

武汉大学

本科毕业论文（设计）

跨分辨率遥感影像 精准信息提取方法研究

姓名：张 铢 琦

学号：2020300004045

专业：计算机科学与技术

学院：计算机学院

指导教师：肖 晶

二〇二四年四月

原创性声明

本人郑重声明：所呈交的论文（设计），是本人在指导教师的指导下，严格按照学校和学院有关规定完成的。除文中已经标明引用的内容外，本论文（设计）不包含任何其他个人或集体已发表及撰写的研究成果。对本论文（设计）做出贡献的个人和集体，均已在文中以明确方式标明。本人承诺在论文（设计）工作过程中没有伪造数据等行为。若在本论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名： 指导教师签名：
日 期： 年 月 日

版权使用授权书

本人完全了解武汉大学有权保留并向有关部门或机构送交本论文（设计）的复印件和电子版，允许本论文（设计）被查阅和借阅。本人授权武汉大学将本论文的全部或部分内容编入有关数据进行检索和传播，可以采用影印、缩印或扫描等复制手段保存和汇编本论文（设计）。

作者签名： 指导教师签名：
日 期： 年 月 日

摘 要

随着遥感技术的飞速进步，我们获得了丰富的高分辨率(HR)和低分辨率(LR)遥感图像资源。HR 图像能提供丰富的细节信息，然而却面临着成本和连续性等限制。易于获取且成本低廉的LR 图像具有很高的应用潜力。然而，如何有效地从 LR 图像中提取难以直观识别的信息，成为了一个挑战。本文针对这一问题，在 CRVC 跨分辨率车辆计数数据集上设计并实现了一种基于注意力机制的深度学习 U 型网络模型，旨在充分利用遥感图像中隐含的空间和时间信息，从而提高目标计数的准确性。

本研究对 CRVC 数据集中的高分辨率和低分辨率图像之间的空间一致性及时间连续性进行了分析，并设计了时间注意力模块和跨分辨率注意力模块来针对性的处理数据。研究展示了注意力机制在提升模型性能中的应用潜力，尤其是在增强模型对多种特征表示的综合能力的方面。本文模型在车辆计数的准确性和模型泛化能力方面取得了显著的进步。通过设计消融实验验证模型设计的合理性，与当前领先的计数方法进行比较验证模型性能。引入了焦点损失来解决类别分布不平衡的问题，在小型货车和大型货车的计数结果上取得了显著的提升。本文的研究成果不仅在车辆计数领域具有重要的应用价值，对于其他涉及到跨分辨率图像以及利用低分辨率图像的问题上，该模型也具有很强的迁移潜力。

关键词：遥感图像，跨分辨率，目标计数，注意力机制，深度学习，U型网络

ABSTRACT

With the rapid advancement of remote sensing technology, we now can acquire a wealth of high-resolution (HR) and low-resolution (LR) remote sensing image resources. HR images can provide rich detail information, but they are limited by cost and continuity constraints. On the other hand, easily obtainable and cost-effective LR images hold great potential for application. However, effectively extracting subtle and non-intuitively recognizable information from LR images poses a significant challenge. Addressing this issue, this paper designs and implements an attention mechanism-based deep learning U-Net model on the CRVC cross-resolution vehicle counting dataset, aiming to fully leverage the implicit spatial and temporal information in remote sensing images to enhance the accuracy of object counting.

This study analyzes the spatial consistency and temporal continuity between HR and LR images within the CRVC dataset and introduces targeted processing through a temporal attention module and a cross-resolution attention module. The research demonstrates the potential of attention mechanisms to boost model performance, particularly in enhancing the model's ability to synthesize diverse feature representations. By comparing with leading counting methods and further optimizing the algorithm through ablation studies, our model shows significant improvements in vehicle counting accuracy and model generalization ability. A focal loss was introduced to address the issue of class distribution imbalance, achieving significant improvements in the counting results for small and large trucks. The research findings presented in this paper are not only valuable for the field of vehicle counting but also possess strong transfer potential for other issues involving cross-resolution images and the use of low-resolution images.

Keywords: Remote Sensing Images; Cross Resolution; Object Counting; Attention Mechanism; Deep Learning; U-Net

目 录

摘要	I
ABSTRACT	II
1 绪论	1
1.1 研究背景与意义	1
1.1.1 高分辨率和低分辨率遥感图像简介	1
1.1.2 研究意义	3
1.2 国内外研究现状	4
1.2.1 跨分辨率遥感影像目标计数现状	4
1.2.2 跨分辨率车辆计数数据集	6
1.3 研究目标与内容	9
1.3.1 研究目标	9
1.3.2 研究内容	9
1.4 章节安排	10
2 目标计数基础理论和技术框架	11
2.1 深度学习基础理论	11
2.1.1 卷积层	11
2.1.2 填充和步幅	11
2.1.3 激活函数	12
2.1.4 池化层	12
2.1.5 权重衰减	13
2.1.6 暂退法	13
2.1.7 批量归一化	13
2.2 目标计数	14
2.3 U-Net 网络	14
2.3.1 模型结构	15
2.3.2 收缩路径	15
2.3.3 扩展路径	15
2.3.4 双线性插值	16
2.3.5 转置卷积	17

2.4	注意力机制	17
2.4.1	注意力机制基本原理	17
2.4.2	注意力机制在图像领域的应用	19
2.4.3	注意力机制处理序列图像	19
2.5	Attention U-Net 网络	20
2.6	Flow1D 网络	21
2.7	CRVC 网络	23
2.7.1	网络设计	23
2.7.2	回归模型	24
2.7.3	模型局限	24
3	基于注意力机制的跨分辨率遥感影像计数	25
3.1	问题分析	25
3.2	网络设计	26
3.3	编码器	27
3.4	多来源注意力机制	28
3.5	解码器	29
3.6	损失函数	30
3.6.1	轿车计数阶段	30
3.6.2	其余类别计数阶段	30
4	跨分辨率遥感影像计数实验及分析	33
4.1	实验设计	33
4.2	模型训练	33
4.3	分割精度测试	34
4.3.1	消融实验	35
4.3.2	实验结果分析	35
4.4	覆盖率估计测试	36
4.5	车辆计数回归结果测试	36
4.6	训练细节	38
4.7	局限	38
5	总结与展望	40
5.1	结论	40

5.2 不足与展望	40
参考文献	41

1 結論

随着遥感技术的快速发展，我们能获得到越来越丰富的遥感观测数据。既有高分辨率的清晰图像，也包括低分辨率的鸟瞰图。这些图像从更宏观的视角为我们提供了许多有价值的数据。然而如何合理的使用这些图像数据，特别是从低分辨率图像中找出更多难以通过人类肉眼直接识别出的有效信息成了一个重要的研究课题。在车辆计数等真实场景中，高分辨率图像虽然能提供丰富的细节，但有着获取成本高昂以及难以获取稳定连续的高频率数据的问题。本文在跨分辨率车辆计数数据集（Cross Resolution Vehicle Counting,CVRC）^[1]上，设计了一种基于注意力机制^[2]的深度学习U型网络模型，旨在提高跨分辨率遥感图像中隐含的空间和时间信息进行更丰富全面的表示，以进一步提升目标计数的准确性。

1.1 研究背景与意义

1.1.1 高分辨率和低分辨率遥感图像简介

在遥感领域，分辨率是用来描述遥感图像细节程度的一个重要指标，根据遥感卫星搭载的传感器不同，可以从多个维度进行描述，包括空间分辨率、时间分辨率、光谱分辨率和辐射分辨率。

空间分辨率指的是传感器能够区分的最小地面单元的尺寸，也称为地面采样距离（Ground Sample Distance,GSD）。例如，如果一个卫星图像的空间分辨率是10米，那么图像上的一个像素代表真实世界中10米×10米的区域。空间分辨率越高（即图像中单个像素所代表的地面面积较小），图像的细节就越丰富，能够观察到更小的对象。

时间分辨率定义为重新访问并获取同一地点数据的所需的时间。重访周期是指卫星完成一个完整轨道周期所需的时间长度。因此，遥感系统第二次以相同视角对完全相同的区域进行成像的绝对时间分辨率等于该周期。然而，由于大多数卫星相邻轨道的成像带存在一定程度的重叠，并且这种重叠随着纬度的增加而增加，地球的某些区域往往会更频繁地重新成像。因此，传感器的实际时间分辨率取决于多种因素，包括卫星/传感器能力、测绘带重叠和纬度。

光谱分辨率描述的是遥感设备能够在电磁光谱中区分不同波长的能力。通常可以使用非常宽的波长范围（可见光和近红外）来区分广泛的类别，例如水和植被。其他更具体的类别，例如不同的岩石类型，就需要高光谱分辨率传感器在更精细的波长范围内进行比较才能将它们分开。

辐射分辨率描述的是遥感设备检测不同能级（亮度）的能力。不同传感器对电磁能大小的敏感度决定了辐射分辨率。传感器的辐射分辨率越精细，它对检测反射或发射能量的微小差异就越敏感。这对于分析物体的热特性、地表材料的性质等非常有帮助。

在讨论高分辨率 (High Resolution, HR) 图像和低分辨率 (Low Resolution, LR) 图像时，通常指的是空间分辨率的差异。空间分辨率的详细分类可以参照^[3]:

1. 低分辨率定义为地面采样距离为 30 m 或更大
2. 中分辨率的地面采样距离范围为 2–30 m
3. 高分辨率的地面采样距离范围为 0.5–2 m
4. 极高分辨率的地面采样距离范围 <0.5 m

1.1.1.1 高分辨率图像

高分辨率图像通常指具有较高空间分辨率的图像，即图像中单个像素所代表的地面面积较小，能够显示更加精细的地面特征^[4]。高分辨率图像使得用户可以观察到较小的地面对象，例如单个车辆、道路标线甚至是行人。虽然“高分辨率”这个术语没有绝对的定义，但在遥感领域，我们把空间分辨率小于 1 米（通常在 0.3 米到 1 米之间）的图像常被认为是高分辨率图像。目前我们可以使用的高分辨率遥感图像来源主要有：航空摄影（搭载高分辨率摄像机或低空高分辨率无人机拍摄的数据）和某些高性能的卫星遥感仪器，例如 WorldView 系列^[5]、GeoEye、QuickBird 等。高分辨率图像中的精细地面特征信息，在城市规划、交通监控、农业监测（如作物健康分析）、详细的地物分类、灾害评估等领域都有广泛的使用。特别是在目标识别领域中，高分辨率的图像可以使用深度学习中多种模型和方法，具有很高的应用价值。

1.1.1.2 低分辨率图像

低分辨率图像指的是空间分辨率较低的图像，即图像中单个像素所代表的地面面积较大，只能显示较为粗糙的地面特征。低分辨率图像难以分辨较小的地面对象，但适合于观察大范围的地表变化。通常空间分辨率大于 10 米（如 10 米、30 米或更大）的图像被认为是低分辨率图像。低分辨率图像主要来源于具有宽幅覆盖能力的卫星遥感仪器，如 MODIS（具有数百米的空间分辨率）、Landsat 系列（15 米到 30 米分辨率）、Sentinel-2（10 米到 60 米分辨率）等。图像中的大范围地表特征，在气候变化研究、大范围土地覆盖变化监测、环境监测、城市发展规划、海洋

和大气研究等领域有着很高的应用价值。

对于不同任务，识别目标的大小不尽相同，对于图像分辨率的要求也随之变化。对于房屋的识别需要 2-5 米的分辨率，识别车辆需要 1 米以内的分辨率，行人的识别则需要 0.3 米或者更高。本文讨论的 CVRC 数据集是车辆识别数据集，将分辨率大于等于 1 米的图像认定为低分辨率图像，小于 1 米的视为高分辨率图像。对于目标识别以及目标计数领域来说，模糊的图像使得目标的轮廓极为模糊，稠密目标轮廓间的重叠更加大了模型识别的难度。

1.1.1.3 重访周期与成本

对于遥感卫星而言，其拍摄影像的分辨率是由传感器在地面采样的间隔决定的，即每个传感器探测元件在地面投影的大小，称为地面采样距离 (Ground Sample Distance, GSD)。在理想状况下，地面采样距离可以通过下式计算：

$$GSD = \frac{dR}{f} \quad (1.1)$$

其中 d 为传感器探测元件的宽度， R 为传感器距离地面高度， f 为光学系统焦距。

通过万有引力定律和开普勒第三定律，可推出遥感卫星绕地球旋转的周期与轨道半径之间的关系。

$$T = 2\pi \sqrt{\frac{r^3}{GM}} \quad (1.2)$$

由式1.1和式1.2，可以分析高低分辨率图像拍摄卫星的轨道条件。高分辨率图像一般由低轨卫星拍摄，具有较小的轨道半径和周期。同时也因此具有较小的视场 (the field of view, FOV) 较小。高分辨率图像分辨率卫星通常使用任务驱动模式进行地球观测。这意味着，如果不提前提交观测任务，一颗卫星需要 6 个月才能完成全球覆盖，获得特定区域的图像的重访周期相当长。此外，高分辨率图像非常昂贵，例如 WorldView-3 的价格为 34 美元/平方公里。作为对比，低分辨率卫星的往往运行在更高的轨道上，虽然空间分辨率有所降低，但视场较大。因此获得同一地点重访周期要短得多，例如 PlanetScope 卫星每天重访一次，价格也低得多，为 1.8 美元/平方公里。不同分辨率卫星的拍摄成本和覆盖周期如下表1.1所示。

1.1.2 研究意义

单独依靠低分辨率图像进行目标计数是十分困难的，而仅通过高分辨率图像进行目标计数，不仅花费巨大，同时还需要面对连续监控数据的缺失。如何通过低成本且具有时间连续性的低分辨率图像进行目标计数及实时监测就成为解决问题

表 1.1 不同分辨率卫星拍摄成本及覆盖周期

名称	分辨率	重访周期	全球覆盖周期	花费
WorldView-3	0.3m	4.5 天	6 个月	34 美元/ km^2
GeoEye-1	0.4m	4 天	6 个月	29.5 美元/ km^2
SuperView-	0.5m	4 天	6 个月	23 美元/ km^2
QuickBird	0.6m	7 天	6 个月	17.5 美元/ km^2
Spot 6/7	1.5m	2 天	1 个月	5.75 美元/ km^2
PlanetScope	3m	小于 1 天	小于 5 天	1.8 美元/ km^2
RapidEye	5m	1 天	小于 1 个月	1.28 美元/ km^2

题的关键。本文提出的方法通过少量高分辨率图像的辅助，在时间连续的低分辨率图像上实现目标计数，具有很高的应用价值。该方法不仅局限于目标计数，更好地利用了低分辨率图像中蕴含的模糊信息，在稠密车流人流识别、智慧城市设计等领域也有很高的应用潜力。

1.2 国内外研究现状

1.2.1 跨分辨率遥感影像目标计数现状

高分辨率图像计数主要有三类方法。基于检测的计数方法通过识别出具体的物体位置来进一步计数。基于回归的计数类方法通过学习出图像与图像中对应物体的数目的对应关系，从而估计出目标数目。基于密度图的计数方法生成图像对应的密度图，通过密度分布求和计算出最终的数量估计。

1.2.1.1 基于检测的计数方法

基于检测的目标计数是目标检测下的一个分支。目标检测作为计算机视觉的一个主要研究方向，主要研究的是如何利用计算机视觉技术识别特定对象并确定其大小和位置。目前随着深度学习技术的发展，利用神经网络进行目标检测已经成为主流方案。主要的目标检测算法包括 RCNN 系列^[6-8]、SSD^[9]、YOLO 系列^[10, 11]等。应用这些方法，可以通过先检测出目标位置，再进一步计数从而满足目标计数的需求。

这些方法虽然在目标识别上极为准确，但对数据集的质量和分辨率要求非常高。它们依赖于精确的边界标注，通常需要大量的人工标注后的高分辨率图像以确保检测的准确性。但是在 CVRC 数据集中，只包含少量具有人工标注的高分辨率图像，其余大量数据均为低分辨率图像。这些低分辨率图像上的目标识别难度

大，以至于即便是人工也难以准确识别出具体的车辆数目。在这种情况下，即使是研究人员也需要借助同一地点不同时间拍摄的高分辨率图像来辅助识别。

除此之外，目标检测算法在处理极度拥挤或遮挡严重的环境时也面临挑战。为了解决这些问题，一些研究开始集中于开发更加鲁棒的检测算法，这些算法可以更好地处理低质量图像和复杂场景、进一步地，为了降低对高分辨率标注数据的依赖，研究人员也在探索半监督和弱监督学习方法^[12]。这些方法通过利用少量标注数据与大量未标注数据，能够有效地提升模型的泛化能力和减少人工标注的工作量。此外，迁移学习技术^[13] 也被广泛使用，先在高分辨率图像上训练得到的模型，使用该模型在低分辨率图像的进行微调来提高目标检测性能。

1.2.1.2 基于回归的计数方法

基于回归的计数方法^[14-16] 从深度神经网络提取的特征中直接进行数目的回归估计，常常用于处理目标密集、相互遮挡严重的场景，如人群计数。贝叶斯泊松回归模型^[16] 考虑数据的内在波动性，为估计值提供置信区间，从而增强模型在各种拥挤场景下的鲁棒性。隐私保护人群监控^[15] 通过模糊处理技术确保在计数过程中不泄露个人信息，适用于对隐私要求严格的应用场景。Shi 等人^[17] 设计的模型通过训练多个回归器并引入负相关性来提高总体预测的准确性和泛化能力。这些方法的共同点在于它们都通过深入分析和学习图像特征来直接预测目标数量，无需进行繁琐的目标检测和识别，从而在处理极度拥挤场景时显示出显著优势。

1.2.1.3 基于密度图的计数方法

基于密度图的计数方法通过估计目标区域的目标密度，从而计算出数量估计。为了生成像素级的密度估计图像，语义分割中的许多方法也被研究人员用以参考。主要的语义分割模型包括全卷积网络 (FCN)^[18]、U-Net^[19]、SegNet^[20]、DeepLab^[21-23] 系列、Mask R-CNN^[24] 和 Attention U-Net^[25]。目前主要的密度图估计方法包括 CSRNet^[26]、MCNN^[27] 和 CrowdNet^[28]。这些方法通过使用不同架构的卷积神经网络来处理稠密场景下的复杂视觉信息，以精确地估计高密度目标如人群的密度密度。CSRNet 采用了带扩张卷积的深层网络，能够在保持有效感受野的同时，更好地捕获人群密集区域的细节信息。而 MCNN 通过多列卷积结构来适应不同尺度的人群密度，这对于处理不同距离拍摄的人群图像特别有效。CrowdNet，结合浅层和深层网络架构，分别识别高低分辨率中的特征信息，来有效捕捉人群图像中的多尺度信息，从而使得模型能够在复杂的人群场景中更准确地估计人数。

尽管基于密度图的方法在稠密目标数量估计问题中显示出了优越的性能，它们依然面临一些挑战。首先，这些方法往往需要大量标记数据来训练深度学习模型，而高质量的标记数据获取成本较高，尤其是在极端环境下的人群场景。此外，这些模型在遇到极端天气条件或光线不佳的情况下，性能可能会显著下降。Liu 等人^[29]提出了一种通过增强数据的方法来提升模型在不同环境下的鲁棒性，该方法通过模拟不同天气和光照条件下的图像，增强了模型对这些变化的适应性。另外，Boominathan 等人^[28]则关注模型泛化能力的提升，他们通过引入域自适应技术减少了模型对特定数据集的依赖，从而提高了模型在未见过场景中的表现。

上述几种方法在大型高分辨率图像数据集下均有着不错的表现。但对于 CRVC 数据集来讲，由于高分辨率图像的稀缺，这些方法不能很好的完成 CRVC 数据集中的任务。如何通过少量高分辨率图像信息指导改进大量低分辨率图像上估计的结果成为模型设计的关键。

1.2.2 跨分辨率车辆计数数据集

跨分辨率车辆计数数据集^[1] (CRVC) 收集了日本常陆那珂港的 192 张极低分辨率图像和 8 张高分辨率图像，日期范围为 2016 年至 2019 年。

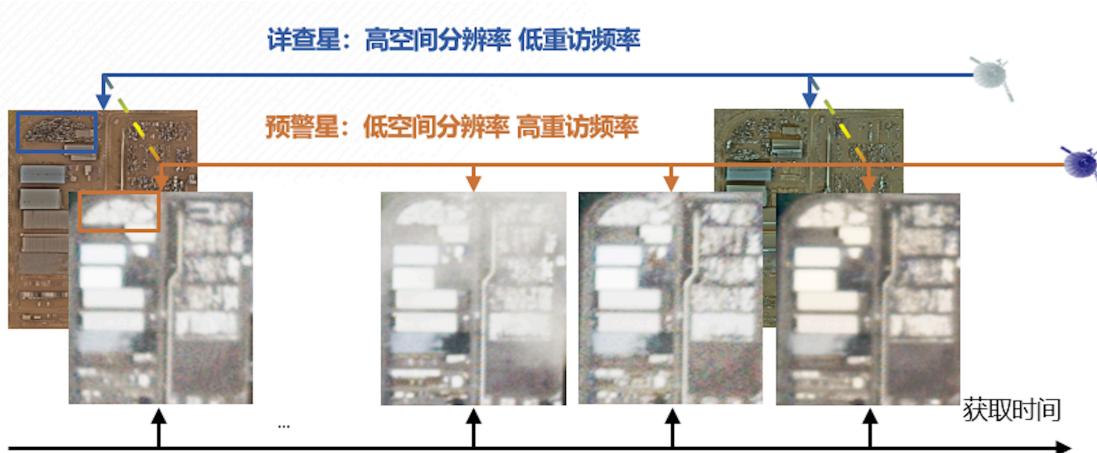


图 1.1 数据集概览

其中 LR 图像是 www.planet.com 下载的，由 PlanetScope 卫星拍摄，地面分辨率为每像素 3m。为了起到监督作用，HR 图像是在根据相应 LR 图像的日期选择的，这些图像是从 WorldView 捕获的，地面分辨率为每像素 30 厘米。

1.2.2.1 数据集分析

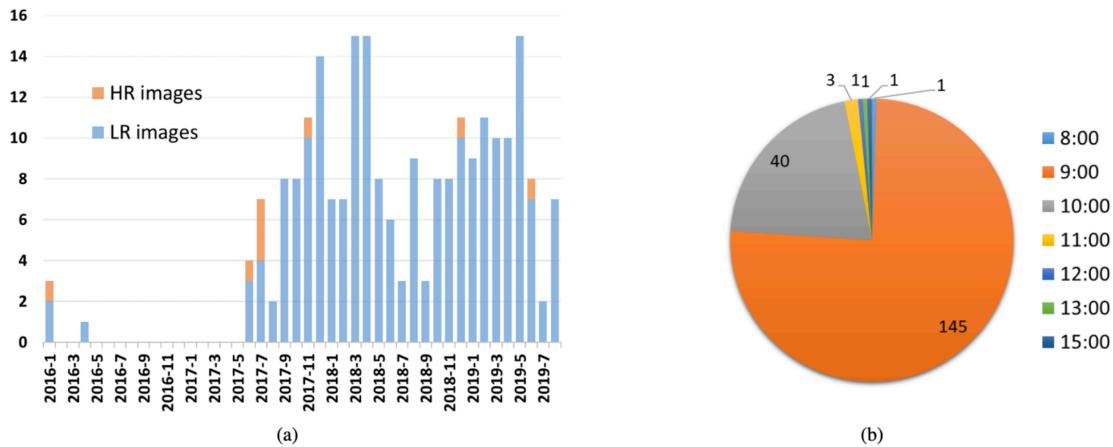


图 1.2 数据的日期和时间分布

图 1.2 显示了数据集中 LR 和 HR 图像采集的日期和时间分布。可以观察到，数据的分布并不平均，主要集中在 2017 年 5 月至 2019 年 7 月之间。就 8 张高分辨率图像的日期而言，它们之间的时间间隔最小为 6 天，最大为 17 个月。如表格 1.2 中数据所示，HR 图像和相应 LR 图像之间的采集时间并不完全一致，平均采集时间差异为 39 分钟。在这种情况下，我们认为短时间内的 HR 和 LR 图像中的车辆数目一致。这一假设和实际情况相符并大大降低了建模难度。

表 1.2 HR 和 LR 图像的获取日期与时间

日期	HR 拍摄时间	LR 拍摄时间
4th January, 2016	10:56	10:36
26th June, 2017	10:24	9:34
2nd July, 2017	10:20	9:36
9th July, 2017	10:34	9:44
15th July, 2017	10:30	9:41
9th November, 2017	10:25	9:42
19th December, 2018	10:46	9:56
6th June, 2019	10:35	10:29

1.2.2.2 高分辨率图像处理

为了进行计数任务，该数据集在 HR 图像上标注了车辆的边界及类别。HR 图像上标注框的数量作为对应日期的 LR 图像的真值。标注的边界则作为停车场位置的空间提示信息。该数据集中总共注释了 37852 个车辆实例，包含四类车辆，包

括轿车、小型卡车、大型卡车和起重机（图 1.3）。不同类别的车辆在尺寸形状上有很大的不同，分类计数有助于提升计数质量。各类车辆数量极不平衡，分别为轿车 35844 辆、小型货车 737 辆、大型货车 1211 辆、起重机 60 辆。

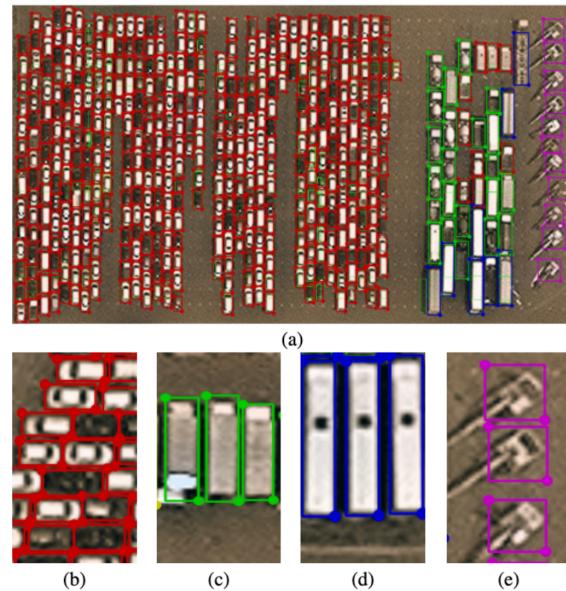
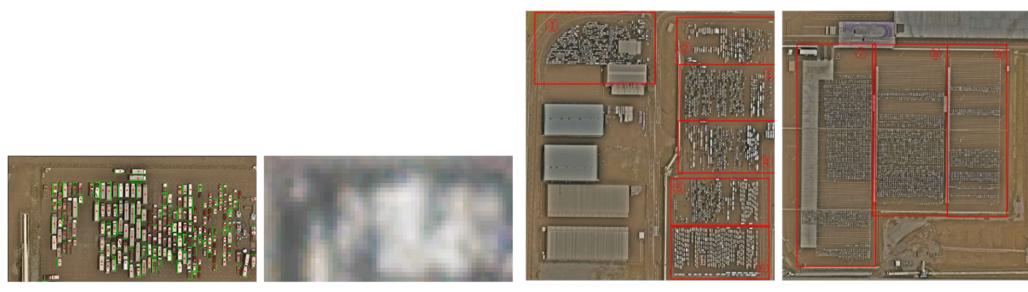


图 1.3 HR 图像车辆标注及分类

1.2.2.3 低分辨率图像处理

数据集中共包含 192 张低分辨率图像。这些图像在 2016 至 2019 年之间被采集，主要分布在 2017 年 6 月至 2019 年 8 月之间，其中 61.5% 的采集间隔在两天以内。如图 1.2 所示，低分辨率图像的采集时间为日间，大部分图像拍摄于上午 9 点到 10 点。我们可以认为这些图像具有相似的拍摄条件。不同于高分辨率图像，低分辨率图像的标注要困难得多。因为难以在低分辨率的条件下辨认清晰的车辆轮廓，所以标注车辆覆盖率是一个更可行的方法。由于低分辨率图像的视场较大，车辆区域只占图像中很小的一部分。因此先在 HR 图像中划出 9 个区域（图 1.4b），在 LR 图像中对应位置进行覆盖率估计（图 1.4a）。



(a) 区域 2 对应的 HR 和 LR 图像

(b) 9 个停车区域

低分辨率图像可以分为两类，有对应高分辨率图像的和没有对应高分辨率图像的。对于前者而言，直接从对应高分辨率标注结果中计算覆盖率即可。那些没有 HR 图像对应的 LR 图像则由多名专家进行视觉标注并取平均值。

1.2.2.4 数据集特性分析

通过上述对数据集的分析，我们可以看到高分辨率图像和与其对应的低分辨率图像之间间隔时间不大，同时对应的是同一位置。因此我们可以认为在它们上进行的目标计数结果应当也相同。两者所映射的空间信息应该具有一致性，这也给后续处理方法提供了思路。如何应用高分辨率图像信息指导改进低分辨率图像上估计的结果成为提高估计精度的关键节点之一。

同时单一的低分辨率图像很难得出合理的目标计数估计，然而由于低分辨率遥感图像的短重访周期，我们能得到一段连续的低分辨率图像。比较相邻图像间的变化或者从多个图像间进行学习，可以补充那些单张图像因低分辨率造成的信息缺失。图像间的时间连续性也是指导改进图像计数估计结果的关键因素。

1.3 研究目标与内容

1.3.1 研究目标

1. 跨分辨率车辆计数算法的开发：开发一种新的基于深度学习的车辆计数算法，该算法能够有效利用有限的高分辨率图像来指导低分辨率图像中的车辆计数。
2. 探究高效利用数据集中空间一致性和时间连续性信息的方式：探究同一时刻高分辨率与低分辨率图像间的空间一致性和连续低分辨率图像间的时间连续性的高效利用方式。研究不同分辨率图像对算法效果的具体影响。
3. 探究注意力机制在跨分辨率目标计数的应用：探索注意力机制在提高跨分辨率车辆计数准确性中的应用，尤其是如何通过注意力机制来增强模型综合多种特征表示的能力。

1.3.2 研究内容

1. 研究背景与意义分析：分析相关遥感技术的发展背景，高分辨率与低分辨率遥感图像的特点及其在车辆计数中的面临困难和挑战。梳理了当前目标检测、语义分割、基于回归和密度图的计数方法等方面的研究进展，特别关注跨分辨率图像处理及车辆计数领域的最新研究成果。

2. 跨分辨率车辆计数数据集分析：对 CRVC 数据集进行分析调研，了解其数据组成分布以及数据中隐含的性质的分析及建模。
3. 基于注意力机制的车辆计数模型设计：设计并实现一种新的基于注意力机制的深度学习模型，用于提高跨分辨率遥感图像车辆计数的准确性。
4. 算法的性能评估及优化：在 CRVC 数据集上评估设计算法的性能，进行消融实验，并与现有的车辆计数方法进行比较。进一步优化算法，以达到更高的计数准确性和更好的泛化能力。

1.4 章节安排

第一章为绪论，整体介绍研究背景及研究内容。第二章为目标计数方法计数框架，详细解释了目标计数领域的方法细节。第三章介绍了基于注意力机制的跨分辨率遥感影像计数方法，详细解释了基于注意力机制的深度学习网络设计细节。第四章设计了多组实验从分割精度、覆盖率估计和回归计数结果这三个方面测试了模型的性能。第五章对本设计做出总结和展望。

2 目标计数基础理论和技术框架

2.1 深度学习基础理论

目前主要的目标识别以及计数方法基于深度学习模型设计的，本小节介绍本文模型涉及到的深度学习相关理论。

2.1.1 卷积层

对于 CRVC 数据集中的图像数据，常使用卷积层^[30, 31]来进行特征提取。卷积层具有的平移不变性和局部性非常适合处理图像数据，可以掌握图像的空间特征。下面给出卷积层的基本定义。

$$[\mathbf{H}]_{i,j} = u + \sum_{a=-\Delta}^{\Delta} \sum_{b=-\Delta}^{\Delta} [\mathbf{V}]_{a,b} [\mathbf{X}]_{i+a,j+b} \quad (2.1)$$

通过使用系数 $[\mathbf{V}]_{a,b}$ 对位置 (i,j) 附近的像素 $(i+a,j+b)$ 进行加权得到 $[\mathbf{H}]_{i,j}$ 。其中 $|a| > \Delta$ 或 $|b| > \Delta$ 约束条件使得该式满足局部性，即只关注于在位置像素 $(i+a,j+b)$ 的小领域范围内的参数，大大减少了参数量。 \mathbf{V} 被称为卷积核 (convolution kernel) 或者滤波器 (filter)，也是该卷积层的权重，通常该权重是可学习的参数。参数 a, b 也对应着卷积核的尺寸 k_h, k_w 。

上式 (2.1) 是单通道情况下卷积层的数学表示，当输入图像的通道数为 c_i 时，那么我们需要构造一个形状为 $c_i \times k_h \times k_w$ 的卷积核。由于输入和卷积核都有 c_i 个通道，我们可以对每个通道输入的二维张量和卷积核的二维张量进行互相关运算，再对通道求和得到一个二维张量。这就是一个输出通道的结果。如果我们需要输出通道数为 c_o 时，只需创建一个卷积核的形状是 $c_o \times c_i \times k_h \times k_w$ 。通道数量可以视作对于不同特征的描述，随着神经网络层数的加深，通常的做法是减少空间分辨率的同时增加通道数量。

2.1.2 填充和步幅

在应用多层卷积时，我们常常丢失边缘像素。填充 (padding) 可以解决这个问题。在输入图像的边界填充一定数量的元素（通常填充元素是 0）。通常，如果我们添加 p_h 行填充（大约一半在顶部，一半在底部）和 p_w 列填充（左侧大约一半，右侧一半），则输出形状将为

$$(n_h - k_h + p_h + 1), (n_w - k_w + p_w + 1) \quad (2.2)$$

这意味着输出的高度和宽度将分别增加 p_h 和 p_w 。在许多情况下，我们可以设置 $p_h = k_h - 1$ 和 $p_w = k_w - 1$ ，这样使得输入和输出具有相同的高度和宽度。假设 k_h 是奇数，我们将在高度的两侧填充 $p_h/2$ 行。如果 k_h 是偶数，通常会在输入顶部填充 $\lceil p_h/2 \rceil$ 行，在底部填充 $\lfloor p_h/2 \rfloor$ 行。同理，我们填充宽度的两侧。

感受野是指卷积网络中某一层输出特征图上的一个元素所对应的输入图像上的区域大小。它表征着特征图能“看到”的区域的大小。我们可以通过连续的卷积来增加感受野，但这会增加参数量。我们还可以通过调整步幅来增大感受野。步幅是卷积操作中卷积核移动的步长。在对图像进行卷积时，卷积核从图像的一个角落开始，按照指定的步幅在图像上滑动，每次移动指定的像素数，直到覆盖整个图像。当步幅大于 1 时，卷积核每次移动多个像素，输出的特征图的尺寸也会随之减小。具体公式如下：

通常，当垂直步幅为 s_h ，水平步幅为 s_w 时，输出形状为

$$\lfloor (n_h - k_h + p_h + s_h)/s_h \rfloor \times \lfloor (n_w - k_w + p_w + s_w)/s_w \rfloor \quad (2.3)$$

如果我们设置了 $p_h = k_h - 1$ 和 $p_w = k_w - 1$ ，则输出形状将简化为 $\lfloor (n_h + s_h - 1)/s_h \rfloor \times \lfloor (n_w + s_w - 1)/s_w \rfloor$ 。更进一步，如果输入的高度和宽度可以被垂直和水平步幅整除，则输出形状将为 $(n_h/s_h) \times (n_w/s_w)$ 。

2.1.3 激活函数

卷积神经网络中常用的激活函数包括 ReLU (线性整流单元)^[32]、Sigmoid^[33]、Tanh (双曲正切)^[34] 等。这些激活函数的目的是在网络中引入非线性特性，使得网络能够学习到更加复杂的数据表示。本文用到的是线性整流函数 ReLU (Rectified Linear Unit) 函数和 Sigmoid 函数。对于给定元素 x ，ReLU 函数被定义为该元素与 0 的最大值。它是目前最常用的激活函数之一。因为它的导数在大于 0 时为 1，小于 0 时为 0，这使得它可以用来缓解梯度消失的问题。

$$f(x) = \max(0, x) \quad (2.4)$$

Sigmoid 函数将输入值映射到 $(0, 1)$ 区间，这常常用于分类预测或者给出概率预测。然而，由于其在输入值绝对值较大时梯度接近 0，可能会导致梯度消失问题。

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.5)$$

2.1.4 池化层

池化 (pooling)^[35] 是卷积神经网络中常见的一种方法，主要用于减少特征图的维度，减少计算量的同时保留重要的一致性信息。与卷积层类似，池化运算也

是通过一个固定形状的窗口滑动来实现的。与之不同的是，池化通过对邻近像素进行统计学操作（如取最大值或平均值）来实现，因此也不包含参数。主要有两种类型的池化：平均池化（Average Pooling）^[35] 和最大池化（Max Pooling）^[36]。池化操作通常有两个参数：池化核的大小 ($K \times K$) 和步幅 (S)。池化核指定了池化操作的邻域范围，步幅定义了池化操作的移动间隔。对于输入大小为 $W \times H$ 的特征图，池化操作后的输出大小 $W' \times H'$ 可以通过以下公式计算：

$$W' = \left\lfloor \frac{W - K}{S} + 1 \right\rfloor \quad (2.6)$$

$$H' = \left\lfloor \frac{H - K}{S} + 1 \right\rfloor \quad (2.7)$$

在卷积网络的实践中，池化层通常有降低特征维度、引入不变性、增加鲁棒性和防止过拟合的作用。

2.1.5 权重衰减

在模型训练时，可能会遇到过拟合的问题，使得模型在已有数据上有着较好的性能，而在测试数据上表现不佳。我们可以使用多种正则化技术来缓解过拟合的问题。权重衰减（weight decay）是最广泛使用的正则化的技术之一，它通常也被称为 L_2 正则化。 L_2 正则化在损失函数中添加模型权重的平方之和作为惩罚项。同时通过一个非负的超参数 λ 来控制正则化的强度。 L_2 正则化正则化修正后的损失函数如下式：

$$L(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (\mathbf{w}^\top \mathbf{x}^{(i)} + b - y^{(i)})^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \quad (2.8)$$

L_2 正则化的目的是鼓励模型学习到更小更分散的权重值，从而提高模型的泛化能力。它对大的权重值施加较大的惩罚，从而防止模型依赖于少数几个可能具有高噪声的特征。

2.1.6 暂退法

Dropout（暂退法）^[37] 在训练过程中以一定几率随机“丢弃”（即暂时移除）网络中的一部分神经元（包括其连接），这有助于模型学习到更加鲁棒的特征，减少神经元间复杂的共适应关系。需要注意的是，在测试时，我们通常不使用 dropout。

2.1.7 批量归一化

批量归一化（Batch Normalization）^[38] 是通过对每个小批量数据进行归一化处理，调整神经网络中间层的输出，使其均值接近 0，标准差接近 1。这可以通过减去它们的均值除以它们的标准差得到。这有助于稳定和加速深度网络的训练过程，

同时也具有一定的正则化效果。批量归一化 (Batch Normalization, 简称 BN) 是一种在深度神经网络中广泛使用的技术，用于加速训练过程并提高模型的稳定性。其基本思想是在网络的每层之后添加一个归一化步骤，这个步骤会对每个小批量数据 (mini-batch) 进行归一化处理，以确保网络中间层的激活分布保持稳定。批量归一化的公式如下：

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \quad (2.9)$$

其中， ϵ 是一个很小的数，用来防止除以零。归一化后的 \hat{x}_i 具有零均值和单位方差。

2.2 目标计数

目前目标计数领域主要有三类方法。一类是检测方法^[6-8, 10, 11, 24]，通过目标检测模型识别出具体的物体位置，之后根据结果来进一步计数。但这类方法对于输入图像的分辨率有着较高的要求，往往需要物体具有明确清晰的边缘特征。在低分辨率下往往表现效果较差。一种是基于回归的方法^[16, 39, 40]，直接拟合出图像特征和目标数目之间的回归模型得到图像中对应物体的数目。但这种方法未能完整利用图像中的空间位置信息和时间序列信息。当输入图像的大小和分布有变化的情况下，往往不具有很强的泛化能力。另一类方法是基于密度图的目标计数方法^[26, 27, 29, 41]。此类方法通常先得出一个目标物体在区域内的一个分布，之后就可以通过密度分布来估计总体的数量。该方法在稠密计数的场景下往往具有较好的效果。在本文使用的跨分辨率车辆计数数据集上，可以把车辆计数视为一个稠密计数场景。使用基于密度图的计数方法相较其余两类方法有着更好的表现。受上述方法启发，本文将跨分辨率车辆计数问题转换为两个子问题，即综合跨分辨率图像信息的图像分割网络和映射分割结果和最终计数目标的回归模型。

2.3 U-Net 网络

U-Net^[13] 是一个广泛被应用的语义分割模型，最初被应用于医学图像的分割问题上。U-Net 是一个具有对称结构的网络，通过使用跳跃连接 (Skip Connection) 来结合低层次的位置信息和高层次的语义信息，从而在细节上进行更准确的预测。

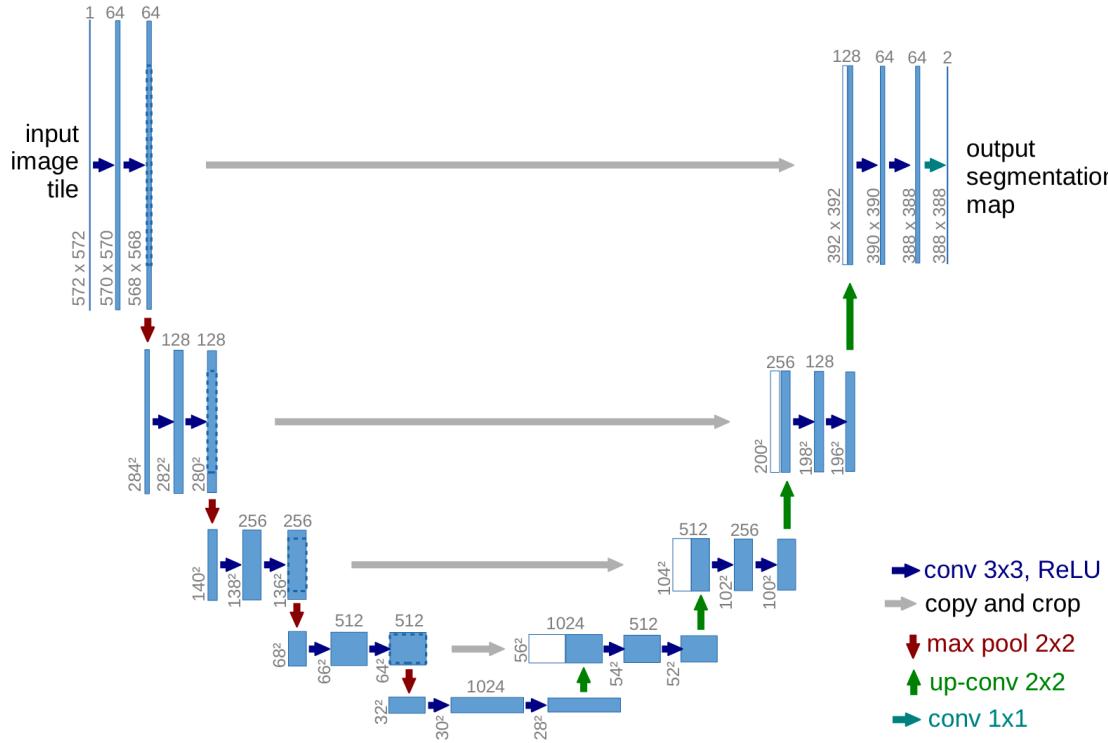


图 2.1 U-Net 网络结构

2.3.1 模型结构

U-Net 网络由一个收缩路径 (contracting path) 和一个扩展路径 (expansive path) 组成。网络的输入是一张 572×572 的图片 (input image tile)。网络最终的输出为同样尺寸的分割图预测。图像经过收缩路径提取综合特征，并保留中间特征信息。在扩展路径中，综合上采样后前一层特征结果与对应尺度的编码特征，得到最终的结果。因为整个网络结构形似字母‘U’，因此称为 U-Net。

2.3.2 收缩路径

收缩路径是由多个卷积层、线性整流函数单元 (ReLU) 和最大汇聚层 (Max Pooling) 构成的一系列降采样操作。论文中将这一部分叫做压缩路径 (contracting path)。压缩路径由 4 个块组成，每个块使用了 3 个有效卷积和 1 个 Max Pooling 进行下采样。每个块处理之后特征图的通道数扩大为 2 倍，特征图的长和宽也有相应缩小。这样的处理使得不同的特征被逐步提取到不同的通道中。最终得到了尺寸为 32×32 的特征图。

2.3.3 扩展路径

扩展路径是相同数量的相似模块组成。不同的是扩展路径中使用了反向卷积和上采样。同时扩展路径通过跳跃连接从收缩路径对应的层中获取特征图，并与

当前层的特征图进行融合。这种结构有助于恢复图像的精细信息，使得在深度网络中消失的某些信息不至被遗忘。在深度学习和计算机视觉中，上采样 (Upsampling) 和反向卷积（也称为转置卷积，Transposed Convolution）是两种常用的技术，用于增加图像或特征图的分辨率。这两种技术常见于像 U-Net 这样的网络结构中，用于从深层特征映射中恢复图像的细节信息，尤其在图像分割和生成模型中十分重要。

2.3.4 双线性插值

上采样是一种用于增加图像或特征图的尺寸的方法。它通过已有数据的插值来增加分辨率，主要有最近邻插值、双线性插值和双三次插值等方法。下面主要介绍双线性插值^[42]。在双线性插值中，输出像素的值是输入像素值的加权平均，权重基于像素之间的距离。假如我们想得到未知函数 f 在点 $P = (x, y)$ 的值，假设我们已知函数 f 在 $Q_{11} = (x_1, y_1)$, $Q_{12} = (x_1, y_2)$, $Q_{21} = (x_2, y_1)$, 及 $Q_{22} = (x_2, y_2)$ 四个点的值。

首先在 x 方向进行线性插值，得到：

$$f(x, y_1) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}), \quad (2.10)$$

$$f(x, y_2) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}). \quad (2.11)$$

然后在 y 方向进行线性插值，得到

$$\begin{aligned} f(x, y) &\approx \frac{y_2 - y}{y_2 - y_1} f(x, y_1) + \frac{y - y_1}{y_2 - y_1} f(x, y_2) \\ &= \frac{y_2 - y}{y_2 - y_1} \left(\frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}) \right) \\ &\quad + \frac{y - y_1}{y_2 - y_1} \left(\frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}) \right) \\ &= \frac{1}{(x_2 - x_1)(y_2 - y_1)} (f(Q_{11})(x_2 - x)(y_2 - y) + f(Q_{21})(x - x_1)(y_2 - y) \\ &\quad + f(Q_{12})(x_2 - x)(y - y_1) + f(Q_{22})(x - x_1)(y - y_1)) \\ &= \frac{1}{(x_2 - x_1)(y_2 - y_1)} \begin{bmatrix} x_2 - x & x - x_1 \end{bmatrix} \begin{bmatrix} f(Q_{11}) & f(Q_{12}) \\ f(Q_{21}) & f(Q_{22}) \end{bmatrix} \begin{bmatrix} y_2 - y \\ y - y_1 \end{bmatrix}. \end{aligned} \quad (2.12)$$

如果先在 y 方向插值、再在 x 方向插值，其结果与按照上述顺序双线性插值的结果是一样的。由上式我们不难看出，双线性插值由两个线性函数的积构成，因此为网络带来了非线性。

2.3.5 转置卷积

转置卷积^[43]是一种更复杂的上采样技术，它通过神经网络来试图学习一种更有效的插值方式。它不仅增加了特征图的尺寸，还可以学习在上采样过程中引入新的信息。它通过反转卷积操作的流程实现，因此被称为转置卷积。标准卷积操作是将卷积核应用于多个输入上，得到一个输出，实际上就是建立了一个多对一的关系。对于转置卷积而言，我们实际上是想建立一个逆向操作，也就是建立一个一对多的关系。对于标准卷积，我们有：

$$Y = CX \quad (2.13)$$

转置卷积其实就是要对其进行逆操作，求出 X

$$X = C^T Y \quad (2.14)$$

假设输入特征图大小为 $W \times H$ ，卷积核大小为 $K \times K$ ，步长为 S ，填充为 P ，输出特征图大小可以通过以下公式计算：

$$W' = S(W - 1) + K - 2P$$

$$H' = S(H - 1) + K - 2P$$

这里 W' 和 H' 分别是输出特征图的宽度和高度。

2.4 注意力机制

注意力机制 (Attention Mechanism)^[2] 是一种模仿认知注意力的机制。在认知科学中，由于信息处理的瓶颈，人类会选择性地关注信息中的某一一部分，同时忽略其他可见的信息。上述机制通常被称为注意力机制。随着该机制在 Transformer^[2]、BERT^[44]、GPT^[45] 等 NLP 领域的成功，该机制及应用又成为了研究的热点话题。目前在计算机视觉领域，ViT^[46]、Flow1D^[35] 等网络也都基于注意力机制进行设计。从注意力的形式来分类的话，可以分为软注意力 (soft attention) 和硬注意力 (hard attention)。其中软注意力机制是可微可导的，本文中主要探讨的也是软注意力机制。

2.4.1 注意力机制基本原理

如图 2.2 所示，注意力机制主要涉及到 3 类数据，分别是键 (key)、值 (value) 和查询 (query)。当一个查询值到来时，计算查询和键的相似度，得到权重，并进

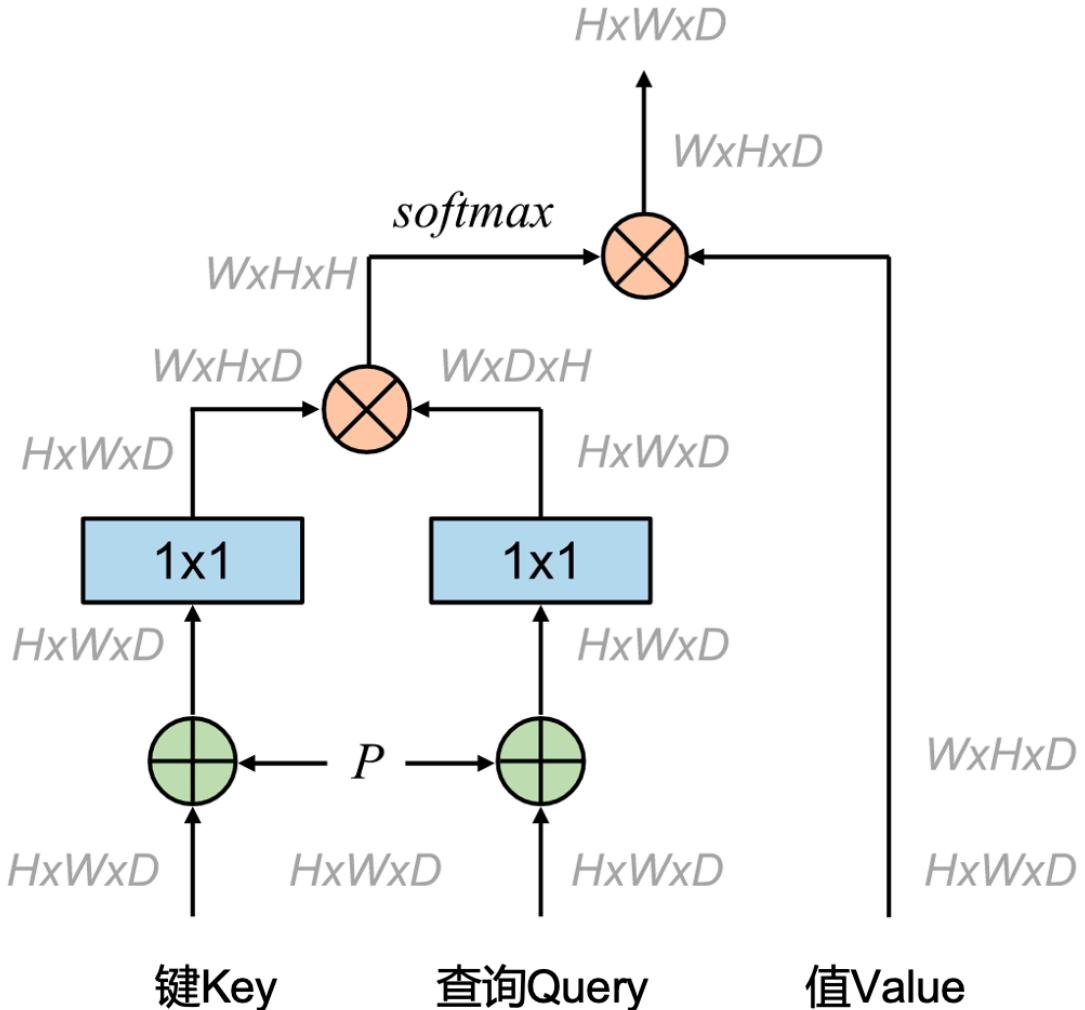


图 2.2 注意力机制

行归一化处理。再将得到的权重和值加权求和得到我们最终的注意力结果。首先计算查询与每个键之间的相似度。这一步通常使用点积 (dot product) 或者缩放点积 (scaled dot product) 来实现。具体来说，对于每个查询，通过计算它与所有键的点积，得到一个相似度分数：

$$\text{score}(Q, K) = QK^T \quad (2.15)$$

接下来使用 Softmax 函数对上一步得到的相似度分数进行归一化，以确保所有的权重加起来等于 1。：

$$\text{Attention Weight} = \text{Softmax}(\text{score}(Q, K)) \quad (2.16)$$

用归一化后的权重对值进行加权求和，得到最终的注意力输出：

$$\text{Output} = \text{Attention Weight} \cdot V \quad (2.17)$$

将上述步骤合并，注意力机制的输出可以通过以下公式计算：

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.18)$$

其中， d_k 是键的维度，这个因子用于缩放点积，避免在维度很高时计算结果过大，导致 Softmax 函数处于饱和区，从而缓解梯度消失的问题^[2]。

2.4.2 注意力机制在图像领域的应用

在图像处理领域中，使用注意力机制可以显著提升模型的性能，尤其是在图像分类、目标检测和图像分割等任务中。根据任务需要不同，常用的注意力机制有以下几种：

1. 空间注意力 (Spatial Attention)：关注图像的特定区域，通常用于增强模型对图像中重要部分的感知能力。可以用来替代传统的卷积网络，找到目标区域。
2. 通道注意力 (Channel Attention)：关注不同通道的相关性，可以帮助模型识别哪些特征是更加重要的。
3. 自注意力 (Self-Attention)：通过计算图像内所有位置之间的关系，可以捕捉更广泛的上下文信息。在时间序列模型中，自注意力机制可以保证长序列中的所有位置的信息有参与后续计算的可能。在图像领域中，对图像数据自身使用自注意力机制使得输出中每一位置均含有输入图像中所有位置的加权信息。

在图像领域实践中，同时还使用以下几种训练策略：

1. 多尺度注意力：使用多尺度注意力可以帮助模型同时关注图像的粗略和详细特征，这在处理具有不同尺寸和形状的对象时特别有效。
2. 融合不同的注意力机制：同时使用空间和通道注意力，或者将传统的注意力机制与自注意力结合起来，可以提取更丰富的特征并提高模型的性能。
3. 注意力正则化：添加注意力正则化可以防止模型对某些特征过度依赖，从而提高模型的泛化能力。使用如残差连接等设计可以训练更深层的网络，防止训练过程中的信息丢失。

2.4.3 注意力机制处理序列图像

在处理序列图像，如视频帧、时间序列的医学图像或连续的监控遥感数据时，我们不仅要考虑图像中的空间信息，也需要考虑图像间的序列信息。对于注意力机制的设计使用有着更高的要求。自注意力机制和多图像帧的相互注意力机制常常用来捕获时间和空间上的复杂关系。以下是这些注意力机制在序列图像处理中

的一些常见应用方式：

自注意力机制可以用于分析序列图像中的时间依赖性，这对于识别视频中的动态事件或时间序列图像中的变化特别有效。

1. 时间自注意力：在处理视频或其他序列图像时，可以在时间维度上应用自注意力，以识别不同时间点图像帧之间的关键依赖关系。在视频帧序列中，模型可以学习到哪些帧之间具有高度相关性，这对于动作识别、事件检测等任务非常有用。
2. 空间自注意力：在单个图像帧内部，可以应用空间自注意力来分析图像中不同区域之间的相互作用，对于解决目标检测和图像分割等任务有很大的帮助。
3. 时空自注意力：结合时间和空间自注意力，可以同时考虑空间位置和时间演变的关系。这种方法可以用于复杂场景的动态解析，如多物体交互的场景。

这些方法在处理动态场景解析和增强特征表示上有着不错的表现。在动态变化的场景中，模型可以使用多目标间的相互注意力来预测未来的状态。而通过计算不同目标之间的相互关系，可以获得更丰富的场景表示，这对于场景分类、事件检测等任务非常有帮助。

2.5 Attention U-Net 网络

Attention U-Net^[25] 是一种结合了注意力机制的 UNet 网络，最初被应用于医学图像的分割问题上。它在 U-Net 的架构上增加了 Attention Gate 注意力门使得模型能更好的聚焦在目标区域。

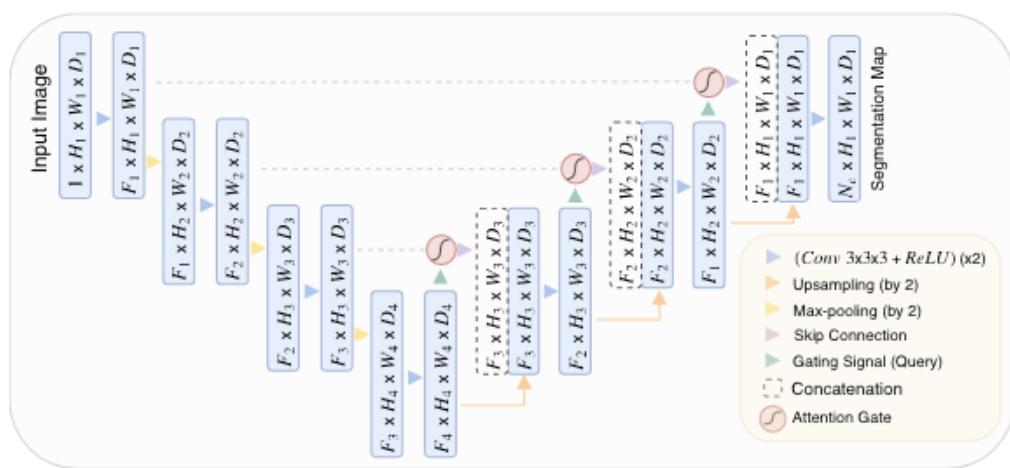


图 2.3 Attention U-Net

如图 2.3 所示，Attention U-Net 沿用了 U-Net 的基本架构，包括编码器（逐步下采样）和解码器（逐步上采样）两部分，以及跳跃连接（skip connections）来保留多尺度的特征。值得注意的是在每个跳跃连接处，新引入了注意力门控模块。这些模块对来自编码器的特征图进和解码器的相应特征图进行注意力计算。这使得网络能够聚焦于那些对最终分割任务更为重要的区域。

该方法将来自解码器的特征图作为查询，将来自编码器的特征图作为值和键作为注意力门的输入。注意力系数是通过一个小型的卷积网络学习到的，该网络计算当前解码器特征和对应编码器特征之间的相关性。

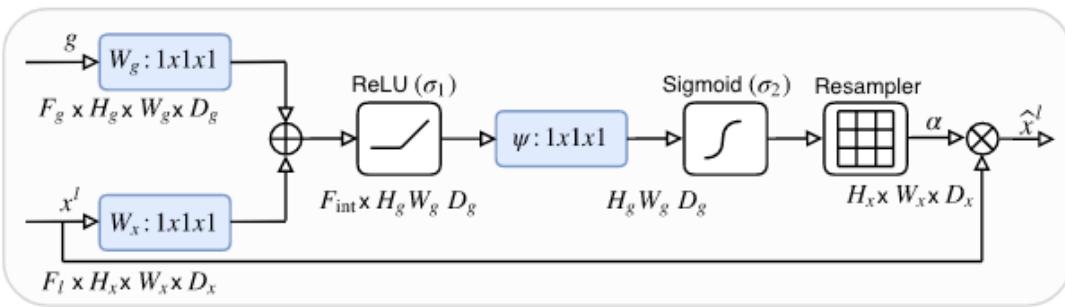


图 2.4 Attention gate

在图 2.4 中展示的是一个注意力门结构。注意力门接收两组输入，一组是来自上一下采样层的特征图 (g)，作为查询。另一组是来自跳跃连接的特征图 (x^l) 键和值。两组特征图首先通过一个 $1 \times 1 \times 1$ 的卷积层（表示为 W_g 和 W_x ），这一步用于减少通道的数量，以降低后续计算复杂度。接着，两组卷积后的特征图相加，并通过 ReLU 激活函数，得到 σ_1 。经过 ReLU 激活的特征图再次经过一个 $1 \times 1 \times 1$ 卷积层，通常标识为 ψ ，然后通过 Sigmoid 激活函数得到 σ_2 ，此时每个特征的激活值位于 $[0, 1]$ 区间，代表了特征的重要性权重。将 Sigmoid 输出的权重与跳跃连接的特征图 (x^l) 相乘。在这个过程中，三个 $1 \times 1 \times 1$ 卷积层包含了我们需要学习的参数，也赋予了该模块掌握关键权重的能力。通过注意力门，我们得到了在解码器特征图做查询的情况下加权编码器特征图。利用我们新得到的特征图来进行下一步解码，比原本单纯接受编码器输入获得了更丰富的信息。

2.6 Flow1D 网络

Flow1D^[35] 网络是一个基于注意力机制的光流估计网络。光流估计是计算机视觉中的一个基本问题，它旨在估计一幅图像上的每个像素点在时间序列中的运动，这在视频处理、运动分析、超分辨率、3D 重建和自动驾驶等众多领域中都有

广泛应用。光流估计是计算机视觉领域中的一个核心问题，光流是图像中像素点在时间维度上的瞬时运动速度和方向的场。光流是从连续的视频帧中估计出来的，这些连续的图像不仅具有时间上的连续性，光流也是从这些图像的空间关系中估计出来的。本文探讨的跨分辨率车辆计数问题，需要从同一时间的高分辨率低分辨率图像中找到空间一致性的关联，同时在连续的低分辨率图像中也要找到时间上的连续性关系。这和光流估计对于连续图像数据的利用有着不少相同之处。

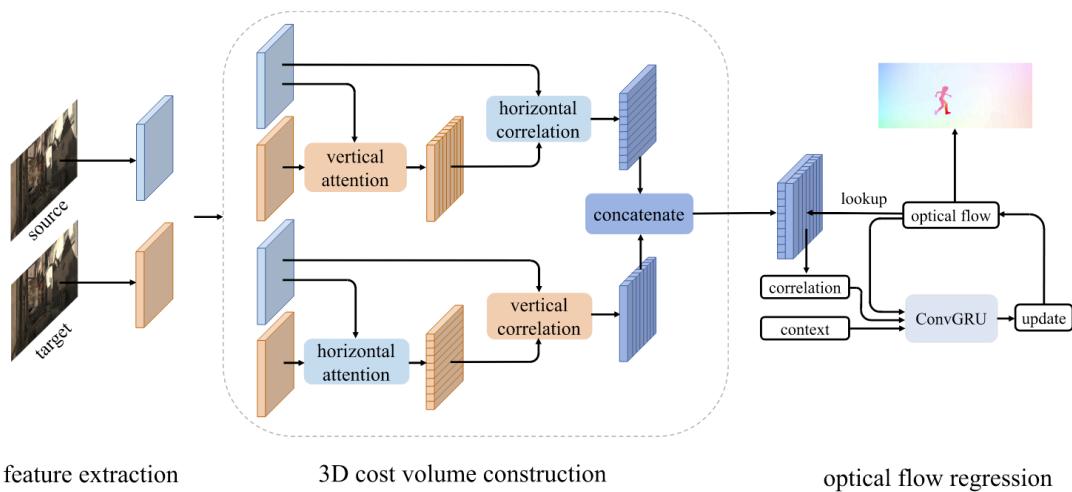


图 2.5 Flow1D 网络

图 2.5展示了模型的基本框架。对于源和目标两个图像，先分别进行特征提取，然后利用注意力机制计算 3D cost volume。最后通过门控循环单元，通过相关性特征和初始提取出的特征，进行隐状态的计算。反复迭代计算出光流。其中 3D cost volume 的设计充分利用了注意力机制的全局观察能力，通过两个一维的注意力操作，表征三维的光流状态。在水平竖直方向分别进行自注意力计算和相互之间的注意力计算。

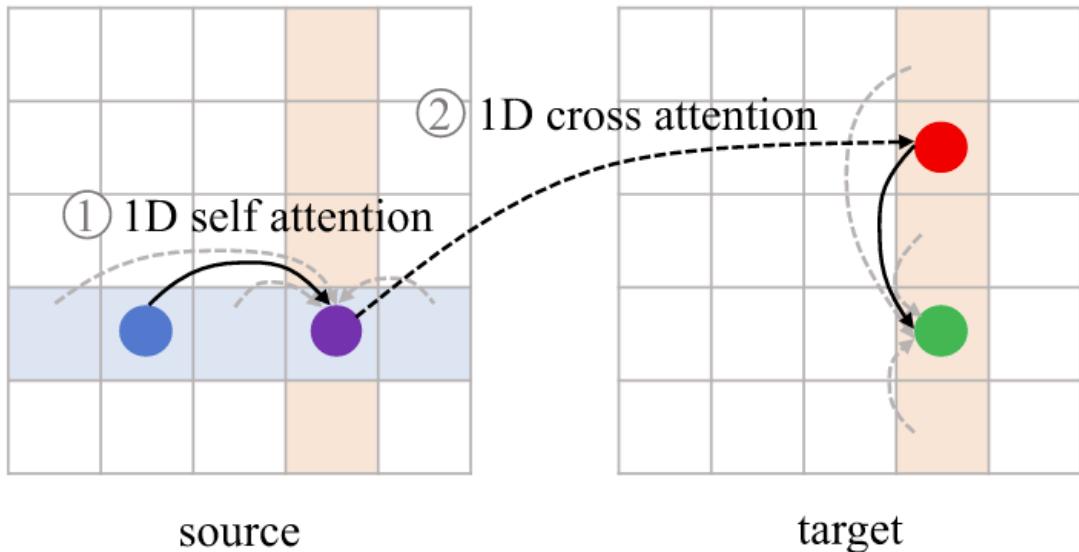


图 2.6 self attention 和 cross attention

图 2.6是对注意力机制的一个很直观的展示。如果要计算源和目标的相关度，直接进行 cross attention 是不能得到红点与蓝点之间的相关关系的。因此需要再源上先进行 self attention，使得每一个列向量包含着原先该列的一种加权分布。然后再进行 cross attention 操作，综合不同图像不同位置的信息。这是一种非常有效的策略。

2.7 CRVC 网络

CRVC 网络是针对 CRVC 数据集设计的深度学习模型。它以 U-Net 模型为骨架，针对跨分辨率空间 CRVC 数据集中的数据特性设计了两个分支来提取跨分辨率空间信息和时序信息。通过上述网络估计出密度分布后，使用线性回归模型得出最终目标计数结果。

2.7.1 网络设计

图 2.7展示了模型的基本框架。模型接受 4 个输入，分别是高分辨率图像输入 I^{HR} ，对应低分辨率图像输入 I^{LR} ，与 LR 图像时间间隔较近的 I_{near}^{LR} 和与 LR 图像时间间隔较远的 I_{far}^{LR} 。模型包含两个独立学习的编码器 HR encoder 和 LR encoder，前者用来提取高分辨率图像的特征，后者用来提取 3 个低分辨率图像的特征。提取出的低分辨率特征和高分辨率特征作差来综合更高精度的信息。之后通过带有跳跃连接的 decoder 完成分割图的生成。

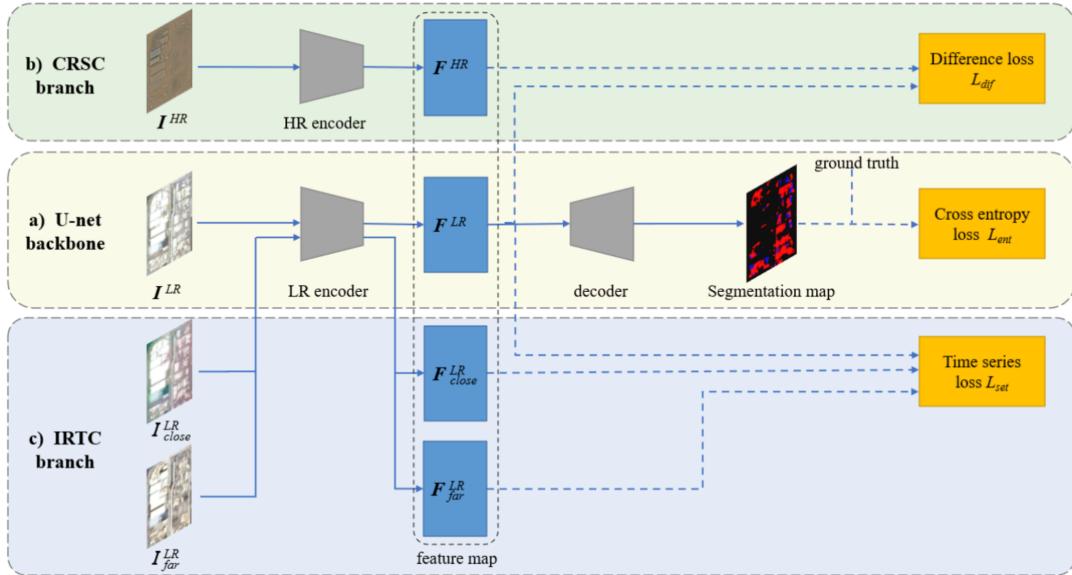


图 2.7 CRVC 网络架构

2.7.2 回归模型

在 CRVC 数据集中，车辆在目标区域是密集停放的，而且不存在重叠。因此从分割图上估计出的覆盖率和最终的车辆数目具有线性关系。通过 CRVC 网络得到的覆盖率分割图，根据低分辨率图像的缩放比例，可以估计出实际面积。通过线性回归，找出参数 k 来拟合

$$Number_i = k_i \dots Area_i + b_i \quad (2.19)$$

其中 i 表示第 i 类车辆。这些参数通过高分辨率图像的真值计算得到。

2.7.3 模型局限

CRVC 网络针对 CRVC 数据集的特点进行了设计，在各项评价指标上均优于传统目标计数方法。然而模型对于样本数量较少的类别的计数效果不是很理想。CRVC 网络通过引入两个额外分支来分别提取高分辨率的空间信息和低分辨率的时序信息，然而，对于数据集中的时间连续性和空间一致性信息的挖掘，网络的表现仍有进一步优化的空间。本文针对上述问题进行了针对性设计。

3 基于注意力机制的跨分辨率遥感影像计数

在本节中，将详细介绍一种全新的基于注意力机制的跨分辨率遥感影像计数方法。通过分析 CRVC 数据集中数据的时间连续性和空间一致性特性，在现有网络基础上，设计了跨分辨率空间注意力和低分辨率下的时空注意力模块，充分结合了图像中的不同尺度，位置信息。

3.1 问题分析

由于直接从极低分辨率的图像中识别出车辆目标具有相当大的难度，本文通过将车辆计数问题转化为图像分割和回归问题来解决这一挑战。这种转换使得处理变得更为可行，是因为不论图像分辨率如何变化，同一类别的车辆数量与车辆实际占用面积之间的线性关系是恒定的。通过数据集中高分辨率图像的计数结果作为同一天同一地点的低分辨率图像的真值是有效的，同时解决了无法直接在低分辨率图像上人工计数得到真值的问题。车辆面积的计算需要基于车辆覆盖率（在图像中的车辆区域百分比）。为了计算车辆覆盖率，我们首先需要通过图像分割技术准确地从低分辨率图像中提取出车辆区域。因此，本文提出的车辆计数大致流程如下：

1. 跨分辨率图像处理：选择性地输入高分辨率图像以及与该高分辨率图像同一天的低分辨率图像，同时还包括距离该日期较近和较远的两张低分辨率图像。这些图像被输入到精心设计的分割网络中，该网络专门针对从极低分辨率图像中有效分割车辆区域进行了优化。
2. 车辆覆盖率的计算：通过应用先进的图像处理算法，网络将输出车辆的分割图。这些分割图将用于计算每个车辆的覆盖率，即每辆车在图像中所占的面积与整个图像面积的比例。
3. 车辆面积的转换与回归分析：将得到的车辆覆盖率转换为实际车辆面积，这一步是通过与车辆实际占用面积和数量之间的已知线性关系相结合完成的。我们将使用从高分辨率图像中得到的数据来计算回归模型的系数，这些系数反映了车辆面积与车辆数量之间的关系。然后，这些回归模型系数被应用到低分辨率图像上，以估计出车辆的数量。

通过上述流程处理，即使这些低分辨率图像在视觉上难以分辨车辆细节，我们也能够参照高分辨率图像在极低分辨率图像中有效地进行车辆计数。此方法不仅提高了车辆计数的准确性，还为处理其他低分辨率图像分析任务提供了可能的方法

论指导。

3.2 网络设计

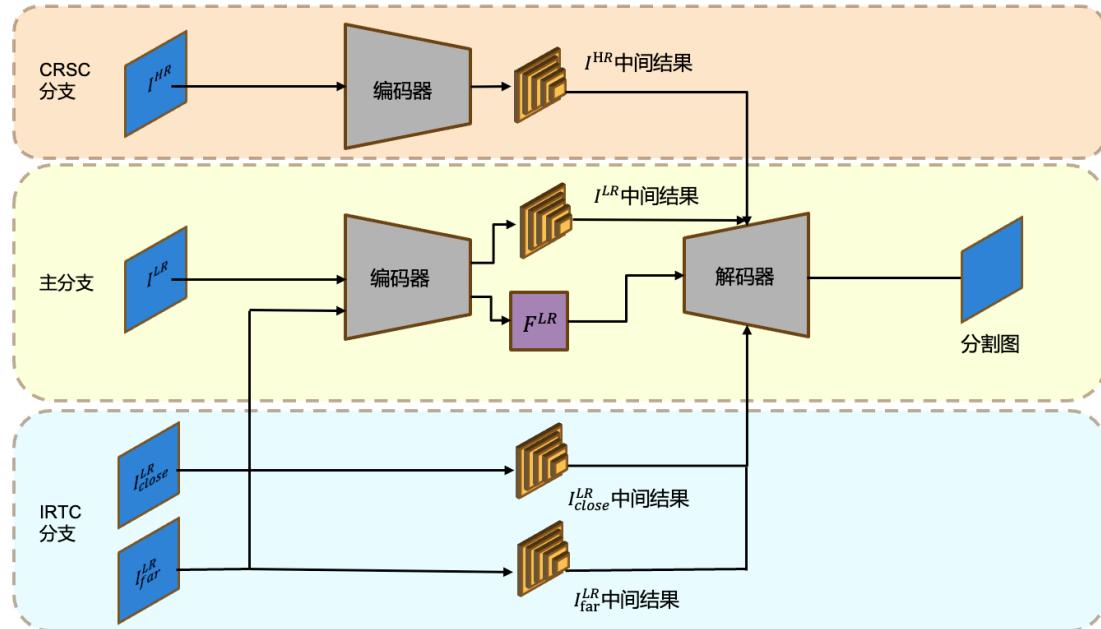


图 3.1 网络结构

网络设计如图3.1所示，网络基于CRVC-Net骨干网络搭建，采用U-Net编码器解码器结构。对于目标低分辨率图像，通过多层卷积网络组成的编码器提取出特征，用作后续解码器输入的一部分。在该编码器结构中，同时保留各层的中间结果，以便后续解码器中对应层使用。

为了利用来自高分辨率图像的先验空间信息和来自其他低分辨率图像的时间连续性约束，引入了两个监督分支跨分辨率空间一致性分支CRSC(crossresolution spatial consistency)和同分辨率下时间一致性分支IRTC(intraresolution time continuity)。CRSC分支采用和主干编码器相同的架构，提取来自相应高分辨率图像的特征，以使同一天高分辨率图像和低分辨率图像的提取特征尽可能相似。IRTC分支共享主分支的模型及参数，提取来自距离该日期较近和较远的两张低分辨率图像的特征，后续将比较他们和主分支输出的差异。这是因为编码器应具有提取同一分辨率下所有图像的能力，而对于不同分辨率图像来说，像素密度不同导致图像细节不同，采取独立训练的编码器能更好表征不同细节特征，避免造成高分辨率图像细节的丢失，从而影响其指导能力。解码器接受主分支特征作为输入，同时在每一层解码器处，使用对应层的主分支及近处和远处低分辨率图像和高分辨率图像的中间结果作为key，设计了专门的注意力门，综合多渠道输入，进行解码。

逐层操作直至生成分割图像。

3.3 编码器

本文设计的网络中，主要通过编码器进行特征提取。网络共接受同样尺寸的4张图像，第一张图像是低分辨率图像 I^{LR} ，第二图像为相应日期的高分辨率图像 I^{HR} ，第三张和第四张低分辨率图像 I_{close}^{LR} 和 I_{far}^{LR} ，选取与 I^{LR} 日期最接近的图像和与其日期间隔较远的图像。其中 I^{LR} 、 I_{close}^{LR} 和 I_{far}^{LR} 三个低分辨率图像使用同一个编码器，参数共享； I^{HR} 使用结构相同但不共享参数的另一个编码器。这两个编码器用于分别学习同样尺寸的高分辨率和低分辨率图像输入的特征。编码器结构如图3.2所示。

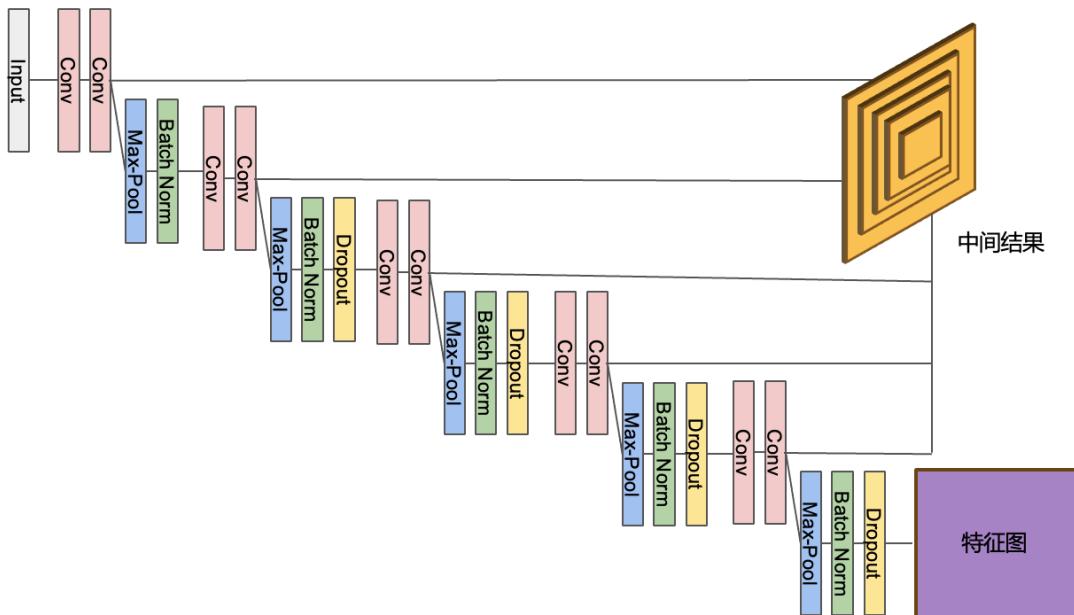


图 3.2 编码器

编码器被设计为通过逐步减小图像的空间维度（高度和宽度）的同时，增加其特征通道数量，来提取图像的特征，共含有5层。每层结构相同，且保留中间结果用作后续跳跃连接输入。

在每层编码器模块中，首先先通过两个卷积层，每个卷积层都使用 3×3 的卷积核，激活函数是 ReLU，并使用填充，确保输出和输入有相同的空间尺寸，这样可以在不改变空间尺寸的情况下，连续两次增强特征的提取。接着通过一个池化层进行最大池化减半特征图的尺寸，实现下采样。除第一层外，后续连接 dropout 层以防止过拟合。最后进行批量归一化操作。特征通道的数量在每个编码器模块

之后都会增加。在最后一层编码器输出后不进行池化，而是再次通过一个卷积块，为解码器提供特征。

3.4 多来源注意力机制

注意力机制有助于掌握时序信息，同时对于综合不同特征间相关关系有着很强的综合能力。本文中就针对 CRVC 数据集的特点，设计了自注意力门 SAG (Self Attention Gate)、跨分辨率注意力门 CAG (Crossresolution Attention Gate) 和时间序列注意力门 TAG (Timeseires Attention Gate) 三种注意力门模块，来充分利用数据集中的时间一致性和空间连续性特征。网络中使用的注意力门结构如图3.3所示

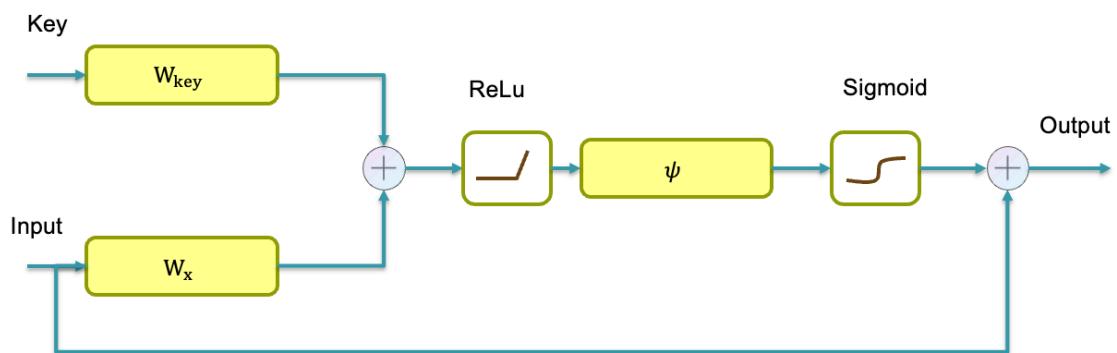


图 3.3 注意力门结构

其中 W_{key}, W_x 和 ψ 是三个线性变化的参数矩阵，通过这三个矩阵，注意力的权重就可以通过反向传播进行学习。该门接受两个输入，一个是 key，另一个作为 query 和 value。输出的结果是根据 key 加权得到的新的 value，满足下述公式：

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3.1)$$

自注意力门，用主分支编码器对应中间层的结果作为 key，解码器上一层上采样结果作为 query 和 value 进行计算。跨分辨率注意力门，用 CRSC 分支编码器对应中间层的结果作为 key，解码器上一层上采样结果作为 query 和 value 进行计算。时间序列注意力门，用两个 IRTC 编码器对应中间层的结果分别作为 key，解码器上一层上采样结果作为 query 和 value 进行计算。这样最终得到四个加权后的解码估计，将这些值连接起来作为整体门模块的输出供后续模块使用。

本文设计的网络将主分支、CRSC 分支和 IRTC 分支编码器的对应层的中间结果分别用作注意力机制中的 key，而对应的 query 和 value 则均来自上一层解码器的输出。这种设计允许模型在每一步解码过程中都重新评估各个编码器分支的重

要性，更加动态地融合信息。此方法允许模型根据解码器的当前状态动态调整各个分支的权重，在反向传播中加以修改优化。同时通过考虑来自不同时间和空间信息的多个分支，模型能够更全面的利用数据集中的全部信息。

3.5 解码器

本文设计的网络中，解码器是编码器的对称部分，负责将压缩后的特征图逐步上采样回原始图像的尺寸，并通过注意力门跳跃连接恢复细节信息以生成精确的分割结果。与编码器相同，解码器也包含五个解码器模块。解码器部分的具体细节如图3.4所示。

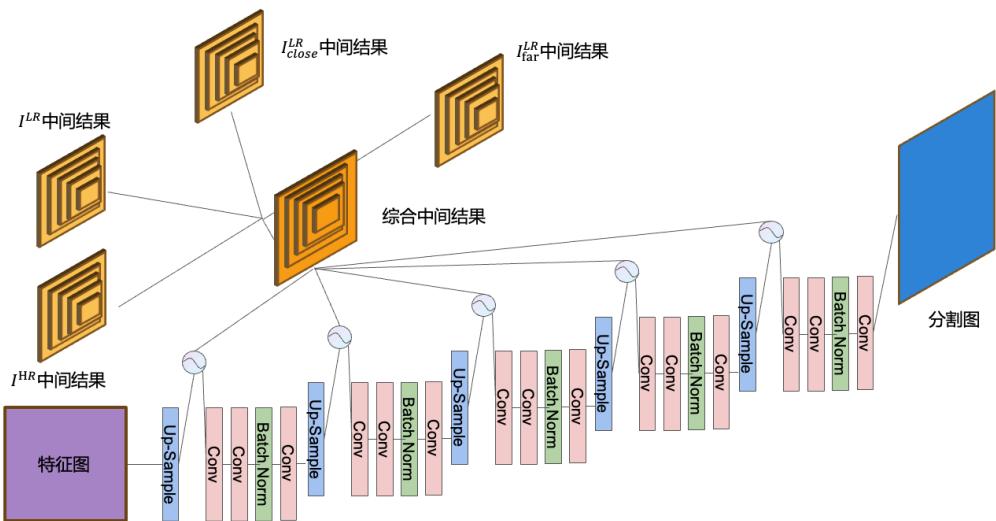


图 3.4 解码器

与编码器相同，解码器也包含五个解码器模块。从最深层的特征开始，逐步应用解码器模块，每一步都将特征图的尺寸增大，直到恢复到原始图像的尺寸。将特征图尺寸扩大的关键是

每个解码器模块的输入是前一个块的输出和主分支、CRSC 分支和 IRTC 分支编码器层对应层的输出。首先将前一块输出进行上采样使其和当前层特征尺寸一致。可以使用卷积转置或双线性插值进行上采样。接着将上采样后的输入和经过注意力门的跳跃连接输出连接起来，使用两个 3×3 卷积层来进行特征提取。最后会通过一个批量归一化和一个卷积块来进一步处理合并后的特征。重复上述操作五次，最终的输出经过一个以 SIGMOD 函数为激活函数的卷积层后得到最终的输出结果。

最后的结果与初始输入图像的尺寸相同，但通道数为四，分别代表四个汽车

类别。

3.6 损失函数

由于 CRVC 数据集中不同类别间样本数量存在较为显著的差异，本文采用分段多重损失函数来平衡不同类别对模型的影响。

3.6.1 轿车计数阶段

本阶段采用三个损失函数。第一个损失函数为交叉熵损失函数，这个函数常用来衡量概率之间的距离。网络经 `sigmod` 输出的分割图也是一种概率分布。用交叉熵衡量其偏差程度有着不错的效果。

$$\mathcal{L}_{\text{ent}} = -\frac{1}{n} \sum_i y_i \ln a_i \quad (3.2)$$

对于高分辨率和低分辨率图像间的空间一致性，它们之间的差异越小，模型对于真实情况的把握就越好。使用如下的损失函数进行约束。

$$\mathcal{L}_{\text{dif}} = \sum_i |F_l^{LR} - F_l^{HR}|^2 \quad (3.3)$$

网络接受 3 个低分辨图像，分别是有高分辨率对应的低分辨率图像，距离这个低分辨率图像时间较近的图像和距离这个低分辨率图像时间较远的图像。由于时间连续性的约束，日期接近的图像间的差异应该小于日期相隔较远的差异。因此这部分损失函数的设计要求相邻日期图像间的差异更小，日期间隔较远的差异较大。下面的公式同时满足上述条件，且符合最小化损失的要求

$$\mathcal{L}_{\text{ser}} = \sum_{l=1}^m \frac{|F_{\text{closel}}^{LR} - F_l^{LR}|}{|F_{\text{far } l}^{LR} - F_l^{LR}|} \quad (3.4)$$

这三个损失函数的加权和作为模型训练第一阶段的损失函数。

3.6.2 其余类别计数阶段

为了解决类别不平衡问题，在第一阶段的训练后，采用焦点损失（Focal Loss）作为交叉熵函数的补充。

焦点损失函数（Focal Loss）^[47] 最初是为解决目标检测中的类别不平衡问题设计的，尤其是在目标识别中背景与前景类别之间的不平衡问题。这种损失函数的设计原理使其同样适用于广泛的多分类问题，特别是在存在明显类别不平衡的情况下。对于包含多种车辆类型的 CRVC 数据集，各类车辆数量极不平衡，其中包括轿车 35844 辆、小型货车 737 辆、大型货车 1211 辆、起重机 60 辆。如果数据

集中某些类型的车辆比其他类型少得多，使用焦点损失函数可以带来明显的优势。由于轿车数量显著多于其他几类车辆，采用交叉熵函数（公式 3.2）时，由于其中的轿车样本占大多数，将显著影响损失函数整体的值，而缩小其他类别对于损失函数的贡献。因此在最终得到的网络中，对于其他几类的分类效果也因此受到影响。

焦点损失函数是交叉熵损失函数的一种改进。焦点损失函数通过重新设计交叉熵损失，可以调节不同数量样本之间的权重，从而使模型对于少量样本的表示能力更强，从而提高模型整体性能。

焦点损失函数的定义如下：

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (3.5)$$

其中： p_t 是模型对当前样本的预测概率，对于正样本 $p_t = p$ ，对于负样本 $p_t = 1 - p$ 。 α_t 是平衡正负样本权重的系数，通常设置为一个小于 1 的值，用来增加少数类的重要性。 γ 是调整易分类样本对损失的影响的聚焦参数， $\gamma \geq 0$ 。当 $\gamma = 0$ 时，Focal Loss 退化为标准的交叉熵损失。 γ 的值越大，对易分类样本的惩罚就越大。

通过引入调制因子 $(1 - p_t)^\gamma$ ：当一个样本被错误分类，并且错误程度很大（即 p_t 很小）时， $(1 - p_t)^\gamma$ 接近 1，损失不受影响。当一个样本被正确分类，且分类器置信度很大时（即 p_t 很大）时， $(1 - p_t)^\gamma$ 接近 0，这使得这类样本对总损失的贡献大大降低，这样可以让模型集中精力学习那些难以分类的样本。

在多分类问题中，焦点损失函数的应用类似于其在二分类中的用法，但需要一些调整来处理多个类别。多分类版本的焦点损失函数通常表示为：

$$FL(p_t) = - \sum_{c=1}^C \alpha_c (1 - p_{t,c})^\gamma \log(p_{t,c}) \quad (3.6)$$

其中： C 是类别的总数。 $p_{t,c}$ 是模型对于每个类别 c 的预测概率。如果样本属于类别 c ，则 $p_{t,c}$ 是该类别的预测概率；否则为 $1 - p_{t,c}$ 。 α_c 是针对类别 c 的平衡系数，用于调节不同类别间的不平衡。 γ 是聚焦参数，用来减小易分类样本的损失贡献，增加难分类样本的影响。

通过调节 α_c 和 γ 参数，焦点损失函数可以帮助模型更好地学习那些样本数量较少的类别。这是通过增加这些类别样本的损失贡献来实现的，从而使模型在训练过程中更加关注它们。通过 $(1 - p_{t,c})^\gamma$ 这一调节项，焦点损失函数提高了那些模型难以正确分类的样本的损失权重，从而激励模型改进这些区域的预测性能。

对于一个批次中的所有样本，焦点损失函数完整的表达式通常写作：

$$\mathcal{L}_{\text{FL}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C \alpha_c (1 - y_{i,c})^\gamma \log(y_{i,c}) \quad (3.7)$$

其中： N 是批次中样本的总数。 C 是类别的总数。 i 是批次中的样本索引。 $y_{i,c}$ 是第 i 个样本的为类别 c 的预测概率， α_t 是一个调制因子，用于平衡正负样本之间的影响。 γ 是一个调整参数，用于减少易分类样本的权重。

最终的损失函数可以将两个阶段各个损失函数加权得到，即

$$\mathcal{L} = \omega_1 \mathcal{L}_{\text{ent}} + \omega_2 \mathcal{L}_{\text{dir}} + \omega_3 \mathcal{L}_{\text{ser}} + \omega_4 \mathcal{L}_{\text{FL}} \quad (3.8)$$

其中 $\omega_1, \omega_2, \omega_3, \omega_4$ 为各个损失函数的权重，是设定好的超参数。

4 跨分辨率遥感影像计数实验及分析

4.1 实验设计

本设计探究的是跨分辨率遥感影像的计数问题，在跨分辨率车辆计数数据集上，训练本文设计的网络，测试其性能。数据集中共包含了拍摄于同一位置不同时间的 192 张极低分辨率图像和 8 张高分辨率图像，低分辨率图像中包含 8 张和高分辨率图像再同一天拍摄的图像。如何通过少量的高分辨率图像的计数结果，得到可以在低分辨率图像上做出准确计数估计的模型就是本设计的目标。本设计基于 CRVC-Net 骨架网络，通过独特设计的 3 个注意力门，更好的综合了从高分辨率及其余两个时间的低分辨率图像的特征信息，逐层与解码器输出进行计算，改进输出效果。

4.2 模型训练

在训练阶段，四张图像输入到网络中。第一输入是 I^{LR} ，低分辨率图像用于在骨架中生成车辆分割。第二输入是相应的高分辨率图像 I^{HR} ，起到空间引导的作用。 I^{LR} 和 I^{HR} 应该在 8 对相应的低分辨率和高分辨率图像中选取。我们使用其中的 7 对进行训练，1 对进行测试。第三和第四输入是一对低分辨率图像，即 I_{close}^{LR} 和 I_{far}^{LR} ，以强调时间连续性。 I_{close}^{LR} 选为与 I^{LR} 日期最接近的图像，而 I_{far}^{LR} 在图像集中随机选取，与 I^{LR} 的时间差超过 50 天。由于低分辨率和高分辨率图像对数量较少，为了增强鲁棒性，将训练图像裁剪为 32×32 的补丁后，通过旋转 90 度、180 度和 270 度，上下左右翻转，将训练集扩大至 4000 个图像补丁。测试集扩增到 1000 个图像补丁。在解码器的上采样过程中，使用 U-Net 结构将上采样的图像与相应比例的特征图跳跃连接。网络输出与原始图像补丁大小相同的 32×32 的最终分割图。由于训练集是通过旋转和翻转方法生成的，最终的分割结果应该是来自同一图像的各种增强图像的平均值。然后将 32×32 的补丁拼接成所需的图像大小，以获取整个停车场区域的图像。在测试阶段，只需将 I^{LR} 输入到分割骨架中，输出即为车辆分割图。考虑到起重机的数量太少，对分割结果影响不大，我们排除了这一类别，训练和测试阶段只涵盖轿车、小型货车、大型货车和背景四类。

本方法先得到从低分辨率图像得到的分割图，然后据此估计出覆盖率，再由覆盖率通过回归模型估计出最总计数结果，因此主要从以下三个维度度量模型性能：

1. 测试模型在低分辨率 (LR) 图像上的分割结果，与相应的高分辨率 (HR) 图像的真实标注进行比较。
2. 在所有低分辨率图像上测试的车辆覆盖率结果，与人工标注进行比较。
3. 测试模型在低分辨率图像上的计数结果，与相应高分辨率图像的真实标注进行比较。

4.3 分割精度测试

以同一日期的高分辨率图像的分割图作为标注，在同一位置生成低分辨率图像的分割图如图 4.1 所示。

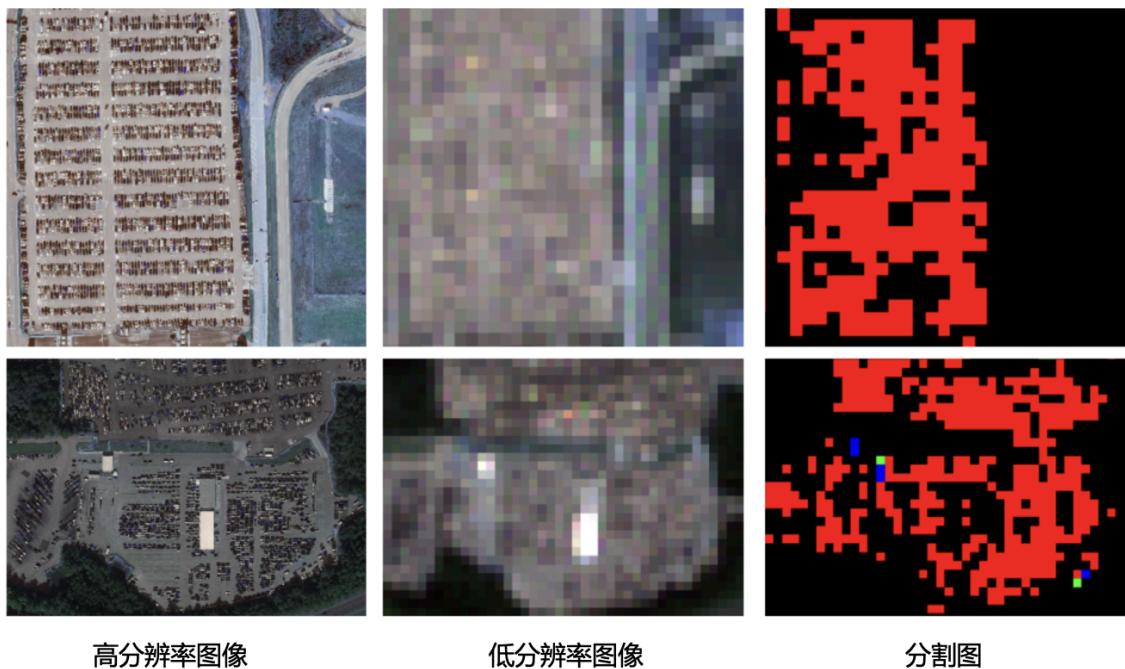


图 4.1 高低分辨率图像及对应分割图

本设计在各种网络模块的组合上测试了分割精度。采用像素精度 (Pixel Accuracy, PA) 来定量评估分割精度。它表示所有像素中正确分类的比例。

$$PA = \frac{\sum_i^k p_{ii}}{\sum_i^k \sum_j^k p_{ij}} \quad (4.1)$$

假设有 $k + 1$ 个类别 (包括 k 个目标类别和 1 个背景类别)， p_{ij} 表示属于类别 i 但被预测为类别 j 的像素数量。因此 p_{ii} 代表类别为 i 预测结果也为 i 的真阳性 (TP)， p_{ij} 代表类别为 i 预测结果也为 i 的假阳性 (FP)， p_{ji} 代表类别为 i 预测结果也为 i 的假阴性 (FN)。基于真阳性 (TP)、假阳性 (FP) 和假阴性 (FN)，为了更全面的度量该方法在分割方面的性能，通过使用了如召回率 (Recall)、精确度

(Precision) 和 F1 分数等指标来进一步指示估计精度。

$$\text{Recall} = TP / (TP + FN) \quad (4.2)$$

$$\text{Precision} = TP / (TP + FP) \quad (4.3)$$

$$F1 = 2TP / (2TP + FP + FN) \quad (4.4)$$

结果列在表 4.1 中。本文的方法在召回率、精确度和 F1 分数上均有一定提高。

表 4.1 召回率、精确度和 F1 分数

模型	真阳性	假阴性	假阳性	召回率	精确度	F1 分数
CRVC-Net	2027	2194	1037	48.0%	66.2%	0.556
本文模型	2134	2149	975	49.8%	68.6%	0.577

4.3.1 消融实验

为了验证并解释模块设计的合理性和因果性，针对 3 种注意力门设计了消融实验。

1. 仅使用 CRVC-Net 骨架进行训练。
2. 仅使用自注意力门的 CRVC-Net 进行训练。
3. 仅使用跨分辨率注意力门的 CRVC-Net 进行训练。
4. 仅使用时间序列注意力门的 CRVC-Net 进行训练。
5. 使用自注意力门和跨分辨率注意力门的 CRVC-Net 进行训练。
6. 使用自注意力门和时间序列注意力门的 CRVC-Net 进行训练。
7. 使用自注意力门、时间序列注意力门和跨分辨率注意力门的 CRVC-Net 进行训练。

实验的结果如下表所示：

4.3.2 实验结果分析

从上表 4.2 我们可以发现，单独增加自注意力门可以提升像素估计精度，但单独增加跨分辨率注意力门和时间序列注意力门却对估计精度有负面影响。这可能是因为单独引入高分辨率分支特征虽然在训练集上有助于提升最终估计效果，但对于低分辨率的图像的泛化性能却有所下降。单独引入时间序列注意力门虽然丰富了解码器的输入，但没有高分辨率参照下的其余低分辨率特征，并不能很好的指导模型做出正确的估计，反而影响了模型的性能。同时引入自注意力门和跨分辨率注意力门比较明显的提升了模型的性能，同时引入自注意力门和时间序列注

表 4.2 注意力门消融实验

模型	自注意力门	跨分辨率注意	时间序列注意	像素精度
	SAG	力门 CAG	力门 TAG	
CRVC-Net	×	×	×	87.5%
base+SAG	√	×	×	87.7%
base+CAG	×	√	×	86.3%
base+TAG	×	×	√	84.8%
base+SAG+CAG	√	√	×	88.2%
base+SAG+TAG	√	×	√	87.4%
base+SAG+CAG+TAG	√	√	√	88.4%

注意力门也较先前模型有了明显的提升。同时使用三种注意力门单元的模型取得了最好的像素精度估计，也证明了时间序列注意力门单元设计的有效性。虽然单独添加 TAG 门的效果不佳，但其作为完整模型的一部分有效的弥补了对于不同时间的低分辨率图像的识别能力的不足。

4.4 覆盖率估计测试

在上一步骤中，通过网络模型可以得到低分辨率图像的分割图像估计，据此可以得出车辆的覆盖率。但是由于只有少量低分辨率图像有对应的高分辨率图像作为标注，大部分低分辨率图像采用人工专家标注的方法来判断模型估计的覆盖率结果。低分辨率图像的覆盖率以 5% 的间隔手工标注。考虑到人工标注可能存在一些错误，我们根据两个标注者的标注结果的差对标注进行了一些预处理：差值不超过 5% 的标注被认为是准确的，而差值位于 10% 左右被认为是可接受的。如果差值大于 15%，则认为发生了来两者标注中可能存在错误，将该值排除并由其他专家进行重新标注。上述规则确保了标签的可信度。然后，使用所有人的平均标签值来评估车辆覆盖估计的准确性。下图 4.2 展示了模型估计覆盖率和人工标注覆盖率。

4.5 车辆计数回归结果测试

通过车辆覆盖率，可以通过低分辨率图像的大小来计算出车辆区域面积。通过高分辨率图像提供的真值计数信息，可以进行回归分析。本文中对四个车辆类别分别进行了车辆区域面积和车辆数目的回归分析，得到的结果如下图所示。

对于小汽车、小型货车、大型货车和起重机四个类别，回归模型预测的斜率分

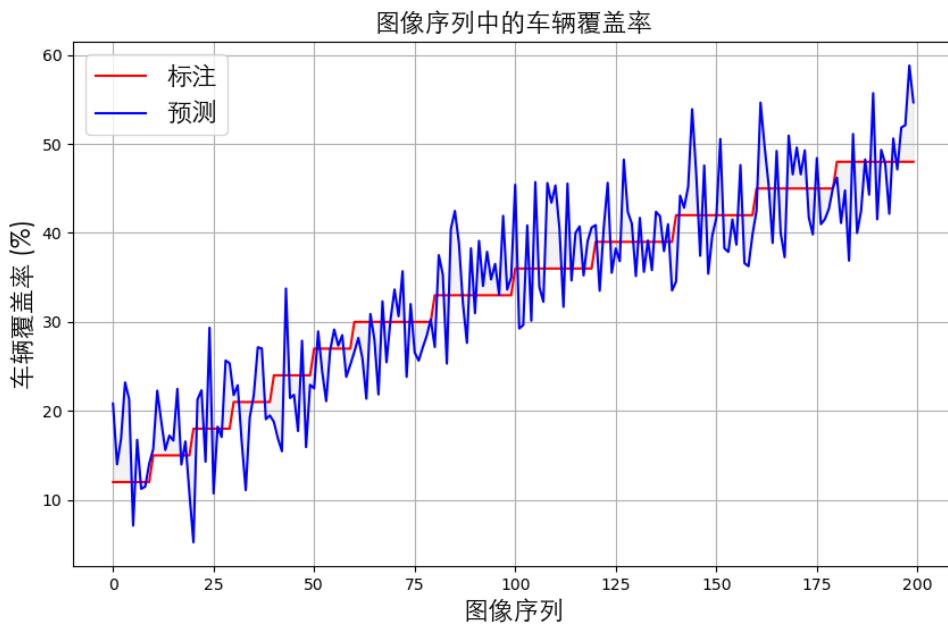


图 4.2 覆盖率估计

别为 10.353、18.354、42.468 和 27.323, R^2 分别为 0.9867、0.9821、0.9854 和 0.9376。 R^2 大于零且接近一，具有很强的正相关性。

使用回归模型，可以得到最终的估计计数结果。本文使用平均绝对误差 (Mean absolute error, MAE) 和均方根误差 (root-mean-square error, RMSE) 来衡量预测值和高分辨率图像中真实计数结果的差异。本文比较了使用焦点损失函数和未使用焦点损失函数的模型在不同类别上这两个指标上的表现。如表 4.3 所示：可以观察到使用了焦点损失的模型在小型货车和大型货车类别上均有远超未使用焦点损失的模型的预测结果，而在轿车类别的结果上，与未使用焦点损失的模型相差不大。

表 4.3 模型在不同类别上的均方误差和均方根误差

是否使用焦点损失	类别	平均绝对误差	均方根误差
无焦点损失	轿车	9.85	9.97
焦点损失	轿车	9.87	9.98
无焦点损失	小型货车	29.73	32.34
焦点损失	小型货车	25.61	29.77
无焦点损失	大型货车	39.67	42.12
焦点损失	大型货车	32.71	38.62

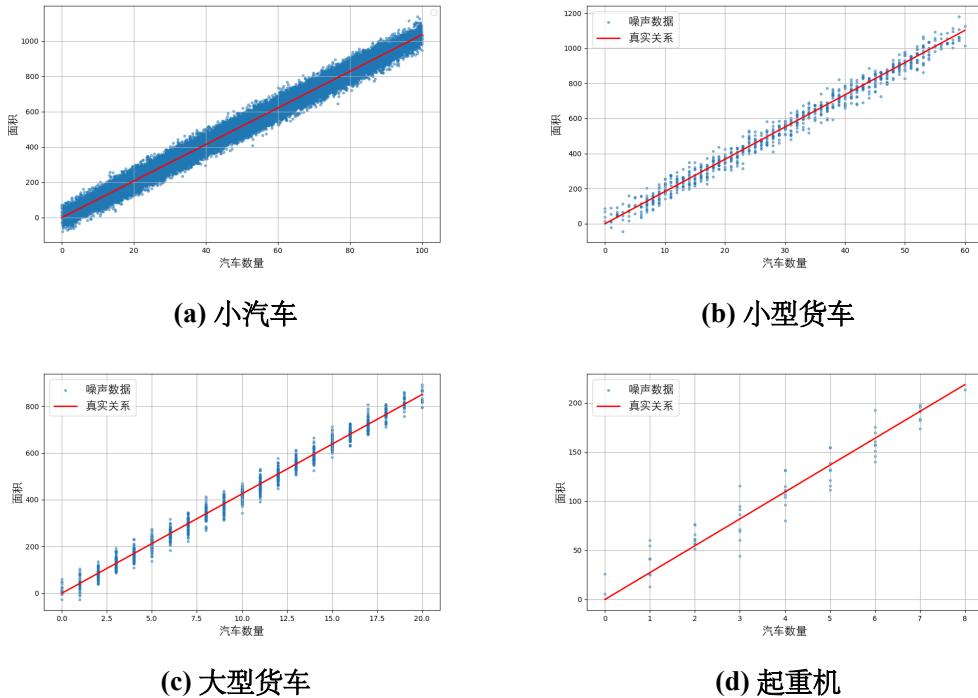


图 4.3 四种汽车回归计数的结果

4.6 训练细节

在训练过程中涉及到多个超参数的选择，其中比较重要的是损失函数的权重。模型使用的损失函数如下：

$$\mathcal{L} = \omega_1 \mathcal{L}_{\text{ent}} + \omega_2 \mathcal{L}_{\text{dir}} + \omega_3 \mathcal{L}_{\text{ser}} + \omega_4 \mathcal{L}_{\text{FL}} \quad (4.5)$$

为了能够让模型能够先掌握车辆的整体结构及更好的适应数量占据大多数的轿车类别，在前 10000 个轮次中，将 ω_4 设置为 0，将 ω_1 设置为 0.5，将 ω_2 设置为 0.4，将 ω_3 设置为 0.1，以避免对于轿车分类的惩罚。在后续的 5000 个轮次中， ω_1 设置为 0， ω_2 和 ω_3 ， ω_4 设置为 0.5，这时使用焦点损失函数专注于学习大型货车和小型货车类别。这有助于更好的识别小样本类别，同时对于背景类别的也能有更好的区分

4.7 局限

本模型采用的多头注意力门结构确实可以增加模型对数据的理解深度，但同时也会带来一系列局限性和挑战。

首先多头注意力机制通常会导致显著的参数增加。这种参数增加意导致在训练和推理过程中需要更多的内存来存储额外的参数和中间计算结果。这种增加的

内存需求可能会限制模型能够处理的最大批量大小。当该方法应用于其他更大规模的数据集上时，参数量及训练时间的影响将更为显著。下表展示了不同模型的参数数量对比，可以直观的观察出参数量的增长。

表 4.4 不同模型的参数量

模型	总参数量	可训练参数量	内存占用
CRVC-Net	37,922,787	37,916,773	144.66MB
CRVC-Net+AG	50,689,801	50,677,897	193.37MB
本文模型	134,969,540	134,956,548	514.87MB

参数数量的增加，尽管可以帮助模型学习更复杂的特征，但也可能使模型更容易过拟合，特别是在如 CRVC 这样的小型数据集上训练时。即使使用图像增广技术，如翻转、缩放和平移，这些操作可能不足以生成模型需要的多样化数据来充分学习并泛化到未见过的数据。同时大量的参数和复杂的模型结构可能导致优化过程中出现问题，如梯度消失或梯度爆炸，对于参数的调整和优化算法的选择上有着更高的要求。

5 总结与展望

5.1 结论

本文设计了一种基于多头注意力机制的 U-Net 网络模型，同时使用焦点损失平衡类别数目不平衡的问题。本设计测试了该方法在跨分辨率遥感影像车辆计数问题上的准确性，尤其是对于稀少类别的估计准确率有不小的提升。模型通过实现不同的注意力门（自注意力门、跨分辨率注意力门、时间序列注意力门），展现了模型在处理复杂图像数据时的灵活性。通过设计消融实验，验证了各注意力门的重要性和影响，显示单独和组合使用各注意力门对模型性能的不同影响。最终实现的三种注意力门组合提供了最高的像素精度，说明了多头的注意力机制门的在综合时间连续性和空间一致性上的有效性。在训练过程中分阶段采用不同的损失函数，特别是引入了焦点损失，有效解决了类别分布不平衡的问题，在小型货车和大型货车的计数结果上取得了显著的提升。本文的研究成果不仅在车辆计数领域具有重要的应用价值，对于其他涉及到跨分辨率图像以及利用低分辨率图像的问题上，该模型也具有很强的迁移潜力。

5.2 不足与展望

本模型虽然在 CRVC 数据集上取得了不错的计数估计结果，但也仍存在一些不足之处。多头注意力机制虽然增强了模型的性能，但也大幅增加了参数数量，这导致更高的内存需求和可能的存在过拟合问题。高参数量和复杂的结构可能引起优化过程中的问题，如梯度消失或梯度爆炸，这要求更精细的参数调整和优化策略。同时 CRVC 数据集较小的数据量可能限制了模型的泛化能力，对于未见过的更大或更复杂的数据集，模型可能需要进一步的调整和优化。

本研究表明，虽然设计的模型在跨分辨率图像的车辆计数任务中表现出了良好的性能，但对于未来的工作，探索参数效率更高的模型架构、增强模型的泛化能力以及优化内存和计算资源的使用，将是进一步提升模型应用实用性的关键方向。进一步发掘高低分辨率图像间的隐含关系也将影响模型预测性能。

参 考 文 献

- [1] ZHAO Q, XIAO J, WANG Z, et al. Vehicle counting in very low-resolution aerial images via cross-resolution spatial consistency and intraresolution time continuity [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-13.
- [2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [3] NAVULUR K. Multispectral image analysis using the object-oriented paradigm[M].
- [4] SHIMIZU K, SHIMIZU F. Laser induced fluorescence spectra of the a $3\Pi_u - X\ 1\Sigma_g^+$ band of Na₂ by molecular beam[J]. J Chem Phys, 1983, 78: 1126-1131.
- [5] Scitor Corporation. Project scheduler[CP/DK]. Sunnyvale, Calif.: Scitor Corporation, 1983.
- [6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
- [7] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- [8] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
- [9] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector [C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016: 21-37.
- [10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
- [11] REDMON J, FARHADI A. Yolo9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
- [12] KARIMIJAFARBIGLOO S, AZAD R, MERHOF D. Self-supervised few-shot learning for semantic segmentation: An annotation-free approach[C]//International Workshop on PReditive Intelligence In MEdicine. Springer, 2023: 159-171.
- [13] CHENG D, LAM E Y. Transfer learning u-net deep learning for lung ultrasound

- segmentation[J]. arXiv preprint arXiv:2110.02196, 2021.
- [14] SHEIKH A, ITKELWAR R, MASKATI A, et al. Crowd detection and analysis for surveillance videos using deep learning[J]. AIP Conference Proceedings, 2022, 2424(1): 080001.
- [15] CHAN A B, LIANG Z S J, VASCONCELOS N. Privacy preserving crowd monitoring: Counting people without people models or tracking[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-7.
- [16] CHAN A, VASCONCELOS N. Bayesian poisson regression for crowd counting [C]//Proceedings of the IEEE International Conference on Computer Vision. 2009: 545-551.
- [17] SHI Z, ZHANG L, LIU Y, et al. Crowd counting with deep negative correlation learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 5382-5390.
- [18] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
- [19] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer, 2015: 234-241.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution

- for semantic image segmentation: arXiv:1706.05587[M]. arXiv, 2017.
- [24] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
- [25] OKTAY O, SCHLEMPER J, FOLGOC L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.
- [26] LI Y, ZHANG X, CHEN D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1091-1100.
- [27] ZHANG Y, ZHOU D, CHEN S, et al. Single-image crowd counting via multi-column convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 589-597.
- [28] BOOMINATHAN L, KRUTHIVENTI S S, BABU R V. Crowdnet: A deep convolutional network for dense crowd counting[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 640-644.
- [29] LIU L, WANG H, LI G, et al. Crowd counting using deep recurrent spatial-aware network[J]. arXiv.org, 2018.
- [30] MINAEE S, BOYKOV Y, PORIKLI F, et al. Image segmentation using deep learning: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(7): 3523-3542.
- [31] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems: volume 25. Curran Associates, Inc., 2012.
- [32] CHEN Y, DAI X, LIU M, et al. Dynamic relu[C]//European Conference on Computer Vision. Springer, 2020: 351-367.
- [33] HAN J, MORAGA C. The influence of the sigmoid function parameters on the speed of backpropagation learning[C]//International workshop on artificial neural networks. Springer, 1995: 195-201.
- [34] APICELLA A, DONNARUMMA F, ISGRÒ F, et al. A survey on modern trainable activation functions[J]. Neural Networks, 2021, 138: 14-32.
- [35] ZAFAR A, AAMIR M, MOHD NAWI N, et al. A comparison of pooling methods for convolutional neural networks[J]. Applied Sciences, 2022, 12(17): 8643.

- [36] NAGI J, DUCATELLE F, DI CARO G A, et al. Max-pooling convolutional neural networks for vision-based hand gesture recognition[C]/2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). 2011: 342-347.
- [37] WU L, LI J, WANG Y, et al. R-drop: Regularized dropout for neural networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 10890-10905.
- [38] BJORK N, GOMES C P, SELMAN B, et al. Understanding batch normalization [J]. Advances in neural information processing systems, 2018, 31.
- [39] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]/2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2547-2554.
- [40] RYAN D, DENMAN S, FOOKE C, et al. Crowd counting using multiple local features[C]/2009 Digital Image Computing: Techniques and Applications. 2009: 81-88.
- [41] TAYARA H, GIL SOO K, CHONG K T. Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network[J]. IEEE Access, 2018, 6: 2220-2230.
- [42] BARR J R, SOBEL M, THATCHER T. Upsampling, a comparative study with new ideas[C]/2022 IEEE 16th International Conference on Semantic Computing (ICSC). 2022: 318-321.
- [43] DUMOULIN V, VISIN F. A guide to convolution arithmetic for deep learning: arXiv:1603.07285[M]. arXiv, 2018.
- [44] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]/BURSTEIN J, DORAN C, SOLARIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.
- [45] YENDURI G, M R, G C S, et al. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions[J]. arXiv.org, 2023.
- [46] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16

- words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. 2021.
- [47] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

