分割结果图。第一行中的汽车轮廓明显以及面积较大，本模型精准的分割出了汽车类别，但是右边墙壁上有类似柱体的边缘，模型将墙壁边缘误判为柱子。第二行中的道路和植被

的边缘都被精准的分割，但远处的柱子边缘断续不连贯，远处的骑行者过小与汽车重叠，导致分割不完整，但是也识别出了骑行者的类别。第三行中准确预测了汽车与骑行者的轮廓，但是柱子和交通指示牌处于强光下导致对比度高，无法分割出原始形状。第四行中的植物边缘分割较为模糊，行人在暗处密集，导致行人中间空隙也被识别为行人，由于图像画面复杂，植物和柱子重叠在一起容易被误判。

本文方法对于目标轮廓清晰，光线正常，图像简单的图片可以实现精确分割。如果兑现处在强光或者暗处，存在相似的重叠部分以及目标在远处较小时，精确度会有一定的下降，但是本方法能精确的识别出前景物体，分割出绝大部分类别且轮廓清晰。



(a) 输入图像　　(b)标签图像　　(c)本节方法

图 3-7 在 Camvid 数据集上的分割结果

Figure 3-7 Segmentation effect on camvid dataset

图 3-7 展示了 L2RS-Net 在 Cityscapes 数据集上分割结果，其中第一列是输

入图像，第二列是真实标签，第三列展示的是本节方法产生的分割结果图。第一行中的道路和汽车得到了相对精确的分割，但是交通标志并未被识别到，树由于间隙比较大，分割边缘并不准确。第二行中精确的识别了建筑类别，但是对于重叠性较高的建筑和车辆，以及交通标志都未被识别出。第三行中由于汽车轮廓清晰光线良好，所以精确的分割出了汽车，但是右边的树枝在天空中可见度不高，所以被误判为天空。第四行中的交通指示拍颜色较黑，与建筑的颜色重叠，但是本方法有效识别出了交通指示牌的类别，但是分割的不够完整。本方法在能够实现较好的分割效果，证明了本方法良好的特征提取能力与模型分割能力。

3.4 基于嵌入式 TX2 的道路场景实时语义分割实现

本文采用的嵌入式平台验证提出的轻量化网络，在算力有限的设备上实现实时语义分割，利用 TensorRT 加速推理，在不降低分割精度的同时提高模型计算速度。

3.4.1 环境安装与配置

表 3-10 Jeston TX2 硬件参数
Table3-10 Jeston TX2 hardware parameter

GPU	Pascal™ 架构，含有 256 个 NVIDIA CUDA
CPU	双核 NVIDIA Denver，四核 ARM Cortex-A57
内存	8GB 128bit LPDDR4
储存空间	32GB eMMC5.1
网络连接	板载 Wifi 和千兆以太网
摄像头	12 通道 MIPI CSI-2 D-PHY 1.2(30 Gbps)
接口	HDMI 2.0/eDP1.4/2x DSI/2x DP1.2
尺寸	87mm×50mm
规格尺寸	配有热转印板（TTP）的 400 针连接器

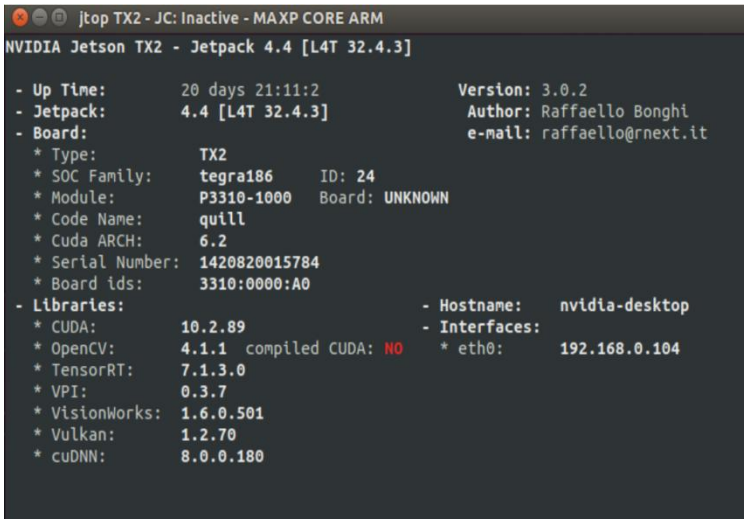


图 3-8 Jeston TX2 软件参数

Figure 3-8 Jeston TX2 software parameters

TX2 采用 JetPack（Jetson Development Pack）安装软件环境。JetPack 简化了软件之间的兼容性问题，操作简单高效。首先在 PC 端安装 Ubuntu 系统，然后在 PC 端执行 Jetpack 安装脚本，选择需要安装的组件。需要一个交换器或者路由器，确保 TX2 与主机通过网线连接在同一交换机上，使其在同一网段中。

安装完组件后，使用 HDMI 线连接到显示器上，使用 Micro USB 线将 TX2 与 PC 相连接，插上电源适配器，按下 Power 开机，然后按住 Recovery 键后点击 Reset 键，松开 Recovery 键后在主机终端输入 lsusb 命令，当显示 Nvidia Corporation 时则表示 TX2 与 PC 连接成功，此时 TX2 进入刷机模式，等待一段时间以完成安装。

3.4.2 模型构建与部署

图像语义分割模型在 TX2 上的构建部署流程如下图 3-9 所示。模型构建阶段如图 3-9 左边所示。首先建立一个 builder 文件，在文件中添加 network 对象。然后构造解析过程解析网络模型，并在 network 对象中标记网络模型输出层，接着设置 batch size 和工作空间。最后优化网络模型，在 builder 中建立 cudaEngine，合并网络层数，低精度校准数据。模型构建完成后，进入模型部署阶段，其流程图如图 3-9 右边所示。首先建立一个 context 文件，启动 CUDA 核心并获取输入输出 TensorRT 索引。然后分配 GPU 显存，将数据上传到 GPU 中。最后，将 GPU 显存中计算得到的结果下载到 CPU 中。这样的部署流程确保了模型在 TX2 上的

高效运行和准确预测。通过 TensorRT 功能，可以充分利用 TX2 的 GPU 资源，实现快速而精准的推理过程。

图 3-9 分割模型构建部署流程图

Figure 3-9 Segmentation model construction and deployment flowchart

3.4.2 实验结果和分析

本章实验采用的是嵌入式设备 Jetson TX2，其中使用的 TensorRT 版本为 7.13，操作系统为 Ubuntu 18.04，以 Python 为编程语言。实验采用道路场景数据集 Cityspaces，图像分辨率设定为 512×1024。在 Jetson TX2 上实验时，采用 Max-N 模式，即最大功率模式。TX2 配置了 8GB 的虚拟内存以确保内存充足。TX2 的算力大约为 RTX 1080Ti 的 10 倍，实验结果如表 3-11 所示。

表 3-11 Jeston TX2 中实验结果

Table 3-11 Experimental results in Jeston TX2

数据类型	Tensor	FPS	耗时/ms	MIoU
FP(32)	×	6.5	153.8	72.2
FP(32)	√	9.8	102.1	72.2
FP(16)	√	12.9	77.5	72.1

从表 3-1 中可以看出，经过 TensorRT 加速后的轻量化道路场景实时语义分割方法计算速度得到了提升，相较于非 TensorRT 加速单精度（FP32），TensorRT 加速后半精度（FP16）的 FPS 提升了 6.4，并且分割性能基本不变。通过在 Jetson TX2 上进行实验，能够验证所提出方法的有效性，并评估其在嵌入式环境下的性能表现。这些实验结果为进一步的研究提供了有力支持，也为未来在实际应用中的部署提供了重要参考。

3.5 本章小结

本章详细论述了一种基于轻量化网络的道路场景实时语义分割网络 L2RS-Net。首先，将 DeepLabv3+的特征提取网络替换为 MobileNetv2，同时轻量化 MobileNetv2，用以减少网络的参数量达到实时效果。然后构建了一种轻量化金字塔池化模型 LPPM，金字塔分支特征图非固定通道数量，以减少模型计算量，并且使用非对称卷积减少计算负担。其次，在网络中加入坐标注意力机制、优化焦点交叉熵和编码器-解码器多次拼接，以弥补轻量化网络带来的精度损失。然后在 Cityscapes 数据集上的消融实验验证了 MobileNet 和 LPPM 的有效性。最后在嵌入式 Jeston TX2 上对该模型开展模型转换与模型部署。

本章提出的轻量化道路语义分割网络计算速度大幅提升，实现了实时语义分割。本章在 DeepLabv3+网络的基础上，在网络中添加不同的模块，进行组合，对比试验，实验结果表明，模块结合后的网络层数较浅，参数量较少，计算速度较快；将 DeepLabv3+、ICNet、SegNet、SwiftNet 和本文方法进行对比试验，并从分割精度和计算速度的角度分析四种网络模型，综合上述性能，本文使用的语义分割网络模型在完成交通场景图像分割任务时表现较好。

在嵌入式平台 Jeston TX2 开展轻量化道路场景语义分割试验，使用 FP32 和 FP16 模型参数进行对比试验。试验结果表明使用 FP16 进行 TensorRT 加速推理后的模型，几乎在不损失分割精度的前提下提高计算速度，其单张图片分割耗时为 77.5ms，取得了较为理想的分割效果。

第 4 章　基于知识蒸馏的道路场景实时语义分割优化方法研究

4.1 引言

在第三章中详细介绍了基于轻量化网络的道路场景实时语义分割方法，通过在 Cityscapes 和 Camvid 数据集上验证 L2RS-Net 的网络性能与计算速度，展示了其在分割速度方面表现出色，但分割精度仍有待提升。针对这一问题，本文对第三章的网络进行了系统的优化。在保证不影响分割速度的前提下，通过对网络结构进行调整，针对性地提升了网络的推理精度以提高其在道路场景语义分割任务中的性能表现。

在道路场景语义分割领域，通常需要大型数据集训练语义分割模型，以此实现高精度语义分割。但是数据量大、分布均匀的优良数据集相对稀缺，这是导致目前图像分割精度难以进一步提升的问题，故制作质量高的数据集是图像语义分割精度的重中之重。然而，这一过程却充满了挑战与困难。制作高质量数据集的复杂性主要体现在大量的人力和时间资源投入，以及确保标记和注释的准确性。针对轻量级网络图像语义分割计算速度与分割精度难以平衡的问题，当降低网络模型参数量和复杂度的同时，模型的分割能力也同时降低。如何提高轻量级网络模型的精度也是实时语义分割任务中的一大难题。在面对制作质量高的数据集困难的挑战和语义分割精度和计算速度无法平衡的问题下，采用知识蒸馏技术显然是一种创新而有效的解决方案。在知识蒸馏的框架下，首先通过训练一个大型复杂的教师模型，该模型在数据集上能够表现出色。然后，通过一个小型简单的学生模型学习教师模型的知识，实现高阶知识迁移。这种方式不仅有效地解决了数据集规模有限的问题，还在不损失推理速度的情况下提升了推理精度。

所以本节在第三章提出的 L2RS-Net 网络基础上，结合知识蒸馏方法，该知识蒸馏融合对抗性网络，使用结构化提炼方案：像素级蒸馏，关注像素点之间的类别差异；成对蒸馏，可以传输局部像素块以外所有像素对之间的关系；整体蒸馏，使用 WGAN 网络传输捕捉高阶知识的整体关系。具体来说，通过将

DeepLabV3+网络模型作为教师网络训练学生网络 L2RS-Net 网络，提升模型实际应用能力。

4.2 基于知识蒸馏的道路场景实时语义分割优化

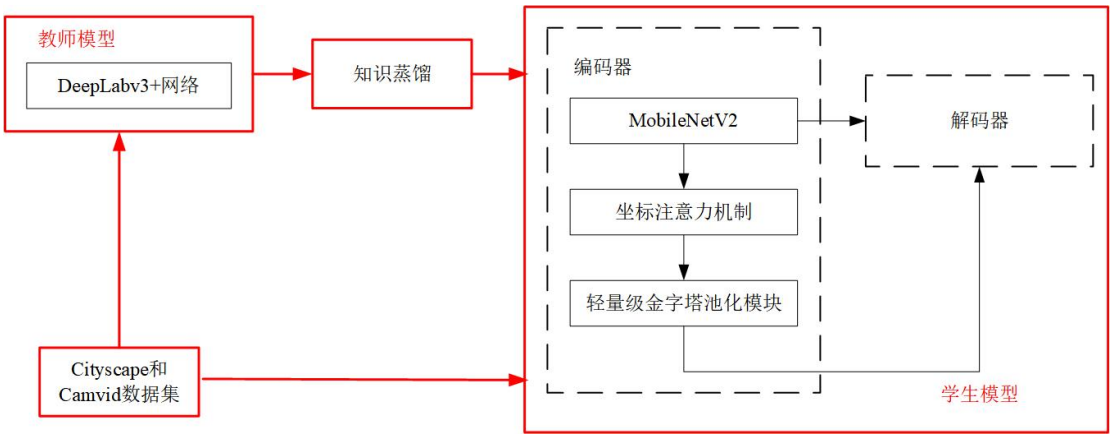


图 4-1 优化算法整体结构

Figure 4-1 Overall structure of the optimization algorithm

在第三章的基础上，本节设计了基于知识蒸馏的道路场景实时语义分割优化算法整体结构如图 4-1 所示。DeepLabv3+网络作为教师模型经过充分的预训练，在吸收大量数据的基础上积累了丰富的知识和经验，将模型在处理不同场景、对象和特征时所学到的高层抽象表示的知识，在结构化知识蒸馏的框架下，将这些知识迁移至学生网络 L2RS-Net 中。

4.2.1 整体蒸馏框架

本节提出的知识蒸馏策略，将结构化知识从繁琐网络转移到紧凑网络。图 4-2 展示了这一流程。由图可知，图像分别输入教师网络和学生网络生成特征图，在特征图之间进行成对蒸馏，可以提高模型的稳定性，并且帮助学生模型避免过拟合和不稳定的训练过程；紧接着进行像素分类，在得分图之间进行像素级蒸馏，学生模型可以学习到更细粒度的特征表示，泛化能力提升，增加模型鲁棒性；再将学生网络的输出作为 WGAN-GP 对抗性网络中的生成器，分别得到生成器损失和鉴别器损失，而生成器损失即为学生网络的整体蒸馏损失，生成样本质量能够得到提升；输入图像预测概率图与实际标签之间产生交叉熵损失，该损失不仅能有效避免梯度问题，还能准确量化概率分布差异，进一步提高分割准确性。

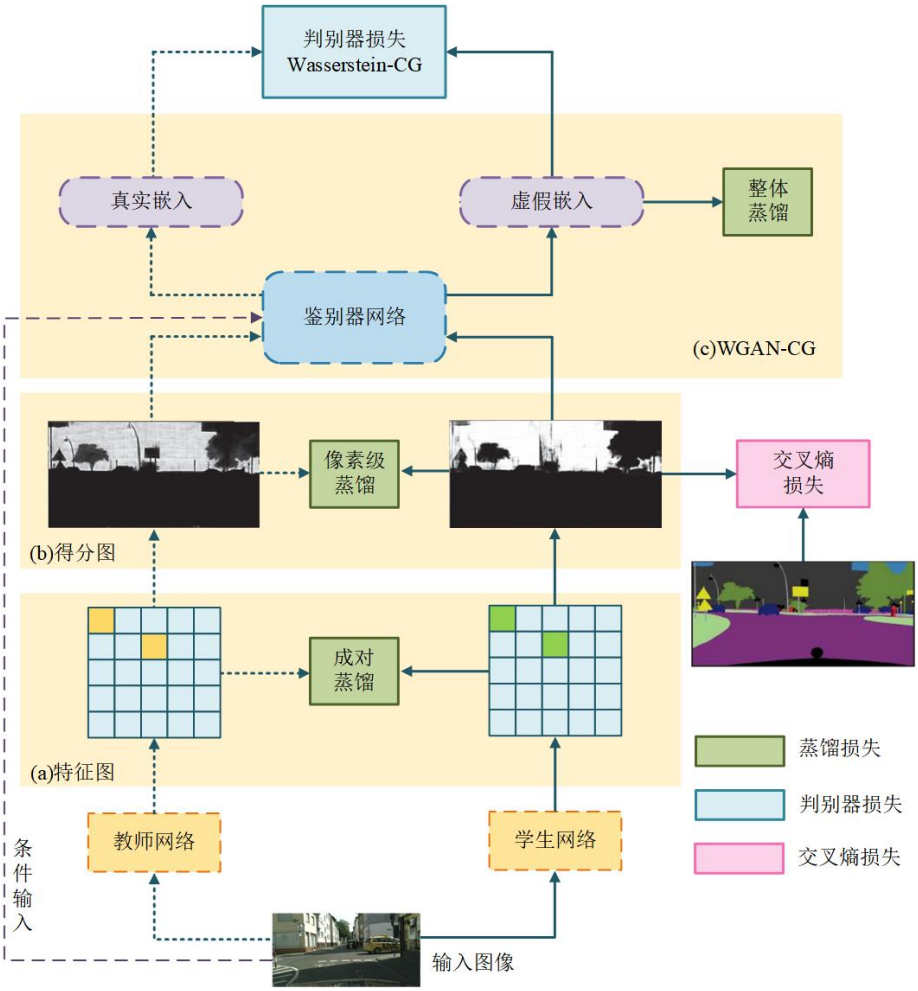


图 4-2 结构化知识蒸馏

Figure 4-2 Structured knowledge distillation

4.2.2 最优化目标函数

（1）像素级蒸馏（Pixel-wise Distillation，PI）

将学生模型和老师模型的每个像素点进行点对点知识蒸馏，老师模型所产生的每个像素点的类概率为软标签来训练学生模型，所产生的 Loss 是：

$$\ell_{pi}(S)=\frac{1}{W'\times H'}\sum_{i\in R}KL(q_i^s \| q_i^t) \tag{4-1}$$

其中，q_i^s 表示学生网络 S 的第 i 个像素点的类概率，q_i^t 为教师网络 T 的第 i 个像素点的类概率，$KL(\cdot)$ 是代表这个类概率之间的 Kullback-Leibler 散度，其中 $R=\{1,2,...,W'\times H'\}$ 代表所有像素点。

（2）成对蒸馏（Pair-wise Distillation，PA）

成对蒸馏是描述教师网络的不同像素点之间像素对与学生模型之间像素对的相似度，采用平方差来描述像素对蒸馏损失，所产生的 loss 为：

$$\ell_{pa}(S)=\frac{1}{(W'\times H')^2}\sum_{i\in R}\sum_{j\in R}(a_{ij}^{s}-a_{ij}^{t})^2 \tag{4-2}$$

其中 a_{ij}^{t} 表示学生网络中第 i 个像素和第 j 个像素之间的相似度，a_{ij}^{t} 是指老师网络中第 i 个像素和第 j 个像素之间的相似度。

（3）整体蒸馏（Holistic Distillation，HO）

本节使用条件生成对抗性网络来表述整体蒸馏问题，将学生网络的预测概率图作为生成器，并将预测概率图视为虚假样本，教师网络的概率预测图被视为真实样本，期望假样本尽可能与真样本相似。采用 Wasserstein 距离和梯度约束（gradient penalty）来评估真假本和假样本之间的差异，产生的 loss 为：

$$\begin{aligned}\ell_{ho}(S,D)&=E_{Q^s\sim p_s(Q^s)}[D(Q^s\mid I)]\\&-E_{Q^t\sim p_t(Q^t)}[D(Q^t\mid I)]\\&+\lambda\underset{\hat{Q}\sim P\hat{Q}}{E}[(\|\nabla_{\hat{Q}}D(\hat{Q})\|_2-1)^2]\end{aligned} \tag{4-3}$$

$$\hat{Q}=\varepsilon Q^s+(1-\varepsilon)Q^t,\varepsilon\sim U(0,1) \tag{4-4}$$

其中，$E[\cdot]$为期望算子，$D[\cdot]$为嵌入网络，在 WGAN 中做鉴别器，将 Q 和 I 一起投射到整体嵌入得分中。D 是一个具有五个卷积的全卷积神经网络。在最后三层之间插入了两个自监督注意力模块，以捕捉结构信息。这样的判别器能够产生一个整体嵌入，代表输入图像和分割图的匹配程度。$\hat{Q}$ 取自某一个训练小批次（mini-batch）中的全部 Q^s 和 Q^t，以及他们的随机“混合”部分。目标函数包括多类交叉熵损失 $\ell_{mc}(S)$，像素级蒸馏和结构化蒸馏：

$$\ell（S,D）=\ell_{mc}(S)+\lambda_1(\ell_{pi}(S)+\ell_{pa}(S))-\lambda_2\ell_{ho}(S,D) \tag{4-5}$$

其中 λ_1 和 λ_2 设为 10 和 0.1 使这些损失值范围具有可比性，根据学生网络 S 的参数使目标函数最小化，同时根据判别器 D 的参数使目标函数最大化：训练判别器 D 等同于最小化 $\ell_{ho}(S,D)$，D 的目标是为来自教师网络的真实样本提供高嵌入分值，为假样本提供低嵌入分值；训练学生网络 S 是最小化多类交叉熵损

失和蒸馏损失。

$$l_{mc}(S)+\lambda_1(\ell_{pi}(S)+\ell_{pa}(S))-\lambda_2\ell^{s}_{ho}(S) \tag{4-6}$$

$$\ell^{s}_{ho}(S)=E_{Q^s\sim p_s(Q^s)}[D(Q^s\mid I)] \tag{4-7}$$

4.3 实验结果与分析

4.3.1 实验环境及参数配置

本章实验基于第三章的实验环境与参数配置，在 Cityscapes 和 Camvid 数据集上进行实验，验证基于知识蒸馏的道路场景实时语义分割优化方法的有效性。

4.3.2 实验结果分析

在本章进行了一系列实验，旨在验证本节提出结构化知识蒸馏优化方法在不影响推理速度的前提下提高了精确度。实验在 Cityscapes 和 Camvid 训练集上进行，在 Cityscapes 验证集测试的图片分辨率为剪裁后的512×1024像素，消融实验结果如表4-1所示。在Camvid验证集测试的图片分辨率为剪裁后的360×480像素。评价指标包括 MIoU、网络的运行速度和网络的参数数量。其中教师网络使用的 DeepLabv3+，学生网络使用的是本文第三章提出的 L2RS-Net，知识蒸馏方法使用三种蒸馏项，分别是像素级蒸馏，成对蒸馏，和整体蒸馏。

表 4-1 Cityscapes 数据集上的消融实验

Table 4-1 Ablation experiments on Cityscapes dataset

方法	MIoU（%）
教师网络(DeepLabv3+)	78.8
学生网络(L2RS-Net)	72.2
+PI	72.5
+PI+PA	73.7
+PI+PA+HO	74.3

由表可知，知识蒸馏可以有效提高学生网络的性能，有助于学生网络更好的学习。学生网络加上所有的知识蒸馏项得到了最好的分割结果，MIoU 达到了

74.3%。可以观察到每个蒸馏方案都能获得更高的 MIoU 分数，这意味着这三种知识蒸馏相互互补，有助于训练学生网络。但是再加入成对蒸馏项后，MIoU 得到了显著提高，提高了 1.2%,，是三个蒸馏项中对分割精度贡献最大的一项。其中，PI 为像素级蒸馏，PA 为成对蒸馏，HO 为整体蒸馏。

表 4-2 和表 4-3 为 L2RS-Net 通过结构化知识蒸馏后在数据集 Cityscapes 和 Camvid 上的实验结果。结果表明经过优化后的网络模型参数量不变，FLOPs 也不变，但是平均交并比比原本模型 L2RS-Net 分别提升了 2.1%和 3.4%，这表示本节提出的基于知识蒸馏的道路实时语义分割优化效果显著，在计算能力有限的情况下，可以不增加计算负担的同时增强分割精度。

表 4-2 优化后网络在 Cityscapes 数据集上性能实验

Table 4-2 Performance experiment of the optimized network on the Cityscapes dataset

方法	参数量（M）	FLOPs（G）	MIoU（%）
L2RS-Net	4.65	11.82	72.2
L2RS-Net+Distillation	4.65	11.82	74.3

表 4-3 优化后网络在 Camvid 数据集上性能实验

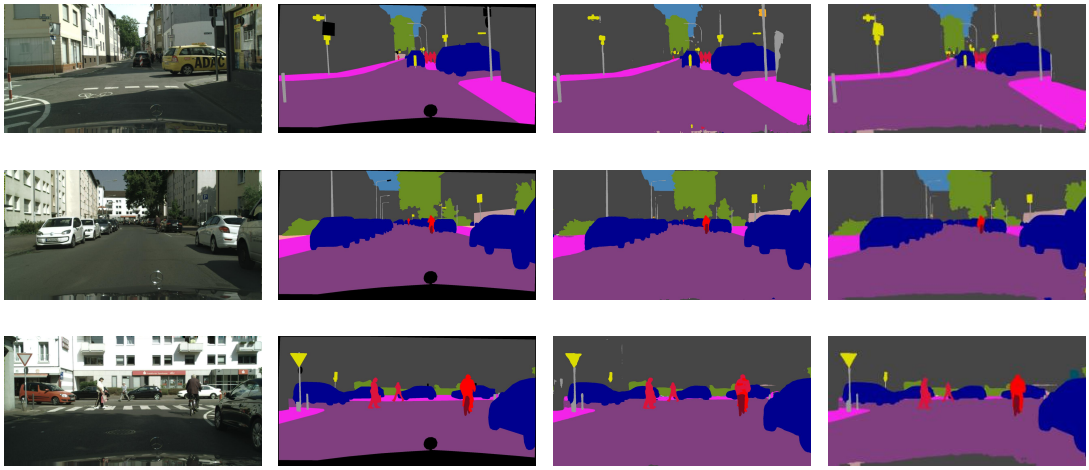
Table 4-3 Performance experiment of optimized network on Camvid dataset

方法	参数量（M）	FLOPs（G）	MIoU（%）
L2RS-Net	4.65	11.82	53.8
L2RS-Net+Distillation	4.65	11.82	57.2

4.3.3 可视化结果分析

L2RS-Net 和优化后的网络在 Cityscapes 和 Camvid 数据集上的分割图如图 4-3 和图 4-4 所示。

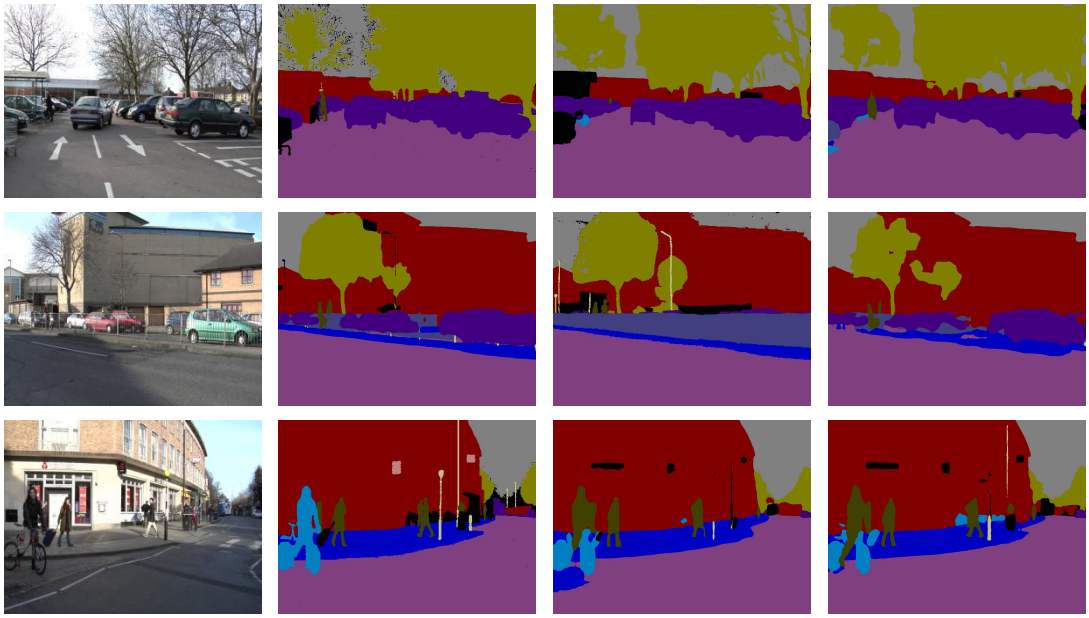
图 4-3 是轻量级网络及使用知识蒸馏后在 Cityscapes 上的分割图。其中图 4-3（a）是输入图像，图 4-3（b）是真实标签，图 4-3（c）是 L2RS-Net 的分割图，图 4-3（d）是使用知识蒸馏后的 L2RS-Net 的分割图。第一行明显可以看出使用知识蒸馏优化后的网络分割精度提升显著，处于图像右边原本误判为柱子的区域被识别出是建筑。第二行中优化后网络的分割结果中柱子的边缘更加清晰连贯。第三行中优化后的网络可以识别到车头的部分非道路，左边原本被识别成柱子的建筑也被识别成正确的类别。



（a）输入图像　　(b)真实标签　　(c)L2RS-Net　　(d)知识蒸馏

图 4-3 L2RS-Net 以及优化后网络在 Cityscapes 数据集上的分割图

Figure 4-3 L2RS-Net and segmentation of optimized network on Cityscapes dataset



(a)输入图像　　(b)真实标签　　(c)L2RS-Net　　(d)知识蒸馏

图 4-4 L2RS-Net 以及优化后网络在 Camvid 数据集上的分割图

Figure 4-4 L2RS-Net and the segmentation of the optimized network on Camvid dataset

图 4-4 是轻量级网络及使用知识蒸馏后在 Camvid 上的分割图。其中图 4-4（a）是输入图像，图 4-4（b）是真实标签，图 4-4（c）是 L2RS-Net 的分割图，图 4-4（d）是使用知识蒸馏后的 L2RS-Net 的分割图。第一行中优化后的网络成功的识别出左侧车辆，以及分割出较完整的树的部分；第二行中优化后的网络可以分割出汽车轮廓，以及识别出误判的建筑区域；第三行中优化后的网络对行人分割更加完整，柱子也被识别出来。

4.4 本章小结

本章详细论述了基于知识蒸馏的道路场景实时语义分割优化方法，提出了一种融合对抗性网络的结构化知识蒸馏。该方法包括像素级蒸馏，结构化蒸馏，以及整体蒸馏。

像素级蒸馏关注像素点与点之间的类别概率差异，提高学生模型精度节约计算成本；结构化知识蒸馏通过对齐像素对差异获取高阶知识进行迁移，提高泛化性能，提高模型对于训练数据的鲁棒性；整体蒸馏是生成器的 Loss，学生网络的输出作为对抗性网络的生成器，教师网络的输出作为条件进行判别，在对抗性网络中使用 Wasserstein 距离和梯度约束作为损失函数，整体蒸馏可以有效的全局知识传递，关注任务中的复杂关系和模式。并且融合交叉熵损失，最优化目标函数。

本章对知识蒸馏中的每一个蒸馏项进行消融实验，证明每个蒸馏项都互补增益，对网络模型有着显著性能增强；验证整体蒸馏项对于整体网络的影响，证明 GAN 网络的有效性；对第三章中提出的 L2RS-Net 网络进行知识蒸馏优化，明显提高原轻量化网络的推理精度，并且对推理速度不产生影响。

第 5 章　总结与展望

5.1 工作总结

道路场景实时语义分割作为图像语义分割领域的一个重要研究方向，在现实的自动驾驶领域中发挥着巨大的应用价值。与关注精度的大型图像语义分割网络相比，道路场景实时语义分割需要解决由于轻量化网络层级浅，参数量低导致的推理精度低的问题。在真实的道路场景中，实现语义分割的实用性和扩展性成为研究的重中之重。本研究提出的一种基于轻量化网络的道路场景实时语义分割方法是为了实现实时性和高精度的平衡，但实时分割的同时必然带来了精度的损失。所以对提出的道路实时语义分割方法进行优化，对轻量化模型使用对抗性的结构化知识蒸馏进行优化，在不损失计算速度的前提下提升推理精度。为此，论文针对道路场景实时语义分割领域，主要围绕轻量化网络构建和知识蒸馏优化两方面展开研究，具体的研究成果如下：

（1）提出了基于轻量化网络的道路场景实时语义分割方法。采用大型高精度网络模型作为基础模型，替换其特征提取网络为轻量化网络，并对轻量化网络进行网络压缩，优化网络推理速度，减少网络参数量；构造了轻量化金字塔池化模块替换 ASPP 模块，使用非对称空洞卷积减少模型参数量，并且对分支通道使用压缩因子降低通道数，提升计算速度；在网络中融合注意力模块，弥补精度损失；优化交叉熵损失为焦点损失函数，以此解决样本不平衡问题；编码器-解码器之间特征图多次拼接，连接上下文信息，提高推理精度。

（2）提出了基于知识蒸馏的道路场景实时语义分割的优化方法。提出一种融合对抗性网络的结构化知识蒸馏，通过结构化知识蒸馏迁移高阶知识至学生网络，而且对抗性网络能够提炼整体知识，提高轻量化网络的精度并且不影响推理速度。而且目前数据量大、均匀分布及优良的数据集相对稀缺，构建优秀数据集需要耗费大量的人力物力，知识蒸馏可以显著的平衡该问题。

5.2 工作展望

本文提出了道路场景实时语义分割方法，虽然分割效果较好，但仍存在以下方面有待改进，在后续的研究中，主要从以下几个方面进行深入研究。

1.为了实现精度高的语义分割效果，通常需要庞大的数据集来训练语义分割模型，而在真实的交通场景中，道路场景有各种不同的情景出现，目前的道路场景数据集不能完全概括，故制作质量高的数据集有利于提升图像分割的精度。

2.在自动驾驶系统应用中，仅仅依靠优化图像语义分割算法是远远不够的，还需结合激光雷达点云目标检测等其他算法进行改进，并将这一复杂问题用在移动设备上达到实用价值的目的。

致谢

随着毕业论文的完成，我的硕士生涯也即将画上圆满的句号。站在这个人生的分岔路口，回首过去的岁月，心中充满了感激与不舍。在此，我要向所有在我求学路上给予我帮助、支持和关爱的师长、同学和亲友们致以最诚挚的谢意。

首先，我要衷心感谢我的导师杨会成教授。您的严谨治学、深厚学识和无私奉献的精神深深地影响了我。从选题到论文的撰写，您总是耐心地指导我，给予我许多宝贵的建议。在我遇到困难时，您总是鼓励我，让我有信心面对挑战。您的教诲和指导，不仅让我在学术上取得了进步，更让我在人生道路上受益匪浅。也要感谢胡老师对于我们三年的帮助，不仅仅是老师更像是朋友一般，还有徐老师，在我写论文的时候给予了很多帮助，很感谢所有的恩师。

其次，我要感谢我的家人。你们一直是我最坚实的后盾，你们的支持和鼓励是我前进的动力。在我求学的道路上，你们始终默默付出，为我提供物质和精神上的支持。你们的爱，让我感受到了无尽的温暖和力量。

同时，我也要感谢我的同学和朋友们。林园园和帅真同学，我们一起度过了无数个挑灯夜战的夜晚，一起分享过学术上的喜悦和挫折。你们的陪伴和鼓励，让我在求学路上不再孤单。我们共同度过的时光，将成为我人生中最宝贵的回忆。

最后，我要感谢所有在我人生道路上给予我帮助和支持的人。你们的善良和爱心让我感受到了人间的温暖和美好。我将永远铭记你们的恩情，并以此为动力，不断前进、追求卓越。

至陪伴了我七年的张广源，无需多言，会一直到老。

硕士阶段的学习经历是我人生中一段宝贵的时光。在这里，我不仅学到了知识，更学会了如何面对挑战、如何与人相处、如何成为一个更好的人。这些宝贵的经验和财富将伴随我走过未来的人生道路。

再次感谢所有在我求学路上给予我帮助和支持的人。愿你们在未来的日子里一切顺利、幸福安康！

参考文献

[1] GB/T 40429-2021，汽车驾驶自动化分级[J]．北京：中华人民共和国全国汽车标准化技术委员会，2022.

[2] Janai J, Guney F, Behl A, et al. Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art[J]. 2017:1-3.

[3] Zhou Q, Yu C, Wang Z, et al. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4081-4090.

[4] Araslanov N, Roth S. Self-supervised Augmentation Consistency for Adapting Semantic Segmentation[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15384-15394.

[5] Minaee S, Boykov Y, Porikli F, et al. Image segmentation using deep learning: A survey[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3523-3542.

[6] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[7] Li K, Tao W, Liu L. Online semantic object segmentation for vision robot collected video[J]. IEEE Access, 2019, 7: 107602-107615.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005 IEEE computer society conference on computer vision and pattern recognition, 2005: 886-893.

[9] LiangKai H, MaoJiun W. Image thresholding by minimizing the measures of fuzziness[J]. Pattern recognition, 1995, 28(1):41-51.

[10] Hong L, Wan Y, Jain A. Fingerprint image enhancement: Algorithm and performance evaluation[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(8):777-789.

[11] Wang H， Chen Y， Cai Y， et al.SFNet-N：an improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes [J]. IEEE Transactions on Intelligent.Transportation Systems, 2022, 23(11): 21405–21417.

[12] 陆剑锋, 林海, 潘志庚. 自适应区域生长算法在医学图像分割中的应用[J]. 计算机辅助

设计与图形学学报, 2005, (10):28-33.

[13] BoykovY Y, JollyM P. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images[C]. Proceedings of the eighth IEEE international conference on computer vision, 2001:105-112.

[14] Ou Z H, Wang Z F N, Xiao F R, et al. AD-RCNN: Adaptive Dynamic Neural Network for Small Object Detection[J]. IEEE Internet of Things Journal, 2022, 10(5): 4226-4238.

[15] Long J, Shelhamer E, Darreel T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2015:3431-3440.

[16] 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, 2015:234-241.

[17] 陈婷，姚大春，高涛，等.基于 PReNet 和 YOLOv4 融合的雨天交通目标检测网络[J]. 交通运输工程学报, 2022, 22(03): 225-237.

[18] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J].arXiv preprint arXiv:1511.07122.2016.

[19] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2017:2881-2890.

[20] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2017:1925-1934.

[21] Yang M, Yu K, Zhang C, et al. DenseASPP for semantic segmentation in street scenes[C]. Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition,2018: 3684-3692.

[22] Cui M H， Gong G L， Chen G， et al. LC-YOLO: A Lightweight Model with Efficient Utilization of Limited Detail Features for Small Object Detection[J]. Applied Sciences, 2023, 13(5): 3174-3189.

[23] Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation[J]. arXiv preprint arXiv:1805.10180, 2018.

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2018:7132-7141.

[25] Han S, Mao H, DallyW J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv preprint arXiv:1510.00149, 2015.

[26] 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.

[27] 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.

[28] Zhao H, Qi X, Shen X, et al. Icnet for real-time semantic segmentation on high-resolution images[C]. Proceedings of the European Conference on Computer Vision, 2018:405-420.

[29] 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.

[30] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]. Proceedings of the European conference on computer vision, 2018:325-341.

[31] Chen W, Zhu X, Sun R, et al. Tensor low-rank reconstruction for semantic segmentation [C]. Proceedings of the European conference on computer vision. 2020: 52-69.

[32] Zhu Z, Xu M, Bai S, et al. Asymmetric non-local neural networks for semantic segmentation [C].Proceedings of the IEEE international conference on computer vision. 2019:593-602.

[33] Verhelst M, Moons B. Embedded deep neural network processing: Algorithmic and processor techniques bring deep learning to iot and edge devices[J]. IEEE Solid-State Circuits Magazine, 2017, 9(4):55-65.

[34] 罗顺心，张孙杰. 基于卷积神经网络的回环检测算法[J]. 计算机与数字工程, 2019, 47(5): 1020-1026, 1048.

[35] Wang X W，Zhao Q Z，Jiang P，et al. LDS-YOLO: A lightweight small object detection method for dead trees from shelter forest[J]. Computers and Electronics in Agriculture, 2022,

198: 107035-107045.

[36] Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning[J]. arXiv preprint arXiv:1603.07285, 2016.

[37] 李鹏举,郑方坤,吕建兵, 等. 基于大数据迁移学习的灰岩地区排水孔淤堵自识别技术[J]. 现代隧道技术,2021,58(4):37-47.

[38] Li Y， Yu F.CDMY: A Lightweight Object Detection Model Based on Coordinate Attention[C].Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Chongqing: IEEE, 2022: 1258-1263.

[39] 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.

[40] 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.

[41] ChenL C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. Proceedings of the European conference on computer vision, 2018:801-818.

[42] 谢新林，罗臣彦，续欣莹，等. 双注意力引导的跨层优化交通场景语义分割[J]. 交通运输系统工程与信息, 2023, 23(01), 236-244.

[43] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2017:1251-1258.

[44] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2016: 2818-2826.

[45] IandolaF N, Han S, MoskewiczM W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and<0.5 MB model size[J]. arXiv preprint arXiv: 1602.07360, 2016.

[46] Zhang X, Zhou X, Lin M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2018: 6848-6856.

[47] Ma N, Zhang X, ZhengH T, et al. ShuffleNetv2:Practical guidelines for efficient CNN

architecture design[C]. Proceedings of the European conference on computer vision, 2018: 116-131.

[48] HowardA G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv: 1704. 04861, 2017.

[49] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2018: 4510-4520.

[50] Howard A., Sandler M., Chu G., et al. Searching for MobileNetv3[C]. Proceedings of the IEEE International conference on computer vision, 2019: 1314-1324.

[51] Molchanov P, Mallya A, Tyree S, et al. Importance estimation for neural network pruning[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019:11264-11272.

[52] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning[C]. trained quantization and huffman coding.2016: 1-14.

[53] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C].Advances in Neural Information Processing Systems. 2015: 1135-1143.

[54] Lecun Y, Denker J, Solla S. Optimal brain damage[C].Advances in Neural Information ProcessingSystems.1989: 598-605.

[55] YangT J, Howard A, Chen B, et al. Netadapt: Platform-aware neural network adaptation for mobile applications[C].Proceedings of the European Conference on Computer Vision.arXiv preprint arXiv: 1804. 03230, 2018.

[56] 凌志，李幸，张婷，等. 基于多层次知识蒸馏的连续图像语义分割方法[J]. 计算机集成制造系统, (2021-11-29)[2023-03-15].

[57] NeillJ O, Dutta S, Assem H. Aligned weight regularizers for pruning pretrained neural networks[J].ArXiv preprint arXiv: 2204.01385, 2022.

[58] Jonathan F, Michael C. The lottery ticket hypothesis: Finding sparse, trainable neural networks[C].International Conference on Learning Representations. 2019: 1-13.

[59] Li H, Yue X, Meng L. Enhanced mechanisms of pooling and channel attention for deep learningfeature maps[J]. PeerJ Computer Science, 2022, 8:1161.

[60] Kuang J, Shao M, Wang R, et al. Network pruning via probing the importance of filters[J]. InternationalJournal of Machine Learning and Cybernetics, 2022, 13(9): 2403-2414.

[61] Li H, Yue X, Wang Z, et al. Optimizing the deep neural networks by layer-wise refined pruning andthe acceleration on FPGA[J]. Computational Intelligence and Neuroscience, 2022, 2022: 1-22.

[62] 艾青林，张俊瑞，吴飞青，等. 基于小目标类别注意力机制与特征融合的 AF-ICNet 非结构化场景语义分割方法[J]. 光子学报. 2023, 52(01): 189-202.

[63] Li H, Yue X, Wang Z, et al. A survey of convolutional neural networks—from software to hardwareand the applications in measurement[J]. Measurement: Sensors, 2021, 18: 100080.

[64] Yu F, Chen H, Wang X, et al. BDD100k: A diverse driving dataset for heterogeneous multitask learning[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 2636-2645.

[65] Liu Y, Chen K, Liu C, et al. Structured knowledge distillation for semantic segmentation [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019:2604-2613.

[66] Xie Q, LuongM T, Hovy E, LeQ V. Self-training with noisy student improves imagenet classification[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10687-10698.

[67] 孟宪法, 刘方, 李广, 等. 卷积神经网络压缩中的知识蒸馏技术综述[J]. 计算机科学与探索, 2021: 1-20.

[68] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2016: 770-778.

[69] 梁延禹, 李金宝. 多尺度非局部注意力网络的小目标检测算法[J].计算机科学与探索, 2020, 14(10):1744-1753.

攻读硕士学位期间的研究成果

1 学术论文

[1] Real-time road scene segmentation based on knowledge distillation， ACM International Conference Proceedings Series，EI 收录。