

电子科技大学
UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

专业学位硕士学位论文

MASTER THESIS FOR PROFESSIONAL DEGREE



论文题目 基于语义分割的变电站巡检机器人

环境感知技术研究

专业学位类别 工程硕士

学 号 201822090416

作 者 姓 名 陈前

指 导 教 师 左琳 教授

分类号 _____ 密级 _____

UDC ^{注1} _____

学 位 论 文

基于语义分割的变电站巡检机器人 环境感知技术研究

陈前

(作者姓名)

指导教师 左琳 教授

电子科技大学 成都

(姓名、职称、单位名称)

申请学位级别 硕士 专业学位类别 工程硕士

工程领域名称 软件工程

提交论文日期 2021.03.22 论文答辩日期 2021.05.08

学位授予单位和日期 电子科技大学 2021 年 06 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号

Research of Environment Sensing of Substation Inspection Robot Based on Semantic Segmentation

A Master Thesis Submitted to

University of Electronic Science and Technology of China

Discipline:	Master of Engineering
Author:	Qian Chen
Supervisor:	Prof.Lin Zuo
School:	School of Information and Software Engineering

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名: 陈彪 日期: 2021 年 3 月 15 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将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、汇编学位论文。

(保密的学位论文在解密后应遵守此规定)

作者签名: 陈彪 导师签名: 方利

日期: 2021 年 3 月 15 日

摘要

随着社会逐步进入智能化时代，变电站人工巡检方式因消耗巨大且效率低下已开始呈现逐步被巡检机器人替代的趋势。在巡检机器人执行巡检任务过程中，巡检机器人需要解决自主定位、路径规划等问题。环境感知技术是解决以上问题的基础。基于激光的感知技术难以达到巡检机器人准确理解周围环境的目标，而基于场景语义分割的感知技术可以更精确地辅助巡检机器人完成路径规划、躲避障碍物等任务。但是，实际情况下，变电站场景中存在识别物类别长宽比例相差较大以及各识别物在变电站出现频次差异大等问题。特别地，在巡检任务过程中，巡检机器人可能会遇见未知识别物类别，这些因素将严重地影响语义分割准确率。因此，针对上述问题，本文对变电站场景的语义分割展开研究，提升变电站巡检机器人环境感知的识别精度。本文研究内容如下：

(1) 针对变电站环境感知中存在识别物的类别间形状差异大、部分类别长宽比例悬殊以及部分识别物类别出现的频次较少等问题，提出了一种基于多视角注意力机制的语义分割模型。该模型采用空洞空间金字塔结构和注意力结构，多角度地从图像特征中提取不同的局部视觉特征，并加强同一类别像素之间的关联。同时，该模型设计多尺度特征融合结构，来解决全局特征和局部特征对齐及边缘信息丢失问题，降低了特征融合阶段出现阴影现象的可能，提升了变电站场景识别的精度。

(2) 针对变电站环境感知中存在的对未知识别物的辨识问题，提出了一种基于零样本学习的语义分割模型。为了学习未知识别物的视觉特征，该模型设计了一个基于自编码器的特征生成方法，用以生成未知类的视觉特征，同时，该方法基于生成对抗网络技术，减小了生成的视觉特征与真实的视觉特征之间的差距。该模型通过同时使用这两种视觉特征进行训练，提高了预测未知识别物类别的准确率。

(3) 针对变电站巡检机器人语义感知的需求，基于本文提出的多视角注意力机制的语义分割模型和基于零样本学习的语义分割模型，设计并实现了基于语义分割的巡检机器人语义感知系统。

关键词：多视角注意力机制，零样本学习，语义分割，环境感知系统

ABSTRACT

With the era of intelligence coming, inspection robots begin to replace humans to complete inspection tasks. Because manual inspection is costly and low-efficient. In the process of completing the inspection task, the inspection robot needs to solve problems like autonomous positioning and path planning. Environment sensing technology is the basis for solving the above problems. It is difficult for the inspection robot to accurately understand the surroundings with laser-based sensing technology, while the environment sensing based on scene semantic segmentation can more accurately assist the inspection robot to complete tasks such as path planning and avoiding obstacles. However, there are some challenges in substations. For example, there are categories with large ratio difference between length and width, and the frequency of each category is significantly different. Meanwhile, during the inspection task, the robot can encounter unseen categories. These factors will seriously affect the accuracy of semantic segmentation. In this paper, semantic segmentation of substation scenes is studied to improve the accuracy of predicting seen and unseen categories. The main contents include:

(1) proposing a semantic segmentation model based on multi-view attention mechanism to address some problems. For example, the shape of each category is different, the ratio of length to width is different between categories, and the number of some categories are also small. The model adopts atrous spatial pyramid pooling structure and attention structure to extract different local visual features from image features and strengthen the relationship between pixels of each category. At the same time, a multi-scale feature fusion method is designed to solve the problem of alignment and edge information loss, which reduces the possibility of shadow phenomenon in the feature fusion stage and improves the accuracy of substation scene recognition.

(2) proposing a zero-shot semantic segmentation model to predict unseen categories. In order to learn the visual features of the unseen categories, the model contains a feature generation method based on the autoencoder to generate the visual features of the unseen class. At the same time, this method uses generative adversarial network technology to reduce the gap between the generated visual features and the real visual features. The model improves the accuracy of predicting unseen categories by simultaneously using these two visual features for training.

ABSTRACT

(3) In order to meet the environment sensing needs of the substation inspection robot, a semantic segmentation system based on the semantic segmentation model in this thesis is designed and implemented.

Keywords: multi-view attention mechanism, zero-shot learning, semantic segmentation, environment sensing system

目 录

第一章 绪论	1
1.1 研究背景及意义	1
1.2 研究现状.....	2
1.2.1 巡检机器人研究现状.....	2
1.2.2 环境感知研究现状.....	2
1.3 本文研究内容.....	6
1.4 主要结构安排.....	6
第二章 基于语义分割的环境感知技术.....	8
2.1 语义分割模型.....	8
2.2 零样本语义分割.....	13
2.3 语义分割评价指标	16
2.4 本章小结.....	17
第三章 基于多视角注意力机制的语义分割模型	18
3.1 多视角注意力机制语义分割模型	18
3.2 多视角注意力结构	20
3.2.1 多视角学习	20
3.2.2 注意力网络结构	22
3.3 多尺度特征融合方法	24
3.4 实验结果.....	25
3.4.1 实验数据.....	25
3.4.2 实验设定	27
3.4.3 实验结果与分析	28
3.5 本章小结.....	34
第四章 基于零样本学习的语义分割模型	35
4.1 基于零样本的语义分割网络模型	35
4.2 面向零样本学习的特征生成方法	37
4.2.1 自编码器	37
4.2.2 生成对抗网络	38
4.2.3 基于自编码器的特征生成方法	41
4.3 实验结果.....	44

4.3.1 实验数据	44
4.3.2 实验设定	44
4.3.3 实验结果与分析	44
4.4 本章小结	49
第五章 变电站巡检机器人语义感知软件系统设计与实现	50
5.1 需求分析	50
5.1.1 系统功能需求	50
5.1.2 系统主要用例图	51
5.1.3 系统性能需求	51
5.1.4 系统运行环境需求	51
5.2 系统总体设计	52
5.3 系统的详细设计与实现	54
5.3.1 系统功能设计与实现	54
5.3.2 数据库设计	58
5.4 系统测试	59
5.4.1 测试环境	59
5.4.2 测试结果	59
5.5 本章小结	61
第六章 总结与展望	62
6.1 全文总结	62
6.2 后续工作展望	62
致 谢	64
参考文献	65
攻读硕士期间取得研究成果	71

第一章 绪论

1.1 研究背景及意义

在变电站场景下，人工巡检的方式消耗巨大以及巡检效率低。随着科技的发展，智能巡检机器人逐步应用在变电站巡检任务之中。变电站智能巡检机器人主要的功能包括定位导航、路径规划、故障检测、远程控制等。上述功能需要依赖智能巡检机器人的三大核心能力：控制能力、环境感知能力以及动作行为能力。其中，环境感知能力是智能巡检机器人研究的关键点和难点。环境感知能力主要指巡检机器人通过传感器获取周围环境信息的能力。比如使用激光雷达进行环境建模^[1]；实时地利用陀螺仪输出智能机器人导航信息^[2]；利用视觉仪器对设备、环境进行识别^[3,4]。巡检机器人利用环境感知技术能够更好地理解周围的环境从而进行避障和路径规划。巡检机器人的环境感知技术包括基于激光的感知方法和基于视觉的感知方法等。目前，巡检机器人常用的环境感知方法是基于激光的方法^[5,6]。这类方法可以使得巡检机器人快速获得周围物体的信息，但是却不能使其理解身边的环境情况，如前方障碍物类别、前方道路的具体情况等。这说明基于激光技术的巡检机器人环境适应力相对较弱。随着图像算法研究的不断深入，基于视觉的感知技术被应用在巡检机器人上^[7,8]。国内外学者对图像分割和识别算法在巡检机器人的应用展开了深刻的研究^[9,10]。基于这些研究的巡检机器人可以完成更高级的任务，如仪器仪表的识别、故障设备的检测。然而，变电站巡检机器人往往在复杂的环境中执行任务。这些复杂的环境包括：巡检机器人行动时会遇见行动不定的人、突如其来的障碍物、气候的变化导致路况环境的变化等。面对这些复杂的变电站场景，使用传统图像分割和识别算法虽然具有一定的效果，但是还不能完全满足巡检任务的需求。随着计算机计算能力和存储能力的提高，越来越多的学者开始研究以深度学习为代表的人工智能算法。在这样的时代背景下，深度学习在文本翻译、自然语言处理、语音识别、计算机视觉等领域的研究日新月异。尤其是在计算机视觉领域中，一些图像相关的复杂问题逐步被解决。比如，AlexNet^[11]在 ImageNet 大规模视觉识别挑战赛上崭露头角；ResNet^[12]以增加网络深度为创新点在 ILSVRC 比赛中获得冠军。基于深度学习的算法能够极大地增强智能巡检机器人的环境感知能力——提高识别与检测任务的精度、完成更为复杂的任务。其中移动场景的语义分割是巡检机器人环境感知算法中最有挑战的一项任务。语义分割主要完成对目标图片的逐像素分类。语义分割的结果图可以帮助巡检机器人理解场景以及辅助它完成规划路径或者紧急避障。比如，鲜开义^[13]等提出一种基于全卷积的语义分割模型。该

模型将语义分割结果图转化为巡检机器人前方道路信息以辅助机器人完成避障任务。

1.2 研究现状

1.2.1 巡检机器人研究现状

变电站环境危险、包含设备种类多且复杂。这为人工巡检带来了巨大的难度。因此，国内外企业、高校非常重视对变电站巡检机器人的研究。在智能巡检机器人领域，国外研究开展较早。在 2005 年，A.Birk 等^[14]研制出一种带有红外测温设备和彩色摄像头的变电站巡检机器人，该机器人能够有效地对变电站电气设备进行温度测量，并且在美国西部电力公司投入使用。John-Young 等^[15]提出了一种基于视觉的高压电线绝缘端子检测方法，该方法适用在巡检机器人上且在现场实验证明了其有效性。相比于国外，国内智能巡检机器人研究起步较晚。国内智能巡检机器人的研究于 20 世纪 90 年代开始至今，通过国家的政策支持、国内诸多高校以及龙头企业的不懈努力，现已取得显著的成果。2005 年，我国第一台变电站智能巡检机器人在长春投运。矫德余等^[16]针对工业巡检场景为智能巡检移动机器人设计了基于嵌入式的平台软件。该软件的功能包含巡检控制、路径规划等等。王建元等^[17]依据巡检机器人任务的特点结合图论思想提出一种基于传递闭包理论的电力巡检机器人路径规划方案，该方法能够解决巡检机器人搜索路径过程中陷入死循环、占用大量资源的问题，降低了检测的所消耗的时间。高青等^[18]研究 500kV 久安站的巡检机器人并设计实现了一种变电站巡检机器人系统。通过配置红外设备、可见光电耦合原件等传感器元件，该系统可以高效地辅助人工完成变电站高压变电设备的巡检任务。鲁能智能技术有限公司一直致力于研制巡检机器人。该公司研制的一款携带红外成像仪和可见光摄像头的变电站室内巡检机器人能够在变电站环境下完成设备测温、仪表识别等任务，并且在江苏沙洲正式投入使用。总的来说，在过去的时间段，国内外对于巡检机器人的研究取得相当大的进展。在一些场所内，巡检机器人开始逐步地替代人工作业，极大提高了效率。随着智能化的发展，巡检机器人将会应用到更广的范围。

1.2.2 环境感知研究现状

环境感知技术是机器人实现路径规划以及执行特定任务的基础和关键。环境感知是指机器人利用自身配置的一系列传感器对周围环境信息进行捕捉获取，然后通过特征提取和处理，建立周围环境的数学模型表达^[19]。根据不同的传感器，机

器人会获得周围信息的不同表示。同时，依据传感器的类型，机器人环境感知方法分为：基于激光雷达的感知方法、基于声纳的感知方法、基于视觉的感知方法。上述的方法都可以帮助巡检机器人完成路径规划的任务。其中基于语义分割的感知方法是基于视觉的感知方法中的研究重点和难点。下面将分别介绍环境感知技术的研究现状，并且将会说明基于语义分割的感知方法的优势。

声纳采用脉冲或者连续波的形式对物体进行探测从而达到感知周围环境的目的。李江等^[20]对 AS-RE 机器人的声纳传感器、红外传感器进行了研究，并提出了一种基于声纳和红外传感器的机器人自主运动方法。Thongchai 等^[21]利用声纳传感器建立周围环境的数学模型，辅助机器人完成避障任务。段丙涛等^[22]认为超声波传感器能够有效地为移动机器人提供环境感知信息，并在此基础上提出了两种避障算法。这些算法降低了移动机器人自主移动的危险性。Jie 等^[23]根据人类听觉系统的优先效用模型研究了一种装有实时声音定位且包含声纳系统的移动机器人。实验结果表明——通过声纳定位，该机器人能够感知周围环境，并且能够在没有碰撞的情况下接近目标物体。因为声纳在水下传播速度快且不易受水下介质变化的影响，所以基于声纳的环境感知方法可以用于水下机器人。比如，黄东武等^[24]研究了基于声纳系统的水下机器人，并分析不同水下声成像的特征和规律，提高了水下机器人检测和识别精度。但是基于声纳的方法具有一定的缺点：声纳传感器在发射超声波的过程中，其能量强弱会受到传播距离远近的影响以及容易受到镜面反射的影响。

基于激光的感知方法可以快速获取较远距离物体的信息——位置、大小等。与声纳相比，激光具有传播速度快、抗干扰性强的特点。杨明等^[25]提出了一种基于激光雷达的环境建模和避障方法。该方法能够客观地对机器人周围的环境进行描述，并提高机器人运行的安全性和可靠性。Nuchter 等^[26]采用三维激光点云技术并提出了一种具有六个自由度的建图方法。该方法能够使得机器人更好地感知地下矿道环境。Serafin 等^[27]针对三维激光点云匹配困难的问题，提出了一种适用于空旷的环境且基于主成分分析的特征提取方法。Steux 等^[28]提出一种基于激光定位的方法来建图。然而，基于激光的感知方法也具有一定的缺点：虽然它能够很好地感知周围环境的几何信息——物体的大小、形状等，但是并不能理解周围物体的种类，并且不同种类的障碍物会对激光测量的效果和准确性有着不同的影响。

随着科学技术的发展，基于视觉的感知方法逐步展示其优越性。首先视觉传感器具有价格优势，其次传统图像算法和深度学习不断发展使得各类视觉任务的准确率进一步提高。基于视觉的环境感知方法主要包含检测方法和语义分割方法。两种方法都能辅助巡检机器人理解周围的环境。对于检测方法，杨涛等^[29]提出了一

种基于视频对齐的背景差分技术。该技术解决了由于巡检机器人运行速度不稳定而导致的视频无法对齐问题，使得巡检机器人检测异物的精度和效率变高。为了解决光照变化对巡检机器人提取导航线的影响，薛阳等^[30]提出了一种基于朴素贝叶斯分类的导航线检测方法。该方法首先使用朴素贝叶斯算法对图像进行分类，然后结合导航线的边缘信息和颜色信息进一步识别导航线。针对基于单目视觉的变电站巡检机器人导航问题，赵坤等^[31]设计了直行、拐弯和停止的路面标识来引导巡检机器人行动。该方法先后对获取图像进行色彩空间变换、灰度化、二值化和中值滤波操作，然后再进行边缘检测、角点检测和畸变还原，最后进行光学字符识别。针对变电站复杂的环境，Wei 等^[32]提出来一种基于视觉的巡检机器人导航控制方法。该方法将巡检机器人采集的图像转换到 HSV 空间以减小图像受强光、雨天的影响，然后应用模板匹配算法识别巡检道路。

从上述研究可以看出，巡检机器人可以通过检测方法来实现周围环境的感知并且制定下一步行动计划。但是检测方法只能使得巡检机器人粗略地理解场景，而语义分割方法因其逐像素分类的优势可以更加精确地让巡检机器人理解周围的环境。因此，许多学者对语义分割方法展开研究。机器人通过视觉传感器采集路况信息——道路图像，然后对图像进行逐像素点的分类从而辨别周围环境的具体情况。语义分割不仅要区分高度相似的不同物体，而且也要克服同类物体因光线、角度和状态等不同而产生的差异性。此外，语义分割的实际场景往往是复杂多样的，不同物体之间常常伴有交错、遮掩等情况，这进一步增加了语义分割的难度。在深度学习兴起之前，语义分割是采用传统的图像处理方法。景晓军等^[33]研究了一种基于区域特征的分割技术——一种二维最大类间方差的图像分割算法。该算法既考虑图像元点的灰度分布信息又考虑它们之间的空间信息。这种算法具有计算量小、消耗存储空间低的特点。区域增长法也是一种计算机视觉十分重要的图像分割方法。它将区域作为图像的处理对象且通过考虑区域之间的同异性来找到图像中各类物体的边界。Pohle 等^[34]提出一种具有强健壮性的基于区域增长的图像分割方法，该方法在 CT 图像和人造图像上具有较好的效果。Zhang 等^[35]设计了一种不需要种子点的自动分割算法。该算法对初始种子点的选取顺序等区域生长算法关键问题不敏感。马范援等^[36]设计了一种基于金字塔结构的区域分割算法并在医学图像上取得较好的分割效果。从上述研究可以看出，传统图像语义分割的方法效率高、速度快，但是在面对复杂环境的情况下，传统图像分割算法没有表现出较强的鲁棒性和智能化的特点。随着深度学习的流行，图像语义分割准确率有了进一步的提升。目前已有很多学者和专家展开研究。如：Buettner 等^[37]提出了一个可以使机器人在危机情况下识别路线的深度学习模型。Asadi 等^[38]为了克服语义分割模型的计算复杂

度过大而提出了一种新的深度学习网络模型。该模型可以在嵌入式平台上实时运行，并且有益于未来的自主机器人系统的发展。Ummenhofer 等^[39]提出了一种基于单目相机的神经网络来评估位姿和深度信息。Zhou 等^[40]针对无结构的视频序列提出了一种无监督的端到端的学习模型来估计每一帧图像的深度。Badrinarayanan 等^[41]提出了一种新的深度全卷积神经网络架构——SegNet。这个架构的核心是编码器和解码器。编码器网络的架构在拓扑上类似于 VGG16^[42]网络的架构。解码器网络的作用是实现从低分辨率的编码器结果图到输入分辨率大小的原始图的还原。SegNet 的新颖之处在于使用解码器对较低分辨率的特征图进行上采样。具体而言，经过解码器计算出的特征图和相应编码器的特征图进行拼接之后再执行非线性上采样，然后将上采样的结果进行卷积操作以生成密集的特征图。相比于 FCN^[43]、DeepLab-Large FOV^[44]以及 DeconvNet^[45]模型，该模型不仅展示出良好的分割性能，而且达到了内存消耗与准确率之间的平衡。与其他网络相比，SegNet 是一种可以进行端到端的训练且可训练参数较少的网络模型，更易于在工程项目中使用。Zhao 等^[46]认为上下文关系对复杂场景的理解十分重要。因此，他们提出一种适用于复杂场景理解的语义分割模型——PSPNet。该模型使用金字塔模块和金字塔场景解析结构聚合基于不同区域的上下文信息。与大部分模型相比，这种方式能够更好地利用全局场景中的类别信息。同时，该模型能够很好地解决部分基于 FCN 的模型因没有捕捉充足的上下文信息而导致的分割错误的问题。Lin 等^[47]提出了一种能利用下采样的所有信息来实现高分辨率预测的多路径强化网络。该网络提出三个结构来解决下采样过程中信息损失的问题。该网络的核心结构包括链式残差池化模块、残差卷积模块和多分辨率融合模块。其中链式残差池化模块能够从较大的图像区域捕获背景上下文，多分辨率融合模块目的是融合多种分辨率的特征图。Hou 等^[48]认为在基于深度学习的语义分割任务中，大部分损失函数的定义都是基于像素的，并且对像素进行分类不易于学习类别的形状信息。因此，他们将待分割区域看作为一类需要学习的形状，在此基础上提出了一种新的形状预测网络和一种形状域内误差的损失函数。该方法对图像噪声不敏感性且具有较强的鲁棒性。由于注意力机制在语音、文字识别等领域的流行^[49,50]，不少研究者逐步研究注意力机制在语义分割领域的应用^[51,52]。Fu 等^[53]提出了一种包含两种注意力模块的网络——通道注意力模块和位置注意力模块。通道注意力模块模拟图像特征通道之间的依赖性；位置注意力模块学习图像特征不同位置之间的关系。通过增加两种注意力模块，该网络能自适应地捕捉局部特征和全局特征的依赖关系。在图像领域中，使用自注意力机制需要生成一张较高分辨率的特征关系图且每一个像素的特征关系图都需要进行全图计算，这个过程具有相当大的空间复杂度和时间复杂度。面对这一问题，

Li 等^[54]提出一种期望最大化的注意力机制模型。期望最大化注意力机制并没有在全图上不断地进行注意力计算，而是使用期望最大化算法(EM)^[55]逐步迭代出一组的基，在这组基上运行注意力机制从而大大降低了复杂度。其中，E 步更新注意力图，M 步更新这组基，E、M 交替执行。

1.3 本文研究内容

本文的主要研究内容是基于变电站场景的语义分割问题。本文针对变电站场景设计了基于多视角注意力机制的语义分割模型和基于零样本学习的语义分割模型，同时设计和开发了一套变电站巡检机器人语义感知系统。研究内容主要包括一下几点：

1. 设计一种基于多视角注意力机制的语义分割模型。在变电站场景中，存在识别物的类别之间形状差异较大且部分识别物类别长宽比例悬殊等问题。针对识别物的类别之间形状差异较大且类别长宽比例悬殊的问题，模型以不同大小的感受野为切入点进行多尺度的图像视觉特征提取。针对像素占比比较低的问题，模型采用注意力机制去加强图像中各个像素点之间的联系。同时，为了减少语义分割中不断下采样导致的图像信息丢失，模型包含一个多尺度特征融合模块——融合全局特征和多视角下的局部特征。在变电站场景下，该语义分割模型识别准确率有所提高。
2. 设计一种基于零样本学习的语义分割模型。传统语义分割任务中，模型只能识别已知类别。但在实际识别任务中，模型会接收含有未知识别物的图像作为输入。对于未知识别物类别，模型一般无法给出正确的预测结果。虽然可以使用大量公开数据集训练模型来增加模型识别已知类别的数量，但是公开数据集也不能包含现实生活中所有的类别。因此，本文提出一种基于零样本学习的语义分割网络。该模型分为两个部分：语义分割模块和特征生成模块。特征生成模块目的是生成未知识别物的视觉特征。将生成的视觉特征和真实的视觉特征合并一起训练语义分割模型从而使模型能够一定程度上识别未知识别物类别。
3. 设计并实现了一个变电站场景下的语义感知系统。该系统能够对巡检机器人视觉传感器采集的图像进行语义分割任务并且展示周围环境的识别结果。

1.4 主要结构安排

本论文总共分为六章，具体的章节安排如下：

第一章，绪论。在本章中介绍巡检机器人环境感知技术的研究背景和意义，阐述巡检机器人和环境感知的发展现状，同时也介绍了本文主要研究内容。

第二章，基于语义分割的环境感知技术。本章介绍基于语义分割的环境感知技

术，并且分析语义分割模型的特点和语义分割任务所面临的困难与挑战。同时，本章也介绍了零样本学习的方法和基于零样本学习的语义分割技术。

第三章，基于多视角注意力机制的语义分割模型。本章提出了一种基于多视角注意力机制的语义分割模型。该模型采用空洞空间金字塔结构多角度地学习局部特征，同时利用注意力机制增强模型对变电站场景类别的识别能力。

第四章，基于零样本学习的语义分割模型。本章提出了基于零样本学习的语义分割模型。该模型使用特征生成的方法合成未知识别物的视觉特征并学习这些合成特征，能够一定程度上识别未知类别。

第五章，变电站巡检机器人语义感知软件系统设计与实现。本章基于本论文提出的语义分割模型设计并实现一种变电站场景的语义分割系统。本章分别介绍系统的需求分析、总体设计、功能设计和测试结果。

第六章，总结与展望。本章概括和总结本文对巡检机器人语义感知的工作，并对存在的不足进行分析，同时对下一步工作和研究内容进行阐述。

第二章 基于语义分割的环境感知技术

2.1 语义分割模型

图像语义分割是图像理解的基础^[56,57]。它根据不同的语义或形状信息将整幅图像划分成不同的区域块，并且逐像素地推理出这些区域块的类别，最终生成一幅具有逐像素语义标注的分割图像。如图 2-1 所示，(a) 为待分割的原始图片，(b) 为语义分割之后的结果图。可以看出经过语义分割操作之后，(a) 被分成了公路、车辆、树木等区域，对于智能巡检机器人来说，这样的图片可以帮助其规划路径、躲避障碍物。

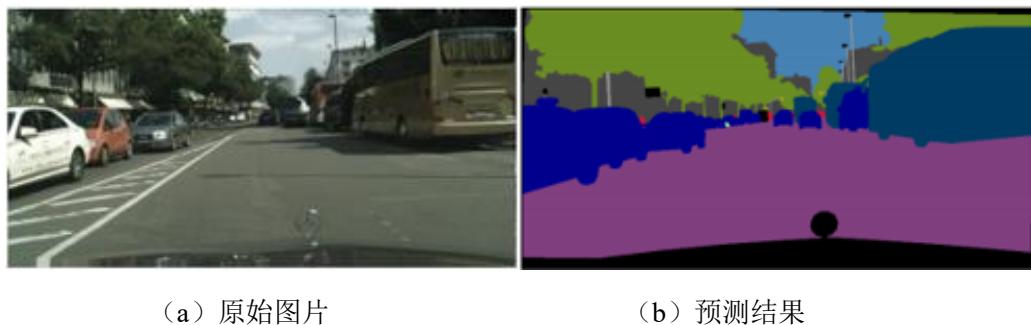


图 2-1 语义分割示意图

目前，语义分割任务面临许多问题。一是数据集问题。首先，由于语义分割任务是对输入图像进行逐像素的分类，那么输入图像所对应的目标图像就需要逐像素进行标记。这会消耗大量的人力与时间。其次，数据量较少导致无法从零开始训练语义分割模型。二是语义分割算法的实时性问题。为了满足自动驾驶等任务的实时性，语义分割算法需要在相对较短的时间内计算出结果，然后为路径规划、智能避障提供辅助信息。然而，目前的语义分割模型大部分都使用含有空洞卷积的全卷积语义分割框架。这导致模型训练和预测的时间变长。三是图像中类别的精确化预测。首先，在图像识别任务中，算法往往只用提取相应特征就可以识别出具体类别。但是在语义分割任务中，算法不仅需要提取相应特征，而且还需要对这个类的形状大小进行预测。因为所有类别在输入图像中出现的频率是不同的且它们的像素占整个图像的总像素比例也不相同，所以需要平衡不同类别之间的这种差异。其次，类别边缘和小物体的预测较为困难。四是语义分割不仅要区分高度相似的不同物体，而且也要克服同类物体因光照、位姿和状态等不同而产生的差异性。

在卷积神经网络应用在图像语义分割任务之前，基于传统方法的语义分割技

术是主流技术。基于传统方法的语义分割技术主要包括基于阈值的图像分割技术、基于边缘图像分割技术、基于区域的图像分割技术和基于特定理论的图像分割技术。基于阈值的图像分割技术主要是利用图像像素的灰度值进行分类。通过把图像中每一个像素的灰度值与每个类别的灰度阈值进行匹配从而进行分类。这种方式的优势在于其易于理解、算法的时间和空间复杂度较低。但是它没有考虑到像素与像素之间的关系以及局部阈值和全局阈值之间的关系。基于边缘的图像分割技术主要是依赖图像的边缘检测方法。通过分析像素值之间的变化情况来进行图像的分割。比如 canny、sobel、prewitt 算子采用求一阶导的方式进行图像分割，laplacian 算子采用求二阶导的方式进行图像分割。这类方法有很强的数学基础，但是受图像质量的影响很大。因此，在使用这些算法之前，常常需要进行图像的预处理。基于区域的图像分割技术是通过一些相似性准则对图像进行区域划分，如纹理类似准则和颜色类似准则等。传统图像分割的方法需要针对不同场景分别进行设计且其准确率受图片的质量影响较大。与传统图像分割方法相比，基于深度学习的图像语义分割有较强的鲁棒性和更高的精度。深度学习利用其复杂的拟合能力可以准确地学习每一副图像中复杂的语义。下面介绍三种经典的语义分割模型以及它们的优劣势。

1.全卷积网络

Jonathan Long 等^[43]首次使用全卷积神经网络进行图像的语义分割。其网络结构如图 2-2 所示。

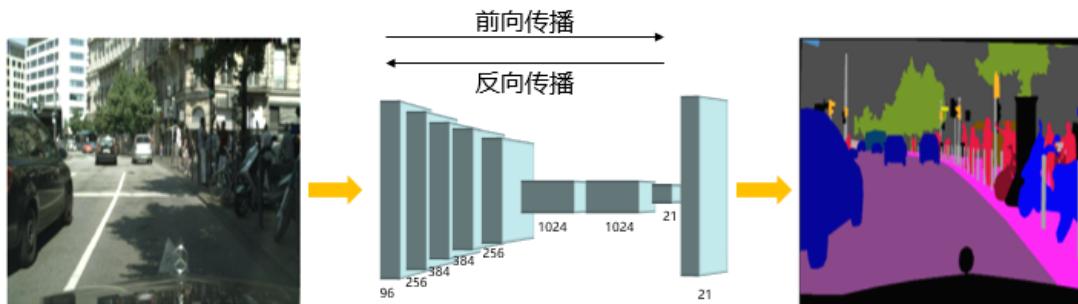


图 2-2 语义分割网络

该模型包含卷积层、池化层和上采样层。卷积层和池化层的作用是精确化地学习原始图像特征，而上采样层的作用是将学习后的特征大小还原至原始图像大小。如果仅使用最后一层池化层的结果进行上采样，那么所得的结果图比较粗糙和模糊。因此，为了得到更加精细化的结果图，该模型还将前几层所得到的输出特征和

最后一层的输出特征进行融合。如图 2-3 所示，该模型含有三种不同采样步长的上采样层，分别为 32、16 和 8。比如采样步长为 16——将 pool4 与 conv7 进行融合并且上采样。在他们的实验结果中，采样步长为 8 的分割效果最好。总体来看，尽管 FCN 模型分割精度比传统图像分割方法要高很多，但是 FCN 模型也具有不足之处。首先，没有解决特征融合时特征的对齐问题；其次，没有考虑特征图中像素与像素之间的关系；最后模型不能够充足地捕获上下文信息。

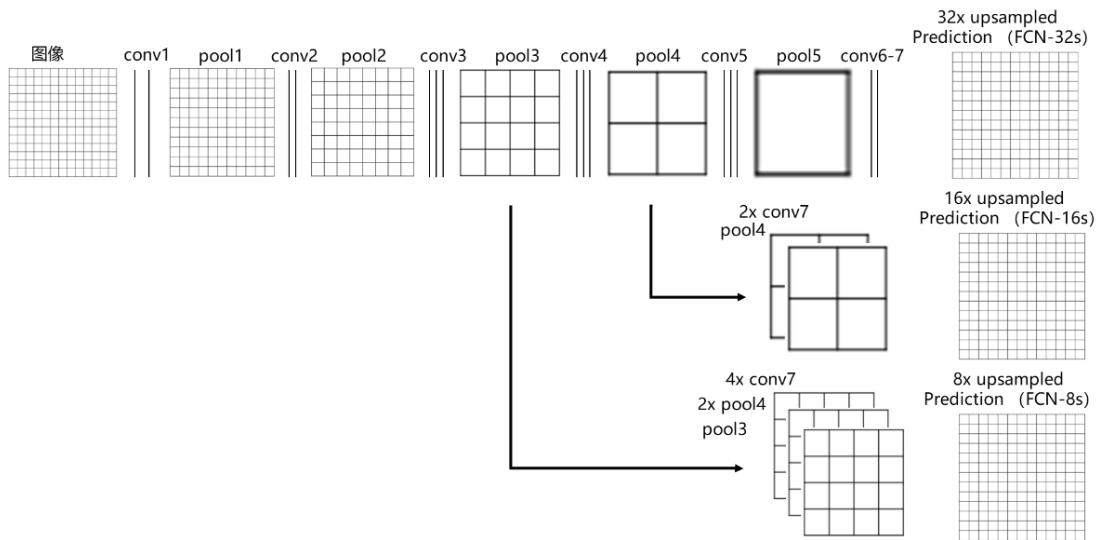


图 2-3 全卷积网络模型

2.Unet

Unet^[58]是被提出来解决生物医学图像分割问题的模型。Unet 延续了全卷积网络的思路——采用特征融合的方式。在语义分割任务中，随着卷积层和池化层的增多，图像分辨率降低且图像信息逐步由浅层信息变成深层信息。若仅由深层信息进行上采样得到结果图，那么结果图会因流失浅层信息变得模糊。Unet 之所以在生物医学图像分割上面有很好的效果，主要得益于它特殊的特征融合方式——深层信息和浅层信息的融合。深层信息是指经过多次下采样后的低分辨率信息。它表示类别的高级语义信息，有助于识别图像中的物体类别；浅层信息是指经过少量卷积层和池化层的高分辨率图像信息。它能提供更加精细的特征，如纹理，边缘等。如图 2-4 所示，红色箭头代表 2x2 的最大池化操作；蓝色箭头代表 3x3 的卷积操作；灰色箭头表示复制和剪切操作；绿色箭头代表 2x2 的上采样操作。为了解决同一层左边的图像特征的分辨率与右边的图像特征的分辨率不一致的问题，Unet 使用了剪切技术。Unet 包含编码部分和解码部分。在编码部分，模型采用多次卷积、最大池化操作对原始图像进行编码，最后得到深层信息；在解码部分，模型以深层

信息为基础不断融合浅层信息从而获得高分辨率的结果图。从图 2-4 中可以看出，Unet 尽管继承了 FCN 特征融合和上采样的思想，但是也针对其面对的医学图像问题进行了一定的改进。首先，FCN 在恢复原图像分辨率中采取的是转置卷积操作，而 Unet 采取的是将上采样与卷积的串联操作；其次，FCN 探索了不同浅层特征和深层特征融合的效果，而 Unet 引入编码器与解码器使得浅层特征和深层特征有所对应，更易于图像细节的还原。同时，Unet 采用剪切的操作解决了深层特征和浅层特征分辨率不匹配的问题；最后，FCN 的特征融合方式是将两种特征进行相加；而 Unet 则是采用拼接的方式。相比于 FCN，这样的方式可以减少特征融合所带来的对齐问题。Unet 相比于 FCN 模型有了相当大的改进，但是 Unet 仍然存在一些不足。比如在面对形状极小的类别时，Unet 分割效果会有所下降；没有考虑到像素与像素之间的关系。

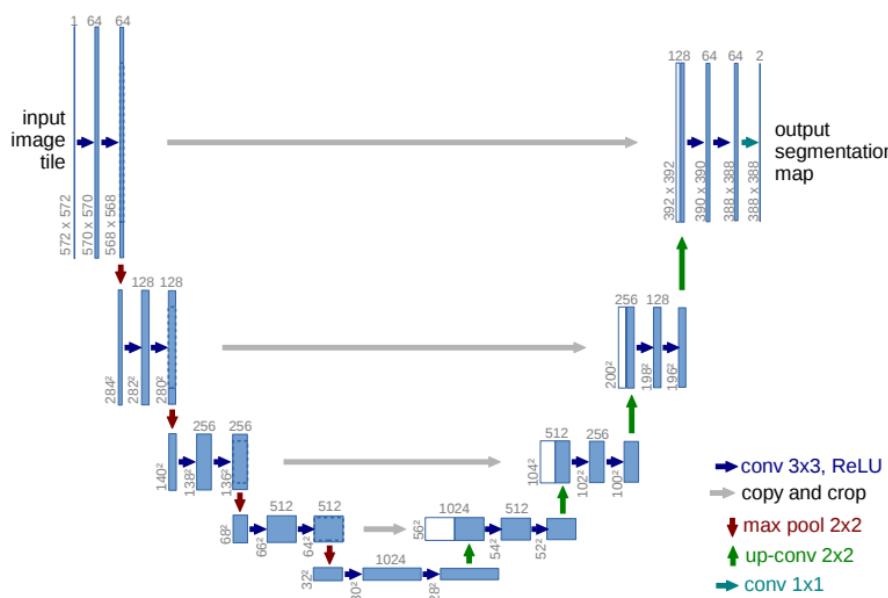


图 2-4 Unet 模型

3.DANet

在语义分割任务中，模型的设计需要考虑如何区分一些容易混淆的类别以及一些形状相似的类别。随着 Bahdanau 等^[59]提出注意力机制之后，越来越多的研究人员展开相关研究。特别是 Vaswani 等^[49]认为只需要注意力结构就能够很好地完成自然语言处理领域相关的任务而提出的一种新的注意力机制模型。如图 2-5 所示，Fu 等^[60]提出一种基于双重注意力机制的语义分割模型。该模型采用注意力机制捕获上下文信息来解决类别尺度不一样以及像素占比不一样的问题，提高了语义分割的准确率。

如图 2-6 和 2-7，该模型包含两种注意力机制结构——位置注意力结构和通道注意力结构。图 2-6 为位置注意力结构，位置注意力结构主要是捕获特征图的任意两个位置之间的空间依赖关系。即每一个位置的数值的计算都与其他位置的数值有关系。图 2-7 为通道注意力结构，通道注意力结构主要是捕获任意两个通道之间的相互依赖关系。即根据其他通道的数值计算下一通道的数值。

从图 2-6 和 2-7 可以看出，位置注意力结构和通道注意力结构都是通过矩阵相乘的方式分别建立像素之间、通道之间的依赖性。对于位置注意力结构，首先，将图像特征 A 分别经过三个卷积操作得到特征 B、C、D；然后将特征 B、C、D 进行维度变换，将变换后的 B、C 进行矩阵相乘和 softmax 操作得到注意力图 S；紧接着将特征 D 与注意力图进行矩阵运算获得中间结果；最后将中间结果和原始图像特征 A 进行拼接。

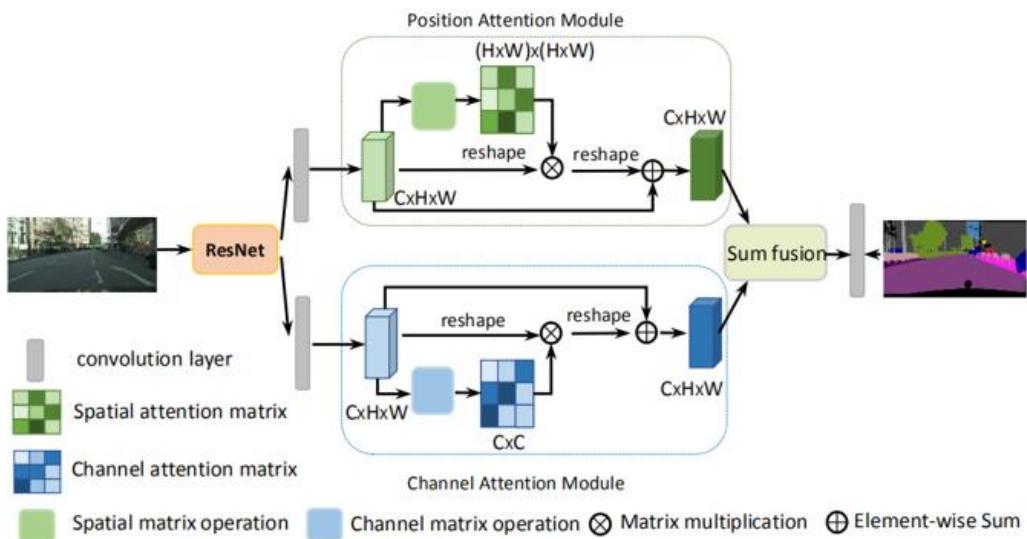


图 2-5 DANet 模型

与位置注意力结构相比，通道注意力结构较为简单——去除了卷积操作。位置注意力结构公式如（2-1）和（2-2）；通道注意力结构公式如（2-3）和（2-4）。

$$S_{ij} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^N \exp(B_i \cdot C_j)} \quad (2-1)$$

$$E_j = \alpha \sum_{i=1}^N (s_{ji} D_i) + A_j \quad (2-2)$$

$$x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^C \exp(A_i \cdot A_j)} \quad (2-3)$$

$$E_j = \beta \sum_{i=1}^C (x_{ji} A_i) + A_j \quad (2-4)$$

其中， $A \in \mathbb{R}^{H \times W \times C}$ 和 $\{B, C, D\} \in \mathbb{R}^{C \times W \times H}$ 都表示图像特征， $S \in \mathbb{R}^{(H \times W) \times (H \times W)}$ 表示注意力图，即每个像素之间的关系， s_{ji} 表示第 i 个位置的像素对第 j 个位置像素的影响， α 、 β 是平衡参数， x_{ji} 表示第 i 个通道对第 j 个通道的影响。

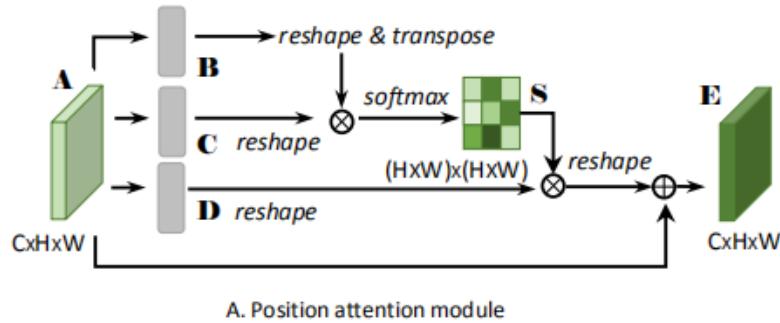


图 2-6 位置注意力结构

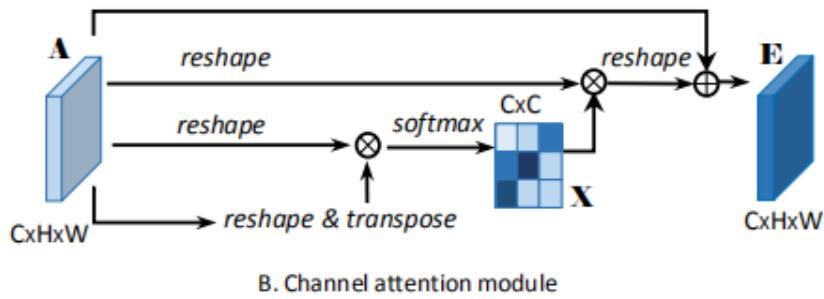


图 2-7 通道注意力结构

相比于 FCN、Unet 模型提取特征的方式，这样的方式有利于模型捕捉更多的局部特征和上下文信息。但是矩阵相乘的方式会极大增加运算时间。

2.2 零样本语义分割

大部分深度学习模型都是通过使用有标签的数据集以有监督的方式训练的。这类深度学习模型主要完成在训练集中已有类别的分类任务。然而在现实场景中，任何模型都不可能学习过所有类别的样本。对于训练集中未包含类别的分类任务，这些模型的识别效果相对较差。零样本学习就是为了解决这一问题而出现的研究

方向^[61-63]。该类方法假设是训练实例所涉及的类别与测试集中所涉及的类别不完全相同或者完全不相同。模型需要学会知识迁移的能力——根据学习过的类别样本获得先验知识，然后利用这些知识去识别一种未知类别。

如图 2-8 所示，该图描述了零样本学习的一种方式。图左上方标记为可见类别的图片集合表示训练集，图左下方标记为不可见类别的图片则表示测试集。在训练阶段，模型需要使用可见类别数据和相对应的语义属性数据进行有监督训练。监督训练使得模型建立语义属性和类别视觉特征的映射关系。在预测阶段，将不可见类别语义属性和图像作为模型的输入从而获得分类结果。如图 2-8 所示，首先模型从马的图片中学习视觉特征；然后将马的语义特征与其视觉特征进行关联学习；最后模型将斑马的语义属性和马的视觉特征相结合进行推理完成识别任务。

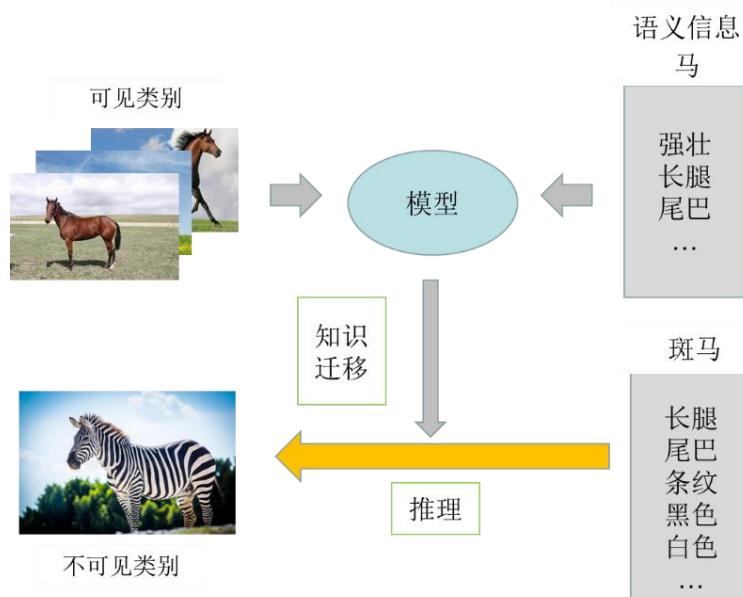


图 2-8 零样本学习思路

目前，零样本语义分割有两类研究方法：基于嵌入的方法和基于生成模型的方法。基于嵌入的方法的主要目标是学习一个映射函数——将语义属性空间和图像视觉空间同时映射到一个公共空间。这个公共空间可以是语义属性空间、也可以是原本图像的视觉特征空间或者重新学习的新空间。

比如，使用语义属性空间作为公共嵌入空间的方法。在训练期间，模型通过使用已知类的数据集学习从视觉空间到语义空间(即词向量/语义嵌入)的投影函数。在测试期间，首先将含有未知类别的图像输入至已经过训练的模型中并获得相应的语义特征向量；然后在语义属性空间中使用相似性搜索算法寻找与该语义特征向量相似的语义属性向量；最后将相似的语义属性向量所对应的标签作为输入图

像的输出标签。

如图 2-9 所示，该图是由 Xian 等^[64]提出的基于嵌入的零样本语义分割网络。首先将输入图像通过语义分割特征提取网络从而得到维度为 $a \times b \times d_w$ 的图像视觉语义特征向量。然后将该特征向量作为语义投影操作的输入，语义投影操作将视觉空间转化为语义空间。具体操作是将视觉语义特征向量与 W_{tr} 矩阵进行运算。最后，生成一个维度为 $a \times b \times |S|$ 的新特征，并且使用该特征进行语义分割的损失计算。

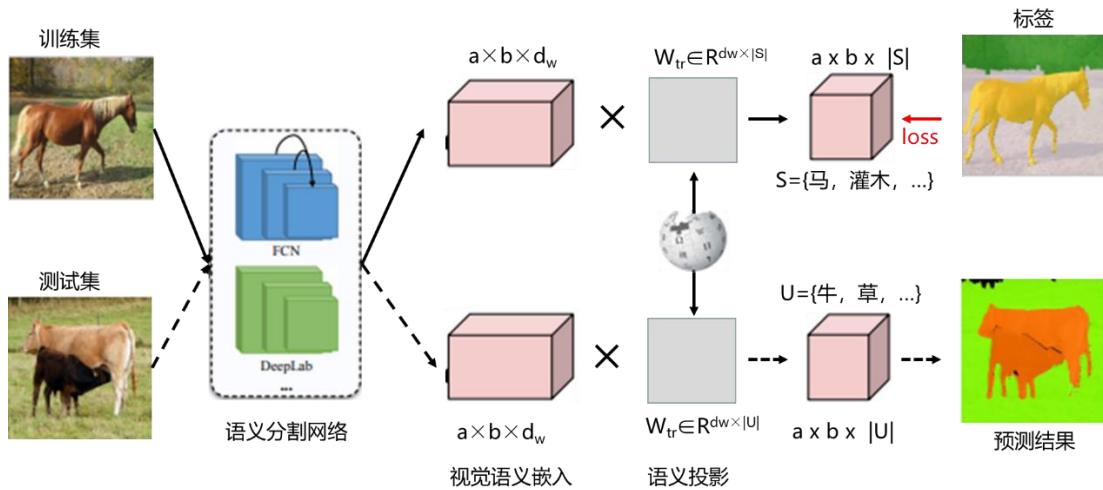


图 2-9 基于嵌入的零样本语义分割网络

相比于基于生成的方法，基于嵌入的方法存在域移位和偏差的问题。模型会在预测过程中偏向于用已知类别进行类别预测，因为模型的投影函数仅仅是通过使用可见类别训练学习来的。此外，在预测过程中，因为模型没有使用未知类别的样本进行训练，所以模型学习的投影函数不一定能完全将未知类别的视觉特征正确地映射到相应的公共空间上。为了克服这个缺点，基于生成的方法模型应运而生。基于生成的方法模型主要是生成未知类别的视觉特征从而使得识别或分割模型可以在已知类和未知类上进行训练。其中，生成方法旨在使用语义属性生成未知类别的视觉特征。

图 2-10 是一个基于生成的零样本学习示意图。在训练生成模型阶段，首先将原始图像输入至特征提取网络中获取图像视觉特征向量；然后将属性向量作为生成器的输入，生成器的输出则是一个合成的伪特征向量；最后将视觉特征向量和伪特征向量都作为判别器的输入，判别器则判断输入特征的真假。重复以上步骤训练生成器和判别器直至模型损失稳定。伪特征向量会逐步逼近由特征提取网络所提取的真实视觉特征向量。在训练图像识别模型阶段，先将未知类别的属性向量输入生成模型获得未知类别的视觉特征；然后将已知类别的视觉特征和未知类别的视

觉特征一起作为训练数据训练识别模型。

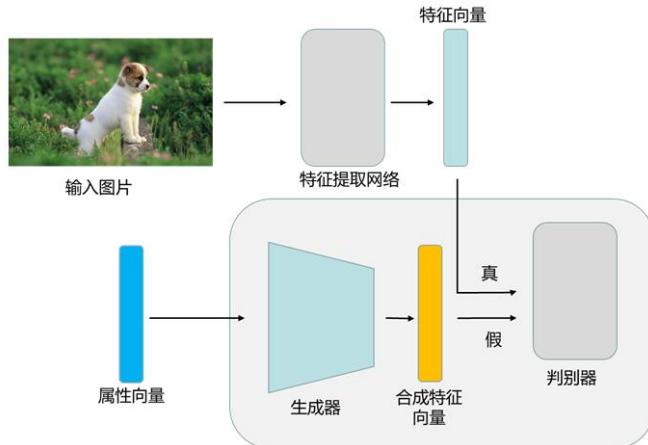


图 2-10 基于生成的零样本学习示意图

2.3 语义分割评价指标

随着语义分割研究的不断深入，越来越多的指标被提出来用于衡量模型准确率。总体而言，语义分割评价指标基本都是交并比、像素准确率及它们的变体。下面介绍常用的四种评价指标^[65]——平均交并比、像素准确率、像素平均准确率和加权平均交并比。

(1) 平均交并比

平均交并比是用来描述预测正确区域与目标区域的比值大小，其公式如(2-5):

$$MIOU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN + FP + TP} \quad (2-5)$$

其中， k 为类别数量， TP 表示实际值为正例且预测结果为正例， FN 表示实际值为正例且预测结果为负例， FP 表示实际值为负例且预测结果为正例。

(2) 像素准确率

像素准确率是用来描述正确预测的类别像素数量与图片标签全部像素数量的比值，其公式如 (2-6):

$$PA = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \quad (2-6)$$

其中， k 表示类别数量， p_{ii} 表示真实像素类别为 i 的像素被预测为类别 i 的数量， p_{ij} 表示真实像素类别为 i 的像素被预测为类别 j 的数量。

(3) 像素平均准确率

像素平均准确率指标先分别计算每个类被正确识别的像素数量，然后累加求平均，其公式如(2-7):

$$mPA = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij}} \quad (2-7)$$

其中， k 表示类别数量， p_{ii} 表示真实像素类别为 i 的像素被预测为类别 i 的数量， p_{ij} 表示真实像素类别为 i 的像素被预测为类别 j 的数量。

(4) 加权平均交并比

加权平均交并比指标根据每个类别出现的次数对每个类的交并比进行加权求和，其公式如(2-8):

$$FWIoU = \frac{1}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \sum_{i=0}^k \frac{\sum_{j=0}^k p_{ij} p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (2-8)$$

其中， k 表示类别数量， p_{ii} 表示真实像素类别为 i 的像素被预测为类别 i 的数量， p_{ij} 表示真实像素类别为 i 的像素被预测为类别 j 的数量。

2.4 本章小结

本章主要包含两部分内容：语义分割技术基础和零样本学习技术基础。首先本章先后阐述了语义分割的定义、语义分割的经典模型及其蕴含的设计思路；然后本章引入零样本学习的概念并介绍基于零样本学习的方法和基于零样本学习的语义分割技术思路；最后本章介绍语义分割的评价指标。以上的详细介绍为本文接下来算法章节提供理论基础。

第三章 基于多视角注意力机制的语义分割模型

影响语义分割模型准确率的因素包含特征学习阶段的图像分辨率降低以及图像特征上采样阶段的特征对齐。为了解决上述问题，本章分别提出了多视角特征提取结构和多尺度特征融合方法。本章节首先介绍整体模型框架，然后分别介绍模型框架中具体的算法细节，最后对该网络进行实验以及结果分析。

3.1 多视角注意力机制语义分割模型

为了解决局部特征与全局特征对不齐、局部特征提取模式单一、类别形状差异大以及部分类别像素占比低的问题，本章提出一种新的语义分割模型。如图 3-1 所示，该模型分别包括全局特征提取模型、多视角特征提取网络以及多尺度特征融合网络。全局特征提取模型包含少量卷积操作，其目的是从原始图像中提取全局特征。多视角特征提取网络包含多视角结构和注意力结构，其目的是尽可能地学习图像中不同的类别特征并且加强这些特征的联系。注意力机制的作用不仅是加强图像特征中同一类别像素之间的联系，而且也能够使得模型更加注意类别像素占比较低的类别。多尺度特征融合网络主要采用拼接、卷积、上采样等操作以及使用边缘损失函数完成从分辨率低的图像特征到分辨率高的原图像还原任务。

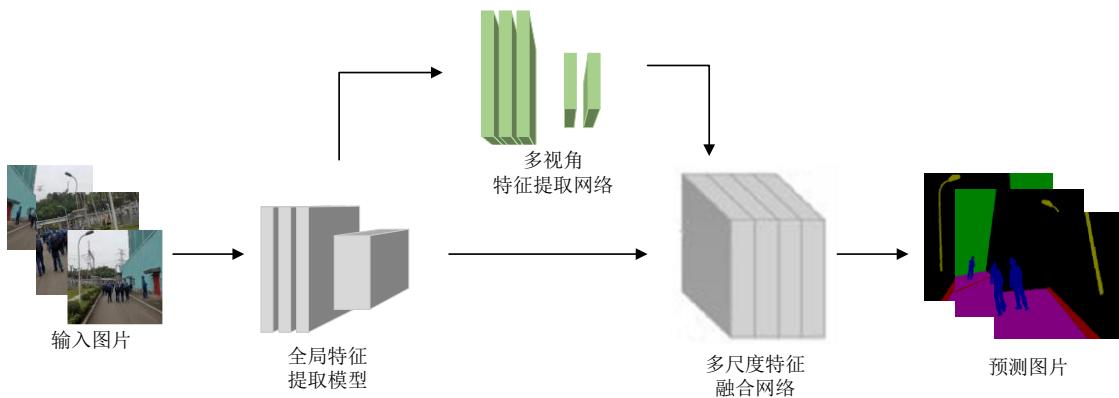


图 3-1 多视角注意力机制语义分割模型

本模型采用 Googlenet^[66]作为基础模型。如表 3-1 所示，该表是 Googlenet 的模型结构表，这个模型共计 22 层卷积，其中包含卷积层、池化层、线性层、Inception 层以及 softmax 层。Inception 层作为该模型最重要的一层，其结构如图 3-2 所示。每一次 inception 操作具体为：将前一层卷积层的结果分别输入 4 种不同参数的卷积层获得四种特征结果，然后将这四种特征结果拼接成一个特征并作为下一层的

输入。Inception 结构使用了 1×1 卷积核，一方面，在尽可能减少参数的同时增加模型的深度；另一方面， 1×1 卷积核有利于把相关性高、在同一位置空间但在不同通道的特征结合起来。

表 3-1 Googlenet 模型结构

Type	Patch size/stride	Depth	#1x1	#3x3 reduce	#3x3	#5x5 reduce	#5x5	Pool proj
Convolution	$7 \times 7 / 2$	1						
Max pool	$3 \times 3 / 2$	0						
Convolution	$3 \times 3 / 1$	2		64	192			
Max pool	$3 \times 3 / 1$	0						
Inception(3a)		2	64	96	128	16	32	32
Inception(3b)		2	128	128	192	32	96	64
Max pool	$3 \times 3 / 2$	0						
Inception(4a)		2	196	96	208	16	48	64
Inception(4b)		2	160	112	224	24	64	64
Inception(4c)		2	128	128	256	24	64	64
Inception(4d)		2	112	144	288	32	64	64
Inception(4e)		2	256	160	320	32	128	128
Max pool	$3 \times 3 / 2$	0						
Inception(5a)		2	256	160	320	32	128	128
Inception(5b)		2	384	192	384	48	128	128

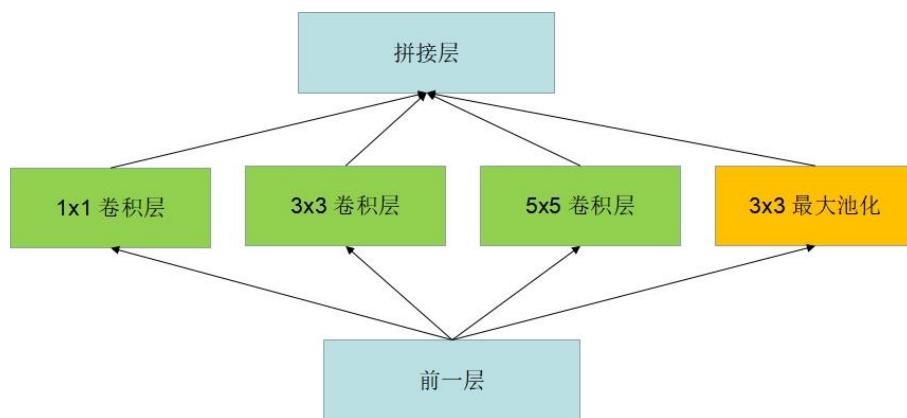


图 3-2 inception 结构

3.2 多视角注意力结构

3.2.1 多视角学习

多视角注意力结构是多视角特征提取网络的核心。如图 3-3 所示，多视角注意力结构是由多视角结构和注意力机制组成。其中多视角结构主要目的是从全局图像特征中多角度地学习局部特征，而注意力机制主要目的是捕捉从多视角结构中获得局部特征图中同一类别的像素，并且加强像素之间的联系、使得模型关注这些像素。如图 3-3 所示的多视角注意力结构，其中 $A \in R^{h*w*c}$ 代表输入的全局特征——全局特征提取模型的输出， $B_i \in R^{h*w*c_i}$ 是经过多视角结构所得出来的局部特征结果。它包含了描述全局特征 A 的多种局部特征， B_i 经过注意力机制得到输出 $C_i \in R^{h*w*c_i}$ ，最后将 C_i 进行拼接形成全局特征 A 的所有局部特征。

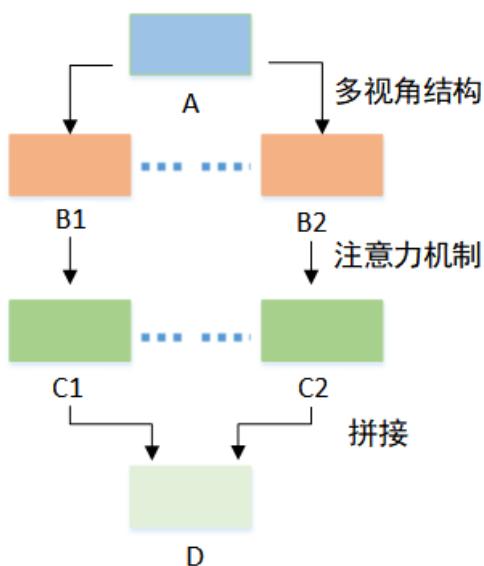


图 3-3 多视角注意力结构

如上段所述的多视角局部特征，这些局部特征的提取主要依靠空洞空间金字塔结构。空洞空间金字塔结构如图 3-4 所示，对于同一个特征，使用带有不同空洞比率的相同大小卷积核（如 $3*3$ 卷积核）会得到不同的特征。基本原理是不同感受野会提取不同的图像特征。改变感受野大小的方式有两种：一是直接调节卷积核大小，二是使用不同扩张率大小的空洞卷积。空洞卷积如图 3-4 所示，从左到右分别是扩张率为 1、2、3 的卷积核。从感受野角度来看，以 $3*3$ 卷积核为例子，扩张率为 2 的卷积核对应的感受野和 $5*5$ 卷积核的感受野相同，扩张率为 3 的卷积核对应的感受野与 $7*7$ 卷积核的感受野相同。

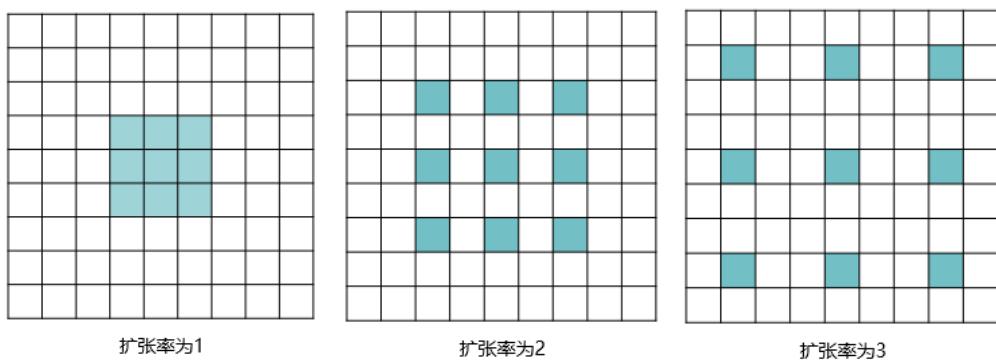


图 3-4 空洞空间金字塔结构

如图 3-4 所示,如果空洞空间金字塔结构选用长宽相同的卷积核,那么该卷积操作将以等长宽比例的方式提取图像特征。这类卷积核容易从长宽比例相差不大的类别中提取出特征。根据变电站场景的特殊性——包含长宽比例过大的类别,如电线杆、避雷针等。如果一个类别的长宽比例相差过大,经过一系列空洞空间金字塔结构之后,该类别可能会逐步从图像特征中消失或者退化为几个像素点。这样不利于上采样阶段图像的还原。因此,如表 3-2 所示,本章节模型将卷积核的长宽大小设计为不同比例,为了从图像特征中捕捉不同比例大小的类别特征。本章节模型一共选取 13 种不同类型的卷积核进行特征提取。首先,这里没有选取较大的卷积核,如 $5*5$ 卷积核、 $7*7$ 卷积核和 $11*11$ 卷积核。其主要原因有以下两点:一是采取卷积核大小较大的卷积核会增加计算复杂度;二是较大卷积核的使用会以减少图像特征分辨率的方式来减少网络模型的深度。但是网络模型越深,网络模型效果越好。此外这里也没有选择长宽相同的卷积核,其原因是前一步提取特征已经使用这种方式的卷积核。其次,在卷积核扩张率的选择方面,本章节以没有公约数为原则进行选择(不包含 1 的公约数)。若选用含有公约数扩张率的卷积核,则会出现图像特征的像素遗漏问题。

表 3-2 卷积核和扩张率选取

卷积核 扩张率	$1*1$	$1*3$	$3*1$	$3*5$	$5*3$
1	1	1	1	1	1
2	-	2	2	3	3
3	-	1	1	1	1

如图 3-5 所示,(a) 为选择扩张率为 2 的卷积核进行堆叠的效果,而(b) 图分别为选择扩张率为 1、2 和 3 的卷积核的效果。从图中可以看出,相比于使用

扩张率为 2 的卷积核，使用无公约数的扩张率卷积核能够捕捉更大的图像特征，即更大的感受野。扩张率含有公约数则会导致出现感受野不连续的问题。最后，扩张率为 1 和 3 的卷积核的数量设置为 1，而扩张率为 2 的卷积核的数量设置为 2 或者更多。在叙述其原因之前，先介绍一下卷积分辨率计算公式，见式（3-1）：

$$o = \left\lceil \frac{i + 2p - k - (k-1)(d-1)}{s} \right\rceil + 1 \quad (3-1)$$

其中， i 表示输入图像特征分辨率， p 表示填充值， k 表示卷积核大小， d 表示卷积扩张率， s 表示卷积步长。

由于卷积和池化操作导致特征图分辨率降低。假设特征图分辨率为 76*76。在这样的分辨率下，一些类别可能包含的像素点较少，比如 10 个像素点。如果选取扩张率为 3，那么根据卷积分辨率计算公式 3-1，计算扩张卷积之后的该类别分辨率为 6，而选用扩张率为 2 的卷积核，则分辨率为 8。因此，选择较多的扩张率为 2 的卷积核是为了保留更多的类别信息。

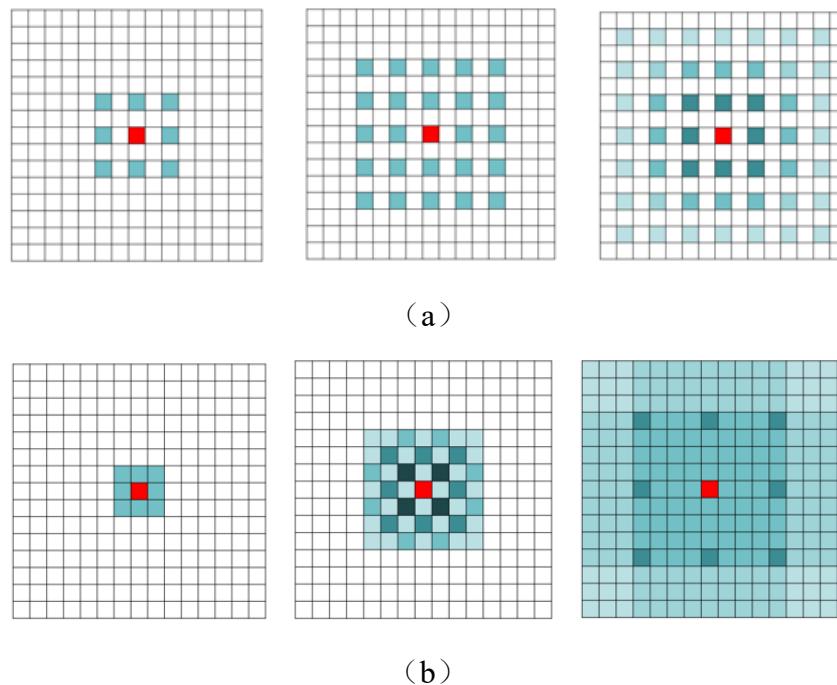


图 3-5 扩张率卷积核感受野

3.2.2 注意力网络结构

注意力机制的本质是使模型关注目标区域从而提高模型识别关注目标的精度。本文在第二章介绍了一种基于双重注意力机制的语义分割模型。该模型分别采用

通道注意力结构和位置注意力结构。其中通道注意力结构捕获特征图中通道之间的依赖关系，而位置注意力结构主要捕获特征图像素之间的依赖关系。虽然该模型添加的两种通道注意力结构能够提高了语义分割的准确率，但却极大增加了模型训练和预测的时间。针对该问题，本章节设计一种注意力网络结构。该注意力网络结构主要作用是捕获局部特征图中同一类别的像素信息，并且该网络结构的时间复杂度较低。

如图 3-6 所示， $A \in R^{h*w*c}$ 表示输入图像特征， $B \in R^{h*w*c}$ 表示经过卷积操作之后的图像特征， $C \in R^{h*w*c}$ 表示经过归一化操作（Batch Normalization, BN）的图像特征， $D \in R^{h*w*c}$ 表示基于注意力机制的图像特征，见式 (3-2)：

$$d_{ij} = \frac{1}{1 + \exp(-C_{ij})} \quad (3-2)$$

其中， C_{ij} 表示图像特征中第 i 行第 j 列的值， d_{ij} 表示输出。

E 表示注意力结构的输出，见式 (3-3)：

$$E_{ij} = d_{ij} * A_{ij} \quad (3-3)$$

其中， d_{ij} 表示图像特征中第 i 行第 j 列的值， A_{ij} 表示原始图像特征中第 i 行第 j 列的值。

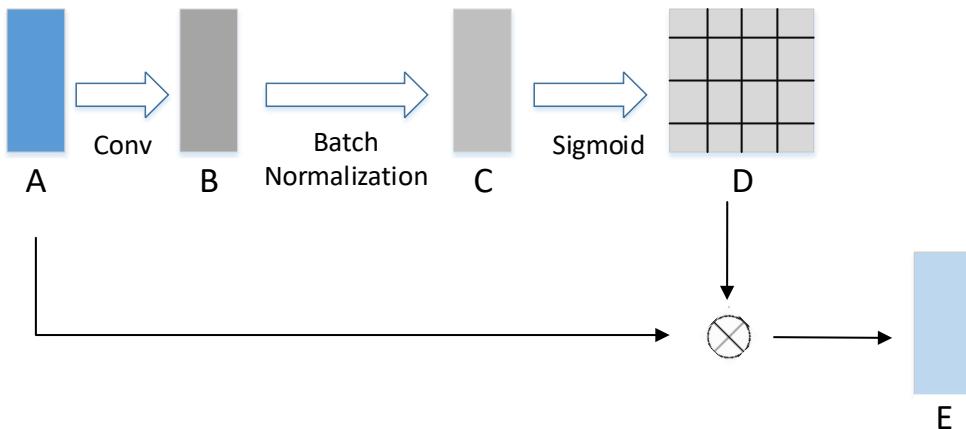


图 3-6 注意力结构

如图 3-6，注意力结构包含两条分支：原始的特征图分支（路径 A 到 E）和注意力分支（路径 A 到 D）。注意力分支较为简单——仅包含卷积层、BN 层和激活层。卷积层是为了进一步提取同一类别的特征；BN 层的主要作用是：对图像特征 B 进行归一化；使注意力结构训练更加稳定；抑制由 sigmoid 函数所导致的梯度消失。激活层选用 sigmoid 函数是为了区分特征图中的像素类别。即属于同一类的像

素靠近 1，而不属于同一类的像素靠近 0。

3.3 多尺度特征融合方法

一般语义分割模型的局部特征提取结构会选用较深的网络，比如 ResNet、Vgg、InceptionNet^[66]等。尽管这些深层次网络能够很好地提取图像的特征信息，但是它们也会给语义分割带来一些问题。深层次的网络拥有较多的卷积层和池化层，不断的卷积和池化操作会使得图像的分辨率降低，比如降低 16 倍或者更多。这会导致两个问题：一是如果一个类别原本形状比较小或者像素占比低，那么它在这一系列操作之后只能剩下少量的像素点或者直接消失；二是即便有很多类别的特征被保留，但是其边缘信息也变得残缺不全。这两点问题会影响语义分割模型的预测精度。因此，语义分割模型需要使用特征融合的方法。

特征融合的方式有很多，比如 FCN 模型先将局部特征上采样的结果和全局特征进行拼接再进行上采样；Deeplabv3^[67]模型对局部特征进行空洞空间金字塔池化操作来获得多个局部特征，通过多个局部特征融合的方式解决图像特征分辨率低而带来的某些类像素点稀少或者消失的问题；Unet 模型则是不断重复以下阶段直至上采样图像特征分辨率为原图像分辨率——首先上采样局部特征，然后将上采样的结果和对应的全局特征相融合。上述特征融合的方式都存在问题——局部特征和全局特征对齐问题。若某类别位于原始图像的位置 A，局部特征上采样之后该类别位于图像的位置为 B。上采样操作一般采用最近邻、线性内插等方法。因为这样的上采样操作并不是卷积和池化的逆操作，所以上采样的结果无法完全还原所有类别的原始位置。即位置 B 相对位置 A 会产生一定的偏移。

针对上面的问题，本章节提出一种多尺度特征融合的结构，如图 3-7 所示。该结构的设计思路是通过边缘损失调整局部特征使其能够和全局特征对齐，并且逐步进行图像对齐和上采样操作。局部特征 C 是经过本章所提出的多视角注意力结构所得到的图像特征，包含多种类别的特征信息和边缘信息。通过监督训练使多视角注意力结构能够尽可能地关注单一类别的全部像素。首先，局部特征 C 经过上采样操作生成局部特征 B，将局部特征 B 进行上采样得到结果图 C，对结果图 C 进行边缘损失的监督训练。对结果图 C 进行边缘损失的监督训练能够矫正类别像素位置的偏移。然后，将矫正的局部特征 B 和全局特征 B 进行卷积和上采样操作生成分辨率较高的局部特征 A。同理，对局部特征 A 进行上采样得到结果图 B 并且进行边缘损失计算。最后，将局部特征 A 和全局特征 A 进行卷积和上采样操作得到输出结果 A。该结果输入交叉熵损失进行监督训练。

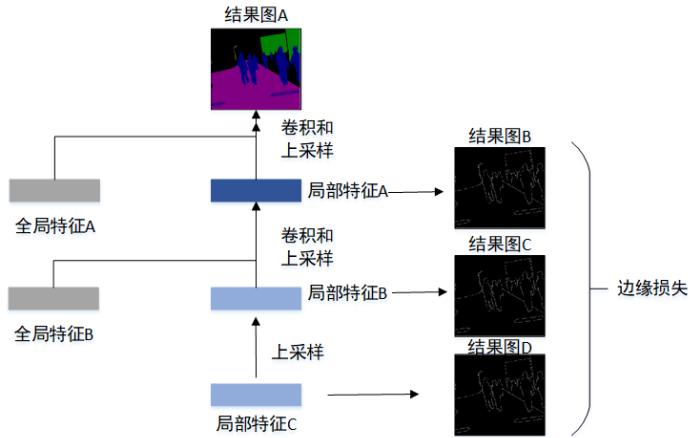


图 3-7 多尺度特征融合结构

交叉熵损失公式如 (3-4):

$$l_{cls} = -\sum_{i=1}^c p_i * \log(q_i) \quad (3-4)$$

其中, c 表示类别的数量, q_i 表示每个像素被预测为第 i 个类别的概率, $p_i \in \{0, 1\}$ 表示每个像素是否属于第 i 个类别。

边缘损失公式如 (3-5):

$$L_{margin} = \sum_H \sum_W \sum_C (\hat{y} - y)^2 \quad (3-5)$$

其中, H 、 W 、 C 分别表示图像特征的长、宽和通道数量, $\hat{y} \in \{0, 1\}$ 表示像素点是否为类别边缘点, y 表示预测某像素点是否是边缘点。

3.4 实验结果

3.4.1 实验数据

随着语义分割技术的研究不断加深, 越来越多的语义分割数据集开始被制作和公开, 比如(Cambridge-driving Labeled Video DataBase, CamVid)^[68,69]、KITTI^[70]、Cityscape^[71]等。然而基于特殊场景的语义分割数据集凤毛麟角, 如变电站场景的数据集。因此, 本论文算法模型所涉及的变电站数据集均为自主采集和制作。

该数据集通过 labelme 工具将其标注为 17 个类别——背景、车行道边缘线、房子、道路路灯、人、公路、隔离开关、公告牌、草坪、电线杆、避雷针、减速带、均压环、石头、圆桶、方桶、楼梯。其中草坪包括人工草地以及道路旁的植物; 石头表示电线杆、避雷针等物体下方的基石; 而其他未被分类的物体均为背景。如图

3-8 所示, (a) 是变电站原始图片, (b) 是经过标签化后的图片。其中不同颜色代表不同类别, 如黄色代表道路路灯、紫色表示公路、蓝色代表人、右上方深绿色表示房子等。具体类别标签如图 3-9 所示。

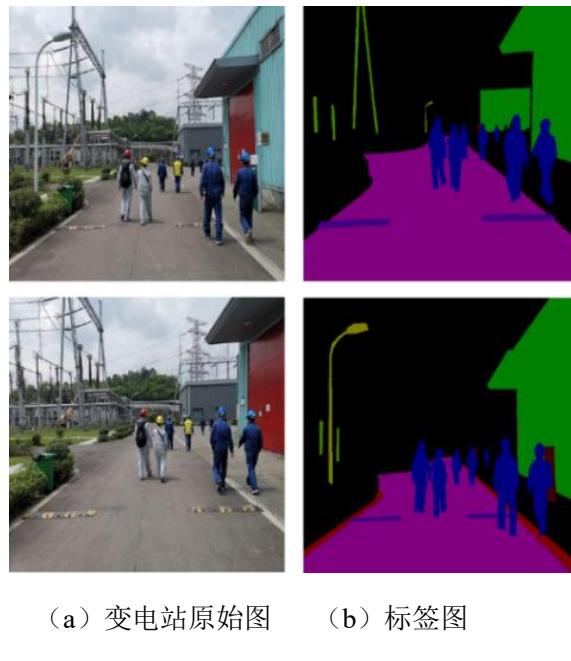


图 3-8 变电站数据集

类别	P模式像素值	颜色
背景	0	[Black]
车行道边缘性	1	[Dark Red]
房子	2	[Dark Green]
路灯	3	[Yellow-Green]
人	4	[Dark Blue]
公路	5	[Purple]
隔离开关	6	[Teal]
...		
楼梯	17	[Brown]

图 3-9 变电站数据集类别标签

如图 3-10 所示, 由于变电站数据集规模相对较小, 本论文采取一些数据增强的方式进行数据集扩充, 比如旋转、对称、裁剪等方式。其中 (a) (c) 分别是旋转和裁剪之后的变电站图片, (b) (d) 是对应的标签。

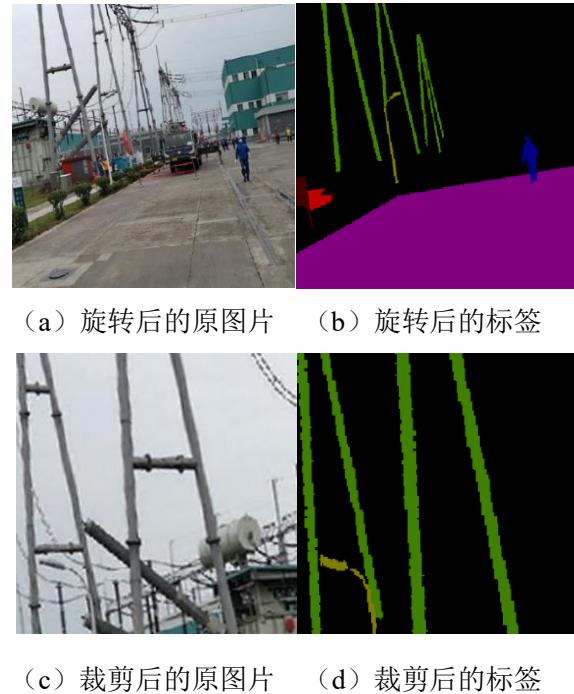


图 3-10 增强后的变电站数据集

3.4.2 实验设定

本章所设计的模型属于深度学习模型，需要强大的软件、硬件支持。在硬件方面，该模型在 Ubuntu 平台上进行实验。在实现方面，该模型使用 python 语言实现并采用 keras 框架；在数据处理与分析方面，我们使用 numpy、matplotlib、pandas 等 python 环境下的第三方包。平台具体参数如表 3-3 所示：

表 3-3 实验环境配置

名称	具体配置情况
内存	DDR4-3200 帧
主板	Z270-AR
硬盘	500GB SSD+2T HDD
CPU	Intel i7-9700k
GPU	GTX 2080Ti 11BG 显存
操作系统	Ubuntu 18.04
CUDA	9.1
Python	3.7
Keras	2.2.4

在该模型中，学习率设置为 0.001，边缘损失超参数 α 设置为 1，训练批次大小设置为 8，采用 Adam^[72]优化器迭代训练 120 次，其中基础学习率为 0.001、 β_1 为 0.9、 β_2 为 0.8。使用经过预训练的 InceptionNet 作为基础模型，上采样层采用线性内插法，模型输入尺寸大小设置为 608*608。对比试验中的各类模型训练设定如下：FCN——训练批次大小为 8，采用 Adam 优化器迭代训练 150 次，其中基础学习率为 0.003、 β_1 为 0.825、 β_2 为 0.99；Deeplabv3——训练批次大小为 4，采用 Adam 优化器迭代训练 130 次，其中基础学习率为 0.001、 β_1 为 0.825、 β_2 为 0.99；Bisenet^[73]——训练批次大小为 3，采用 RMSprop 优化器迭代训练 85 次，其中基础学习率为 0.001、衰减率 rho 为 0.9。根据以上训练细节描述，实验模型训练之后均已收敛。特别地，FCN、Deeplabv3、Bisenet 模型都未针对变电站数据集有相应的优化。为了对比不同的金字塔池化结构在变电站数据集上的效果，Deeplabv3 中的金字塔池化结构也未有相应调整；FCN 模型具体采用 FCN-16s；Bisenet 模型中超参数 α 设置为 1。

3.4.3 实验结果与分析

本章节实验包含 FCN、Deeplabv3、Bisenet 与本章节模型在变电站数据集上的实验结果。实验的评价标准为平均交并比（Mean Intersection over Union, MIOU）。所有实验模型均采用了特征融合的方法。FCN 将局部特征上采样的结果与全局特征融合；Deeplabv3 使用空洞空间金字塔池化进行局部特征提取并将提取的特征进行融合；Bisenet 则是将不同卷积层的特征结果进行融合；本章节模型将多视角特征提取的局部特征和全局特征进行融合。实验结果如表 3-4 所示，最好的 IOU 和 MIOU 结果在表中用加粗表示。本章所提出的基于多视角注意力机制语义分割模型不仅在 MIOU 方面，而且在 IOU 方面都高于其他三种模型。具体而言，在隔离开关一栏中，相比于 FCN 模型，其余模型的预测准确率都有较大的提升。在 FCN 逐步提取局部特征的过程中，因为隔离开关在图像中的长宽比例大，所以 FCN 不能完全提取出隔离开关的特征——图像特征模糊。同时，在特征融合的阶段，FCN 采用直接将局部特征和全局特征融合的方式。这会使得全局特征和局部特征无法完全对齐从而降低预测准确率。Deeplabv3 模型因采用空洞空间金字塔池化结构而能够捕捉更多类别的局部信息，这使得该模型的预测准确率排在第二。但因为该模型采用长宽相同的卷积核的空洞空间金字塔结构，所以在捕捉类似于隔离开关的类别上有些不足。而本章节的模型设计的多视角结构可以很好得提取这种类别的局部特征。

表 3-4 各模型在变电站数据集的实验结果

指标	方法类别	FCN	Deeplabv3	Bisenet	Ours
IOU	车行道边缘线	0.56	0.66	0.65	0.72
	建筑	0.39	0.56	0.63	0.78
	路灯	0.09	0.19	0.20	0.33
	人	0.49	0.63	0.63	0.71
	公路	0.93	0.94	0.94	0.96
	隔离开关	0.09	0.29	0.22	0.41
	公告牌	0.18	0.33	0.31	0.57
	草坪	0.62	0.69	0.72	0.79
	电线杆	0.19	0.33	0.36	0.51
	避雷针	0.31	0.47	0.51	0.69
	减速带	0.19	0.44	0.39	0.56
	均压环	0.16	0.34	0.39	0.75
	石头	0.32	0.40	0.46	0.63
	圆桶	0.14	0.42	0.71	0.82
	方桶	0.44	0.48	0.68	0.80
	楼梯	0.45	0.48	0.60	0.72
MIOU		0.346	0.478	0.525	0.671

在路灯一栏中，尽管本章节提出的模型具有最高的准确率，但是所有实验模型的交并比均小于其他类别的交并比。导致这种情况的原因是：如图 3-11 所示，路灯类别在实验数据集中出现的次数较小且路灯在图像中像素占比不高使得模型对该类别的学习存在欠拟合的现象。这一点从隔离开关和路灯的实验结果中可以看出。隔离开关和路灯形状相似——具有相似的长宽比例，但是隔离开关在数据集中出现的次数要多于路灯出现的次数。因此，Deeplabv3 模型、Bisenet 模型和本章节提出的模型在隔离开关的预测精度上都有一定的提升。如表 3-4 所示，在均压环一栏中，FCN、Deeplabv3 以及 Bisenet 模型识别率比较低，而本章节所提出的模型识别率较高。对于均压环的像素预测，本章节模型与其他实验模型的差距较为突出。造成这实验结果的主要原因是均压环特殊的形状使得通常的卷积操作提取局部特征相对困难，而采用 Deeplabv3 的空洞空间金字塔池化结构所捕获的局部特征的分辨

率较低。这就导致在特征融合阶段或者上采样阶段，模型对该类别的预测变得模糊。本章节提出的模型从多个角度对该类别的局部特征进行提取，能够较大程度为特征融合阶段提供完整的均压环局部信息，提升该类别预测的准确率。

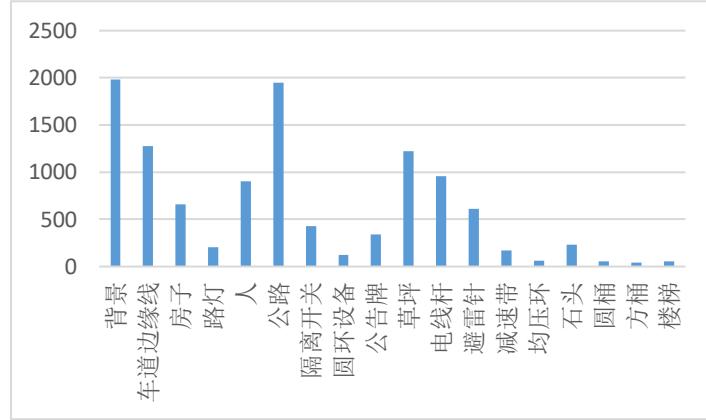


图 3-11 类别数量图

公告牌的预测结果在各个模型上也相差较大，并且公告牌这一类别出现的频率低以及其像素的数量占整张图片的总像素数量较低。但是相比于路灯和隔离开关，所有模型的预测精度在公告牌上的预测精度都有提升，其原因在于：公告牌的长宽比例比前者小，使得实验模型容易通过自己的特征提取方法提取出相关的特征信息。而本章节模型对该类别的预测准确率高于其他实验模型的主要原因是多视角注意力结构能够很好地捕捉这一类型的类别特征。

根据上述实验数据以及实验分析，本章节设计的多视角注意力结构不仅能够提取长宽比例较小的类别的局部特征，而且能够提取长宽比例较大的类别的局部特征。在面对像素占比低的类别时，相比于其他模型，本章节模型的预测精度有所提高。

图 3-12 是语义分割实验的具体结果，从第一行图片来看，相比于 deeplabv3，Bisenet 和本章节的模型能够很好的预测出避雷针；但 deeplabv3 和 Bisenet 模型对于一些像素占比低的区域预测效果没有本章节提出的模型好，比如避雷针下方的人。其原因是本章节提出的模型具有注意力结构，该结构具有针对性的对每张图像中的每一个类别的所有像素进行关注。因此，即使一些类别在图像中的像素较少，也能被关注从而在最后上采样阶段中识别出来。第二、三行图片的内容主要是变电站的电线杆等类别，它们的特征是长宽比例较大。在预测这些类别的时候，因为 deeplabv3 使用了空洞空间金字塔池化结构，所以预测效果比 Bisenet 好。但由于 deeplabv3 主要采用长宽相同的卷积核，因此其不能很好的预测这种长宽比例较大的类别。相比之下，本章的模型采用长宽不同的卷积核能较好地提取电线杆等类别

的特征。同理，在第四行和第五行图片里的长宽比较大的类别——路灯，本章节提出的模型能够很好地显示出大体的轮廓。

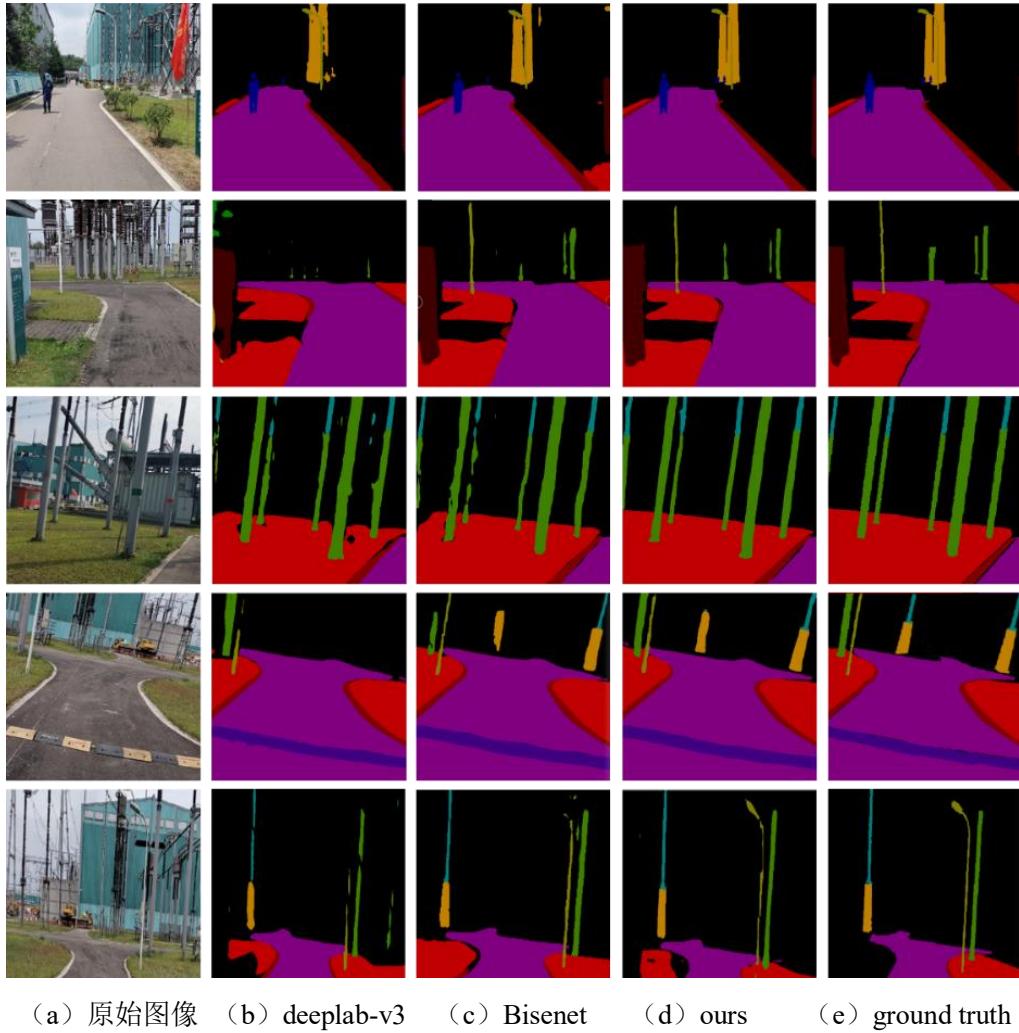


图 3-12 语义分割结果

在上述与其他模型进行实验对比之后，接下来介绍关于注意力机制和空间金字塔卷积具体作用的实验结果。消融实验(ablation experiment)结果如表 3-5 所示，本次实验主要将本章模型拆分为以下几个部分：不含有注意力机制的模型、未含有多视角但含有注意力机制的模型、含有注意力机制且仅含有 $1*3$ 卷积核的多视角模型、含有注意力机制且仅含有 $3*5$ 卷积核的多视角模型、含有注意力机制且含有 $1*3$ 以及 $3*5$ 卷积核的多视角模型。如表 3-5 所示，首先，将不含注意力机制的模型与其他含有注意力机制的模型进行对比。不含注意力机制的模型在公路、避雷针、隔离开关、草坪这四个类别上面的识别率与其他模型相同，仅在建筑、路灯这两个类别略高于包含注意力机制的模型。这说明注意力机制能够捕捉同一类别像

素之间的关系。其次，从多视角结构角度分析，对于人、电线杆等易识别的类别，模型是否含有的多视角结构并没有影响。但是，在其他类别的识别任务中，相比于不含多视角结构模型，含有多视角结构的模型在大部分类别上都有较高的 IOU。

表 3-5 多视角和注意力机制消融实验

指标	方法类别	不含注意力机制	不含多视角结构	仅含 1*3 卷积核	仅含 3*5 卷积核	含 1*3 和 3*5 卷积核
IOU	车行道边缘线	0.71	0.69	0.74	0.71	0.73
	建筑	0.81	0.72	0.72	0.76	0.80
	路灯	0.35	0.32	0.30	0.32	0.34
	人	0.66	0.69	0.57	0.68	0.69
	公路	0.96	0.95	0.95	0.95	0.96
	隔离开关	0.43	0.38	0.38	0.40	0.43
	公告牌	0.51	0.53	0.54	0.50	0.55
	草坪	0.79	0.78	0.76	0.77	0.79
	电线杆	0.49	0.50	0.49	0.50	0.49
	避雷针	0.63	0.60	0.52	0.63	0.58
	减速带	0.51	0.60	0.53	0.60	0.56
	均压环	0.50	0.54	0.16	0.56	0.68
	石头	0.61	0.55	0.61	0.60	0.63
	圆桶	0.82	0.77	0.79	0.83	0.83
	方桶	0.73	0.80	0.25	0.82	0.81
	楼梯	0.68	0.68	0.67	0.69	0.72
MIOU		0.636	0.631	0.561	0.645	0.661

最后，从多视角采用的不同卷积核大小来看，因为 1*3 卷积核形成长条形的感受野，所以仅包含 1*3 卷积核的多视角模型能够很好地捕捉长条状的类别特征，比如车行道边缘线。值得注意的是，相比于其他模型，仅包含 1*3 卷积核的多视角模型的 MIOU 是最低的，低于没有使用注意力机制的模型。这说明如果数据集包含不同比例大小的类别，仅仅使用 1*3 卷积核只能从全局特征中提取出长条状的局部特征，而忽略全局特征中其他比例大小的局部特征。相比于 1*3 卷积核，3*5 卷积核形成长方形状的感受野，所以仅包含 3*5 卷积核的多视角模型对于长方

形状的类别有很好的辨识度，比如避雷针和减速带。仅含 3×5 卷积核的多视角模型的 MIOU 排在第二位。它没有与仅含 1×3 卷积核的多视角模型具有一样低的准确率，其原因是大部分类别都可以由 3×5 卷积核所形成感受野进行捕捉。相比于仅含 1×3 卷积核或者仅含 3×5 卷积核的模型，包含 1×3 卷积核且包含 3×5 卷积核的多视角模型既能够拥有长条状的感受野也能拥有长方形的感受野。在所有类别的识别任务中，该模型取得了一个很好的折中。因此，使用多视角注意力机制且含有不同卷积核的模型具有最好的识别效果。

图 3-13 是实验结果图。(a) 表示原始输入图片，(b) 表示模型中没有使用注意力机制，(c) 表示模型中使用注意力机制但没有使用多视角结构，(d) 表示模型同时使用注意力机制和多视角结构，(e) 表示原始输入图片对应的标签。

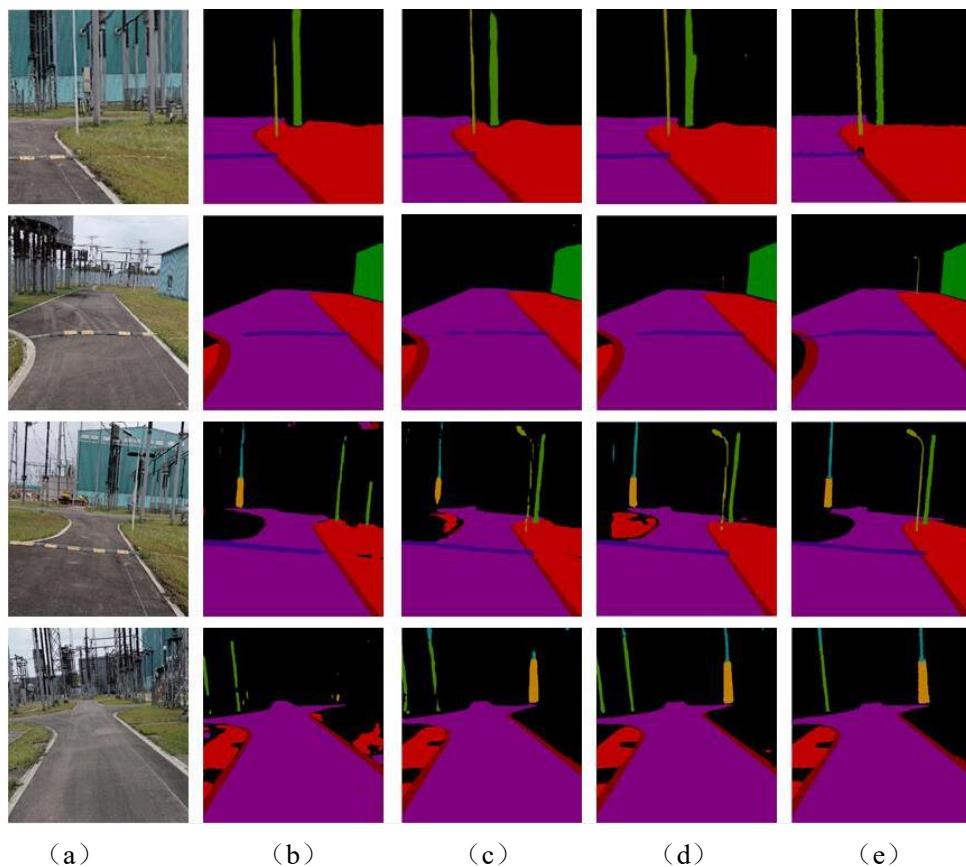


图 3-13 消融实验结果图

如图 3-13，从第一行的实验结果分析，在使用了多视角结构的情况下，相比于没有使用注意力机制的模型，含有注意力机制的模型能预测跟多的像素点。然而，在使用注意力机制的情况下，没有使用多视角结构的模型在电线杆的预测上有些许差异，但并不明显。总体来说，无论是多视角结构还是注意力机制都有助于提高模型准确率。从第二行的实验结果来看，没有多视角结构的模型预测的减速带像素

点并不连续。这可能是由于无多视角结构的模型在提取图像特征的过程中，存在分辨率降低导致类别像素丢失的问题。

基于上面对实验结果的分析，可以得出本章节模型包含的多视角结构通过调整感受野的大小来捕捉不同大小的类别进而提升预测精度，比如对电线杆、路灯类别的预测。注意力机制可以通过加强图像特征中同一类别的像素之间的联系进而提升类别像素预测的完整性。同时，将多视角和注意力机制结合能够进一步提升模型的精度。

本章节模型设计了多尺度特征融合模块来解决全局特征和局部特征的对齐问题。其中， α 是用来平衡边缘损失和分类损失的超参数。表 3-6 是探究超参数 α 对模型效果影响的结果。

表 3-6 超参数 α 对 MIOU 的影响

超参数 α 数值	MIOU
0.1	0.662
0.5	0.655
1	0.671
5	0.653

从表 3-6 可知，当 α 等于 1 时，MIOU 取最优值。即当边缘损失和分类损失权值相同时，模型效果最好。在其他的情况下，模型的 MIOU 稍微下降。当边缘损失比分类损失重要时，模型 MIOU 下降更多。

3.5 本章小结

本章首先给出基于多视角注意力机制的语义分割模型框架，然后分别介绍多视角结构、注意力结构以及多尺度特征融合网络。接下来，本章详细阐述这些结构的原理以及设计思路。最后，使用变电站数据集对该网络模型进行实验验证，并与其他模型进行对比分析，展示了本网络模型在变电站数据集上的性能。

第四章 基于零样本学习的语义分割模型

前面的章节详细介绍了基于多视角注意力机制的语义分割模型。该模型解决了变电站场景中的一些问题，比如全局特征和局部特征对齐问题、部分类别像素占比较低的问题。然而，在变电站巡检的任务过程中，巡检机器人可能会遇见一些未识别物，比如鸟、新设备等。因此，本章节采用零样本学习的方法，使得语义分割模型能够预测未识别物。本章节首先介绍基于零样本学习的语义分割模型框架，然后介绍自编码器、生成网络模型以及特征生成网络，最后对提出的模型进行实验对比和结果分析。

4.1 基于零样本的语义分割网络模型

在传统语义分割的任务中，在面对含有未识别物的图像作为输入时，语义分割模型无法生成相对应的图像特征或者生成错误图像特征。这都使得模型无法完成识别类别的任务。零样本语义分割网络主要解决模型如何识别没有见过的实例类别的问题。在变电站场景中，使用基于零样本学习的语义分割模型能够更好地辅助巡检机器人避障以及规划路径从而达到提高效率的目的。以训练的方式来分类，零样本学习方法分为两类：基于分类器的方法和基于实例的方法。基于分类器的方法主要通过一些方式训练分类器，使得分类器能够识别未见过的实例类别；而基于实例的方法可以生成大量的样本从而直接对分类器进行监督训练。因为目前没有公开的变电站数据集以及本章节所采用的变电站数据集规模小，这些因素严重影响基于分类器的方法模型效果，所以本章节提出一种基于特征合成的语义分割模型。该模型属于基于实例的方法。图 4-1 为基于零样本学习的语义分割网络。该模型除了拥有传统语义分割网络的部分，还拥有一种特征生成网络。特征生成网络以自编码器的方式来生成未知实例类别的图像特征。这些图像特征用于辅助语义分割网络完成识别任务。基于特征合成的方法本质在于如何利用已有的类别信息与未见过实例类别的部分信息（语义信息，属性信息等等）来生成图像特征。

在零样本学习中，尽管一些类别的图像样本无从得知，但是我们可以得到这一类别的信息描述或者这一类别的词向量信息。以类别鸟为例，虽然模型无法得知鸟的图像，但是模型却可以知道它具有的一些特征——翅膀、眼睛等等。以此为基础，特征生成网络的输入包含语义信息和来自于正态分布的随机样本点。特征生成网络包含一个自编码器框架和两个判别网络。自编码器的作用是将类别的语义信息进行有效压缩。但是这种压缩仅仅是对类别语义信息进行特征空间转换，而转换

之后的特征空间并不一定是该类别所对应的图像的视觉空间。因此，仅仅使用自编码器的方式并不能够很好地生成视觉空间的特征。鉴于此情况，特征生成网络分别包含两个判别网络。这两个判别网络主要的目的是使得自编码器压缩的特征尽量与图像的视觉特征相一致。

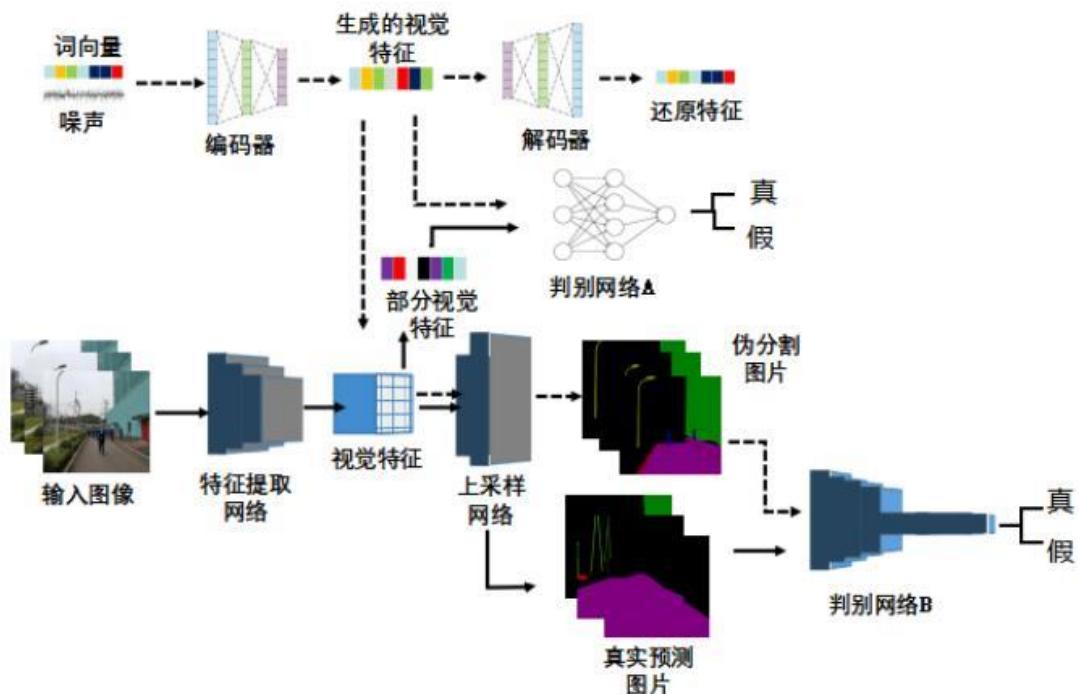


图 4-1 零样本语义分割网络

上述基于零样本的语义分割网络包含多种损失函数——分类损失、重建损失、对抗损失和边缘损失。其中分类损失如公式(3-4)，主要用于训练模型完成识别任务；边缘损失如公式(3-5)，主要用于解决上采样过程中的特征对齐问题；对抗损失在4.2节将会详细介绍，主要用于减小生成的视觉特征和真实的视觉特征之间的距离。重建损失主要用于约束编码器的编码过程，使得编码器的输出保留词向量的基本特征从而能经过解码器还原回原始词向量特征。重建损失如公式(4-1)：

$$l_{recon} = \sum_{i=1}^N \sum_{j=1}^M (x - x')^2 \quad (4-1)$$

其中 N 表示样本数量，M 表示输入特征的维度， x 表示输入特征， x' 表示由解码器还原的特征。

4.2 面向零样本学习的特征生成方法

4.2.1 自编码器

自编码器^[74-76]是一种无监督训练的、对高维度的数据进行压缩或者降维的模型。随着深度学习研究的深入，自编码器也逐步演化为一种的特征提取方法，应用于深度神经网络的预训练之中。此外，自编码器还可以作为一种生成模型用以生成与训练样本相似的数据。比如，如果使用街景数据集来训练自编码器，那么经训练过的自编码器可以生成新的街景图片。自编码器可以通过直接将输入复制到输出的方式来训练。但是为了减小模型过拟合风险，自编码器需要增加不同程度的约束。比如，可以限制隐藏层神经元的数量，或者向训练数据集中增加噪声数据并添加新的约束条件。这些限制条件避免自编码器简单地将输入复制到输出并使自编码器学习更加有效的编码方式。

本小段主要介绍自编码器的原理。如图 4-2 所示，自编码器主要分为编码器（Encoder）和解码器（Decoder）。编码器——将输入数据投影到另一个潜在空间，可以看作是一种空间变化。解码器——重构来自潜在空间的特征。自编码器公式如(4-2)和(4-3):

$$x' = g(f(x)) \quad (4-2)$$

$$x \approx x' \quad (4-3)$$

其中， $f(x)$ 表示编码器， $g(h)$ 表示解码器， x 表示原始输入， x' 表示解码器还原后的结果， h 表示潜在空间表征。

公式(4-3)是一种约束条件——使解码器的输出尽可能得与原始数据相似。这是为了使得自编码器在学习输入数据到潜在空间的表征时，不丢失原来的信息。

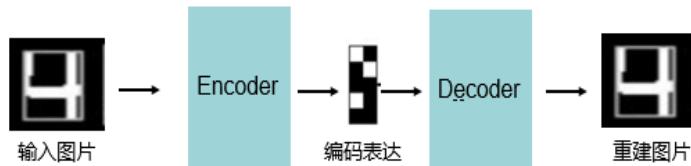


图 4-2 自编码器框架

本论文采用编码器结构如表 4-1 所示，解码器结构如表 4-2 所示。在表 4-1 中，编码器输入维度大小为 256，其中包含维度大小为 128 的类别词向量和维度大小为 128 的服从正态分布的随机 0-1 序列。在表 4-2 中，解码器输出维度大小为 128。

在公式(4-1)重建损失中, x 表示解码器的类别词向量输入, 而 x' 表示解码器的输出。

表 4-1 编码器结构

Type	Input Size	Output Size
Dense	256	100
Dense	100	100
Dense	100	64

表 4-2 解码器结构

Type	Input Size	Output Size
Dense	64	100
Dense	100	100
Dense	100	128

4.2.2 生成对抗网络

前小节介绍了自编码器可以作为生成器进行数据的生成。但是, 在图像领域中, 自编码器生成的图像数据存在一些缺陷, 比如生成的图像存在模糊的情况。这是因为自编码器的约束一般是最小均方误差, 其约束远不能满足生成高质量图像的要求。因此, Goodfellow 提出了一种针对图像生成的方法——生成对抗网络^[77]。

如图 4-3 所示, 生成对抗网络主要包含两个部分: 生成器和判别器。

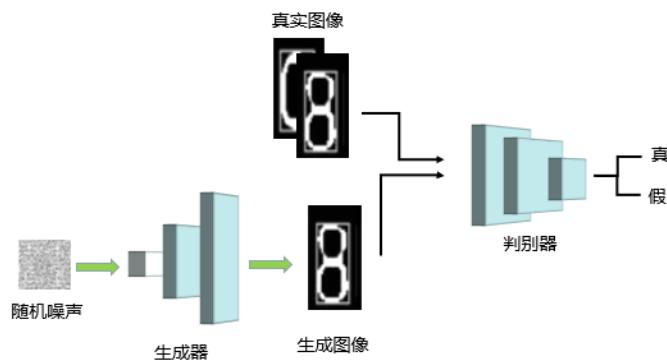


图 4-3 生成对抗网络框架

生成器的目的是生成一副伪图像。生成器的输入一般选择服从某一分布的随机数。判别器的输入既包含真实图像又包含生成器生成的伪图像。判别器的目的是辨别输入图像是否是真实样本。下面将从数学上介绍生成对抗网络的原理。

假设有一个图像数据集 $\{x(1), x(2), \dots, x(n)\}$, 它们的分布都是 $P_{data(x)}$ 并且它们之间是独立的。只要知道它的分布 $P_{data(x)}$, 就可以直接从 $P_{data(x)}$ 中采样生成数据。然而在很多情况下, $P_{data(x)}$ 是无法获取的。要想得到 $P_{data(x)}$ 的分布, 可以通过以下方法。首先, 从一个已知分布 $P_{G(x;\theta)}$ 采样得到一组数据集样本 $\{x(1), x(2), \dots, x(m)\}$, 其中 θ 是分布的参数。然后, 只要让这 m 个样本在 $P_{G(x;\theta)}$ 中同时出现的概率最大, 即如公式(4-4)。最后, 为了寻找 $P_{G(x;\theta)}$ 与 $P_{data(x)}$ 的关系, 将公式(4-4)化简为公式(4-5), 根据极大似然理论, 要想最大化上面的联合概率, 最好的 θ 应为公式(4-6):

$$\theta = \arg \max_{\theta} \prod_{i=1}^m P_G(x^{(i)}; \theta) \quad (4-4)$$

$$\theta = \arg \max_{\theta} \int_x P_{data}(x) \log P_G(x; \theta) dx - \int_x P_{data}(x) \log P_{data}(x) dx \quad (4-5)$$

$$\theta = \arg \min_{\theta} KL(P_{data}(x) \| P_G(x; \theta)) \quad (4-6)$$

根据公式4-6可知, 要想计算 $P_{data(x)}$ 和 $P_{G(x;\theta)}$ 之间的KL散度, 就需要知道它们的表达式。但是只有 $P_{G(x;\theta)}$ 的表达式是已知的。为了解决这个问题, 可以设计一个判别器来衡量两个分布之间的差异。它们之间的差异越小, 说明 $P_{data(x)}$ 和 $P_{G(x;\theta)}$ 越接近。具体做法如下: 从分布为 $P_{data(x)}$ 采样得到数据 $\{x(1), x(2), \dots, x(m)\}$, 从分布为 $P_{G(x;\theta)}$ 采样得到数据 $\{z(1), z(2), \dots, z(m)\}$ 。用判别器去衡量这些采样数据之间的差异, 衡量方式可以看作是一个经典的分类问题, 真实样本类别为1, 生成样本类别为0。判别器最后经过一个sigmoid激活函数, 输出0-1间的值, 见式(4-7)。当输入生成样本的时候, 判别器需要给出一个靠近0的输出值; 当输入真实样本的时候, 判别器则需要给出一个靠近1的输出值。它的目标函数见式(4-8):

$$f(x) = \frac{1}{1 + e^x} \quad (4-7)$$

$$V(G, D) = E_{X \sim P_{data}} [\log D(X)] + E_{x \sim P_G} [\log (1 - D(X))] \quad (4-8)$$

其中, $E_{x \sim P_{data}}$ 表示服从 P_{data} 分布的期望值, $E_{x \sim P_G}$ 表示服从 P_G 分布的期望值, $D(\cdot)$ 表示判别函数。

因此, 问题就转换为了生成器和判别器参数的优化问题。最优的判别函数 D 表示如公式(4-9), 最优的生成函数 G 表示如公式(4-10):

$$D^* = \arg \max_D V(G, D) \quad (4-9)$$

$$G^* = \arg \min_G E_{x \sim P_G} [\log(1 - D(X))] \quad (4-10)$$

本章节模型包含两个判别网络，判别网络 A 判断生成视觉特征和真实视觉特征的真假；判别网络 B 判断还原的语义分割结果图和真实语义分割结果图的真伪。判别网络 A 结构和判别网络 B 结构分别如表 4-3 和表 4-4 所示。

表 4-3 判别网络 A 结构

Type	Input Size	Output Size
Dense	64	50
Dense	50	32
Dense	32	1
Sigmoid	1	1

表 4-4 判别网络 B 结构

Type	Kernel size	Stride/padding	Channel
ChannelNet(2a)	3x3	1/1	32
Max pool	2x2	1/1	
ChannelNet(2b)	3x3		64
Max pool	2x2	1/1	
ChannelNet(3c)	3x3	1/1	128
Max pool	2x2		
ChannelNet(4d)	3x3	1/1	256
Max pool	2x2	1/1	
ChannelNet(5e)	3x3	1/1	512
Max pool	2x2	1/1	
Flatten			
Dense-500			
Dense-1			
Sigmoid			

如表 4-4，判别网络 B 包含 ChannelNet(**)结构，括号中数字表示该结构含有卷积层个数，而字母表示该结构的名字。如 ChannelNet(2a) 表示含有 2 个卷积层。

Dense-500 则表示 Dense 层输出神经元个数为 500。

4.2.3 基于自编码器的特征生成方法

在传统的图像识别领域中，每一个类别拥有自己的类别特征。因此，模型可以根据类别特征来识别原图像属于什么类别。比如，在基于深度学习的图象识别中，一张图片经过一个特征提取网络之后的图像特征为 $F \in \mathbb{R}^{h \times w \times c}$ 。其中 h 、 w 、 c 分别是高、宽和通道数量。 $F_i \in \mathbb{R}^{h \times w \times c_i}$ 表示第 i 个通道的图像特征。 F_i 也代表特征提取网络所学得的图像局部特征。模型根据不同的 F_i 组合就能分辨出来这一张图片属于哪一类。比如输入含有一只鸟的图片， F 可能包含翅膀、爪子、鸟喙等等局部特征。基于此，一些图像生成方法——自编码器、生成对抗网络，能够通过生成类别特征 F 来合成高质量的图像。

虽然本章节的模型也采用生成图像类别特征的方式，但与上述生成方式有些许不同。由于语义分割任务是逐像素分类，所以使用上述的特征生成方式会面临上采样阶段的特征对齐问题。其原因主要如下：一是，尽管生成模型能够很好的生成类别特征 F ，但是这个类别 F 可能会有一定的偏移。同时，无论类别特征 F 发生何种偏移，识别网络模型都能够识别出来。因此，这一点对于图象识别任务影响较小。然而相比于识别模型，语义分割模型对类别特征的位置是有一定要求的。若类别特征 F 偏移量过大，模型会在上采样时把不相对应的类别信息还原到同一个位置上面。比如，原始图像是一只鸟，其鸟喙在图像的左上方、爪子在图像的下方，而生成的鸟喙类别特征 F 在图像的左下方。这样导致上采样过程中，鸟喙和爪子重叠从而影响模型分割效果。二是，对于零样本学习来说，将语义信息转化为视觉信息是比较困难的。具体来说，具有相同语义信息的类别的视觉表现形式可能不同。因此，本章节模型并不采用直接生成 $F_i \in \mathbb{R}^{h \times w \times c_i}$ 这种形式，即以通道为单位进行类别特征生成，而是采用生成 $F \in \mathbb{R}^{h \times w \times c}$ 的形式。这种生成特征方式是为了保证类别特征位置对齐。

根据上面介绍的自编码器以及生成对抗网络。本章节设计的特征生成网络如图 4-1 所示。其中虚线代表伪图像视觉特征，实线代表真实的图像视觉特征。该网络是基于自编码器框架并结合了生成对抗网络的原理所设计的。从自编码器的角度看，该网络的输入是类别的词向量和随机噪声，网络的输出是类别视觉特征。其中网络的约束：类别特征维度小于输入维度并且建立重建损失。从生成对抗网络角度来看，该网络一共包含两个生成对抗网络。如图 4-4 所示，由编码器生成的特征不仅经过解码器，而且还经过判别网络 A。这个判别网络 A 的目的是判断编码器生成的类别视觉特征是否是真实的类别视觉特征。因此，编码器可以看作是一个生

成网络——用于生成类别的视觉特征。它和判别网络 A 组成生成对抗模型。对抗训练的方式使得生成的类别特征与真实类别特征相似，从而使得编码器变成一个空间变换函数——将词向量空间转换到视觉空间。

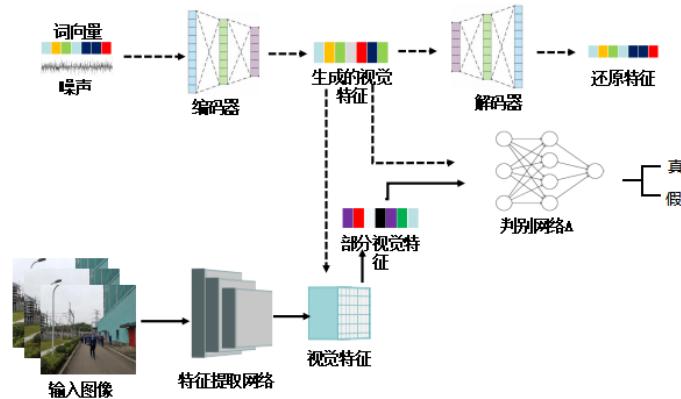


图 4-4 自编码器网络

上面的模型仅仅是学习已知类别词向量和其视觉空间的转换函数。对于未知类别，该模型只能通过推理的方式输出这些类别的视觉特征。这种情况下，生成的未知类别特征就会存在一定的偏移。因此，特征生成模型还包含另一个判别网络 B。如图 4-5 所示，如果将判别网络 B 作为生成对抗网络的判别器，那么编码器和图像上采样网络就组成了生成器。根据图 4-5 的网络结构，编码器能够生成已知类别的特征，使得判别网络 A 不能分清真假。如果是未知类别，编码器也许会生成与真实视觉特征存在偏差的视觉特征。为了解决这种问题，判别网络 B 负责判断生成的图像是否为真实图像，从而对编码器的参数进行进一步训练。最终的目的是使得编码器生成的未知类别的视觉特征也可以靠近真实的视觉特征。

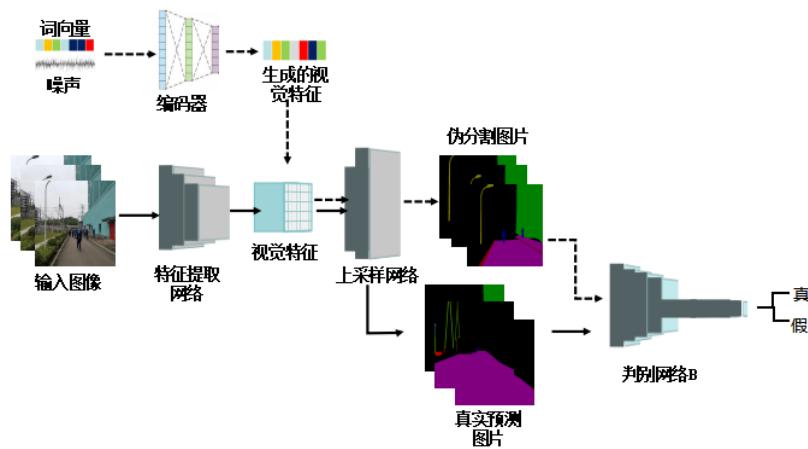


图 4-5 生成对抗网络

训练方式的选择很大的影响生成对抗网络中真实特征分布的拟合。比如，如果判别网络训练得过好，那么会使得整个对抗生成网络崩溃。本章节模型训练流程伪代码如图 4-6 所示。在本章节模型种，生成对抗网络的训练方式是先训练若干次判别器，再训练一次生成器。由于存在两个判别网络，所以将进行两次对抗的训练。对于判别网络 A，先进行 k1 次的判别网络训练。对于判别网络 B，则先进行 k2 次的判别网络训练。

算法·特征生成模型。

输入：类别词向量集合 $A_{embedding}$ ，图像数据集合 $A_{data(x)}$ ，随机分布 B_{noise} 。
输出：无。

- 1: 初始化参数 θ_{d1} 、 $\theta_{encoder}$ 、 θ_{d2} 、 $\theta_{decoder}$ 和迭代次数 k 、 $k1$ 、 $k2$ 。
- 2: **while** $k \neq 0$ **do**
- 3: **while** $k1 \neq 0$ **do**
- 4: 从数据分布 $A_{data(x)}$ 中采集 m 个样本 $\{x_1, x_2, \dots, x_m\}$ 。
- 5: 将 $\{x_1, x_2, \dots, x_m\}$ 输入语义分割网络得到 $\{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m\}$ 。
- 6: 从词向量集合 $A_{embedding}$ 选取与 $\{x_1, x_2, \dots, x_m\}$ 存在的类别所对应的词向量 $\{w_1, w_2, \dots, w_m\}$ 。
- 7: 从随机分布 B_{noise} 中采集 m 个样本 $\{z_1, z_2, \dots, z_m\}$ 。
- 8: 将词向量 $\{w_1, w_2, \dots, w_m\}$ 与 $\{z_1, z_2, \dots, z_m\}$ 输入生成器，得到输出集 $\{\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_m\}$ 。
- 9: 更新判别网络 A 参数 θ_{d1} ：

$$\tilde{V} = \frac{1}{m} \sum_{i=1}^m \log(D(\tilde{x}_i)) + \frac{1}{m} \sum_{i=1}^m \log(1 - D(\tilde{z}_i))$$

$$\theta_{d1} \leftarrow \theta_{d1} + \eta \nabla \tilde{V}(\theta_{d1})$$
- 10: 更新解码器参数 $\theta_{decoder}$ 。
- 11: **end**。
- 12: 更新编码器参数 $\theta_{encoder}$ ：

$$\tilde{V} = \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z_i)))$$

$$\theta_{encoder} \leftarrow \theta_{encoder} - \eta \nabla \tilde{V}(\theta_{encoder})$$
- 13: **while** $k2 \neq 0$ **do**
- 14: 更新判别网络 B 参数 θ_{d2} ：

$$\tilde{V} = \frac{1}{m} \sum_{i=1}^m \log(D(\tilde{x}_i)) + \frac{1}{m} \sum_{i=1}^m \log(1 - D(\tilde{z}_i))$$

$$\theta_{d2} \leftarrow \theta_{d2} + \eta \nabla \tilde{V}(\theta_{d2})$$
- 15: 更新编码器参数 $\theta_{encoder}$ ：

$$\tilde{V} = \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z_i)))$$

$$\theta_{encoder} \leftarrow \theta_{encoder} - \eta \nabla \tilde{V}(\theta_{encoder})$$
- 16: **end**。
- 17: **end**。

图 4-6 特征生成网络训练流程

4.3 实验结果

4.3.1 实验数据

本章节模型的实验数据与第三章实验数据一致。

4.3.2 实验设定

本章节模型中的语义分割模型是第三章节的语义分割模型。因此，在实验软件和硬件方面配置如表 3-3 所示。软件系统依然使用 linux 系统，使用 TensorFlow 开发框架以及大量依赖 python 环境的工具库，如 numpy、pandas 等。

在该模型中，语义分割网络学习率设置为 0.001，边缘损失超参数 α 设置为 1，训练批次大小设置为 8，采用 Adam 优化器迭代训练 85 次，其中 β_1 为 0.9、 β_2 为 0.9。自编码器网络学习率设置为 0.003，采用 SGD^[78]优化器迭代训练 30 次，其中 momentum 为 0.9、decay 为 0.001，判别网络 A 学习率设置为 0.001，采用 Adam 优化器迭代训练 30 次，其中 β_1 、 β_2 和语义分割网络的优化器设置相同， k_1 设置为 3。判别网络 B 学习率设置为 0.005，采用 Adam 优化器迭代训练 30 次，其中 β_1 、 β_2 和语义分割网络的优化器设置相同， k_2 设置为 1。使用经过预训练的 InceptionNet 作为基础模型，上采样层采用线性内插法，模型输入尺寸大小设置为 608*608。对比方法 ZS3^[79]实验设定如下：基础语义分割模型 deeplabv3——采用 SGD 优代器训练 50 次，其中基础学习率为 0.001、momentum 为 0.9、decay 为 0.001；特征生成模型——采用 Adam 优化器迭代训练 30 次，其中基础学习率为 0.002、 β_1 为 0.9、 β_2 为 0.85。经过相应迭代训练，上述零样本语义分割模型均已收敛。

4.3.3 实验结果与分析

实验模型包括 ZS3 以及本章节模型。实验的评价标准是平均交并比(MIOU)。因为本章节模型是为了完成零样本学习的任务，所以实验分别选取不同的类别作为未知类别且将包含有该类别的图片从训练集中剔除。实验结果如表 4-5 所示，其中第一列表示实验选取的未知类别，已知类 MIou 表示模型在仅包含已知类别的数据集上的平均交并比，未知类 MIou 表示模型预测未知类别的平均交并比。首先，对比不同的方法，从表中可得知本章节的模型在大部分预测结果上要好于 ZS3 模型，比如路灯、隔离开关、公告牌等。其原因主要是：ZS3 模型的特征生成网络较为简单，拟合语义空间到视觉空间的转换函数的效果不如本章节的特征生成模型；ZS3 模型使用的 deeplabv3 模型，在变电站场景下，deeplabv3 模型的平均交并比也稍微低于本章节的语义分割模型。其次，将第三章的实验结果和表 4-5 实验结果进

行对比。选取不同未知类别的模型的已知类 MIou 都有所下降，且下降幅度略有差异。比如，选取未知类别为路灯时，相比于第三章最优的 MIou 是 0.67，本章节模型的 MIou 下降至 0.56；当选取未知类别为公告牌时，本章节模型的 MIou 仅仅下降了 0.02。这是由于变电站数据集中包含类别的数量差距较大所导致的。本章节同时选取了两个类别同时作为未知类别——隔离开关和避雷针。本章节实验限定了选取未知类别的数量主要受限于以下两点原因。一是未知类别越多就意味着已知类别越少，模型所能学到的知识越有限。一个极端的例子就是，如果模型只见过一种水果，那么它几乎不可能识别出其他所有的水果。二是本次实验数据集规模较小，如果选取较多类别作为未知类别，那么有可能导致语义分割模型过拟合且特征生成模型难以训练。从表中可以计算出，仅仅选取一个未知类别的模型的 MIou 为 0.286，而选取两个未知类别的模型的 MIou 是 0.19。

表 4-5 实验结果

未知类别	ZS3		Ours	
	已知类 MIou	未知类 MIou	已知类 MIou	未知类 MIou
路灯	0.47	0.10	0.56	0.16
隔离开关	0.58	0.18	0.60	0.22
公告牌	0.63	0.14	0.65	0.17
石头	0.45	0.37	0.49	0.41
楼梯	0.57	0.47	0.62	0.47
隔离开关 和避雷针	0.59	0.15	0.63	0.19

如图 4-7 所示，ZS3 和本章节模型在变电站数据集的实验结果。在图 4-7 中，(a) 表示原始图像，(b) 表示 ZS3 模型预测结果，(c) 表示本章节模型的预测结果，(d) 表示标签。第一、二、三行实验结果所选择的未知类别分别是路灯、避雷针、公告牌。其中房子为绿色，路灯为黄色，避雷针为橙色，公告牌为褐色。从第一行和第二行来看，ZS3 模型和本章节模型都不能完全预测出路灯和避雷针的所有像素。相比于 ZS3 模型，本章节模型能够预测更多的该类别像素。正如前文所说，由于选取未知类别会影响训练数据集的规模大小从而导致已知类别的预测精度下降。这一点在第一行的路灯预测可以充分展示出来。在第一行图像集中，ZS3 模型对房子的预测精度下降的幅度比本章节模型大。部分原因是由于 ZS3 选取的 deeplabv3 模型在变电站场景上效果稍差。当减小一部分数据集之后，deeplabv3 模

型的预测效果会因为远远没达到拟合数据的程度而下降更多。从第二行的图像来分析，尽管两个模型都能够预测出避雷针，但是它们有不同的特点。原图像包含两个避雷针类别。ZS3 模型只能很好的预测出其中一个，而本章节模型能够大体上预测出两个。但是，对于靠近右边的避雷针，本章节模型也没能很好的预测出来。从原始图像上分析，没有预测出来的像素包含车这一类别，而数据集并未标记该类别。因此，特征生成模块生成的避雷针特征离真实的避雷针特征有一定的距离，导致模型无法识别出避雷针。从第三行图像集来看，两个模型都能很好的预测公告牌。但是两个模型预测的结果有些许差别。ZS3 模型所预测的公告牌像素要比真实标签中的公告牌像素多一些，而本章节模型所预测的公告牌形状要更靠近真实形状。

基于上面对实验结果的分析，可以得出本章节模型的特征生成模块能够有效地辅助语义分割模型实现未知类别的预测。对于一些长宽比例悬殊以及像素占比低的类别，尽管语义分割模型预测准确率有所下降，但该模型仍然具有捕捉与识别其视觉特征的能力。

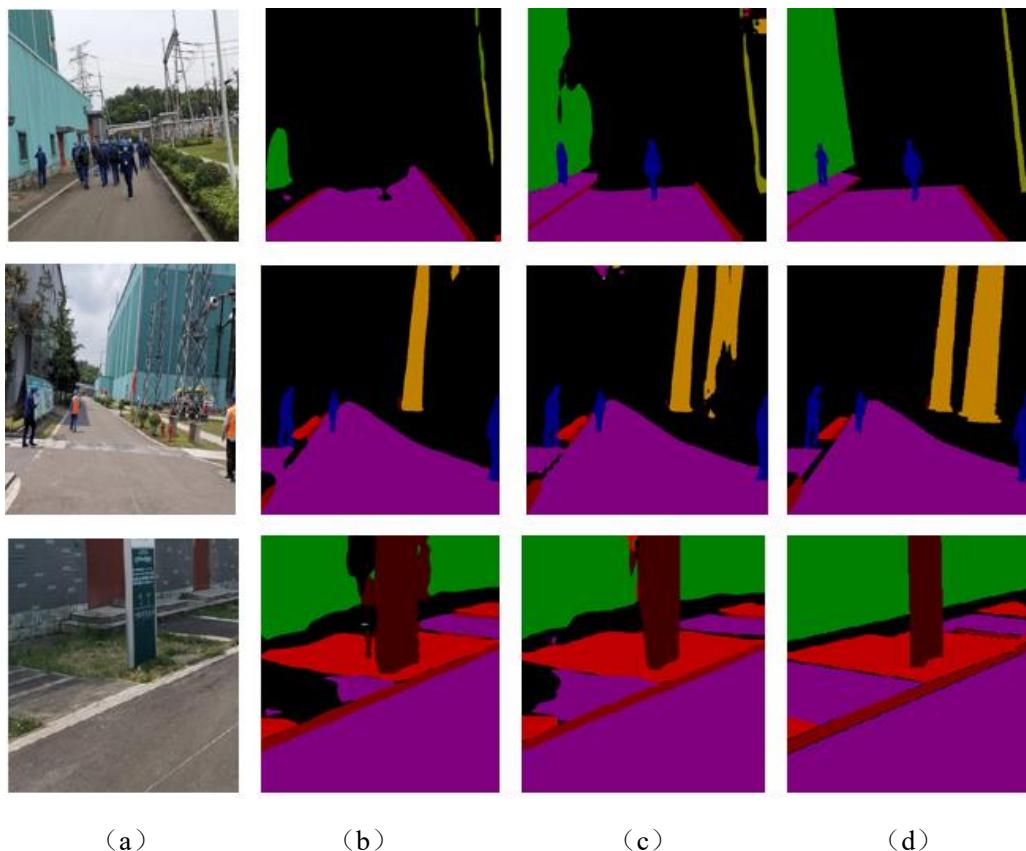


图 4-7 零样本语义分割模型对比

上述实验将本章节方法和其他方法进行了比较。由于本章节模型的语义分割

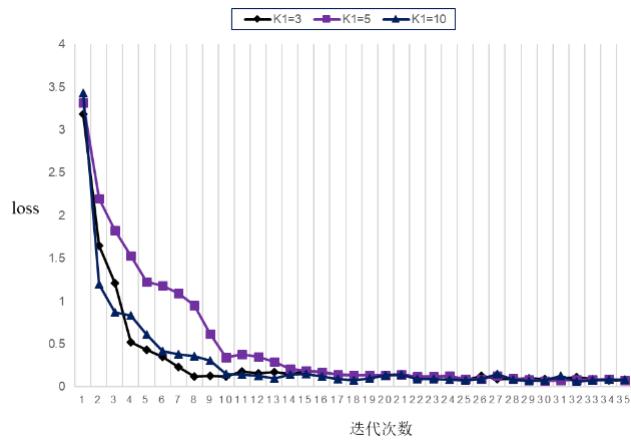
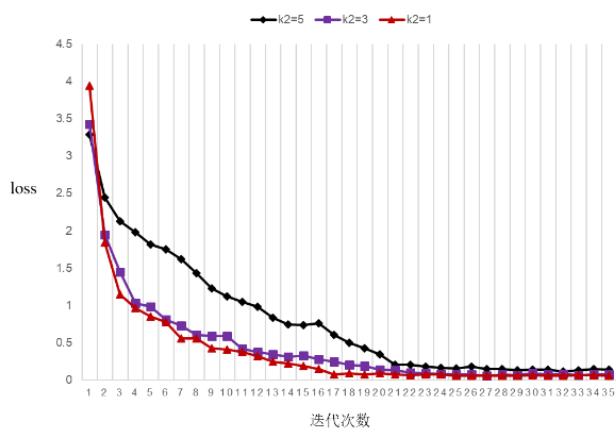
部分是基于第三章的模型，所以本章节也会探究不同视角对零样本学习的影响。如表 4-6 所示，本次实验对比了三种不同模型的实验效果——仅包含 $1*3$ 卷积核的模型、仅含 $3*5$ 卷积核的模型以及含有 $1*3$ 卷积核且含有 $3*5$ 卷积核的模型。下面将从两个角度分析结果。从整体来看，除去路灯这一类别，含有 $1*3$ 卷积核和 $3*5$ 卷积核的模型效果要好于其余两种模型的效果。这有是因为路灯的长宽比差距大导致很适合使用 $1*3$ 卷积核来提取特征。而其他长宽比例相对较小的类别，仅含有 $1*3$ 卷积核的模型平均交并比都小于最优值。对于仅包含 $3*5$ 卷积核的模型，尽管没有优于最优值，但是大部分类别的预测准确率都超过了仅使用 $1*3$ 卷积核的模型。上述这些既说明语义分割网络的精度会影响零样本语义分割网络的整体精度，又说明多视角也能够影响零样本语义分割网络的精度。将第三章表 3-5 与表 4-6 联系起来看，由于选取未知类别之后，导致训练数据集规模减小，所以表 4-6 中的已知类 MIou 均有所降低。同时，选取不同类别的平均交并比下降程度也不相同。比如，相比于公告牌作为已知类别的 0.55，其作为未知类别的 MIou 低于 0.20；路灯则由 0.35 下降至 0.16；隔离开关由 0.43 下降至 0.22 等等。这些下降的幅度不同可能是因为：不同类别在数据集中的数量不一样。数量不同导致去除该类别之后剩下的训练数据集规模不同。这既影响语义分割模块准确率，又影响特征生成模块所生成的特征与真实特征的相似度。如果模型见识的样本太少，那么会导致语义空间到视觉空间的映射关系太少。特征生成模型主要目标是实现语义空间到视觉空间的转换。特征生成模型需要足够的样本数量，才能减少语义空间转换到视觉空间的偏移。

表 4-6 多视角消融实验

未知类别	仅含 $1*3$ 卷积核		仅含 $3*5$ 卷积核		含 $1*3$ 与 $3*5$ 卷积核	
	已知类 MIou	未知类 MIou	已知类 MIou	未知类 MIou	已知类 MIou	未知类 MIou
路灯	0.53	0.17	0.58	0.14	0.56	0.16
隔离开关	0.57	0.12	0.59	0.18	0.60	0.22
公告牌	0.59	0.16	0.61	0.14	0.65	0.17
石头	0.43	0.29	0.44	0.35	0.49	0.41
楼梯	0.56	0.23	0.57	0.31	0.62	0.47
隔离开关 和避雷针	0.52	0.08	0.59	0.15	0.63	0.19

基于上面对实验结果的分析，可以得出第三章的多视角结构仍然可以通过调整感受野的大小来捕捉已知类别的视觉特征。同时，多角度地提取已知类别的视觉特征给予特征生成网络更丰富的视觉样本去学习真实的视觉空间。这使得特征生成网络生成的视觉特征更加靠近真实的视觉特征。此外，将同一语义信息转化为不同的视觉信息可以提升语义分割模型对未知类识别的准确率。

生成对抗模型的准确率受训练方式和迭代参数的影响。因此，本章节对基于自编码器的特征生成模型的迭代参数 k_1 、 k_2 进行实验。如图 4-8 所示，本章分别将 k_1 设置为 3、5 和 10。如图 4-9 所示，本章分别将 k_2 设置为 1、3 和 5。

图 4-8 参数 k_1 损失函数图图 4-9 参数 k_2 损失图

从图 4-8 可以得出以下结论：一是，当 k_1 等于 3，模型收敛速度最快；二是，参数 k_1 的大小对于模型的收敛速度有影响，但是对模型最终的收敛结果影响较小。从图 4-9 看出，当 k_2 取值变小时，模型损失下降变快且收敛数值变低。当判别网

络和生成网络等比例训练的时候，模型的损失值最低。

4.4 本章小结

本章首先介绍基于零样本语义分割的模型框架，然后对该框架中的特征生成模块的设计原理和思路进行详细阐述并且给出模型训练方式。针对生成视觉特征与原视觉特征的偏移问题，本章利用生成对抗网络使生成视觉特征更靠近原视觉特征。最后使用该模型在变电站数据集进行实验，并与其他模型实验结果进行对比和分析。

第五章 变电站巡检机器人语义感知软件系统设计与实现

随着深度学习的发展，语义分割也逐步应用在巡检机器人视觉系统中。在变电站智能巡检机器人执行巡检任务过程中，语义分割可以辅助巡检机器人避障或者重新规划路径。比如，将道路中的障碍物进行像素标记以提醒巡检机器人前方存在障碍物等。语义感知系统是巡检机器人环境感知系统的一部分。本章节基于第三章节的语义分割模型以及第四章的基于零样本学习的语义分割模型，设计并实现一个语义感知系统，包含系统需求分析、系统总体设计、系统功能设计和系统测试。

5.1 需求分析

5.1.1 系统功能需求

本小节主要介绍语义感知系统的主要功能需求。

语义感知系统主要功能分为：用户管理模块、语义分割模型管理模块、数据管理模块和可视化模块。系统功能如图 5-1 所示，用户管理模块主要负责系统用户及其信息的管理，包括用户注册、登录以及信息修改功能。语义分割模型管理模块主要涉及模型的部署、模型训练相关的管理，具体包含：语义分割模型的选择、超参数设置、模型参数保存以及训练日志存储等。数据管理模块主要涉及数据集的管理，比如数据集的上传、数据图片的添加、删除等功能。可视化模块主要是展示输入数据图片以及模型预测结果。

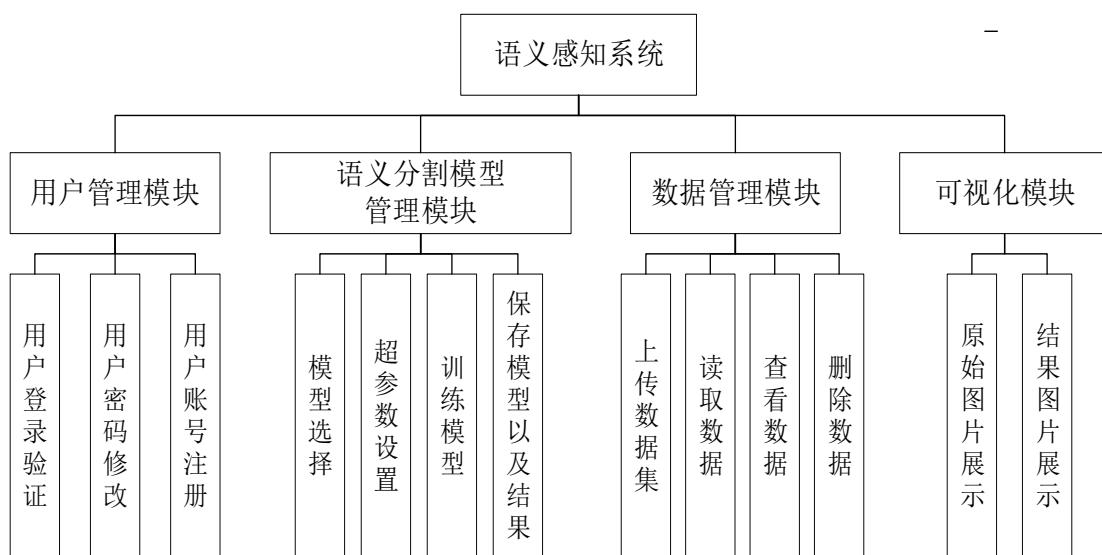


图 5-1 系统功能图

5.1.2 系统主要用例图

系统用例为：

用户：个人信息管理（注册）、系统登录、模型管理、数据集管理；

管理员：注册、系统登录、用户管理、查看日志。

如图 5-2 所示，该用例图的参与者是管理员，管理员主要对用户信息的管理和日志的查询。

如图 5-3 所示，该用例图的参与者是用户，用户主要完成对个人信息的管理、对模型的管理和对数据集的管理。其中用户的注册功能包括注册和修改信息，其中修改信息采用 extend 方式。

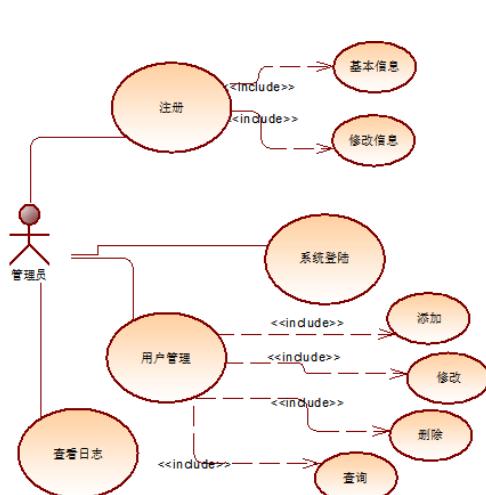


图 5-2 管理员用例图

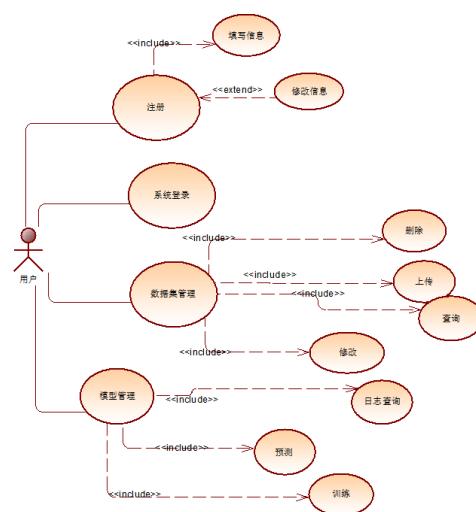


图 5-3 用户用例图

5.1.3 系统性能需求

实时性：从巡检机器人采集图片到服务器识别图片的时间长短主要由采集图像时间、图像传输时间以及图像语义分割时间组成。其中采集图像时间和传输时间由视觉传感器所决定。本论文所搭配的变电站巡检机器人的图像传输过程中会占用大量的带宽。因此，要想提高实时性，降低变电站场景识别的延迟，需要减小图像语义分割的运算时间。在执行图像语义分割的过程中，可以使用 GPU 进行加速以减小运算时间。

准确度：准确度主要体现在语义分割算法上面。精准的语义分割结果能够使得巡检机器人更好地理解周围的环境。

可靠性：系统在一定时间内需要能够完成用户指定的任务并且不能出错。

5.1.4 系统运行环境需求

机器人平台是 husky 并且装有工控机和 AP 天线设备。机器人平台上部是一个移动云台，内部装设高清相机和红外摄像仪，能够执行读表、温度检测等任务。具体机器人硬件参数见表 5-1，服务器硬件环境如表 5-2 所示，软件环境如表 5-3 所示。

表 5-1 机器人硬件参数表

husky 移动 机器人平台	轮径 0.32m; 轮距: 0.555m; 最大速度: 1.0m/s
工控机	处理器: Intel 酷睿 i5-4200M; 主频: 3.15GHz
I/O 口	USB3.0\USB2.0\RS232/422/485
相机	200 万像素\RS-485 通信协议
内存容量	4G DDR3
激光雷达 Sick lms151	最远测量距离: 50m; 扫描频率: 10Hz

表 5-2 服务器硬件环境表

设备名称	具体配置情况
内存	DDR4-3200 帧
主板	Z270-AR
硬盘	500GB SSD+2T HDD
CPU	Intel i7-9700k
GPU	GTX 2080Ti 11BG 显存

表 5-3 软件环境表

软件名称	具体配置情况
操作系统	Ubuntu 18.04
开发语言	Python
开发框架	Keras、Django

5.2 系统总体设计

图 5-4 为本章节的语义感知系统整体框架图。语义感知系统主要分为四层：数据采集层、数据存储层、业务层和展示层。数据采集层主要负责图像数据的提供；

数据存储层主要负责存储采集而来的图像数据和用户上传的图像数据，实现数据的持久化；业务层是整个系统的核心，用来完成系统提供的所有核心功能，如用户管理、模型管理和数据集管理；展示层主要是以界面的形式与用户交互并向用户展示所需要的信息。

下面将详细介绍系统框架图中的每一层：

(1) 数据采集层。该层的目的是提供数据。对于智能巡检机器人来说，可以通过高清相机采集图像、视频数据。同时用户也可以上传数据。

(2) 数据存储层。该层主要功能是存储数据——采集的数据、模型计算的结果数据、模型参数、模型的训练日志等等。存储这些数据能够极大方便用户。比如，通过查询训练日志，用户可以知道目前模型训练的具体情况——迭代次数、训练超参数、预测结果。

(3) 业务层。业务层是整个语义感知系统中最重要的部分。从面向下层的角度来看，业务层可以直接操纵数据存储层；从面向上层的角度来看，业务层则可以提供多个操作的接口。首先，作为系统软件，业务层必须满足用户管理功能——用户注册、用户登录、用户信息修改等。然后，语义分割模型的管理也是该层中最重要的部分。语义分割模型的管理不仅包含与语义分割任务强相关的功能——选择语义分割模型、加载模型参数、保存模型结果等，而且还包含一些辅助的功能——保存训练日志。最后，该层实现对存储数据集的管理——增删等功能。

(4) 展示层。该层主要负责与用户进行交流。用户可以通过填写个人信息登录系统；用户根据自己的需求选择合适的语义分割模型进行变电站场景的分割；用户可以根据存储的训练日志、训练结果等信息进行结果分析。

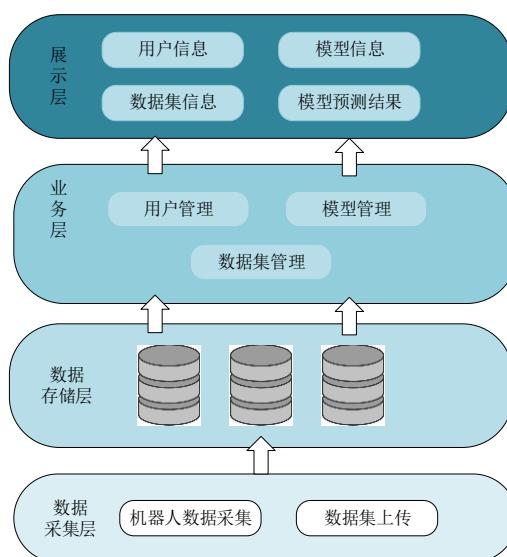


图 5-4 系统框架图

5.3 系统的详细设计与实现

5.3.1 系统功能设计与实现

前面的小节分别介绍了语义感知系统的功能需求和总体设计，本小节主要介绍语义感知系统的功能设计与实现。

1.上传数据功能设计：对于深度学习模型来说，无论是训练还是预测都需要数据。如图 5-5 所示，该流程图表示上传数据功能步骤。首先，用户经过系统身份验证后进入数据集管理界面；然后选择需要上传的数据文件进行上传；最后根据上传结果，将数据文件存储在数据库中。其中，选择数据文件进行上传，若上传成功，则系统显示成功信息并提示用户还可以继续上传；若上传失败，则系统显示失败信息提示用户。若用户选择不重新上传数据文件或者不继续上传数据文件，则结束；若用户需要重新上传数据文件或者继续上传数据集文件，则继续上传。

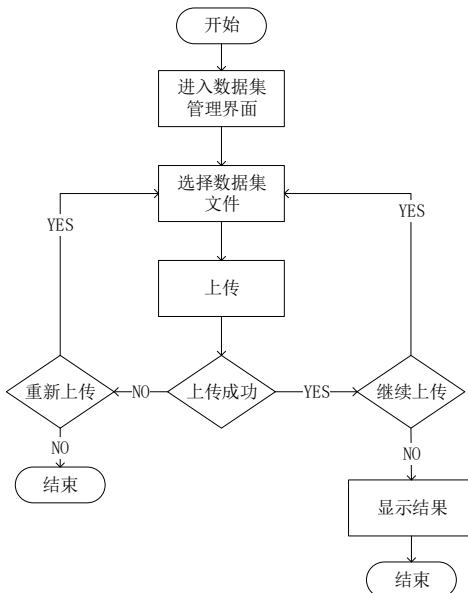


图 5-5 上传数据流程

上传数据功能主要代码如图 5-6 所示。request 获取文件保存路径 savepath 和文件列表 fileList，若文件列表不为空，则循环将文件存入保存路径。

```

savepath=request.POST.get('savepath')
fileList = request.FILES.getlist("filename", None)
if fileList is not None:
    for myFile in fileList:
        destination = open(os.path.join(savepath, myFile.name), 'wb+')
        for chunk in myFile.chunks():
            destination.write(chunk)
        destination.close()
  
```

图 5-6 上传数据代码图

2. 模型训练功能设计：如图 5-7 所示，该图是模型训练的流程图，其中包含模型选取、参数设置、数据集选取等步骤。

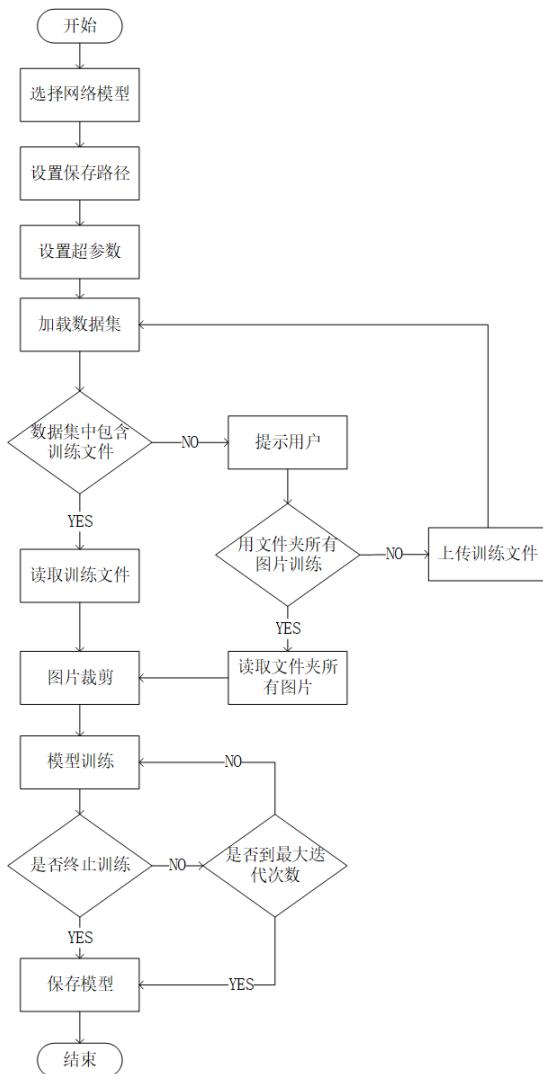


图 5-7 模型训练流程图

关键步骤如下：

- (1) 选择网络模型。网络模型包含 FCN、Deeplab 模型以及前面章节所提出来的模型等。如果是选择前面章节提出来的语义分割模型，那么无需选择基本网络框架。如果选中其他模型，比如 FCN 等模型，可以选择基本网络框架，比如 VGG、ResNet 等等。如果选用不同的基本网络框架，则需要选择相对应的预训练参数。
- (2) 设置保存路径。设置保存路径主要设置模型以及模型参数的保存路径，以便使用该模型预测时进行参数加载。
- (3) 设置超参数。不同的模型有不同的超参数。比如有些模型的损失函数存

在多种，每一种都需要一定的超参数进行约束。待选定模型之后，需要进行模型相应的超参数设置，比如迭代次数、损失权重、学习率以及动量等。

- (4) 加载数据集。用户从文件夹中选择需要训练的图片数据。如果文件夹中没有包含训练图片数据的说明文件 (train.txt)，那么系统需要提示用户是否使用文件夹下所有的图片数据进行训练。
- (5) 迭代训练模型。模型迭代一次包括前向计算结果和后向参数更新。在训练的过程中，会有两种不同的保存模型方式运行。一是，保存模型前一次迭代训练之后的参数；二是，保存迭代到目前为止，训练效果最好的模型参数。这样能保证模型训练过程中因意外中断所带来的时间损失。训练过程直到当前训练迭代次数等于所设置的最大迭代次数或者用户主动停止训练为止。
- (6) 模型训练完之后，再进行一次模型以及模型参数的保存，并且将此次训练模型的超参数、准确率等信息以日志的形式写入相应的保存路径下。

功能实现：训练功能主要代码如图 5-8 所示。首先调用 Model 方法建立网络计算图，然后调用 compile 方法完成方法的优化器和损失函数等参数的加载，其中优化器是 RMSprop，损失函数是 softmax 方法。因为数据集较大不能完全加入内存，所以采用生成器的方式进行加载训练。SegDataGenerator 函数就是一个数据生成器，用于逐步将训练数据加入内存训练。最后调用 fit_generator 函数进行迭代训练，其中通过回调函数 callbacks 中的 checkpoint 函数保存模型参数。

```

model = Model(inputs=[inputs], output=[ans])
def softmax_sparse_crossentropy_ignoring_last_label(y_true, y_pred):...
model.compile(optimizer=keras.optimizers.RMSprop(lr=0.0005),
              loss=softmax_sparse_crossentropy_ignoring_last_label,
              metrics=metrics.)
callbacks = []
checkpoint = ModelCheckpoint(filepath=os.path.join(save_path, 'checkpoint_weights.hdf5'),
                             save_weights_only=True)
callbacks.append(checkpoint)
train_datagen = SegDataGenerator(zoom_range=[0.5, 2.0], zoom_maintain_shape=True, crop_mode='random',
                                  crop_size=target_size, rotation_range=0., horizontal_flip=True,
                                  channel_shift_range=20., fill_mode='constant', label_cval=label_cval)

def get_file_len(file_path):...
steps_per_epoch = int(np.ceil(get_file_len(train_file_path) / float(batch_size)))
model.fit_generator(
    generator=train_datagen.flow_from_directory(
        file_path=train_file_path,
        data_dir=data_dir, data_suffix=data_suffix,
        label_dir=label_dir, label_suffix=label_suffix,
        classes=classes,
        target_size=target_size, color_mode='rgb',
        batch_size=batch_size, shuffle=True,
        loss_shape=loss_shape,
        ignore_label=ignore_label,
    ), steps_per_epoch=steps_per_epoch, epochs=epochs, callbacks=callbacks, workers=4,
    class_weight=class_weight
)

```

图 5-8 模型训练代码图

3. 模型预测功能设计：图 5-9 是模型预测的具体流程图。模型预测是本系统的主要功能之一。模型预测功能中主要包含：预测数据集选择、预测模型选择、模型预测结果保存等。

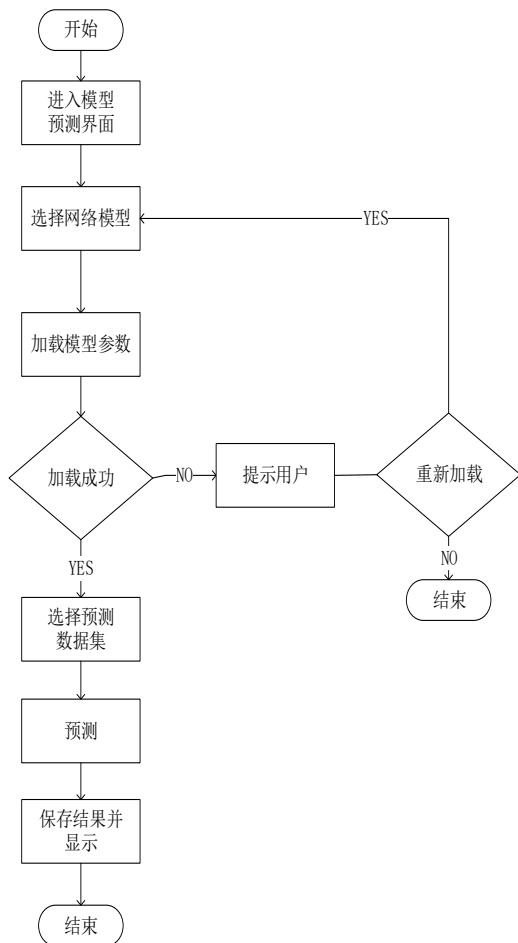


图 5-9 模型预测流程图

关键步骤如下：

加载网络模型参数。从预设置的模型参数路径，加载选中网络模型的参数。在模型参数加载的过程中，有可能加载失败。比如路径不存在、存储参数与所选中模型的参数存在差异等。只有模型参数加载成功，才能进行预测。

功能实现：

预测功能主要代码如图 5-10 所示。首先通过 `load_weights` 方法加载已训练过的模型参数，然后循环预测所需预测的图片集合 `image_list` 中的图片。预测过程具体如下：分别从路径中读取待预测图片 `image` 和对应的标签 `label`；调用 `preprocess_input` 方法对 `image` 进行预处理；使用 `predict` 方法对 `image` 进行预测，并逐像素进行分类；使用标签调色板 `label.palette` 对预测结果 `result_img.palette` 进

行调色（使得预测结果和标签颜色一致）；用 save 方法保存预测结果。

```

model.load_weights(save_path)
for img_num in image_list:
    image = Image.open('%s/%s%s' % (data_dir, img_num, data_suffix))
    image = img_to_array(image)
    label = Image.open('%s/%s%s' % (testlabel_dir, img_num, label_suffix))

    image = np.expand_dims(image, axis=0)
    image = preprocess_input(image)
    result = model.predict(image, batch_size=1)
    result = np.argmax(np.squeeze(result), axis=-1).astype(np.uint8)
    result_img = Image.fromarray(result, mode='P')
    result_img.palette = label.palette
    save_dir = '/home/net/predictImage'
    if save_dir:
        result_img.save(os.path.join(save_dir, img_num))

```

图 5-10 模型预测代码图

5.3.2 数据库设计

系统 E-R 图如图 5-11 所示，其中主要包含三个实体：用户、模型和数据。用户表主要用来存储用户信息，user_id 是主键，user_type 表示用户类型——普通用户和系统管理员，user_remarks 表示用户备注信息；模型表主要用来存储语义分割模型的相关信息，model_id 是主键，base_net 表示模型含有的基础网络，save_path 表示模型参数保存路径，model_param 表示模型超参数。数据表主要负责存储数据集相关的信息，data_id 是主键，data_field 表示数据集类型，save_path 表示数据集保存路径。除实体图转换的数据表之外，多对多关系也需要转化为相应的数据表。因此，系统数据表还需要描述 2 张关系的数据表——上传表和使用表。每张表都将所涉及的实体作为主键和外键。其中使用表新增 train_log_path、test_log_path 字段，分别用来描述训练日志存储路径和测试日志存储路径。

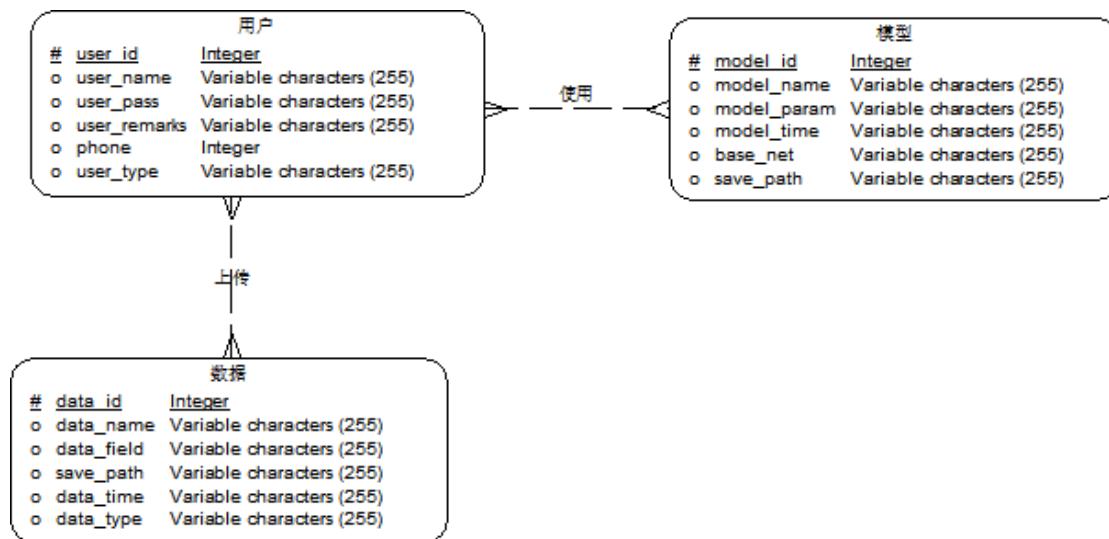


图 5-11 E-R 图

5.4 系统测试

5.4.1 测试环境

系统硬件配置和软件配置情况分别如表 5-2 和表 5-3 所示。巡检机器人如图 5-12 所示。在硬件方面，因为本系统涉及的语义分割算法是基于深度学习的方法——需要较高精度的数字运算和频率较高的矩阵运算，所以硬件平台需要搭载 GPU 以及选用较大的内存。在软件方面，操作系统选择 unBuntu，数据库选用 MySQL；为了简化算法开发，算法部分使用深度学习框架 Keras；开发软件选用 JetBrains PyCharm。



图 5-12 巡检机器人

5.4.2 测试结果

(1) 登录界面。登录时，用户输入用户名、密码进行身份验证。若验证失败，则提示用户错误信息。该界面还为用户提供注册功能，用户注册账号需填写相关信息。系统登录界面如图 5-13 所示，用户注册界面如图 5-14 所示。



图 5-13 系统登录界面



图 5-14 用户注册界面

(2) 主界面。图 5-15 是系统的主界面，该界面总体上分为两个部分：左侧的菜单栏和右侧的显示部分。菜单栏主要包括：基本信息管理、数据集管理、语义分割模型管理等功能。当用户点击菜单栏的功能按钮，右侧部分会显示相应的界面以供用户进行下一步操作。数据集管理界面。图 5-16 为系统数据集上操作界面。该界面包含数据集的一切管理功能，比如数据集上传、数据集修改、数据集删除等。



图 5-15 主界面图



图 5-16 数据集操作界面

(3) 模型选择界面。如图 5-17 所示, 该界面是系统语义分割模型选择界面。用户可以根据需求选择语义分割模型。在每个语义分割模型的右方都有相应的操作, 比如模型参数设置、模型保存路径设置、模型结果图保存路径设置。模型参数设置界面。图 5-18 是模型参数设置界面。当在模型选择界面中点击模型参数设置时, 则进入模型参数设置界面。该界面主要用于设置模型训练相关的参数, 比如最大迭代次数、基础学习率、学习率衰减率、批训练大小和相应模型的超参数。图 5-19 表示模型训练界面, 用户可从该界面获得训练进度、训练损失、模型准确率等信息。同时, 用户可以终止正在训练的模型、用户可以点击模型详细信息以查看正在训练模型的详细信息——超参数设置、基础模型设置、学习率等。



图 5-17 模型选择界面



图 5-18 模型参数设置界面

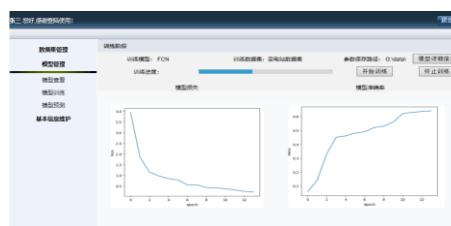


图 5-19 模型训练界面

(4) 模型结果展示界面。图 5-20 模型结果展示界面。该界面总共分为四个部分: 数据集上传部分、模型选择部分、输出模式选择部分以及结果展示界面。数据集上传功能可以选择单张图片、图片集合或者视频。模型选择功能则是选取已经过训练的模型。输出模式选择功能包含三种不同的模式: 视频输出模式、图片输出模

式、视频抓帧输出模式。如果用户输入的数据集是视频类型，则可以选用视频输出模式和视频抓帧输出模式。反之，则是图片输出模式。图 5-21 表示使用模型预测的详细结果。该结果提供了类别、类别数量和类别是否在正前方等详细信息。

类别	数量	正前方	障碍物
电线杆	9	否	是
公路	1	是	否
车行道边缘线	1	否	否
草坪	1	否	否

图 5-20 模型预测界面

图 5-21 详细结果界面

5.5 本章小结

本章设计并实现了一个巡检机器人语义感知系统，可以对巡检机器人采集的图片或视频进行语义分割并显示结果、提示是否为前方障碍物。该系统既包含前面章节设计的语义分割模型，也包含经典的语义分割模型。该系统具有相当地便捷性，能够使用户轻松配置模型和提供语义分割结果。

第六章 总结与展望

6.1 全文总结

随着逐步进入智能化时代，越来越多的行业开始使用智能设备替代人工作业。特别是变电站这样危险的场景，智能巡检机器人逐步开始代替人工进行巡检任务。自主定位、路径规划、环境感知是智能巡检机器人完成巡检任务必不可少的技术。环境感知不仅可以辅助巡检机器人完成路径规划，而且也是巡检机器人完成巡检任务的主要技术。本文主要研究了巡检机器人环境感知中的语义分割任务。主要研究内容总结如下：

(1) 提出了基于多视角注意力机制的语义分割模型。该语义分割模型包含多视角结构、注意力结构和多尺度特征融合结构。这些结构解决了识别物形状差异大、识别物长宽比例悬殊以及局部特征和全局特征的对齐问题。该模型能够从图像特征中以不同方式捕获不同的高级语义信息并有效地融合局部特征和全局特征。在变电站数据集上，该模型取得了不错的效果。

(2) 提出了基于零样本学习的语义分割模型。该模型包含语义分割模型和特征生成模型。特征生成模型借鉴生成对抗网络的思路设计两个判别器，其目的是解决学习投影函数过程中存在的视觉映射偏移问题。在训练阶段，模型的特征生成模块通过样本学习语义空间到视觉空间的投影函数。在预测阶段，该特征生成模块直接将语义特征转化为视觉特征并将视觉特征输入语义分割网络进行预测。

(3) 设计并实现了变电站巡检机器人语义感知系统。根据第三章和第四章所提出的语义分割模型，本文设计并实现了一个变电站场景下的感知语义系统。系统能为巡检机器人所采集的路面信息进行语义分割并提示是否存在障碍物。首先，本文介绍了系统的需求分析、总体设计，然后对系统的功能进行了详细设计，最后采用变电站数据集对系统进行了测试，验证了本文所提出的语义分割模型的可用性。

6.2 后续工作展望

本文对变电站场景的语义分割进行了研究并在分割精度上有一定的提升。但还存在一些不足之处，需要进行下一步研究：

1. 本文研究了一种变电站环境下的语义分割问题。但不同电压等级的变电站具有不同的工况环境，如不同类型的设备。为了探究本论文提出的语义分割方法的鲁棒性，接下来将会把本论文算法应用在不同电压等级的变电站环境下。

2.本文研究了基于零样本的语义分割模型。该模型能够在数据样本较少的情况下识别未知类别。但是数据量与算法精度之间的关系还没有明确的量化。因此，接下来工作是探究该算法的边界。即已知类别数量与算法精度之间的关系、已知类别数量和未知类别相似度对算法精度的影响。

致 谢

时光荏苒，研究生的生涯即将落下帷幕。在这三年的时间里，我不仅从导师、同学以及帮助过我的人身上学会了领域相关的理论知识，而且也从他们身上学会了人生不可或缺的技能。在此，我真诚的向他们表示感谢。

首先，我要向我的导师左琳老师由衷地表示感谢。是她教会了我如何更好地迈入研究生阶段；是她将我引上了科研的道路，每次我遇见科研问题，她都会耐心教导；是她细心地指导我如何对待手上的任务。我研究生阶段的研究方向选择、论文课题研究、项目实践都离不开导师的身影。在老师的悉心培养下，我不仅养成良好的学习习惯，而且也形成了一丝不苟、今日事今日毕的工作态度和踏实的工作作风。

其次，感谢张昌华老师和刘宇老师在科研和项目上对我的帮助。张老师严谨的做事风格和刘老师深厚的专业知识均使我受益匪浅。

然后，我也要感谢实验室的同学。三年的时光说长也长说短也短。记得有一句话：感官上时间的快慢取决于心理是否开心。没有他们的陪伴，也许这三年的时光会相当漫长。他们不仅在生活中关心和帮助我，而且也在我遇见科研或者技术问题的时候，不厌其烦地向我解释。在此，我真诚的感谢他们的陪伴与帮助。同时，也希望我们的感情能够一直延续下去。

我要感谢我的家人。在他们支持和关心下，我才能够安心的求学、完成自己的研究内容和毕业论文。

最后，感谢百忙之中参与本论文评审和答辩的各位专家和老师们！

参考文献

- [1] C. R. Qi, L. Wei, C. Wu, et al. Frustum point nets for 3d object detection from rgb-d data[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 918 - 927.
- [2] 赵晋秀, 刘文杰. 全向底盘机器人智能定位和姿态检测系统——基于正交编码器和陀螺仪[J]. 工业技术创新, 2020, 07(05): 33-37.
- [3] 刘小波, 徐波, 宋爱国, 等. 基于的变电站巡检机器人数字仪表识别算法[C]. 江西省电机工程学会: 江西省电机工程学会, 2019: 6.
- [4] 徐发兵, 吴怀宇, 陈志环, 等. 基于深度学习的指针式仪表检测与识别研究[J]. 高技术通讯, 2019, 29(12): 1206-1215.
- [5] A. E. Assaf, S. Zaidi, S. Affes, et al. Accurate range-free ANN-based localization in wireless sensor networks[C]. In: Proceedings of IEEE International Symposium on Personal, 2017: 1-6.
- [6] O. Kaiwartya, Y. Cao, J. Lloret, et al. Geometry-based localization for GPS outage in vehicular cyber physical systems[J]. IEEE Transactions on Vehicular Technology, 2018, 67(5): 3800-3812.
- [7] 汤义勤, 高彦波, 邹宏亮, 等. 基于机器视觉的室内无轨巡检机器人导航系统[J]. 自动化与仪表, 2020, 35(08): 42-46+76.
- [8] 王吉岱, 郭帅, 孙爱芹, 等. 基于双目视觉技术的高压输电线路巡检机器人在线测距[J]. 科学技术与工程, 2020, 20(15): 6130-6134.
- [9] 郑昌庭, 王俊, 郑克. 基于图像识别的变电站巡检机器人仪表识别研究[J]. 工业仪表与自动化装置, 2020(05): 57-61.
- [10] 臧雪. 巡检机器人自主仪表视觉识别系统的设计与研究[D]. 哈尔滨工业大学, 2016.
- [11] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks[C]. NIPS, 2012.
- [12] K. He, X. Zhang, S. Ren, et al. Deep Residual Learning for Image Recognition[C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770-778.
- [13] 鲜开义, 杨利萍, 周仁彬, 等. 变电站巡检机器人道路语义分割方法及其应用[J]. 科学技术与工程, 2020, 20(15): 6151-6157.
- [14] A. Birk, H. Kenn. An industrial application of behavior-oriented robotics[C]. In: Proceedings of IEEE International Conference on Robotics & Automation, 2006.

- [15] J. Y. Park, J. K. Lee, B. H. Cho, et al. An Inspection Robot for Live-Line Suspension Insulator Strings in 345-kV Power Lines[J]. IEEE Transactions on Power Delivery, 2012, 27(2): 632-639.
- [16] 矫德余. 基于嵌入式系统的智能巡检机器人研制[D]. 中国石油大学(华东), 2010.
- [17] 王建元, 王娴, 陈永辉, 等. 基于图论的电力巡检机器人智能寻迹方案[J]. 电力系统自动化, 2007(09): 78-81.
- [18] 高青. 山西长治久安变电站巡检机器人的应用研究[D]. 华北电力大学, 2012.
- [19] 蔡自兴, 邹小兵. 移动机器人环境认知理论与技术的研究[J]. 机器人, 2004(01): 87-91.
- [20] 李江. 基于红外声纳传感器的机器人自主运动方法[D]. 北京工业大学, 2014.
- [21] S. Thongchai, S. Suksakulchai, D. M. Wilkes, et al. Sonar behavior-based fuzzy control for a mobile robot[C]. Systems, Man, and Cybernetics, 2000 IEEE International Conference on. IEEE, 2000.
- [22] 段丙涛, 杨平, 翟志敏. 基于声纳环传感器的机器人避障研究[J]. 传感器与微系统, 2012, 31(02): 64-66+70.
- [23] H. Jie, T. Supaongprapa, I. Terakura, et al. A model-based sound localization system and its application to robot navigation[J]. Robotics & Autonomous Systems, 1999, 27(4):199-209.
- [24] 黄东武, 李宝森. 水下机器人双频识别声纳系统应用研究[A]. 中国航海学会航标专业委员会测绘学组. 中国航海学会航标专业委员会测绘学组学术研讨会学术交流论文集[C]. 中国航海学会航标专业委员会测绘学组:中国航海学会, 2009: 6.
- [25] 杨明, 王宏, 何克忠, 张钹. 基于激光雷达的移动机器人环境建模与避障[J]. 清华大学学报(自然科学版), 2000(07): 112-116.
- [26] A. Nuchter, H. Surmann, K. Lingemann, et al. 6D SLAM with an application in autonomous mine mapping[C]. In: Proceedings of IEEE International Conference on Robotics & Automation. IEEE, 2004.
- [27] J. Serafin, E. Olson, G. Grisetti. Fast and robust 3d feature extraction from sparse point clouds[C]. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016: 4105-4112.
- [28] B. Steux, O. E. Hamzaoui, tinySLAM: A SLAM algorithm in less than 200 lines C-language program[C]. 2010 11th International Conference on Control Automation Robotics & Vision, 2010, pp. 1975-1979, doi: 10.1109/ICARCV.2010.5707402.
- [29] 杨涛, 李祎, 陈晶华, 等. 基于背景差分的巡检机器人视觉识别方法[J]. 机械与电子, 2020, 38(12): 60-64.
- [30] 薛阳, 江天博, 张晓宇. 基于视觉的变电站巡检机器人导航线提取方法[J]. 广东电力, 2015, 28(12): 13-18.

- [31] 赵坤, 赵书涛. 基于路面标识的变电站巡检机器人单目视觉导航[J]. 电力信息与通信技术, 2014, 12(03): 81-84.
- [32] D. Wei, S. Zhang. A Visual Navigation Method of Substation Inspection Robot[C]. In: Proceedings of IEEE International Conference on Progress in Informatics and Computing, 2016.
- [33] 景晓军, 蔡安妮, 孙景鳌. 一种基于二维最大类间方差的图像分割算法[J]. 通信学报, 2001(04): 71-76.
- [34] R. Pohle, K. D. Toennies. A New Approach for Model-Based Adaptive Region Growing in Medical Image Analysis[C]. In: Proceedings of International conference on computer analysis of images and patterns, 2001: 238-246.
- [35] L. Zhang, J. Jin, H. Talbot. Unseeded region growing for 3D image segmentation[C]. In: Proceedings of International Conference Proceeding Series, 2000, 9: 31-37.
- [36] 马范援, 于水. 一种基于数据融合的医学图像分割方法[J]. 计算机辅助设计与图形学学报, 2001(12): 1073-1076.
- [37] R. Buettner, H. Baumgartl. A highly effective deep learning based escape route recognition module for autonomous robots in crisis and emergency situations[C]. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.
- [38] K. Asadi, P. Chen, K. Han, et al. Real-time Scene Segmentation Using a Light Deep Neural Network Architecture for Autonomous Robot Navigation on Construction Sites[J]. The 2019 ASCE International Conference on Computing in Civil Engineering, 2019.
- [39] B. Ummenhofer, H. Zhou, J. Uhrig, et al. Demon: Depth and motion network for learning monocular stereo[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5038-5047.
- [40] T. Zhou, M. Brown, N. Snavely, et al. Unsupervised learning of depth and ego-motion from video[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1851-1858.
- [41] V. Badrinarayanan, A. Kendall, R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
- [42] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [43] J. Long, E. Shelhamer, T. Darrell. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651.

- [44] L. C. Chen, G. Papandreou, I. Kokkinos, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Computer Science, 2014(4): 357-361.
- [45] H. Noh, S. Hong, B. Han. Learning deconvolution network for semantic segmentation[C]. In: Proceedings of the IEEE international conference on computer vision, 2015: 1520-1528.
- [46] H. Zhao, J. Shi, X. Qi, et al. Pyramid scene parsing network[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2881-2890.
- [47] G. Lin, A. Milan, C. Shen, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1925-1934.
- [48] S. Arif, K. Knapp, G. Slabaugh. SPNet: Shape Prediction Using a Fully Convolutional Neural Network[M]. Medical Image Computing and Computer Assisted Intervention-MICCAI, 2018.
- [49] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
- [50] A. Shen, T. Y. Zhou, G. D. Long, et al. Disan: Directional self-attention network for rnn/cnn-free language understanding[C]. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [51] X. Wang, R. Girshick, A. Gupta, et al. Non-local neural networks[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7794-7803.
- [52] Y. Yuan, J. Wang. Ocnet: Object context network for scene parsing[J]. arXiv preprint arXiv: 1809.00916, 2018.
- [53] J. Fu, J. Liu, H. Tian, et al. Dual attention network for scene segmentation[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3146-3154.
- [54] X. Li, Z. Zhong, J. Wu, et al. Expectation-Maximization Attention Networks for Semantic Segmentation[C]. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9167-9176.
- [55] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22.
- [56] A. Garcia, S. Orts, S Oprea, et al. A survey on deep learning techniques for image and video semantic segmentation[J]. Applied Soft Computing, 2018, 70: 41-65.
- [57] A. Garcia-Garcia, S. Orts-Escalano, S. Oprea, et al. A review on deep learning techniques applied to semantic segmentation[J]. arXiv preprint arXiv:1704.06857, 2017.

- [58] O. Ronneberger, P. Fischer, T. Brox. U-net: Convolutional networks for biomedical image segmentation[C]. In: Proceedings of International Conference on Medical image computing and computer-assisted intervention, 2015: 234-241.
- [59] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [60] J. Fu, J. Liu, H. Tian, et al. Dual attention network for scene segmentation[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3146-3154.
- [61] W Wang, V. W. Zheng, H. Yu, et al. A survey of zero-shot learning: Settings, methods, and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-37.
- [62] Y. Li, J. Zhang, J. Zhang, et al. Discriminative Learning of Latent Features for Zero-Shot Recognition[C]. In: Proceedings of theIn: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 7463-7471.
- [63] Y. Li, D. Wang, H. Hu, et al. Zero-Shot Recognition Using Dual Visual-Semantic Mapping Paths[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 5207-5215.
- [64] Y. Xian, S. Choudhury, Y. He, et al. Semantic Projection Network for Zero and Few Label Semantic Segmentation[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 8248-8257.
- [65] 王波. 基于对抗学习与注意力机制的图像语义分割[D]. 湘潭: 湘潭大学, 2019
- [66] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 1-9.
- [67] L. C. Chen, G. Papandreou, F. Schroff, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.
- [68] G. J. Brostow, J. Shotton, J. Fauqueur, et al. Segmentation and recognition using structure from motion point clouds[C]. In: Proceedings of European conference on computer vision, 2008: 44-57.
- [69] G. J. Brostow, J. Fauqueur, R. Cipolla. Semantic object classes in video: A high-definition ground truth database[J]. Pattern Recognition Letters, 2009, 30(2):88-97.
- [70] A. Geiger, P. Lenz, C. Stiller, et al. Vision meets robotics: the KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11):1231-1237.

- [71] M. Cordts, M. Omran, S. Ramos, et al. The cityscapes dataset for semantic urban scene understanding[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 3213-3223.
- [72] D. Kingma, J. Ba. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.
- [73] C. Yu, J. Wang, C. Peng, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]. In: Proceedings of the European conference on computer vision (ECCV), 2018: 325-341.
- [74] X. Yuan, B. Huang, Y. Wang, et al. Deep Learning-Based Feature Representation and Its Application for Soft Sensor Modeling With Variable-Wise Weighted SAE[J]. IEEE Transactions on Industrial Informatics, 2018: 1-1.
- [75] E. Kodirov, T. Xiang, S. Gong. Semantic autoencoder for zero-shot learning[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 3174-3183.
- [76] D. P. Kingma, M. Welling. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
- [77] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial networks[J]. arXiv preprint arXiv: 1406.2661, 2014.
- [78] L. Bottou. Stochastic gradient descent tricks[M]. Neural networks: Tricks of the trade, 2012: 421-436.
- [79] M. Bucher, T. H. Vu, M. Cord, et al. Zero-shot semantic segmentation[J]. arXiv preprint arXiv:1906.00817, 2019.

攻读硕士期间取得研究成果

一. 科研项目情况

- [1] 国家自然基金：自适应人工智能在线教育关键技术研究，项目编号:G056187709，主研.
- [2] 四川省科技重点研发计划：电缆隧道巡检智能机器人自主导航技术研究，项目编号：2018GZ0396，参研.

二. 已发表论文

- [1] **Qian Chen**, Changhua Zhang, Hao Li, et al. Semantic Segmentation of Substation Scenes Using Attention-based Model[C]. In: Proceedings of IEEE International Conference on Electronics Technology, chengdu, 2021.
- [2] Naijia. Wan, Changhua. Zhang, **Qian. Chen**, et al. A Multi-scene Recognition Model with Multi-dimensional Domain Adaptation[C]. In: Proceedings of IEEE International Conference on Electronics Technology, chengdu, 2021.