广东工业大学
GUANGDONG UNIVERSITY OF TECHNOLOGY

硕士学位论文

定向目标检测方法的研究

作　　者　　姓　　名：张国生

导　　师　　姓　　名：李东

学科（专业）或领域名称：控制科学与工程

论　文　答　辩　年　月：2021 年 6 月

分类号：　　学校代码：11845

UDC：　　密级：　　学　　号：2111804036

广东工业大学硕士学位论文

（工学硕士）

定向目标检测方法的研究

张国生

导师姓名（职称）：李东（副教授）

学科(专业)或领域名称：控制科学与工程

学生所属学院：自动化学院

答辩委员会主席：刘治（教授）

论文答辩日期：2021年5月26日

A Dissertation Submitted to Guangdong University of Technology
for the Degree's Master of Engineering Science

Research on oriented object detection

Candidate: Zhang Guosheng
Supervisor: Li Dong

May 2021
School of Automation
Guangdong University of Technology
Guangzhou, Guangdong, P. R. China, 510006

摘要

目标检测作为计算机视觉的一项基础而极具挑战的任务，在理论研究和工程应用领域逐渐成为研究热点。目标检测是一种基于目标几何和统计特征，将图像中的目标定位和分类统一化处理的图像分割，在定位目标位置的同时，判别对应目标的类别。现阶段目标检测在智能监控、自动驾驶等视觉领域具有重大意义。传统目标检测一直以来以检测目标的水平矩形框为主要任务，即检测目标在图像上左上角位置和右下角位置，最终以一个水平矩形框来表示目标的位置。然而，近期相关研究表明，在特定的场景下，例如文本检测、工业零件检测和航空影像检测等场景，俯视视角拍摄的图像不再具有以水平面作为参考系，图像中的目标呈现任意旋转的特点。当图像中的目标处于拥挤且倾斜状态时，水平矩形框容易出现边框相互交叠情况，使得边框覆盖过多背景区域或者覆盖其他目标区域，这导致水平矩形框不能契合地表示出目标的几何位置。因此本课题对检测任务展开定向目标检测方法的研究，利用定向边框来表示旋转的目标，为目标检测任务提供更完整的几何轮廓与位置信息。本课题的主要研究内容可总结如下：

（1）对目标检测方法的水平边框表示和定向边框表示展开对比分析，发现现有的旋转目标表示方法普遍存在旋转敏感度误差问题。针对此问题，本课题提出了旋转目标的姿态表示，将不同旋转角目标表示成不同姿态，通过检测目标的中心的位置及回归顶点相对坐标来实现旋转目标的检测。同时为了减小模型的复杂度，本课题提出了单阶段无锚框的定向目标检测网络。

（2）针对目标的多尺度特点，利用数据驱动的方式进行多尺度特征选择，提出了自适应特征金字塔网络。在传统特征金字塔网络的基础上，本课题利用可学习权重使网络自动地从多尺度特征中选择更具判别性的特征，从而使得检测网络能够根据目标尺度来自动调整特征融合策略，选择更具判别下的尺度特征，这对多尺度目标检测有较大的性能提升。

（3）结合实际场景，对目标稀疏的数据提出了选择性采样策略，特别是针对超分辨率的图像。利用目标的位置标签提供采样先验来提高的目标占有率，保证样本多样性的同时提升样本的目标占有率，从而有效地提高网络训练效率和缓解了目标检测问

题中出现正负样本比例不平衡问题。

关键词：姿态；旋转；航空影像；目标检测

ABSTRACT

As a fundamental and challenging task of computer vision, object detection has become a research hotspot in theoretical research and engineering applications. Based on the geometric and statistical characteristics of the object, object detection is an image segmentation that unifies the location and classification of objects in image. While locating the object position, it also distinguishes the category of the corresponding object. It has significantly advantages in intelligent monitoring, automatic driving and other visual fields. Traditional object detection has always been to detect the horizontal bounding box of the objects as the main task, that is, to detect the position of the top-left corner and the bottom-right corner of on the image. Finally, the object is represented as the horizontal bounding box. However, recent related studies have shown that the image taken from the bird's eye view no longer has the horizon as the reference coordinate system in specific scenes, such as text detection, industrial parts detection and aerial image detection, and the object usually exhibits rotating in the image. When the object in the image is crowded and inclined, the horizontal bounding box is prone to overlap with each other. This will make the bounding box cover too much background area or other object areas and unable to properly represent the geometric position of the object. Therefore, in this paper, we conduct research on oriented object detection methods for the detection task. The oriented bounding box is used to represent the oriented object, which provides richer geometric contour and position information for the object detection task. The main contents of this paper are as follows:

(1) Through a comparative analysis of the horizontal bounding box representation and the oriented bounding box representation of the object detection, it is found that the existing oriented object representation method generally has the problem of rotation sensitivity error. To tackle this problem, the pose representation of the oriented object is proposed to represent objects with different rotation angles as different poses. The detection of the oriented object is achieved by locating the position of the center of the object and regressing the relative coordinates of the vertices. Meanwhile, to reduce the complexity of the model, we propose

an one-stage and anchor-free oriented object detection network.

(2) Aiming at the multi-scale characteristics of the object, we propose an adaptive feature pyramid network by using a data-driven method for feature selection. Base on the traditional feature pyramid network, we propose to uses learnable weights to make the network automatically select more discriminative features from the multi-scale features. Consequently, the detection network can automatically adjust the feature fusion strategy according to the object scale. It greatly improve the performance of multi-scale object detection.

(3) Combining practical scenarios, a selective sampling strategy is designed for the sparse object data, especially for the extra high resolution images. The bounding box label of the object is used to provide a sampling prior to increase the object occupancy rate, thus effectively improving the efficiency of network training and alleviating the imbalance problem of positive and negative samples.

Key words: pose; orient; aerial image; object detection

目录

摘要 ..... **I**

ABSTRACT ..... **III**

目录 ..... **V**

CONTENTS ..... **VIII**

第一章 绪论 **1**

1.1 研究背景与意义 1

1.2 国内外研究现状 2

1.2.1 传统目标检测算法 2

1.2.2 定向目标检测算法 2

1.2.3 航空影像定向目标检测算法 3

1.3 本课题研究内容及章节安排 4

1.3.1 课题研究内容 5

1.3.2 课题章节安排 5

1.5 本章小结 5

第二章 定向目标检测基础理论 **7**

2.1 深度学习基础理论 7

2.1.1 卷积神经网络 7

2.1.2 深度神经网络 11

2.1.3 深度神经网络的优化算法 13

2.2 目标检测算法基础理论 15

2.2.1 交并比 16

2.2.2 非极大值抑制 18

2.2.3 评测指标 19

2.3 定向目标检测的基本方法 21

2.3.1 基于双阶段的方法 21

2.3.2 基于单阶段的方法 22

2.3.3 定向目标表示问题.....23
2.4 本章小结.....24
第三章 基于关键点的定向目标检测.....26
3.1 基于关键点的目标检测.....26
3.1.1 现有的方法.....26
3.1.2 改进的方法.....28
3.2 定向目标的表示方法.....29
3.2.1 基于边框表示的特点.....29
3.2.2 基于姿态表示的特点.....30
3.3 自适应特征融合网络.....33
3.3.1 特征金字塔网络.....33
3.3.2 多尺度特征自适应融合.....34
3.4 本章小结.....34
第四章 实验设计与结果分析.....36
4.1 实验数据集介绍.....36
4.1.1 DOTA 数据集.....36
4.1.2 VisDrone 数据集.....38
4.2 实验的实现细节.....38
4.2.1 训练策略.....38
4.2.2 均匀采样策略.....39
4.2.3 选择性采样策略.....41
4.3 实验结果与分析.....43
4.3.1 对比实验分析.....43
4.3.2 消融实验分析.....46
4.4 本章小结.....47
结论与展望.....49
参考文献.....51
攻读学位期间取得与学位论文相关的成果.....56

学位论文独创性声明.....57

致 谢.....58

CONTENTS

ABSTRACT(IN CHINESE)..... I

ABSTRACT(IN ENGLISH)..... III

CONTENTS(IN CHINESE)..... V

CONTENTS(IN ENGLISH)..... VIII

Chapter 1 Introduction..... 1

1.1 Background and significance of research..... 1

1.2 Research status at home and abroad..... 2

1.2.1 Traditictional object detection..... 2

1.2.2 Oriented object detection..... 2

1.2.3 Aerial image oriented object detection..... 3

1.3 The chapter arrangement and main contents of this subject..... 4

1.3.1 Contents of this subject..... 5

1.3.2 Chapter arrangement..... 5

1.5 Chapter summary..... 5

Chapter 3 Basic theory of directional object detection..... 7

2.1 Basic theory of deep learning..... 7

2.1.1 Convolutional neural network..... 7

2.1.2 Deep Neural Network..... 11

2.1.3 Optimization algorithm of deep neural network..... 13

2.2 Basic theory of object detection algorithm..... 15

2.2.1 Intersection over Union..... 16

2.2.2 Non-maximum suppression..... 18

2.2.3 Evaluation metric..... 19

2.3 Basic method of oriented object detection..... 21

2.3.1 Based on a two-stage method..... 21

2.3.2 Based on a one-stage method..... 22

2.3.3 The representation of oriented.....23
2.4 Chapter summary..... 24
Chapter 4 Oriented object detection based on key points.....26
3.1 Object detection based on key points.....26
3.1.1 Current method.....26
3.1.2 Improved method..... 28
3.2 Representation method of oriented object.....29
3.2.1 Based on the characteristics of bounding box representation..... 29
3.2.2 Based on the characteristics of pose representation..... 30
3.3 Adaptive feature fusion network..... 33
3.3.1 Feature Pyramid Network..... 33
3.3.2 Multi-scale feature adaptive fusion..... 34
3.4 Chapter summary..... 34
Chapter 4 Experimental design and result analysis.....36
4.1 Introduction to the experimental data set..... 36
4.1.1 DOTA data set..... 36
4.1.2 VisDrone data set..... 38
4.2 Implementation details of the experiment.....38
4.2.1 Training strategy.....38
4.2.2 Evenly sampling strategy..... 39
4.2.3 Selective sampling strategy.....41
4.3 Experimental results and analysis..... 43
4.3.1 Comparative experiment analysis..... 43
4.3.2 Analysis of ablation experiments..... 46
4.4 Chapter summary..... 47
Conclusion and prospect.....49
References..... 51
Publication and patents during study..... 56

Statement of original authorship and copyright licensing declaration.....57

Acknowlegements.....58

第一章 绪论

1.1 研究背景与意义

近些年深度学习的研究，无论是在理论研究领域，还是在工程应用领域都大力推动了计算机视觉的发展，目标检测作为计算机视觉的一项重要且极具挑战性的任务，也成为国内外学者的研究热点。目标检测是一种基于目标几何和统计特征，将图像的检测和识别合二为一的图像分割方法，在完成定位目标的位置的同时，还需对目标的类别进行识别。传统目标检测一直以来以检测目标的水平矩形框为主要任务，即检测目标在图像上左上角位置和右下角位置，最终以一个水平矩形框来表示目标。然而，最近的相关研究表明，当图像中的目标处于拥挤且倾斜状态下，水平矩形框并不能很好的表示出目标的几何位置，如图 2-13 所示，水平边框表示目标容易使边框覆盖太多背景区域或者覆盖其他目标区域，因此最近相关的研究开始针对检测任务提出定向目标检测的研究。定向边框能契合目标轮廓，避免了背景重叠过多的现象，具有更丰富的几何位置信息，能更好的表示目标的几何位置。因此，定向目标检测任务，可以视为传统目标检测任务的一项扩展课题。

传统目标检测旨在检测目标的水平边框，是因为现实中绝大多数图像拍摄于现实生活，且拍摄视角一般水平，所以图像中的目标一般具有水平位置特性。然而，在一些特定应用场景，拍摄视角并不限定于水平视角，例如航空影像和遥感影像等，这些场景下一般为远距离高空视角，同时意味着拍摄图像目标具有任意方向且微小拥挤的特点，传统的水平边框难以契合地表达目标的几何位置，因此，对于非水平视角的影像场景，提出了定向目标检测的扩展任务。随着无人机技术和遥感技术的发展，目标检测在航空遥感影像的应用也逐渐体现其价值，例如无人机视角车辆检测、遥感图像建筑、港口、船只的检测等，这些技术在未来无人机智能巡航，遥感技术的智能监控等领域具有极大推动作用，因此，本课题针对传统目标检测的水平边框不足，提出更加完善的检测方法，用具有任意方向的定向边框来表示目标，使检测目标具有更加丰富发几何位置信息，表明定向目标检测研究意义，同时本文在实验章节将以航空遥感影像中的定向目标检测为例，利用基于卷积神经网络的定向目标检测技术，实现对车辆、船舶、码头等重点目标进行检测，充分展现了本研究课题在推动智能化遥感信息处理

所带来的价值。

1.2 国内外研究现状

1.2.1 传统目标检测算法

采用传统目标检测方法可分为单阶段和双阶段方法，双阶段方法能实现更高的检测性能但需要更大的计算复杂度，而单阶段方法虽然性能次之，但检测速度更快，易于实现实时检测。双阶段方法可以总结为 RCNN 系列[1-3]，第一阶段生成一系列区域提案(Region Proposal)，然后送入第二阶段的分类和回归网络。例如，Fast RCNN[1]在特征图上提取感性区区域 (Region of Interest，RoI)来减小计算量，同时利用 RoIPooling 进行特征的尺寸变换，将原来不同尺寸的特征变换到统一尺寸，以便后层的全连接网络的输入；Faster RCNN[2]提出区域提案网络(Region Proposal Netowrk，RPN)和锚框(Anchor)机制，通过共享卷积特征，从特征图提取 Proposal，而锚框机制为目标尺寸提供了有效的先验信息，进一步提高检测效率和性能；Mask RCNN[3]用 RoIAlign 替换 RoIPooling，利用双线性插值来解决边框的量化误差，从而有效解决因量化误差而造成的精度损失。不同于双阶段方法，单阶段方法省略第一阶段的提案网络，直接进行区域的分类和回归。经典的单阶段方法包括有 YOLO 算法[4-7]，将检测问题直接转化为分类和回归问题，实现单阶段实时检测并得以实际应用[8]，但由于其稀疏监督方式，对小目标检测并不友好；RetinaNet 算法[9]提出 Focal Loss 来解决单阶段训练过程中正负样本不平衡问题，利用难样本挖掘的思想，在损失函数上进行软抑制，减小简单样本的权重比重，从而加强了对难样本的选择，但在样本极端不平衡情况下，依然有所局限。最近单阶段方法尝试利用检测关键点的方法[10-11]来实现目标检测，并实现了能与双阶段方法媲美的检测性能，如 CenterNet[11]通过检测目标的中心点然后进一步在中心点处回归边框来实现单阶段无锚框目标检测，得益于其思路简单与对小目标检测友好等特点，在学术研究领域和工程应用领域也逐渐成为一个热点研究。随着目标检测方法研究的发展，目标检测在检测对象上呈现出多元化和专业化发展，在一些特等的场景发展出针对性的目标检测方法。

1.2.2 定向目标检测算法

定向目标检测方法的研究是首次在光学字符识别 (Optical Character Recognition，OCR) 领域被提出。Liao 等人基于 SSD 方法提出创新的文本检测方法 Textboxes[12]，应

用了不同尺度及宽长比的锚框和1×5的卷积核来实现对长文本目标的检测。Textboxes++方法[13]通过直接回归8个角点实现对Textboxes方法的改进，使其能够检测旋转的文本。Liu等人[14]设计了一种规则来计算定向目标角点的顺序，同时提出了并行IoU计算方法来减小计算时间。旋转区域提案网络[15](Rotation Region Proposal Network, RRPN)在Faster RCNN的RPN阶段提出了(Rotation Region of Interest, RRoI)池化层来实现文本检测。由于文本目标的宽长比的变化很大，基于锚框的方法很难实现对所有目标进行有效覆盖，因此很多方法开始尝试使用无锚框(Anchor Free)的方法。He等人[16]和Zhou等人[17]利用压缩目标掩码生成二进制标签，并在前景像素区域回归目标的角点和旋转角来实现文本的无锚框检测。无锚框区域提案网络[18](Anchor Free Region Proposal Network, AF-RPN)基于Faster RCNN基础上应用了特征金字塔网络（Feature Pyramid Network，FPN），同样采用压缩目标掩码来生成标签并实现无锚框检测。随着检测技术的发展，定向目标检测从OCR技术领域开始向其他俯视场景领域发展，例如航空遥感影像领域。

1.2.3 航空影像定向目标检测算法

受益于传统目标检测算法，航空影像旋转目标检测也得到了相应的研究进展。Ding等人[19]在Faster RCNN基础上提出RoI Transformer来回归水平RoI和旋转RoI的偏移，将RPN输出的水平边框(Horizontal Region of Interest, HRoI)转换为旋转的RRoI，从而实现旋转目标检测；SCRDet[20]通过特征融合和锚框采样的角度出发，设计了一种特征融合结构，利用多维注意力机制（像素注意力和通道注意力）来应对航空影像的复杂背景，并设计了IoU损失函数来进一步提升旋转目标检测的性能；Xu等[21]在水平边框检测基础上，提出Gliding vertex的方法，将旋转边框的顶点转变为水平边框上的现对偏移，通过回归顶点在边框方向的偏移比例实现旋转目标检测。虽然以上方法能实现不错的性能，但是它们都是基于双阶段网络的方法，需要更大的计算代价，不利于方法的实际应用。为此，单阶段的方法开始尝试，RSDet[22]分析了由于角度固有的周期性以及相关的宽度和高度的突然变化导致的损失不连续性，提出了旋转敏感度误差(Rotation Sensitivity Error，RSE)的概念，针对性的引入八参数模型并设计了调制旋转损失函数，有效缓解角度所带来的旋转敏感度误差问题；R3Net[23]提出了可旋转的区域提案网络，通过在特征图上裁剪旋转边框区域来生成旋转的RoI，以上两个单阶段的方

法基于单阶段 RetinaNet 方法，依赖于对锚框的设计，在密集的小目标检测上容易出现漏检问题。DRN[24]尝试无锚框的检测网络的设计，在 CenterNet 的基础上额外增加一个角度变量进行回归，然而忽略了角度的周期性特点会带来的 RSE 问题。

航空影像目标检测的难点主要包括检测背景复杂，目标尺度变化大和成像视角变化大。航空影像主要包括有遥感影像和无人机航拍影像，遥感影像由于其高空视角，成像的目标尺度偏小且背景区域存在许多复杂的干扰信息，增加了识别任务的难度。由于高空视角，成像的分辨率特别大，对于深度神经网络而言，超高分辨率的图像会极大的增加网络的计算量，这也导致深度神经网络计算资源消耗远高于传统机器学习的方法，若对图像简单进行下采样，则会造成原始图像信息的损失，因此针对高分辨率的航空影像设计更为高效的采样方法也尤为重要。更为挑战性的是在俯视视角下目标成像杂乱无章，拥挤等特点，传统的水平边框难以契合的表示出目标的几何位置。而对于无人机航拍影像或者遥感影像，由于拍摄视角具有较高的自由度，视角变化大，目标尺度也呈现多样化，目标往往呈现出细小，密集和杂乱无章的特点，也很大程度加大了检测难度。

综上所述，航空影像定向目标检测是传统目标检测的针对特定任务的一项延伸性工作，大多数工作是由传统目标检测发展而来，因此，本文对传统目标检测进行了相应总结和回顾，同时回顾了定向目标检测的开创性工作，即光学字符识别（OCR）,最后对近期航空影像定向目标检测方法进行总结与分析，在本文第三章也将对部分经典的航空影像定向目标检测方法进行详细介绍。根据对现有的航空影像定向目标检测方法的优劣分析，本文在此明确的立论分析，核心解决的问题可总结如下几点：

（1）为了提高检测方法工程落地的可能性，本文将减小网络的模型复杂度，摈弃复杂锚框检测的设计，并设计基于关键点的单阶段无锚框的检测方法；

（2）解决定向目标检测中因角度周期性而引起的旋转敏感度误差（RSE）问题，提出了旋转目标的姿态表示方法，利用姿态图的预测来解决角度回归问题；

（3）解决对超高分辨率航空影像的训练效率问题，提出了选择性采样，通过提高目标占有率来提高网络的训练效率，同时缓解了目标检测问题中出现前景背景比例不平衡问题。

1.3 本课题研究内容及章节安排

1.3.1 课题研究内容

区别于传统目标检测，本课题主要将针对定向目标检测的难点着手，研究一种创新的定向目标检测方法。本课题的主要研究内容包括：创新性提出定向目标的姿态表示方法，将旋转目标表示为中心点和 4 个顶点构成的不同姿态，有效避免周期性角度变量回归问题，即旋转敏感度误差问题，设计了基于关键点检测的单阶段无锚框定向目标检测网络，利用高分辨率表示的特征提取网络，充分保留图像的空间特征信息；另外本课题针对目标多尺度问题，使用了改进的自适应融合的特征金字塔网络(Adaptive Feature Pyramid Network，AFPN)，利用可学习权重对不同尺度特征进行加权融合，以数据驱动的方式使网络自动地选择更具判别性的尺度特征，整个过程能够实现端到端训练；最后以航空影像中的定向目标检测为例，提出选择性采样(Selective Sample，SS)策略，在超高分辨率的遥感图像进行训练样本选择性采样，相对于一般的滑动窗口均匀采用，选择性采样能有效提高网络的训练效率，同时缓解了训练过程中正负样本不平衡问题，提高了模型的整体性能；本课题通过丰富实验的实验以及相应的对比分析，来证明所设计方法的有效性。

1.3.2 课题章节安排

本文将围绕定向目标检测方法的研究思路，按章节展开。首先本文将在第二章节回顾总结目标检测方法的基础理论，首先介绍在计算机视觉领域应用特别广泛的卷积神经网络，从基础的卷积操作到深度神经网络架构的构成，然后进行特征网络下游网络的设计理论以及相应的优化算法，最后介绍了定向目标检测的基础理论以及典型算法；第三章将详细介绍本课题的所研究的定向目标检测方法的设计思路和创新点，首先介绍基于关键点检测方法的优势即相应的设计思路，然后针对定向目标检测的特点设计了利用关键点表示的定向目标姿态表示方法，最后针对目标多尺度问题对网络结构进行了相应的改进，提升模型检测性能；第四章就所设计的定向目标检测方法进行相应的实验，在实验中提出了数据采样策略，进一步提升模型的性能，并以航空影像数据为例进行实验分析与总结；最后进行本课题的总结与展望。

1.5 本章小结

本章首先介绍了本课题的研究背景和研究意义，指出了定向目标检测在当前研究领域的现状和价值，同时针对传统目标检测和定向目标检测的系列方法进行了回顾，总

结出当前阶段传统目标检测与定向目标检测的关系呈现一种递进发展的关系，同时总结了现有方法存在的问题，从而明确了本文的理论研究内容。本课题也将围绕几个实际问题展开研究，从实用性角度出发提高本课题所研究的现实意义。

第二章　定向目标检测基础理论

本章对目标检测方法所需要的基础理论展开介绍，首先介绍在计算机视觉领域应用特别广泛的卷积神经网络，从基础的卷积操作到深度神经网络架构的构成，然后进行特征网络下游网络的设计理论以及相应的优化算法，最后介绍了定向目标检测的基础理论以及典型算法。

2.1 深度学习基础理论

2.1.1 卷积神经网络

卷积神经网络是由前馈神经网络发展而来，经典的前馈神经网络结构如图 2-1 所示，主要由一层输入层（Input Layer），多层隐藏层（Hidden Layers）和一层输出层（Output Layer）构成。多个神经元构成神经网络的层结构，且每个神经元都连接前一层和后一层的每个神经元，神经元通过权重矩阵对前一层的输入特征向量进行线性变换，随后使用非线性激活函数对变换后的特征进行非线性变换，从而使神经网络具备拟合复杂非线性函数的能力。

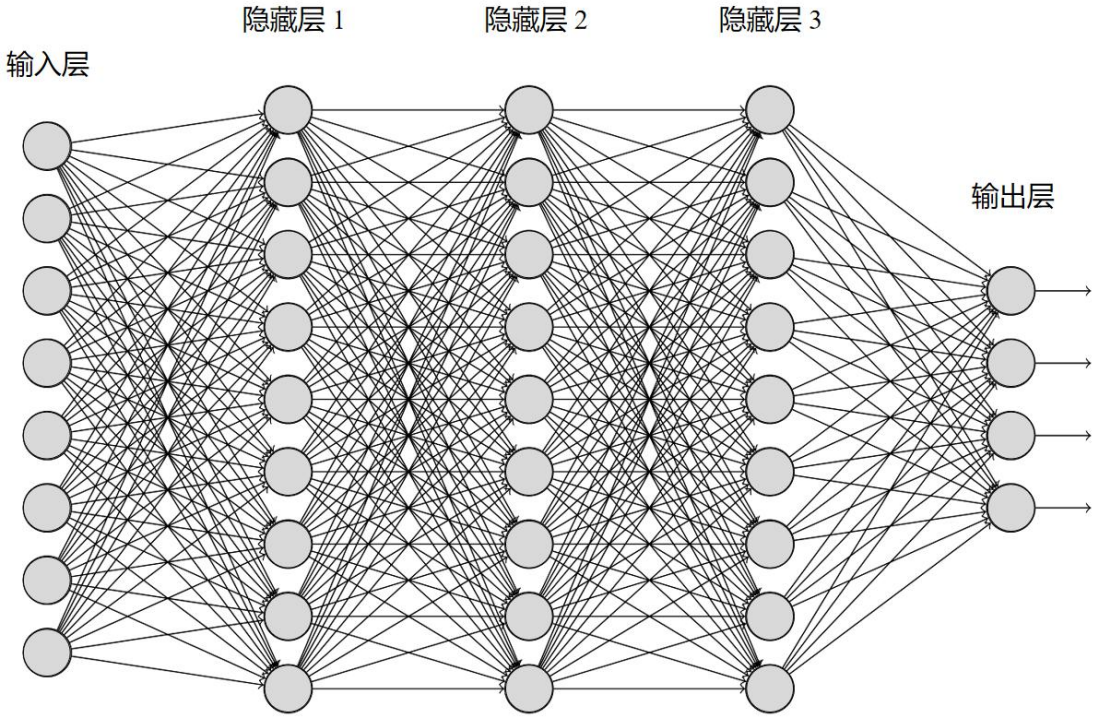

图 2-1 多层前馈神经网络示意图

Fig. 2-1 Schematic diagram of feedforward neural network

然而经典的前馈神经网络采用了全连接结构，各个神经元直接稠密连接，当输入层

的特征数量增大时，例如处理图像时，每层网络的权重参数将大规模增大，导致占用大量的计算资源。同时，前馈神经网络每层神经元与前一层神经元采样一致性连接，忽略了图像中局部相关的语义信息，针对计算机视觉领域图像的特点，研究者提出了卷积神经网络（Convolutional Neural Network，CNN）。

卷积神经网络是一种包含卷积计算的前馈神经网络，具有较强的表征学习（Representation Learning）能力，由于其卷积算子的特点，使其能够按阶层结构对输入信息进行平移缩放不变分类，亦称为平移不变的人工神经网络（Shift-Invariant Artificial Neural Networks，SIANN）。卷积神经网络依靠卷积核计算的特点，与一般的前馈神经网络相比具有局部连接、权重共享的结构特性，即在网络前向传播过程中，每个神经元与下一层网络的神经元进行连接部分连接，同时共享连接权重，由此，卷积神经网络能够降低网络的参数规模。卷积算子带来平移不变性，结合池化算子带来旋转缩放不变性，使其在图像提取局部相关特征带来明显优势，因此，CNN 成功在计算机视觉领域得到了广泛应用。

卷积神经网络输入为多维张量，以计算机视觉领域常用的二维卷积为例，设输入为三维张量 $\boldsymbol{X} \in \mathbb{R}^{M \times N \times D}$，其中 M, N 和 D 分别为输入张量的高度，宽度和深度，若输入为 RGB 图像，则 $D = 3$，另外定义输入特征映射为输入张量的每个切片（Slice）的二维矩阵 $\boldsymbol{X}^d \in \mathbb{R}^{M \times N}$，其中 $1 < d < D$。设相应的卷积输入为三维张量 $\boldsymbol{Y} \in \mathbb{R}^{M' \times N' \times P}$，同样 M', N' 和 P 为输入张量的高度，宽度和深度，定义输出特征映射为输出张量的每个切片的二维矩阵 $\boldsymbol{Y}^p \in \mathbb{R}^{M' \times N'}$，其中 $1 < p < P$。如图 2-2 所示，卷积的输入特征 $\boldsymbol{X}$ 经过卷积核卷积运算得到输出 $\boldsymbol{Y}$，其中卷积核为四维张量 $\boldsymbol{W} \in \mathbb{R}^{m \times n \times D \times P}$，其中切片矩阵 $\boldsymbol{W}^{p,d} \in \mathbb{R}^{m \times n}$，$m, n$ 为卷积核的宽度和高度。

为了清晰说明卷积的计算过程，可以将高维卷积分解为二维卷积的线性叠加，以计算输入特征的其中一个输出特征映射 $\boldsymbol{Y}^p$ 为例，$\boldsymbol{Y}^p$ 由第 p 个卷积核 $\boldsymbol{W}^p$ 与 $\boldsymbol{X}$ 卷积计算得到，如公式（2.1）和公式（2.2）所示，其中 $\otimes$ 为卷积符号，卷积过程将在下文介绍，b^p 为偏置变量，$\boldsymbol{Z}^p$ 为卷积的响应值，其经过非线性激活函数 $f(\bullet)$ 才得到卷积的激活特征映射 $\boldsymbol{Y}^p$。非线性激活函数可以是传统前馈神经网络的激活函数，如 Sigmod 函数，但在卷积神经网络中，一般使用 ReLu 激活函数，进而为了计算整个输出特征映射 $\boldsymbol{Y}$，可以将上述过程重复计算 P 次依次得到输出特征映射 $\boldsymbol{Y}^1, \boldsymbol{Y}^2, \cdots, \boldsymbol{Y}^P$ 即得到完整的输出

特征映射 $\boldsymbol{Y}$。综合以上过程，可以发现一个卷积过程除了需要四维的卷积核 $\boldsymbol{W}$ 以外，还需增加 P 个偏置参数，因此每层卷积核的参数总量为 $P\times D\times m\times n+P$。

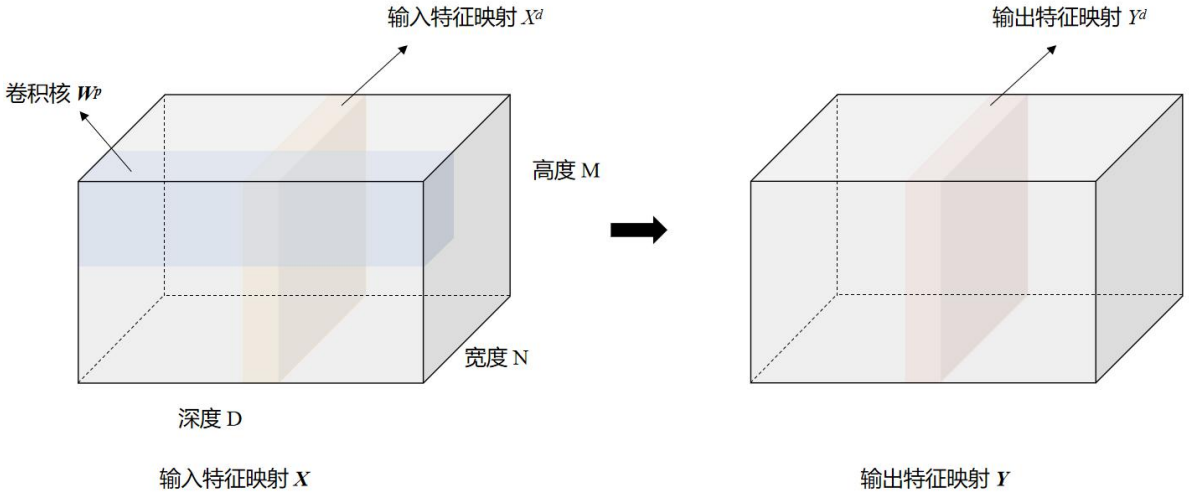

图 2-2 卷积计算示意图

Fig. 2-2 Convolution calculation diagram

b^p 为偏置变量，$\boldsymbol{Z}^p$ 为卷积的响应值，其经过非线性激活函数 $f(\bullet)$ 才得到卷积的激活特征映射 $\boldsymbol{Y}^p$。非线性激活函数可以是传统前馈神经网络的激活函数，如 Sigmod 函数，但在卷积神经网络中，一般使用 ReLu 激活函数，进而为了计算整个输出特征映射 $\boldsymbol{Y}$，可以将上述过程重复计算 P 次依次得到输出特征映射 $\boldsymbol{Y}^1,\boldsymbol{Y}^2,\cdots,\boldsymbol{Y}^P$ 即得到完整的输出特征映射 $\boldsymbol{Y}$。综合以上过程，可以发现一个卷积过程除了需要四维的卷积核 $\boldsymbol{W}$ 以外，还需增加 P 个偏置参数，因此每层卷积核的参数总量为 $P\times D\times m\times n+P$。

$$\boldsymbol{Z}^p=\boldsymbol{W}^p\otimes\boldsymbol{X}+b^p=\sum_{d=1}^{D}\boldsymbol{W}^{p\,d}\otimes\boldsymbol{X}^d+b^p \tag{2.1}$$

$$\boldsymbol{Y}^p=f(\boldsymbol{Z}^p) \tag{2.2}$$

在卷积神经网络中的卷积计算操作可以直观理解为在图像或特征上以滑动窗口（Sliding Window）的方式进行计算，但在现有的深度学习框架中如 TensorFlow，PyTorch 中都将卷积运算转换为矩阵运算来进行加速，通常采样高度优化的通用矩阵乘法（General Matric Multiplication，GEMM）高效实现。而对于滑动窗口的方式，在此定义如下几个概念：滑动步长 S，填充量（Padding，P），如图 2-3 所示，对滑动到的每个窗口进行加权求和计算得到输出映射对应位置的激活值。因此，可以根据步长

和填充量得到相应输出特征映射的尺寸，计算如公式（2.3）所示，其中$\lfloor \bullet \rfloor$为向下取整运算。

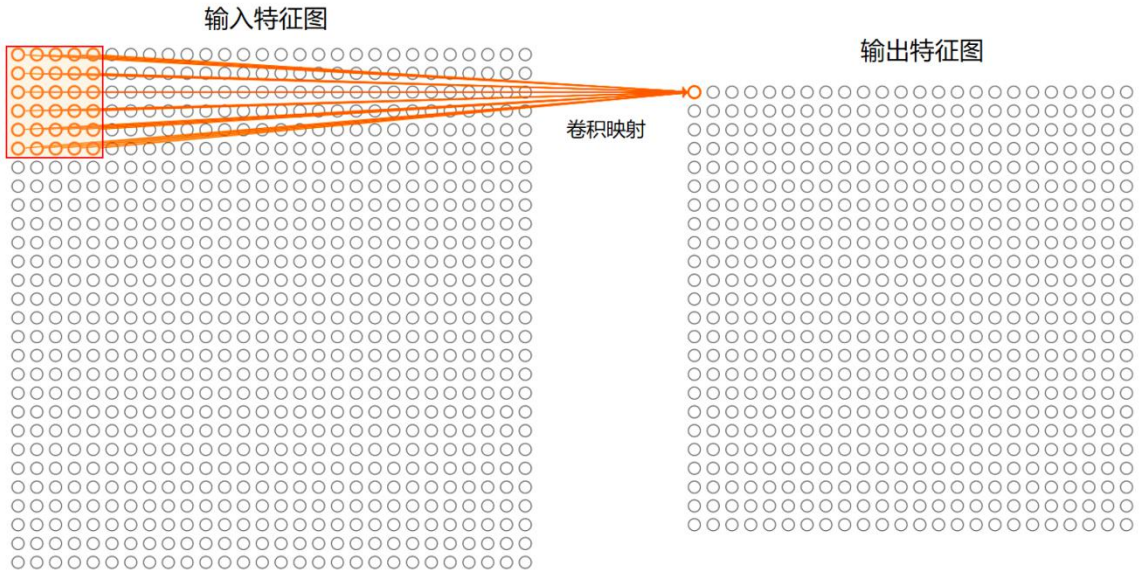

图 2-3 卷积输入特征与输出特征的计算映射关系

Fig. 2-3 The mapping between convolution input features and output features

$$\begin{cases} M' = \left\lfloor \dfrac{M + 2 \times P - m}{S} \right\rfloor + 1 \\ N' = \left\lfloor \dfrac{N + 2 \times P - n}{S} \right\rfloor + 1 \end{cases} \tag{2.3}$$

卷积网络利用网络的稀疏连接来减小网络的参数数量，但每层网络的特征映射的神经元数量并没有减少，因此，为了进一步减小网络的特征数量，在实际应用中通常会在卷积计算后加入池化层（Pooling Layer），池化层一般连接在卷积网络输出特征响应之后，用于进行特征选择，降低特征数量，从而减小网络的计算量，同时能缓解网络过拟合的风险。池化层类似于卷积算子的操作，采用滑动窗口的方式，同样有步长S和核大小m,n，但是池化层仅对输入的特征进行采样，不带任何权重参数。常用的池化层主要分为两种：最大池化（Max Pooling）和平均池化（Mean Pooling）,如图 2-4 所示，最大池化层未来保留特征对定纹理特征，对采样窗口区域的特征取最大值作为对应位置输出的特征值，具有特征选择的作用；而平均池化是对采样窗口区域求平均值作为对应位置输出的特征值，能够保留特征的背景信息。

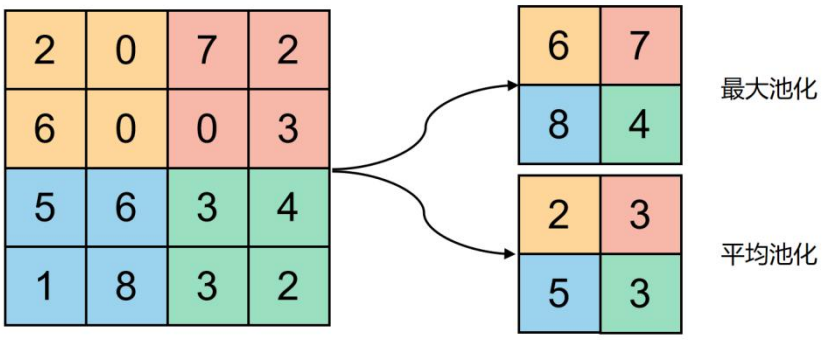

图 2-4 最大池化和平均池化

Fig. 2-4 Max pooling and mean pooling

2.1.2 深度神经网络

深度神经网络在此主要介绍以卷积作为构建单元的深度神经网络，最早利用卷积单元搭建的深度神经网络是由 Yann LeCun 提出的 LeNet[25]，推进了深度学习的发展进程。LeNet 的结构包含了现有卷积神经网络的基本构成元素，包括由卷积层，池化层和线性激活层，并在图像领域取得突破性进展。深度神经网络在最近几年发展颇为迅速，尤其得益于当前计算能力的高速发展，在 2012 年，Alex Krizhevsky 提出 AlexNet[26]并在 ImageNet 大赛取得了冠军，由此深度神经网络进入了研究爆发阶段。

AlexNet 由 LeNet 的基础上发展而来，并提出了多项改进技术，例如将网络的激活函数改用了 ReLu 激活函数，使网络在训练过程中收敛更快，同时有效避免了梯度消失问题；局部响应归一化（Local Response Normalization，LRN）的引入很大程度增强网络的泛化能力；使用了 Dropout 技术，在网络训练过程中以一定概率随机忽略部分神经元，以此来达到缓解过拟合效果。以上的每个改进技术很大程度推动了深度神经网络的发展，牛津大学提出了 VGG 网络[27]，使用了更小的3×3的卷积核，使用多个更小的卷积取代较大的卷积，这能够有效提高网络的感受野（Receptive Field），同时能够引入更多的非线性函数，提高了模型的表达能力。GoogleNet[28]提出了 Inception 模块，如图 2-5 所示，GoogleNet 使用了一系列1×1，3×3和5×5的卷积核并行组合，同时通过使用1×1的卷积来降低特征数，以达到减小计算量的目的。

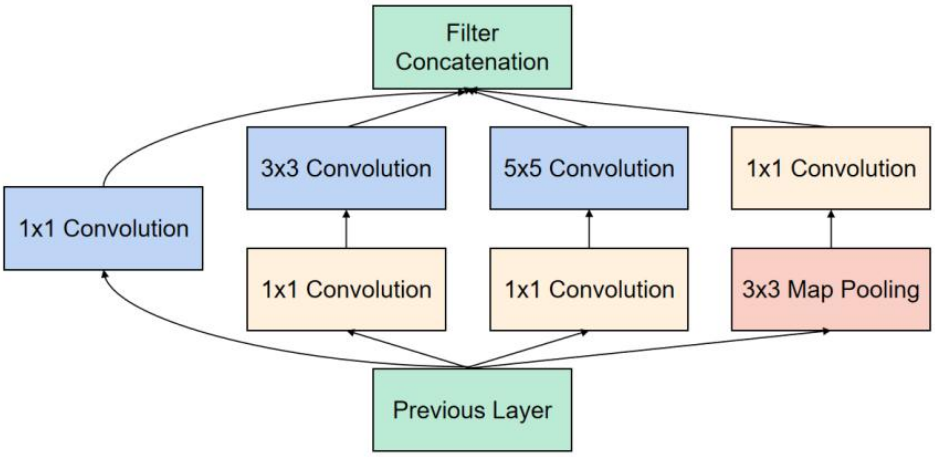

图 2-5 Inception 模块

Fig. 2-5 Inception module

ResNet[29]提出了深度残差学习（Deep Residual Learning）并成为应用最为广泛的网络，通过残差连接促使网络尝试学习输入 $\boldsymbol{X}$ 与输出 $H(\boldsymbol{X})$ 之间的残差 $H(\boldsymbol{X})-\boldsymbol{X}$，在网络加深过程中能够使残差模块学习恒等映射（Identity Mapping）。如图 2-6 所示，在输入和输出之间建立直接关联的通道，通过实验证明，ResNet 网络在深度加深的情况下能够维持网络的性能，恒等映射也很大程度上为梯度反向传播（Back Propagation）提供了路径，解决了网络退化问题，同时残差连接还能缓解了网络训练过程中梯度弥散问题。

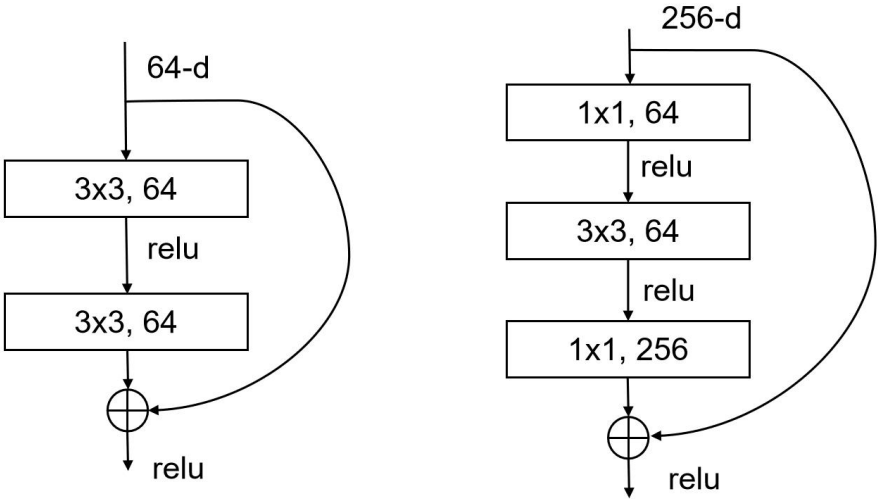

图 2-6 残差模块

Fig. 2-6 Residual module

近几年，深度神经网络开始迈向特定化多样化形式发展，例如适用于检测和分割任务是高分辨率特征表示的网络 UNet[30]，Hourglass[31]，HRNet[32]等。网络结构设计也不

再局限于手工设计，网络架构搜索（Neural Architecture Search，NAS）技术也很大促进了深度神经网络性能的发展，通过给定的搜索空间及相应的搜索策略，使网络自动地设计最佳的结构，较为出色的工作如 EfficientNet[33]，该网络结合网络架构搜索技术提出了模型复合扩张的方法，从网络深度（Depth）,网络宽度（Width）和图像输入分辨率（Resolution）进行协同缩放，并搜索得到一个协同缩放的关系式，在保证模型规模大小的同时，大幅度提升模型性能。

深度神经网络的发展，极大推动了深度学习各项下游任务的发展，尤其是计算机视觉领域，包括分类、目标检测、语义分割和实例分割，而近些年目标检测的发展也很大程度上收益于深度神经网络的发展，特别是利用一个优异的特征提取网络往往为目标检测带来可观的性能提升，例如基于 EfficientNet 作为特征提取网络的 EfficientDet[34]，在目标检测性能上达到最优结果，这充分证明了一个优异特征提取网络对下游任务性能提升的可观性。

2.1.3 深度神经网络的优化算法

局部极小值问题是低维空间的非凸优化问题的主要挑战点，基于一般的梯度下降优化方法容易陷入局部极小值，得到次最优解。深度神经网络的优化其实是在一个超高维空间中的非凸优化问题，因此对优化算法的设计有着更高的要求，下面分别介绍深度神经网络常有优化算法。

小批量梯度下降（Mini-Batch Gradient Descent）算法，是为了缓解深度神经网络进行大规模训练数据训练时，计算整个训练数据梯度计算量大的问题，以小批量数据的梯度方向来估计整体的梯度方向，能有效提高网络训练效率，同时引入的梯度估计带来的有限偏差可以是为随机噪声，能提高网络泛化能力。不失一般性，我们定义 $f(\boldsymbol{x};\boldsymbol{\theta})$ 为深度神经网络，其中 $\boldsymbol{x}$ 为网络输入，$\boldsymbol{\theta}$ 为网络参数，使用批梯度下降进行优化时，每次输入为一个批量样本 $I_t=\{(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(k)})\}_{k=1}^{K}$，其中 t 为第 t 次迭代，$\boldsymbol{y}^k$ 为小批量样本中第 k 个样本的标签。则小批量梯度的梯度计算如公式（2.4）所示。

$$\boldsymbol{g}_t(\boldsymbol{\theta})=\frac{1}{K}\sum_{(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(k)})\in I_t}\frac{\partial L(\boldsymbol{y}^{(k)},f(\boldsymbol{x}^{(k)};\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}, \tag{2.4}$$

其中 $L(\bullet)$ 深度神经网络的损失函数，则根据梯度下降算法进行负梯度更新，t 次迭代的参数更新如公式（2.5）所示。

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \boldsymbol{g}_t(\boldsymbol{\theta}_{t-1}) \tag{2.5}$$

其中α为学习率，在网络优化训练中，学习率大小主要影响网络的优化速度，小批量大小K主要影响梯度估计，即梯度方向的估计，过大的K虽然能得到较好的梯度方向估计，但是影响网络的优化效率，这时可以适当增加学习率，而过小的K则可能带来梯度方向估计过大的方差，从而影响网络收敛，可以减小学习率，实际应用中选择合适的学习率大小和批量大小也对网络优化起关键作用，因此学习率调整也存在由多种策略，例如分段常数衰减（Piecewise Constant Decay）、逆时衰减（Inverse Time Decay）、指数衰减（Exponential Decay）、自然指数衰减（Natural Exponential Decay）和余弦衰减（Cosine Decay）。

针对一般形式的梯度优化方法不足，研究者也提出了系列改进的方法，首先考虑到深度网络中大量参数的作用不同，在梯度更新过程中可能需要不同的学习率，因此Duchi 等人提出了 AdaGrad 算法[35]，类似于正则化，对于梯度较大的参数给予较小的学习率，如公式（2.6）所示，以每个参数梯度的累加平方和来衡量梯度大小，并由此来调整学习率，调整后的梯度更新如公式（2.6）和公式（2.7）所示。

$$G_t = \sum_{\tau=1}^{t} \boldsymbol{g}_\tau \odot \boldsymbol{g}_\tau \tag{2.6}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \frac{\alpha}{\sqrt{G_t + \varepsilon}} \boldsymbol{g}_t(\boldsymbol{\theta}_{t-1}) \tag{2.7}$$

其中ε是为了防止除零而设置的非常小的常数，AdaGrad 算法能动态调整学习率，对于累积梯度较大的参数给予较小的学习率，而对于累积梯度较小的参数给予较大的学习率，但由于累加恒正数，容易是网络学习率过早的调整非常小，导致网络提前停止训练，由此 Geoff Hinton 等人提出了改进方法 RMSprop 算法，如公式所示，以每次迭代梯度平方的指数衰减移动平均来调整学习率，则梯度的动态更新如公式（2.8）所示。

$$G_t = \beta G_{t-1} + (1-\beta)\boldsymbol{g}_t \odot \boldsymbol{g}_t \tag{2.8}$$

其中β为衰减率，一般取值为 0.9，以指数衰减移动平均来代替累积平方的方式，保证能够动态调整学习率同时，又避免了学习率下降过小导致网络停止学习的问题。

除了在学习率调整上改进优化算法，还可以在梯度跟新方向上进行改进，即进行梯

度估计（Gradient Estimation）修正，由于过小的批量大小会带来梯度方向较大的方差，即损失函数下降呈现震荡方式，由此模拟物理中动量（Momentum）的概念，在梯度跟新过程中加入之前时刻的梯度作为参考，如公式（2.9）所示，

$$\Delta\boldsymbol{\theta}_t = \rho\Delta\boldsymbol{\theta}_{t-1} - \alpha\boldsymbol{g}_t(\boldsymbol{\theta}_{t-1}) \tag{2.9}$$

其中ρ为动量因子，一般取值为0.9，由此结合动量的梯度跟新如公式（2.10）所示，

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha\Delta\boldsymbol{\theta}_t \tag{2.10}$$

引入动量进行梯度更新，使得每次梯度更新计算梯度不是仅取决于上一时刻的梯度，而是取决于在最近一段时间内梯度更新的加权平均值，这能有效减小梯度方向上的方差，提高网络学习的稳定性。

结合动量法和RMSprop算法，Kingma等人提出自适应动量估计（Adaptive Moment Estimation，Adam）算法[36]，在利用动量调整梯度跟新方向的同时，又能自适应调整学习率，其梯度更新策略如下公式（2.11-2.15）所示：

$$\boldsymbol{M}_t = \beta_1\boldsymbol{M}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \tag{2.11}$$

$$G_t = \beta_2 G_{t\text{-}1} + (1\text{-}\beta_2)\boldsymbol{g}_t \odot \boldsymbol{g}_t \tag{2.12}$$

$$\hat{\boldsymbol{M}}_t = \frac{\boldsymbol{M}_t}{1-\beta_1} \tag{2.13}$$

$$\hat{G}_t = \frac{G_t}{1-\beta_2} \tag{2.14}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \frac{\alpha}{\sqrt{\hat{G}_t + \varepsilon}}\hat{\boldsymbol{M}}_t \tag{2.15}$$

其中β_1和β_2分别为两个移动平均的衰减率，一般分别取值为0.9和0.99，而$\boldsymbol{M}_t$可以作为梯度的一阶矩，G_t可以作为梯度未减均值的方差，也即二阶矩。Adam优化算法在实际应用中具有出色的表现，很大程度减小网络收敛时间。

2.2 目标检测算法基础理论

近些年目标检测算法以基于卷积神经网络的方法为主导，可以大致分为两类：基于区域提案（Region Proposal）的双阶段方法和基于回归的单阶段方法。双阶段方法的主

要包括 RCNN 系列的方法，例如 Fast RCNN，Faster RCNN 和 Mask RCNN，由于双阶段方法的扩展性，也衍生了相应的多阶段方法，例如 Cascade RCNN[37]。而单阶段方法主要包括有早期的 SSD[38]和 YOLO 系列的方法，以及近期兴起的无锚框（Anchor Free）的方法，例如 FCOS[39]，CornerNet 和 CenterNet 等。对比而言双阶段方法算法复杂度较高，对比早期单阶段方法具有较大的检测性能优势，单阶段方法算法复杂度较低，可实现实时检测，在检测性能上，虽然早期的单阶段方法会劣于双阶段方法，但近期发展的单阶段方法的性能已经可以超越双阶段方法，例如当前检测最佳的方法就是 YOLOv4 和 CenterNet 的改进方法，因此，本可以的研究重点基于单阶段的检测方法 CenterNet 进行改进，下面将对目标基础方法的基础理论进行详细介绍。

2.2.1 交并比

交并比（Intersection over Union，IoU）目标检测中用于表示两个目标之间的交并比，即可以表示预测目标和真实目标之间的相关度，交并比越大表示相关度越高。如图 2-7 所示，传统目标检测中水平边框（Bounding Box，BBox）可以用边框的左上顶点和右下顶点坐标表示 $\boldsymbol{bbox}=(x_1,y_1,x_2,y_2)$,给定两个预测目标和真实目标的水平边框 $\hat{\boldsymbol{b}}=(\hat{x}_1,\hat{y}_1,\hat{x}_2,\hat{y}_2)$ 和 $\boldsymbol{b}=(x_1,y_1,x_2,y_2)$，其交并比的计算如公式（2.16）所示，可见 IoU 取值范围在 0 到 1 之间。

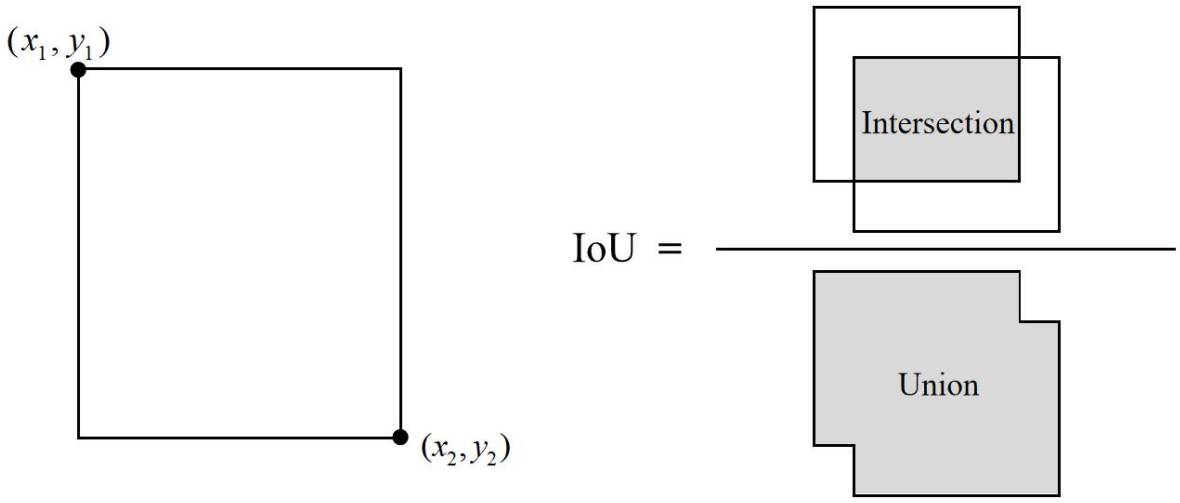

图 2-7 交并比示意图

Fig. 2-7 Schematic diagram of Intersection over Union

$$\begin{cases} inter = \max(0, \min(\hat{x}_2, x_2) - \max(\hat{x}_1, x_1)) \\ \qquad \times \max(0, \min(\hat{y}_2, y_2) - \max(\hat{y}_1, y_1)) \\ union = (\hat{y}_2 - \hat{y}_1) \times (\hat{x}_2 - \hat{x}_1) + (y_2 - y_1) \times (x_2 - x_1) - inter \\ IoU = \dfrac{inter}{union} \end{cases} \tag{2.16}$$

定向目标检测中，由于检测目标是旋转的，所以边框的表示无法通过两个点的坐标来表示，计算交并比的方式与水平边框不同。目标的旋转边框可以一般化为为凸多边形，因此定向边框（Oriented Bounding Box）是通过每个顶点的坐标位置表示，为了一般化计算，定义定向边框为$\boldsymbol{obbox} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_K)$，其中$K$为多边形的顶点总数，而$\boldsymbol{p}_k = (x, y)$为多边形第$k$个顶点坐标。为了计算凸多边形的重叠面积，首先需要计算凸多边形的面积，如图 2-8 所示，将K凸多边形以某一个点作为顶点，将K凸边形划分成K个三角形，则K凸变形的面积有K个三角形面积之和，根据平面几何知识可知，三角形$\triangle \boldsymbol{p}_1\boldsymbol{p}_2\boldsymbol{p}_3$的面积计算如公式（2.17）所示，因此$K$凸边形总面积计算如公式（2.18）所示。

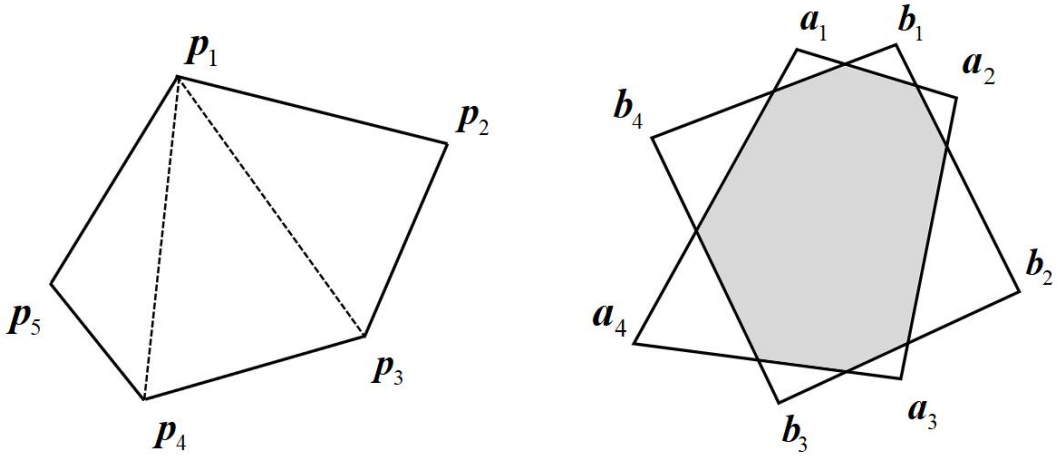

图 2-8 凸多边形重叠面积计算

Fig. 2-8 Convex polygon overlap area calculation

$$S_{\boldsymbol{p}_1\boldsymbol{p}_2\boldsymbol{p}_3} = \frac{1}{2} \bullet \left| \overrightarrow{\boldsymbol{p}_1\boldsymbol{p}_2} \times \overrightarrow{\boldsymbol{p}_1\boldsymbol{p}_3} \right| \tag{2.17}$$

$$S_{\boldsymbol{obbox}} = \frac{1}{2} \bullet \sum_{k=3}^{K} \left| \overrightarrow{\boldsymbol{p}_1\boldsymbol{p}_{k-1}} \times \overrightarrow{\boldsymbol{p}_1\boldsymbol{p}_k} \right| \tag{2.18}$$

凸多边形重叠面积计算相比于水平边框重叠面积计算要复杂，由于本课题研究的定向目标为任意四边形，所以为了直观理解，不妨假设两个四边形 $\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3,\boldsymbol{a}_4)$ 和 $\boldsymbol{B}=(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3,\boldsymbol{b}_4)$，如图所示，两个四边形相交得到阴影部分重叠面积，在此简要介绍求解步骤：1）利用角度和法，求解出 $\boldsymbol{A}$ 位于与 $\boldsymbol{B}$ 中的顶点集合 $\Phi_{\boldsymbol{AB}}$ 和 $\boldsymbol{B}$ 位于 $\boldsymbol{A}$ 中的顶点集合 $\Phi_{\boldsymbol{BA}}$；2）利用跨立法，求解出 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 每一条边的交点集合 $\Theta_{\boldsymbol{AB}}$；3）对集合 $\mathrm{P}=\{\Phi_{\boldsymbol{AB}},\Phi_{\boldsymbol{BA}},\Theta_{\boldsymbol{AB}}\}$ 去除重复点，即可得到重叠区域的顶点集合，由于求得顶点集合无序，因此需要先计算集合中所有顶点的重心，再根据重心与每个顶点的向量夹角来对顶点进行排序，得到有序的重叠区域顶点集合；4）根据公式算出重叠面积，然后再根据重叠面积和总面积求得定向目标的交并比。

算法 2-1 非极大值抑制算法流程图

Algorithm 2-1 Flow chart of non-maximum suppression algorithm

输入：目标候选框集合 $BBox=\{\boldsymbol{box}_1,\boldsymbol{box}_1,\cdots,\boldsymbol{box}_N\}$，$N$ 为候选目标边框总数；
目标候选框的置信度 $S=\{s_1,s_2,\cdots,s_N\}$；
置信度阈值 θ_s；
交并比阈值 θ_{IoU}；
输出：目标边框集合 $BBox^*=\{\boldsymbol{box}_1,\boldsymbol{box}_1,\cdots,\boldsymbol{box}_M\}$，$M$ 为目标边框总数；
1：初始化空集合 $BBox^*$
2：当 $BBox$ 不为空空集：
3： 选出 S 中最大置信度 s_* 与对应 $BBox$ 中的 $\boldsymbol{box}_*$
4： 如果 s_* 小于 θ_s：
5： 返回 $BBox^*$
6： 把 $\boldsymbol{box}_*$ 加入 $BBox^*$，同时从 $BBox$ 中移除 $\boldsymbol{box}_n$，从 S 中移除 s_*
7： 遍历 $BBox$ 中每个 $\boldsymbol{box}_n$：
8： 如果 $\boldsymbol{box}_n$ 与 $\boldsymbol{box}_*$ 交并比大于 θ_{IoU}：
9： 同时从 $BBox$ 中移除 $\boldsymbol{box}_n$，从 S 中移除 s_n
10：返回 $BBox^*$

2.2.2 非极大值抑制

非极大值抑制（Non Maximum Suppression，NMS）算法，顾名思义是对非最优结果进行抑制，传统基于锚框（Anchor）目标检测算法通常会在图像同个区域预测出多个边框，且每个边框都有各自的预测类别和置信度，为了选出最准确的预测边框，需要利用 NMS 算法将冗余的边框去除。NMS 算法在不同应用场合具有不同的表现形式，例如在基于锚框（Anchor）目标检测算法，NMS 同个计算 IoU 来判断是否抑制，最终

预测效果如图 2-9 所示，对于的算法流程如算法 2-1 所示。

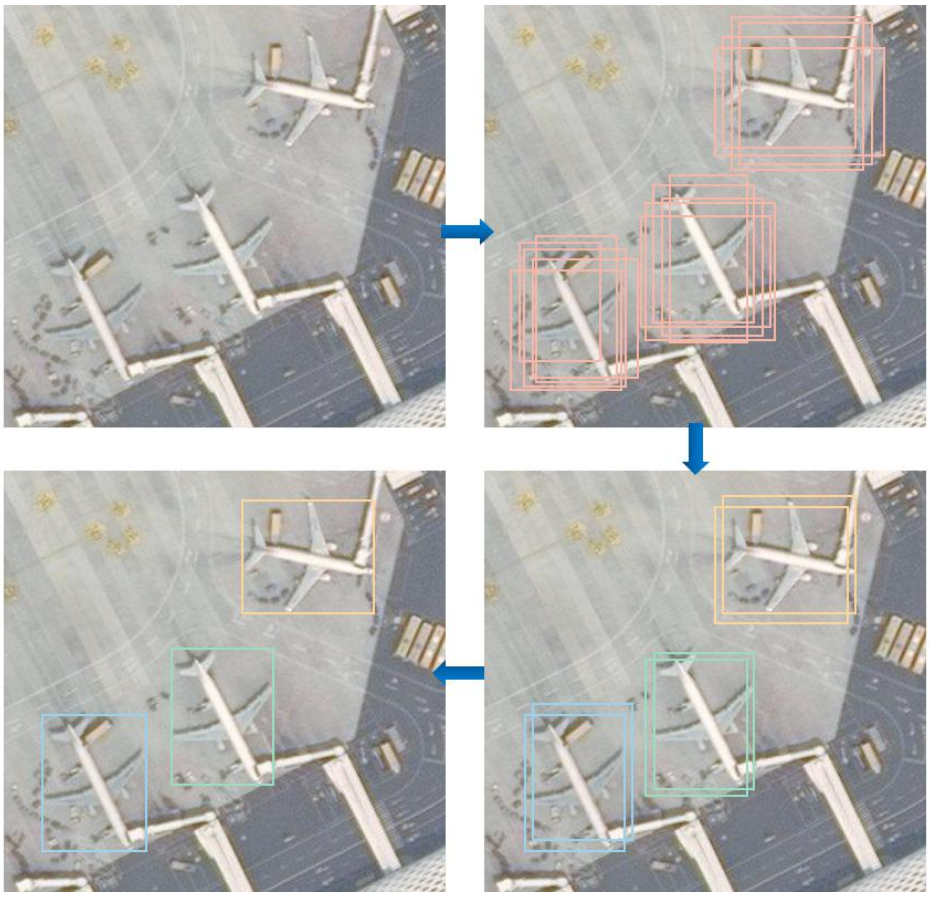

图 2-9 非极大值抑制的效果

Fig. 2-9 The effect of non-maximum suppression

对于无锚框（Anchor Free）目标检测方法中，NMS 算法表现为另一种形式。例如基于关键点的目标基础算法 CenterNet 中，其对目标检测是利用关键点检测的方法，输出为二维高斯热图（Heatmap）,热图上的任意局部极大值均认为一定概率存在目标，所以会导致在一个局部极大值周围存在许多干扰的局部极大值，在此，NMS 算法则可以直接通过最大值池化（Max Pooling）实现，本课题采用了 3×3 窗口大小对输出热图进行最大值池化。

2.2.3 评测指标

目标检测中常用的评测指标为平均精度（Average Precision，AP）和类别平均精度（mean Average Precision，mAP）。首先简要介绍在常用分类任务中的精确率（Precision）和召回率（Recall）,两者的计算如公式（2.19）和公式（2.20）所示。

$$Precision = \frac{TP}{TP + FP} \tag{2.19}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.20}$$

其中精确率为正确识别出来的样本占预测识别出来的总样本的比例，*TP* 为正确识别的正样本，而 *FP* 为被误判为正样本的负样本；召回率则定义为正确识别出来的样本占真实总样本的比例，*FN* 为被误判为负样本的正样本。在模型预测过程中，一般通过一个阈值将模型预测置信度划分的最终的分类结果，而不同阈值是选择则会影响到精确率和召回率，因此为了评估模型在不同阈值下的性能，引入了 PR（Precision-Recall, PR）曲线的，即在 0 到 1 的同阈值下，精确率（查准率）而后召回率（查全率）的关系曲线图，如图 2-10 所示，PR 曲线越往外表示模型性能越好，例如曲线 A 代表的模型性能要优于曲线 C 代表的模型性能，为了区分两曲线相交情形的性能，如曲线 A 和曲线 B，定义了平衡点的，即在两坐标轴对称位置是点来刻画模型性能，因此在平衡点处曲线 A 的性能要优于曲线 B。

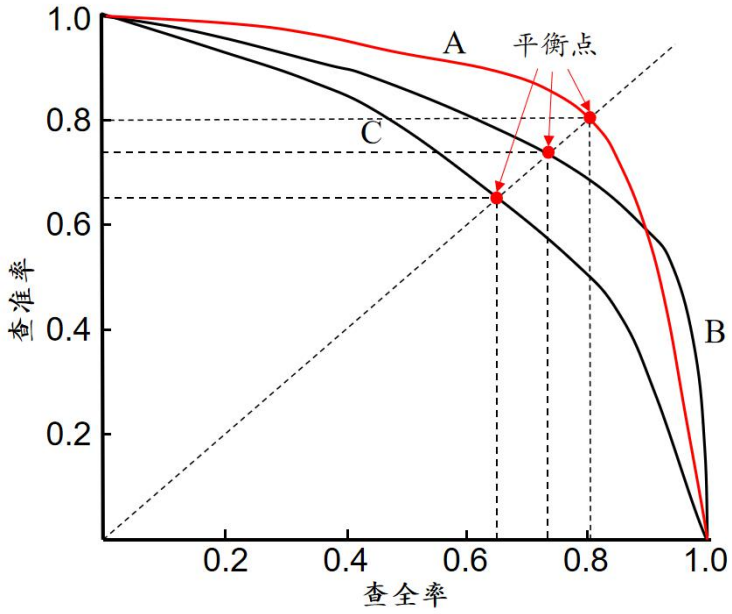

图 2-10 PR 曲线

Fig. 2-10 PR curve

平均精度（AP）即为 PR 曲线的下围面积，AP 的值越大代表模型性能越好，AP 的理想值为 1，即 PR 曲线包围整个坐标轴区域面积。而 mAP 是由于在检测任务中，除了需要正确检测出目标以外，同时还需要正确分出类别，因此目标检测中，会针对

每个类别进行计算平均精度，然后取多个类别的平均即为 mAP。

在通用目标检测任务中，有两个为学术界常用的评测基准，分别定义在两个大型数据集上，分别为 Pascal VOC 数据集[40]和 MSCOCO 数据集[41]。Pascal VOC 最新的一套标准为 2012 年定义的以检测目标和真实目标的 IoU 等于 0.5 来判断目标是否正确检测的标准。而对于更大是 MSCOCO 数据集而言，考虑了不同目标的尺度，因此定义了不同严格程度的评价标准，定了了 IoU 从 0.5 到 0.95 之间，不同 IoU 下的 AP 来评估模型性能，本课题的实验将采用以上两套标准来进行方法评估，具体将在实验章节进行阐述。

2.3 定向目标检测的基本方法

定向目标检测最为常用的两个领域为光学字符识别和航空影像目标检测，两者的共同点来源于图像的成像视角不局限于水平视角，而是更一般以俯视的成像视角出现，因此成像的目标不具有任何水平参考系，也导致了目标多个任意旋转的特点，因此定向目标检测除了完成定位与分类任务之外，还需要具备估计目标的旋转体态，即旋转角度的估计。相比而言，航空影像目标检测由于其复杂的背景以及多样化的目标，检测复杂度会高于纸张背景的字符检测，因此本课题将从航空影像目标检测领域展开研究，以此来设计相应的检测算法。

2.3.1 基于双阶段的方法

定向目标检测方法比较基础的方法就是在传统双阶段目标检测方法的基础上进行改进，Ding 等人在 Faster RCNN 基础上提出 RoI Transformer 来回归水平 RoI 和旋转 RoI 的偏移，如图 2-11 所示，在区域提案网络新增一个额外的旋转 RoI 学习分支，用于将原始的水平提案转成旋转提案框，从而使后面的检测和分类分支能够从旋转区域获取信息。检测流程具体如下：首先水平 RoI 经过一个 10 通道的卷积网络，减少特征维度，然后通过一个全连接网络将特征合并到 10 通道的特征，在经过解码器得到水平 RoI 和旋转 RoI 的偏移量，最后根据解码后的偏移量从原始特征进行旋转特征抽取（Wrapping）,利用此特征进行目标的检测和分类。

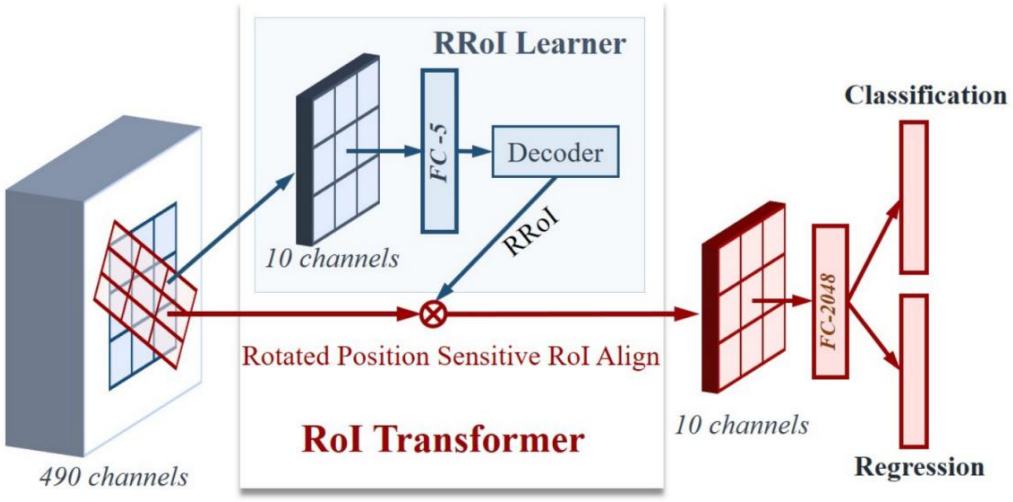

图 2-11 RoI 变换器[19]

Fig. 2-11 RoI Transformer

RoI Transformer 设计创新点主要在于 RRoI Learner 的应用，它回归旋转边框和水平边框的相对偏移，从而不需要增加额外的旋转锚框数量，一定程度上减少了计算量，同时利用 RRoI Warping，类似于 Faster RCNN 的 RoI Pooling 操作，可实现特征计算过程可微，即能实现端到端优化。优化目标如公式（2.21）所示，其中 $(x_r, y_r, w_r, h_r, \theta_r)$ 表示旋转的 RRoI，而 $(x^*, y^*, w^*, h^*, \theta^*)$ 表示旋转边框。

$$\begin{cases} t_x^* = \dfrac{1}{w_r}\left(\left(x^* - x_r\right)\cos\theta_r + \left(y^* - y_r\right)\sin\theta_r\right), \\ t_y^* = \dfrac{1}{h_r}\left(\left(y^* - y_r\right)\cos\theta_r - \left(x^* - x_r\right)\sin\theta_r\right), \\ t_w^* = \log\dfrac{w^*}{w_r}, \\ t_h^* = \log\dfrac{h^*}{h_r}, \\ t_\theta^* = \dfrac{1}{2\pi}\left(\left(\theta^* - \theta_r\right)\bmod 2\pi\right), \end{cases} \tag{2.21}$$

2.3.2 基于单阶段的方法

单阶段的定向目标检测算法最直接的方法就是根据传统单阶段检测网络的基础上增加额外的角度变量，通过直接回归角度来实现定向目标检测。如 Pan 等人提出的图动态改进网络（DRN）,在基于关键点的单阶段检测网络 CenterNet 基础上增加一个额

外的角度回归分支来预测旋转目标，如图 2-12 所示在原始 CenterNet 的基础上再预测分支处新增一个角度回归分支，用于角度预测，如公式（2.22）所示。作者为了解决由于神经元感受野轴向排列的单一结构，提出了动态精细化网络，包括了特征选择模块（Feature Selection Module, FSM）和动态优化分支（Dynamic Refinement Head, DRH），其中特征选择模块类似于可变卷积（Deformable Convolution Network, DCN）能够根据目标的几何形状和方向位置动态地调整神经元的感受野，使其捕获更具判别性的特征，而动态优化分支则能使模块以目标感知的方式动态地实现优化预测，虽然测方法简单易实现，但是角度会因为周期性产生角度旋转敏感度误差（RSE）因此，该方法性能有所局限。本课题也将针对次问题进行相应的设计与改进，利用姿态图检测方式解决旋转敏感度误差。

$$L_{ang}=\frac{1}{N}\sum_{k=1}^{N}\left|\theta-\hat{\theta}\right| \tag{2.22}$$

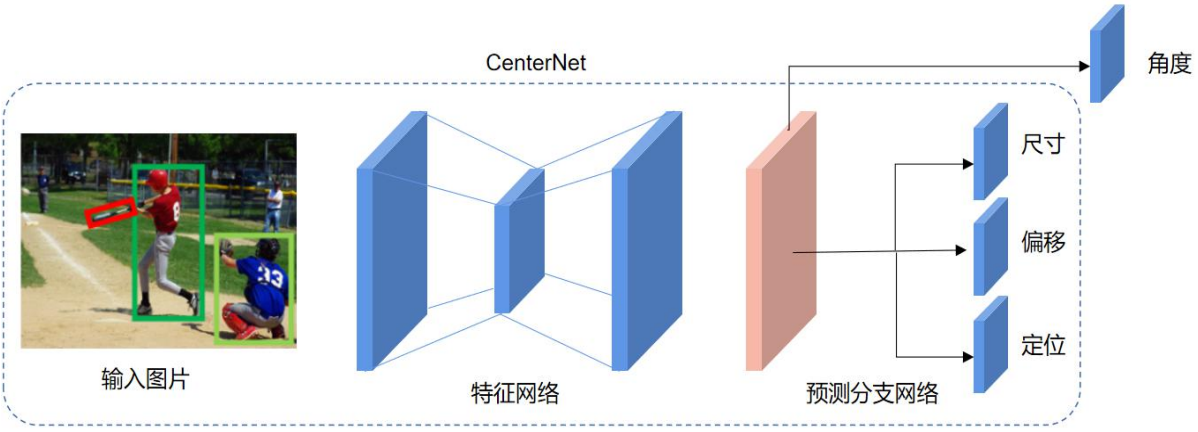

图 2-12 动态改进网络

Fig. 2-12 Dynamic Refinement Network

2.3.3 定向目标表示问题

如绪论中所述，在一些特定场景（航空影像），多变的视角使得目标呈现拥挤，聚集和旋转等特点，如图 2-13 所示，传统的水平边框难以契合的表示出目标的几何位置，特别是对于倾斜拥挤的目标，水平边框会使边框内覆盖过多的背景区域，从而导致目标之间相互耦合，不利于检测位置的表示。相比而言，带额外旋转角度的旋转边框则能够更好的切合这种拥挤且倾斜的目标。因此，在航空影像检测任务中，无论是特定

的遥感船只检测还是遥感通用目标检测，习惯采用旋转的边框来表示目标。当然旋转边框的表示相比水平边框的表示增加的检测难度，因此检测定位任务中，除了需要定位和尺度回归之外，还需要得到一个额外的角度信息。一般来说定向目标的表示方法有两种，一种是直接增加额外的旋转角度变量$bbox=(x,y,w,h,\theta)$，这种方法简单直接，但是会带来前文所述的 RSE 问题，另一种方法则是对角度变量进行解耦，如公式（2.23）所示，利用高维变量(t_1,t_2)来表示一维的角度周期变量θ，从而实现调度周期临界点平滑。下面章节也将针对定向目标的表示问题进行详细分析，并设计更为简单高效的定向目标表示方法。

$$\begin{cases} bbox=(x,y,w,h,t_1,t_2) \\ t_1=\sin\theta \\ t_2=\sin 2\theta \end{cases} \tag{2.23}$$

图 2-13 水平边框与旋转边框

Fig. 2-13 Horizontal bounding box and oriented bounding box

2.4 本章小结

本章从目标检测的基础理论到目标检测的实际应用问题展开了介绍，重点介绍有基于深度学习的目标检测方法基础理论，包括有基础的卷积网络，深度神经网络与其相应的优化方法。针对目标检测以及定向目标检测特有的基础概念，如交并比，非极大

值抑制和常用的评测指标进行了总结。最后针对定下目标检测任务的一般方法和难点展开讨论，并引出本课题研究的关键点。

第三章　基于关键点的定向目标检测

本章将重点介绍基于关键点方法的定向目标检测方法，首先介绍基于关键点检测方法的优势即相应的设计思路，然后针对定向目标检测的特点设计了利用关键点表示的定向目标姿态表示方法，最后针对目标多尺度问题对网络结构进行了相应的改进，提升模型检测性能。

3.1 基于关键点的目标检测

基于关键点的目标检测方法是目标检测领域较新的方向，代表性的方法包括有CornerNet 和 CenterNet。该方法能够使用单阶段的网络达到高于双阶网络的方法检测性能，使得更容易实现实时高性能的检测。传统基于锚框（Anchor）的方法依赖于对锚框的设计，锚框为待检测的目标提高了位置和尺度先验，锚框设计不合理，会严重影响检测性能，同时初始化大量锚框会增加模型的计算量，例如对于图像的一个位置，需要设计不同尺度和宽长比的锚框，特别是在定向目标检测中，如果需要额外增加角度回归，更需要设计不同角度的锚框。对比而言，基于关键点的方法设计只需要图像预测一个二维热图（Heatmap），实现对每个目标的定位，然后再相应位置进行尺度回归和类别分类，极大程度的简化的检测的流程。在当前通用目标检测任务上，基于关键点检测的单阶段方法目前实现最佳性能。

3.1.1 现有的方法

当前基于关键点检测方法较为经典的为 ConnerNet 和 CenterNet，ConnerNet 首次在目标检测领域提出利用关键点检测方法来实现目标检测，如图 3-1 所示，利用 Hourglass 网络对图像进行特征提取，然后分为左上角和右下角两个预测分支，分别预测目标边框的左上和右下两个顶点的位置，从而实现目标边框的预测。其中每个预测分支又划分为三个子分支，分别为二维热图预测，匹配编码预测和量化偏移预测，其中二维热图预测用以定位关键点位置，而匹配编码预测分支用于预测匹配左上和右下两个顶点的匹配，即如果左上和右下两个为同一个目标边框的两个顶点，则两个分支定的匹配编码应该预测出相同的向量，最后的量化偏移预测用于预测因网络下采样造成的量化误差偏移。ConnerNet 方法的另一个创新点为提出了 Corner Pooling 操作，能将目标边框内的激活值巧妙地映射到边框的顶点处，从而解决了因为顶点位置没有目标关联而

导致性能损失问题。

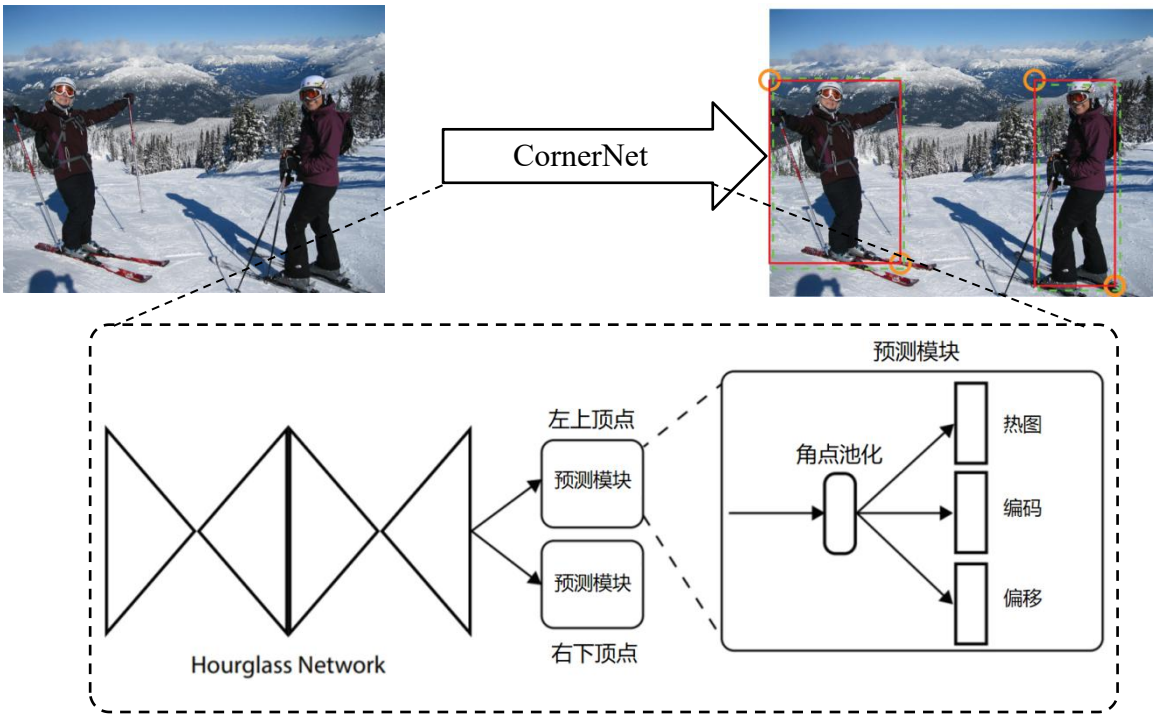

图 3-1 CornerNet 网络结构图

Fig. 3-1 The overview of CornerNet

在 CornerNet 基础上，如图 3-2 所示，CenterNet 提出更为简洁的方法，去除了左上右下两个顶点的匹配预测，直接只对目标的中心的进行定位，然后进行目标边框的尺寸预测以及量化偏移误差的回归。用一中心点来取代两个顶点的预测能有效避免因两个顶点错误匹配而导致的错误边框的出现，其次，在中心点处进行边框尺寸（宽和高）预测，同样能实现目标的无锚框检测，因此基于中心点的目标检测更为简洁实用，在检测性能上也优于 CornerNet。此外，CenterNet 的方法还具有很好的任务迁移能力，如果将其中的尺寸回归替换成其他回归目标，例如人体姿态关键点坐标，那么则可以实现人体姿态估计的任务，同样若将二维的尺寸回归替换成三维尺寸，则可实现相应的三维目标检测任务，本课题也受此激发，将二维的尺寸回归替换成旋转目标的顶点坐标，构成旋转目标的姿态图预测，从而实现旋转目标检测任务。下文将重点介绍基于 CenterNet 改进的定向目标检测方法。

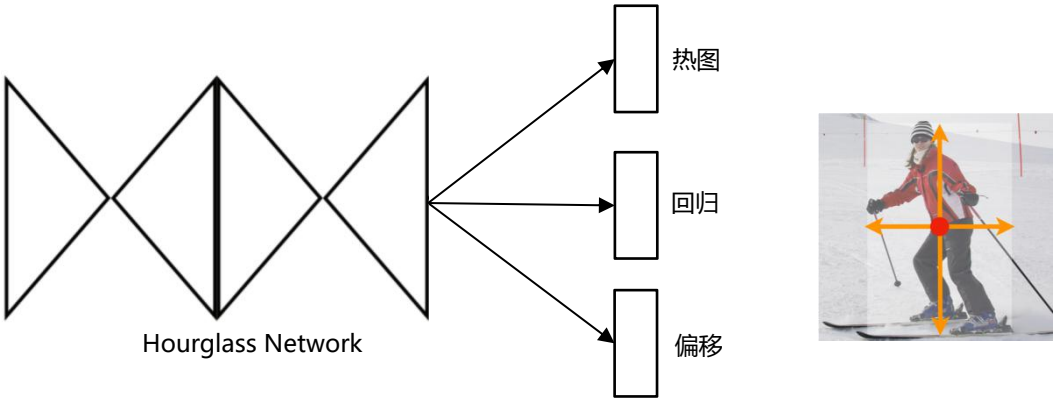

图 3-2 CenterNet 网络结构图

Fig. 3-2 The overview of CenterNer

3.1.2 改进的方法

本课题采样的基于关键点检测方法改进于 CenterNet 的结构，首先，考虑保留输入高分辨率的图像的高分辨率特征表示，如图 3-3 所示，本课题将利用 HRNet 多路并行的高分辨率分支网络对图像进行多尺度特征提取，HRNet 最初设计是用于姿态估计的特征提取网络，区别于传统特征提取网络（VGG，ResNet 等），该网络能够保留图像的高分辨率特征表示，采用递进方式，从高分辨率的子网络逐步增加高分辨率到低分辨率的自网络，并在多个不同分辨率特征表示中通过上下采样进行特征融合，使得特征网络能够获取更丰富多尺度特征，保持特征的高分辨率表示也同时避免了因反复上下采样造成的特征空间信息的损失。

HRNet 提取了多尺度的特征之后，本课题设计了自适应融合的特征金字塔网络，利用可学习的权重参数自底向上对高层语义特征不断进行加权融合，得到了高分辨率的特征表示，这将有效提高下流网络对目标中心的定位的精度。最后两个分支网络是本文提出的姿态表示的旋转目标检测网络，上分支用于目标中心的定位，下分支根据上分支定位中心进行回归顶点偏移，从而实现旋转目标的检测，后续章节 3.2.2 将对以上两个分支网络设计进行详细介绍，并将在第四章进行实验分析，分别用实验证明每个改进点对检测性能的提升。

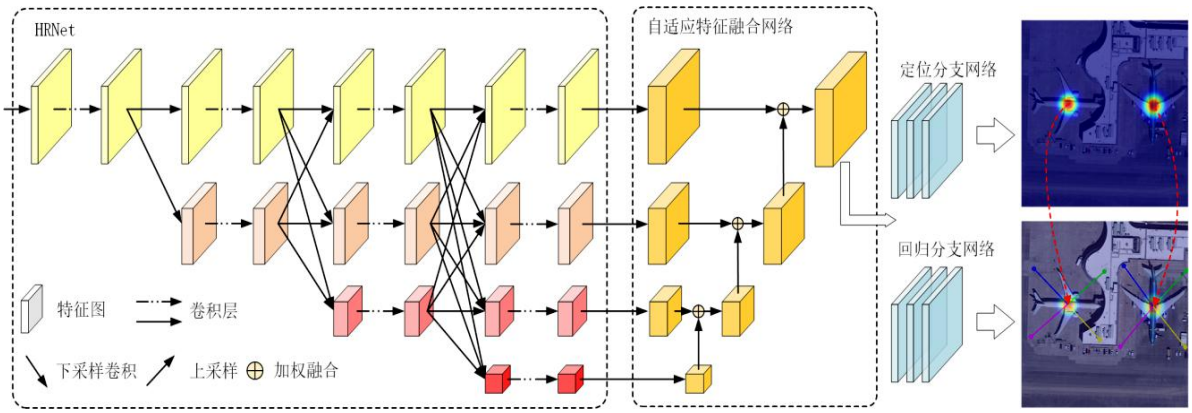

图 3-3 网络整体结构

Fig. 3-3 The overview of proposed network

3.2 定向目标的表示方法

在定向目标检测方法中最为常用的目标表示方法为边框表示，即在传统的水平边框基础上增加一个额外的角度变量，如图 3-4 所示，$\boldsymbol{bbox}=(x_1,y_1,x_2,y_2,\theta)$，其中$\theta$为长边于水平线之间的夹角。基于边框的表示方法一般通过改进传统的基于锚框的检测器，如 Faster RCNN。本课题提出基于关键点的目标表示方法，亦称为姿态表示方法，通过目标的顶点和其重心点来表示目标的位置和几何轮廓，下面将分别讨论两者的特点。

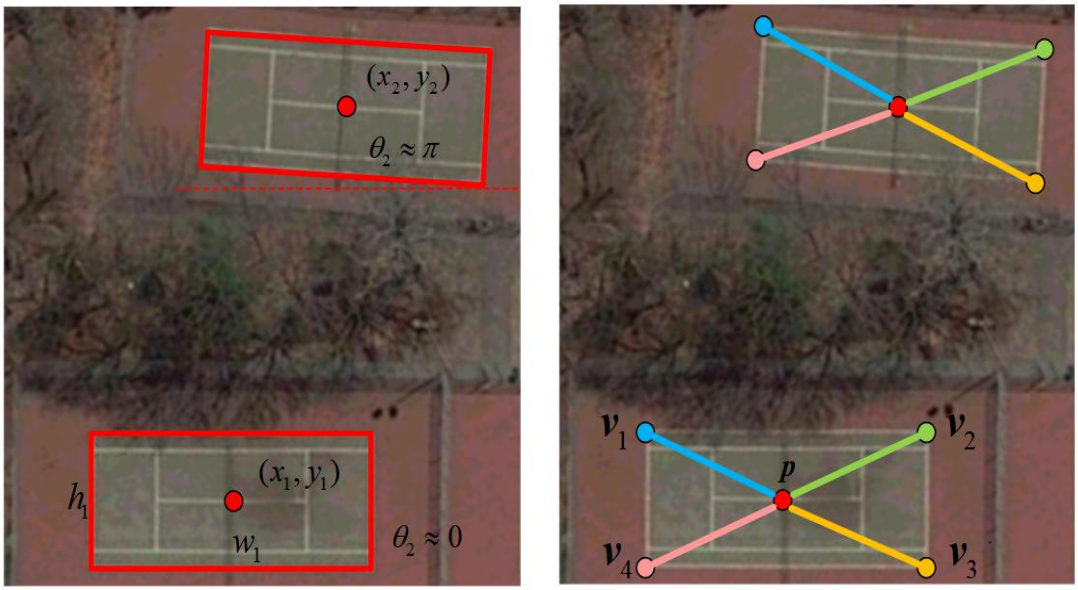

图 3-4 旋转目标的表示方法

Fig. 3-4 Representation of oriented object

3.2.1 基于边框表示的特点

基于边框的表示方法关键点是在传统的水平边框基础上解决角度变量预测的问题，因此许多方法直接在传统的基于锚框的方法上进行改进，增加额外的角度回归分支，

来预测角度变量。然而如分析，这种表示方法会带来旋转敏感度误差[22]（Rotation Sensitivity Error，RSE），如图 3-4 所示，由于角度为周期性变量，角度θ在 0 和π位置具有相似是几何外观，然后需要回归的目标值却不一致，这就导致在模型预测过程中在周期零界点出出现混淆问题，会干扰模型的学习。

为了解决此方法带来的角度敏感度误差问题，一种可行的方法为将一维度的角度变量进行分解，通过在多维空间实现角度平滑表示，如图 3-5 所示，将一维的角度分解成二维分量t_1和t_2，原角度θ可以通过公式（3.1）和公式（3.2）计算得到，其中$\lceil \bullet \rceil$为向上取整。通过角度的分解实现在 0 和π两个位置具有相同的表示t_1和t_2，即t_1和t_2在 0 和π两个位置均为 0，实现了角度的平滑表示，解决了 RSE 问题。虽然通过角度分解能解决 RSE 问题，然而，分解的变量与图像目标的几何外观没有直接的语义关系，深度网络的学习效果欠佳。

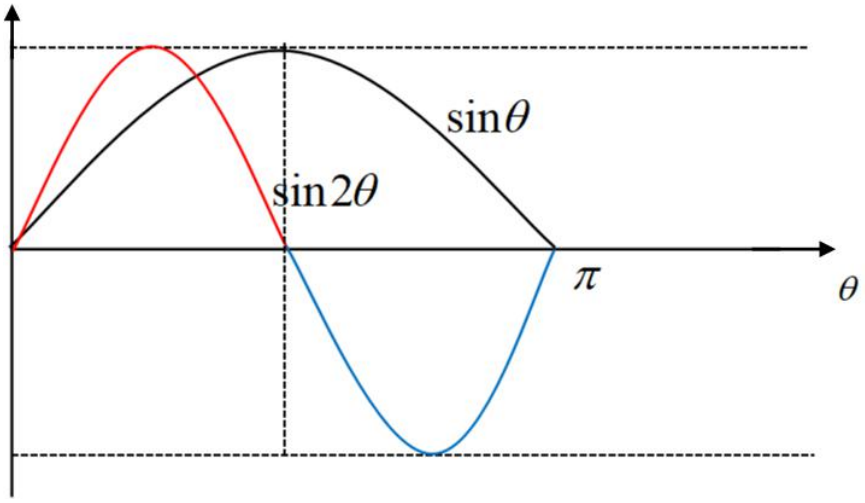

图 3-5 旋转角度解耦合

Fig. 3-5 Rotation angle decoupling

$$\begin{cases} t_1 = \sin\theta \\ t_2 = \sin 2\theta \end{cases} \tag{3.1}$$

$$\theta = \arcsin t_1 + \frac{\lceil -t_1 \rceil \times \pi}{2} \tag{3.2}$$

3.2.2 基于姿态表示的特点

本课题提出基于姿态表示的方法，通过利用绝对的中心的坐标和相对的顶点坐标来构造目标的姿态图。如图 3-4 所示，给定任意旋转目标的顶点绝对像素坐标

$\{(x_i,y_i)|i=1,2,\cdots,K\}$，当目标为任意四边形时 $K=4$，为了统一表示旋转目标的绝对位置坐标，在此额外计算目标的重心点，如公式（3.3）所示，通过外接水平矩形边框，来计算目标的重心点，并以此作为每个旋转目标顶点的参考坐标系原点。

$$\boldsymbol{p}=\left(\frac{\min\limits_i x_i+\max\limits_i x_i}{2},\frac{\min\limits_i y_i+\max\limits_i y_i}{2}\right) \tag{3.3}$$

因此，如图所示，根据重心点坐标，可以得到每个顶点的相对坐标，旋转目标的姿态表示为 $\boldsymbol{pose}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{v}_3,\boldsymbol{v}_4,\boldsymbol{p})$，其中 $\boldsymbol{v}_i=(x_i-p_x,y_i-p_y)$。显然，通过旋转目标的姿态表示能避开周期性角度变量带来 RSE 问题，同时回归的变量与图像目标的几何外观保留有较强的语义关系，有利于模型学习。

因此本课题采样了基于关键点的姿态表示来表示定向目标，并设计了基于关键点的定向目标检测网络，如图 3-3 所示，图像经过特征提取网络之后，分成两个分支，上分支用于目标中心的定位，下分支根据上分支定位中心进行回归顶点偏移，从而实现旋转目标的检测，检测流程如图 3-6 所示。前者输出目标每个类别的定位预测热图（Heatmap）$\hat{\boldsymbol{Y}}\in[0,1]^{W\times H\times C}$ 其中 W,H,C 分别表示热图的宽，高和类别数量，$Y_{xyc}=1$ 表示对应位置（x,y,c）为目标中心，其中 $0<x<W,0<y<H,0<c<C$，而 $Y_{xyc}=0$ 表示对应位置为背景像素。定位分支网络由 3 层 3×3 卷积网络构成，每层网络后接着批归一化层（Batch Normalization，BN）和 ReLu 激活函数，最后一层卷积连接 Sigmod 激活函数，将预测热图归一化到 1。定位分支网络利用训练标签得到标签热图 $\boldsymbol{Y}\in[0,1]^{W\times H\times C}$，用于监督网络训练，使定位分支网络实现中心的定位预测，其中标签热图中的每个位置的值有二维高斯图和标签中心点计算得到，如公式（3.4）所示。

$$Y_{xyc}=\exp\left(-\frac{(x-p_x)^2+(y-p_y)^2}{2\sigma_p^2}\right) \tag{3.4}$$

$$\boldsymbol{p}=\left\lfloor\frac{\boldsymbol{p}}{d}\right\rfloor \tag{3.5}$$

其中 σ_p^2 是与目标尺度正相关的标准差，目标尺寸越大，其值越大，同时由于图像的特征表示相对于原图上存在下采样率 d，因此计算过程需要对原始像素坐标进行相应缩放。最后利用改进的 Focal Loss 进行分支网络的训练，代价函数计算公式（3.6）所示。

$$L_{ct}=-\frac{1}{N}\sum_{xyc}\delta(Y_{xyc})^{\alpha}\left(1-\delta(\hat{Y}_{xyc})^{\gamma}\right)\log(\hat{Y}_{xyc}) \tag{3.6}$$

$$\delta(x)=\begin{cases}x, & x=1\\ 1-x, & otherwise\end{cases} \tag{3.7}$$

其中γ为 Focal Loss 的超参数，超参数α用于控制中心点附近惩罚力度，N为总目标数量，本课题实验采用和 CenterNet 相同的设置$\alpha=4,\gamma=2$。

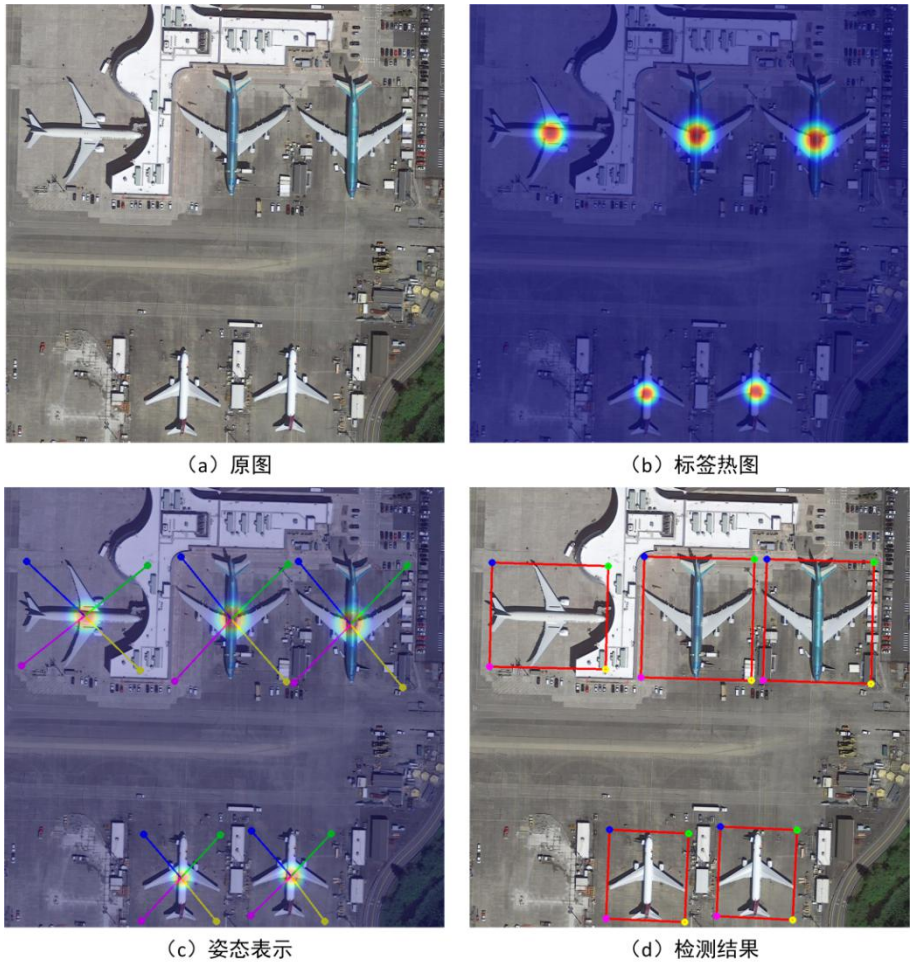

图 3-6 检测流程分析图

Fig. 3-6 Analysis diagram of detection process

偏移回归分支网络通过回归顶点相对于中心点的偏移来实现目标姿态检测。为了避免不同类别目标出现重叠的中心点区域，本课题采用对每个类别输出一个回归预测通道，回归分支网络同样采用了 3 层3×3 卷积网络，不同的是最终输出层连接的激活函数为 ReLu，保证回归变量为正实数。网络回归预测输出表示为$\hat{\boldsymbol{R}}\in\mathbb{R}^{W\times H\times 2(k+1)C}$，其中$k$为顶点个数，额外加 1 是为了进一步回归中心点位置因为网络下采样带来的量化误差。

为了方便表示，将每个回归的偏移向量表示为$\boldsymbol{O}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{v}_3,\boldsymbol{v}_4,\boldsymbol{o}_p)$，其中量化偏移误差如公式（3.8）所示，最后此分支网络采用了 Smooth L1 Loss 来计算网络回归部分的代价函数，如公式（3.9）所示。

$$\boldsymbol{o}_p=\frac{\boldsymbol{p}}{d}-\left\lfloor\frac{\boldsymbol{p}}{d}\right\rfloor \tag{3.8}$$

$$L_{rg}=\frac{1}{N}\sum_{n=1}^{N}\mathrm{SmothL1}\left(\hat{\boldsymbol{R}}_n,\boldsymbol{O}_n\right) \tag{3.9}$$

最终本课题将 2 个网络的代价函数进行加权求和得到整个网络的代价函数，如公式（3.10）所示，

$$L=L_{ct}+\lambda L_{rg} \tag{3.10}$$

其中λ为回归分支网络代价函数的权重超参数，在本课题实验中$\lambda=0.1$。

3.3 自适应特征融合网络

3.3.1 特征金字塔网络

特征金字塔网络[42]（Feature Pyramid Network，FPN）通过融合多尺度特征来解决目标检测中多尺度问题。物体在图像中的成像会因距离远近出现不同尺度问题，在深度卷积神经网络中，在一定深度的网络特征具有一定大小的感受野（Receptive Field），感受野的大小不仅与网络深度有关，还与卷积核的大小和卷积步长有关，不同深度的特征具有不同感受野大小的信息，例如，对于大尺度目标，需要较大感受野特征，因此深层的网络特征更能学习大尺度目标的特征，对于小尺度目标，较小的感受野就能满足小目标的特征学习。同时在深度卷积网络中，由于不同深度网络感受野不一样，能学习的不同层次的特征，浅层网络保留较多的空间信息，学习到的更高是细节的纹理特征，而深层网络空间分辨率低，特征宽度大，学习到的更多的是高层次语义特征。对于目标检测任务而言，既需要对目标定位和尺度预测，也需要对目标的类别进行预测，预测对于细节的纹理特征和高层次的语义特征。

针对传统深度卷积网络只在深层次的网络特征进行最终预测，这使得网络能以学习到较为丰富的低层次纹理特征，因此对小目标检测并不友好。如图 3-7 所示，特征金字塔网络提出将深层网络特征进行上采样与较为浅层网络进行相加融合操作，通过侧边连接使得每层特征都融合了不同层次的纹理特征和语义特征，从而实现对不同尺度

目标的友好检测，同时可以发现 FPN 在原有网络上增加侧边连接就实现了不同特征的融合，无需重新进行特性提取，在实际应用中减小了计算量。

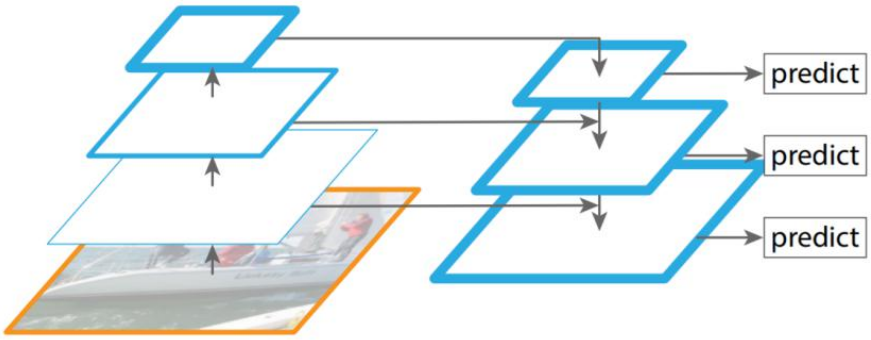

图 3-7 特征金字塔网络[42]

Fig. 3-7 Feature pyramid network

3.3.2 多尺度特征自适应融合

本课题在研究中发现，网络最终的预测特征均匀融合和了各个层次的特征，虽然这样的特征能够很好的应对多尺度目标检测问题，但对于是否能让网络在学习过程中自动地学习哪些层次的特征能提供更丰富的信息。因此区别于传统特征金字塔网络对不同层次特征进行一致性相加融合，本课题提出假设，认为不同层次的特征对最终的融合特征具有不同贡献度，采用数据驱动的方式，使网络在优化过程中利用可学习的权重参数 $\alpha_i \in [0,1]$ 动态的学习不同层次特征的重要性，如图 3-1 所示，对于不同层次的特征 $\{\boldsymbol{F}_i | i=1,2,\cdots,K\}$，其中 HRNet 中 $K=4$，自底向上对不同层次特征进行加权融合，融合策略如公式（3.11）所示。

$$\boldsymbol{M}_i = \mathrm{Conv}\left(\alpha_i \cdot \boldsymbol{F}_i + (1-\alpha_i)\cdot U(\boldsymbol{M}_{i-1})\right) \tag{3.11}$$

其中 $U(\bullet)$ 表示双线性插值的上采样操作，$\boldsymbol{M}_i$ 为第 i 个融合特征，$\mathrm{Conv}(\bullet)$ 为 1×1 的卷积操作。值得注意的是，当可学习的权重设置为常数 0.5 时，本课题提出的自适应特征金字塔网络可以退化为一般形式的特征金字塔网络，实验章节将对此方法进行对比实验验证。

3.4 本章小结

本章通过对基于关键点的目标检测方法进行改进，设计了基于关键点单阶段的定向目标检测网络。首先针对预测网络需求采用了高分辨率特征表示的特征提取网络，结合改进了多尺度自适应特征融合网络，为最终的定位为分支网络和回归分支网络提供

更丰富的信息。此外详细分析了定向目标基于边框表示方法和基于关键点表示方法的特点，提出定向目标的姿态表示方法，利用绝对的中心的坐标和相对的顶点坐标来构造定向目标的姿态图，巧妙避开了 RSE 问题，实现了定向目标检测。

第四章　实验设计与结果分析

本章将对课题的研究内容进行实验验证，为了充分验证本课题所提出的定向目标检测方法的性能，本章节采用两个大型的航空影像 DOTA 数据集[43]和 VisDrone 数据集[44]进行实验，同时进行与现有方法进行对比分析，实验结果也充分体现的本课题所提出方法的有效性。

4.1 实验数据集介绍

4.1.1 DOTA 数据集

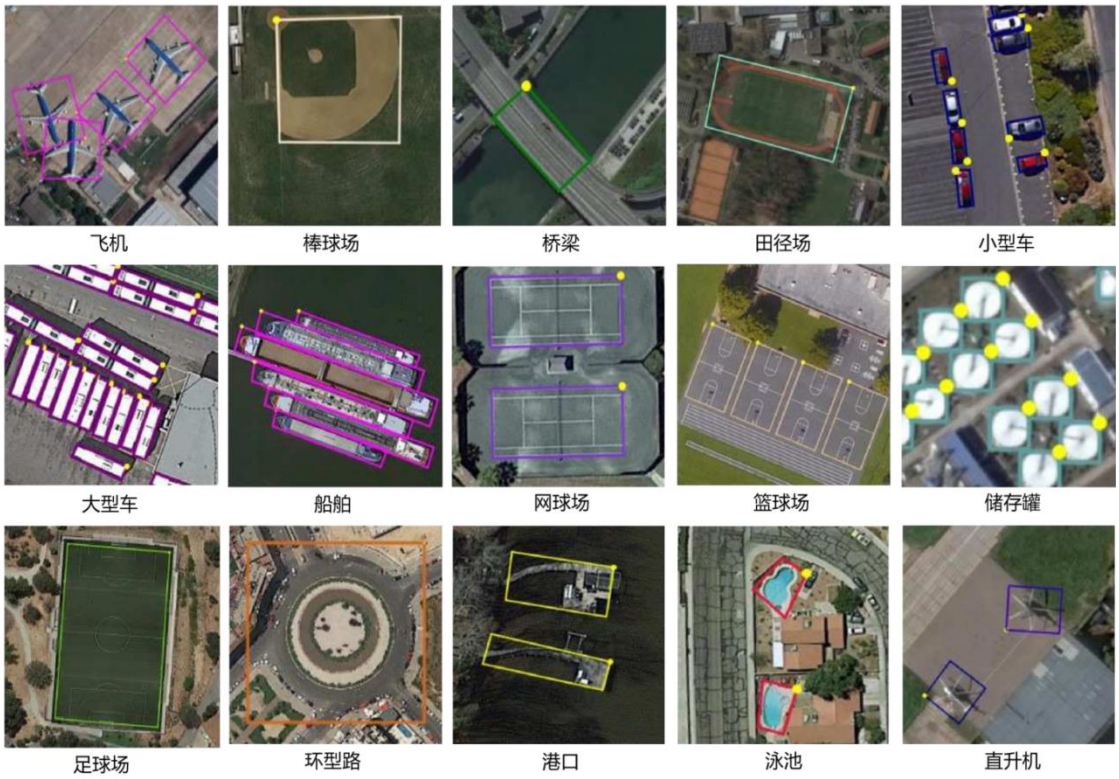

图 4-1 DOTA 数据集

Fig. 4-1 DOTA database

DOTA 数据集是一个大型的遥感图像目标检测数据集，如图 4-1 所示，其主要任务为实现对遥感图像中的目标进行水平和定向检测，本课题将就其中的定向检测任务（检测旋转边框）进行实验。数据集总共包括有 2 806 张高分辨率遥感图像（训练集有 1 409，验证集 548，测试集有 942），高分辨率最大可达 5 000×12 000，如图 4-2 所示。数据集有 15 个目标类别（类别简称对应：飞机-PL，棒球场-BD，桥梁-BR，田径场-GTF，小型车-SV，大型车-LV，船舶-SH，网球场-TC，篮球场-BC，存储罐-ST，足球场-SBF，

环形路-RA，港口-HA，泳池-SP,直升机-HC），总共有 188 282 个标注实例，标注实例微小拥挤极具挑战性，类别统计如图 4-3 所示，每个标注实例均为旋转的四边形，然而之前基于边框表示的方法均为假设四边形为矩形，这个近似假设也会带来精度损失。数据集的实例标注格式为四个顶点的绝对像素坐标，且按顺时针顺序。另外数据集的度量标准是采用经典通用目标检测数据集 PASCAL VOC 一样的度量标准。

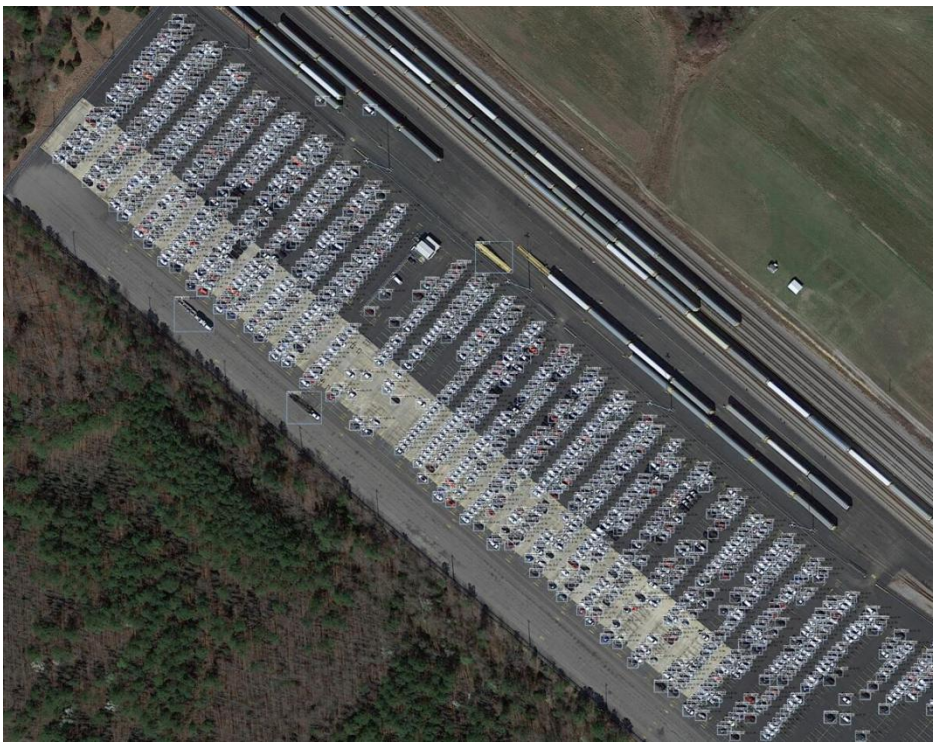

图 4-2 超高分辨率的遥感影像示意图

Fig. 4-2 Schematic diagram of ultra-high resolution remote sensing image

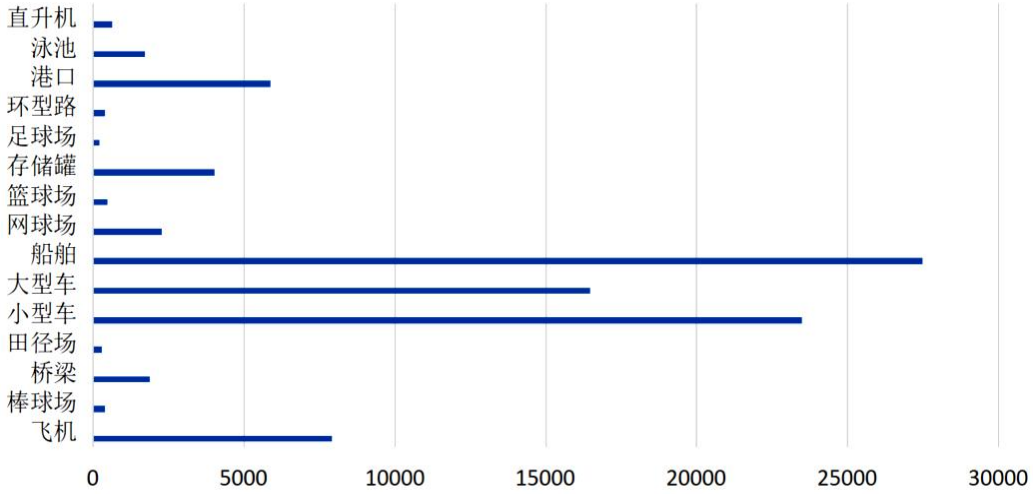

图 4-3 DOTA 数据集类别统计直方图

Fig.4-3 DOTA data set category statistics histogram

4.1.2 VisDrone 数据集

图 4-4 VisDrone 数据集

Fig. 4-4 VisDrone database

VisDrone 数据集由中国天津大学机器学习和数据挖掘实验室的 AISKYEYE 团队收集的一个大型的无人机航拍数据集，如图 4-4 所示，该数据集由各种安装在无人机上的摄像头捕获，涵盖了广泛的方面，包括位置（从中国相距数千公里的 14 个不同城市中拍摄），环境（城市和乡村），场景（白天和黑夜），物体（行人，车辆，自行车等），总共有 10 209 张航拍高清图像（训练集 6 471，验证集 548，测试集 3 190）航拍图像目标高度拥挤和密集，总共包括有约 46 万标注实例，甚至超过通用目标检测基准数据集 MS COCO 数据集。标注类别包括有行人车辆等 10 类。度量标准采用了 MS COCO 数据集计算 mAP 的标准，同时计算不同重叠阈值下的精度 AP50 和 AP75。该数据集也曾在欧洲计算机视觉会议（ECCV）2018 和 IEEE 国际计算机视觉会议（ICCV）2019 上举办了挑战赛，且有许多研究者在此数据集进行实验，因此本课题也即进行相关方法的性能比较。

4.2 实验的实现细节

4.2.1 训练策略

本课题实验的实现配置包括 Python 编程语言、PyTorch 深度学习框架，单块

NVIDIA Tesla V100 32 GB GPU。对于采样区域尺寸，DOTA（VisDrone）数据集图像裁剪成 1 024×1 024(1 024×768)的图像块，网络训练和测试时，为了减小计算量进一步下采样到 768×768(1 024×768)大小。由于测试集没有标签，所以测试本文采用了 512 步长的滑动窗口均匀采样的策略，最后将每个图像块的检测结果合并到原来图像上。网络训练使用的数据增强包括随机裁剪、随机翻转、随机旋转以及随机对比度增强，图像测试仅实用了单尺度图像进行测试，并以滑动窗口形式进行获取图像块。本文骨架网络 HRNet 加载了 ImageNet 预训练的权重，优化器选择了 Adam 优化器，且总共迭代了 8 万次，学习率开始设置为 1e-4 然后再 4 万次迭代之后下降为原来的 10%。最后本课题将 DOTA 数据测试集检测结果提交到数据集官方评测服务器进行评测，VisDrone 数据集由于测试集不公开，在验证集上进行测试且通过 MSCOCO 数据集评测工具进行评测，得到最终实验结果。

4.2.2 均匀采样策略

航空影像包括为遥感图像或者无人机广角镜头航拍影像，一般呈现高分辨率特点，因此在计算资源受限的条件下进行图像检测成为一个棘手点。例如，DOTA 数据集中的遥感图像最大分辨率可达 5000×12 000，如图 4-2 所示。若直接进行图像下采样则会严重损失图像信息，所以一般需要对高分辨率图像进行裁剪训练和测试。如图 4-3 所示，之前的方法一般根据定量步长，利用滑动窗口(Sliding Window，SW)的方式进行图像均匀采样，在图像的宽和高方向进行滑动采样，假设原图正方形且尺寸为H，采样区域尺寸为K，滑动步长为S，则可以计算采样总数如公式（4.1）所示，可见采样总数与图像的边长与滑动步长有关，步长越小，采样总数越大。

$$N=\left(\frac{H-K}{S}+1\right)^2 \tag{4.1}$$

而采样总数与上限重叠度的θ(交并比)的关系可由公式（4.2）和公式（4.3）推导得。

$$\theta=\frac{K(K-S)}{2K^2-K(K-S)}=\frac{K-S}{K+S} \tag{4.2}$$

$$S=K\frac{1-\theta}{1+\theta} \tag{4.3}$$

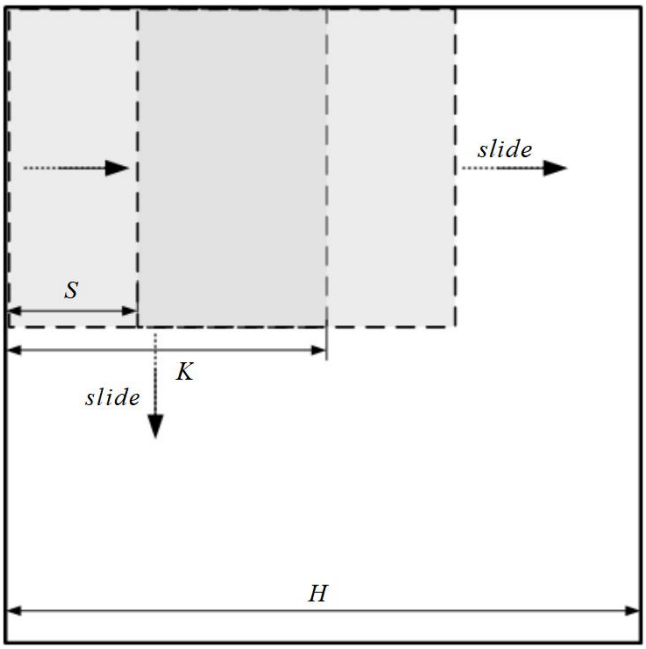

图 4-5 滑动窗口均匀采样

Fig. 4-5 Evenly sampling by sliding window

将公式（4.3）代入公式（4.1）可得到采样总数与上限重叠度的关系如公式（4.4）所示，由图 4-6，实验分析与理论计算基本相符，采样总数随着重叠度上限阈值的增大而增大，本实验总选取了 0.36 的重叠度上限阈值。

$$N=\left(\frac{H-K}{K}\cdot\frac{1+\theta}{1-\theta}+1\right)^2 \tag{4.4}$$

由此可见，这样的滑动窗口均匀采样策略会给网络训练带来如下 2 个问题：1）如图 4-6 所示，当滑动步长很小（上限重叠度很大）时，滑动窗口采样会得到大量的图像块，且其中大多数不包含任何目标，降低了网络训练效率。2）均匀采样策略生成大量包含极少目标的图像块，这些图像块大部分像素为背景，这会给前景背景分类网络训练带来正负样本不平衡问题，因此本实验权衡了训练样本的多样性于训练样本的目标占比选择了一个合适的重叠度上限阈值。

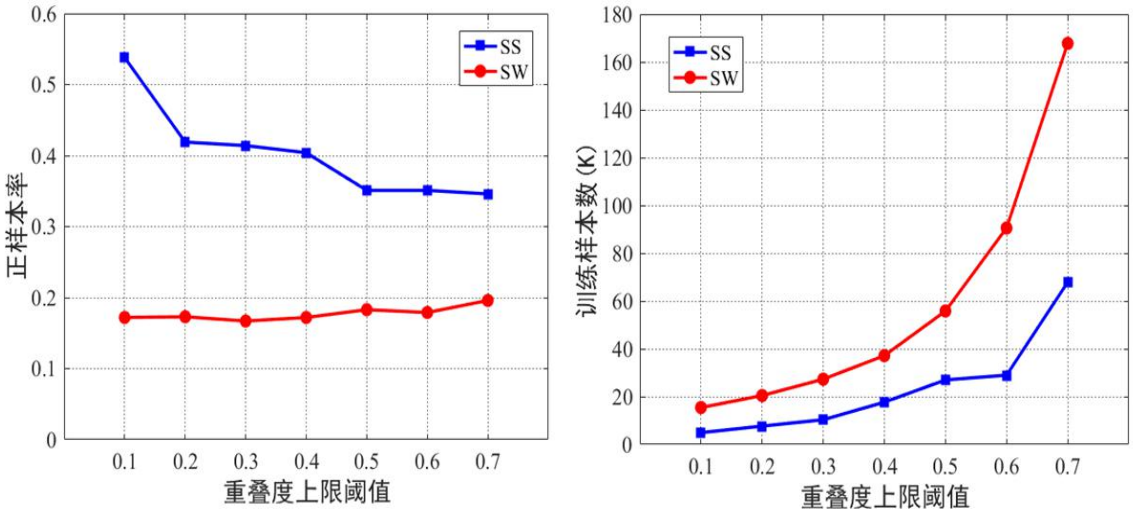

图 4-6 采样方法对比

Fig. 4-6 Comparison of sampling methods

4.2.3 选择性采样策略

虽然本课题的方法采用了 Focal Loss 来解决前景背景网络正负样本不均衡问题，但若训练样本正负比例严重不平衡时效果也是有限的。因此，本课题从原始训练数据的采样策略着手，创新性地设计了选择性采样策略，根据训练样本提供的标签来提供选择依据。具体算法流程如图 4-7 所示，给定输入参数，首先同样采用滑动窗口的形式生成一系列采样候选区域，然后根据采样候选区域中的标签边框总面积来为每个采样候选区域设定分数，最后根据给定的分数对所有采样候选区域进行非极大值抑制(Non-Maximum Suppression，NMS)，选取高分数的采样候选区域作为最终的训练图像块。为了定量衡量采样结果，定义了采样区域的目标占有率 ρ 来刻画训练正负样本比例，即采样区域前景所占有像素面积的比例如公式所示，统计每张图像块中的所以目标边框的像素面积，对边框的像素面积求和，然后除以图像块的总像素面积，得到一张图像块的目标占有率，最后对所以图像块的目标占有率求平均得到采用区域的目标占有率。

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{area}(\boldsymbol{P}_i)}\sum_{j=1}^{N_i}\text{area}(\boldsymbol{B}_j) \tag{4.4}$$

其中 N 为采样图像块的总数，N_i 为第 i 个采样图像块 $\boldsymbol{P}_i$ 内标签边框数量，$\boldsymbol{B}_j$ 为第 j 个边框， area(•) 为计算给定区域像素面积函数。如图 4-6 所示，滑动窗口均匀采样策略无论采样密度多大（滑动步长或上限重叠度多大），采样区域目标占有率几乎不变，

这是因为均匀采样等价于随机抽样过程，所以采样区域目标占有率会等于原图像的目标占有率。相反，选择性采样策略根据样本的真实标签进行选择性采样，可以通过调节采样上限重叠度，得到更高的目标占有率，从而能效缓解训练样本正负比例不均衡问题，同时选择性采用能有效避免冗余低质量的训练样本，及图像块中没有目标或者目标数量极少，如图 4-6 上限重叠度与采用总数的关系可看出，选择性采样能平均减少约 50%的训练样本，而又保证图像块有较高的目标占有比率且不损失样本的多样性，从而大大提供网络的训练效率。下面消融实验部分，将进行实验来证明选择性采样策略的有效性。

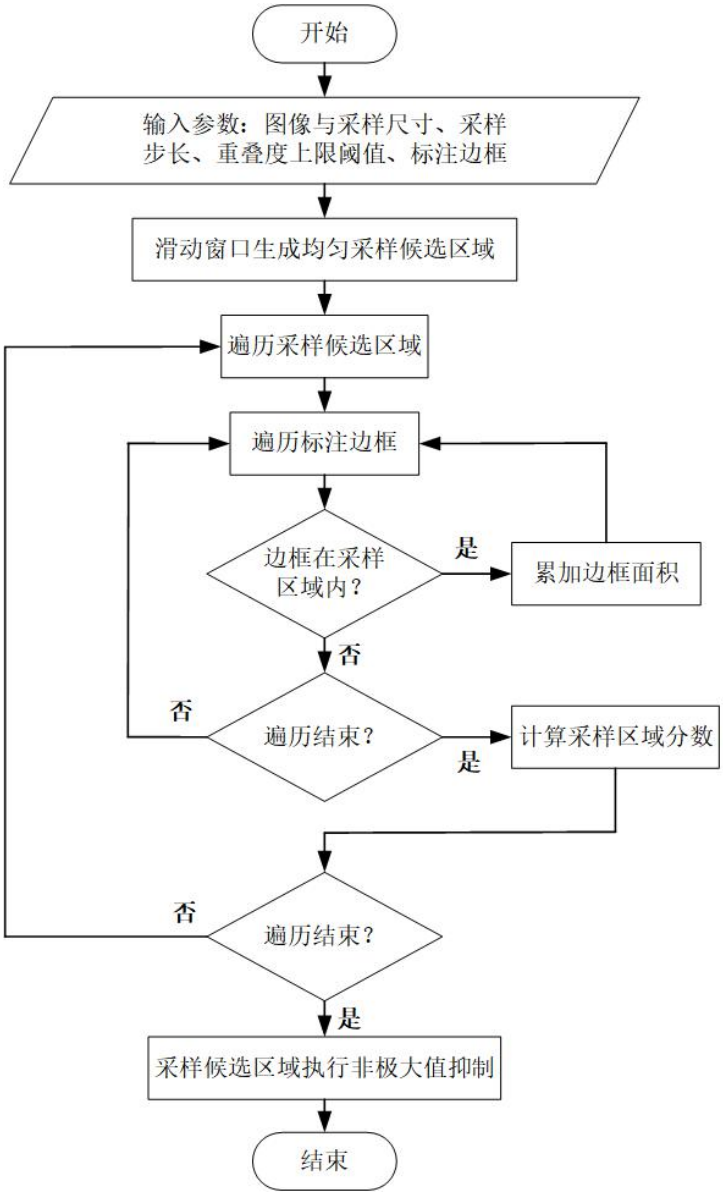

图 4-7 选择性采样算法流程图

Fig. 4-7 Flow chart of selective sampling

4.3 实验结果与分析

4.3.1 对比实验分析

本节将对所提方法在 2 个航空影像数据集的实验结果进行对比分析，通过与现有方法对比发现，所设计的单阶段定向目标检测网络实现了优异性能，甚至超过了大部分二阶段的方法。

表 4-1 展示了在 DOTA 数据集检测的各类别详细结果以及与其他方法的比较，毕竟方法包括有双阶段方法：FR-O[19]，ICN[42]，RoI Trans.[19]，SCRDet[20]和 Glid. Ver.[21]，单阶段的方法包括：IENet[43]，RetinaNet[23]，DRN[24]，O2-Det[47]，R3Det[23]和 RSDet[22]。本课题提出方法实现了 74.9 mAP 的性能，超过了现有大部分单阶段方法及部分双阶段的方法，其中 mAP 计算采样 IoU 为 0.5 的阈值。同时，可以发现本文方法在拥挤、聚集的类别上，如车辆（SV，LV）、船舶（SH）等取得最佳检测效果，这充分说明本文设计的基于关键点无锚框的方法能有效避免因锚框分布密度不足导致密集小目标漏检问题，证明了本方法对拥挤、聚集小目标检测的友好性，也充分体现了无锚框（Anchor Free）方法的优势，但在大目标检测精度上，如球场（BD，GTF，SBF）表现稍微欠佳，分析可能的原因是对目标尺寸直接进行回归，目标在不同尺度下产生的惩罚不一致，从而导致网络不能对每个类别产生较为一致性的优化效果，一个可行的方案是对尺寸回归进行归一化处理，将每个实例的尺寸回归的数值空间归一化到统一的数值空间，或者选择与尺寸无关的损失函数，如 IoU 损失函数，来解决大尺寸目标对尺寸不敏感问题。另外值得强调的是，在对 DOTA 数据集目标类别数量统计中，如图 4-3，发现直升机类别（HC）数量基本数量很少，但从实验结果可知，在直升机类别检测中，即便直升机类别样本在整个训练样本中比例小，也即处于严重类别不平衡情况，但本文方法依然能取得最好的检测精度 AP 达到了 71.0，说明能很好应对类别不平衡问题，也同时说明了车辆（SV，LV）和船舶检查精度高不是因为训练样本实例数量多导致的。

在方法上比较可发现，DRN[24]，O2-Det[47]同样采用了基于关键点检测的单阶段方法，但是其检测精度欠佳，主要是因为其没有彻底解决因旋转角度引起的旋转敏感度误差（RSE）问题。R3Det[23]和 RSDet[22]虽然考虑了旋转 RSE 问题，但其基于锚框的机制在小目标检测精度上欠佳，从而影响其整体性能。本课题综合考虑了多种因素造成的精度损失问题，从而针对性的设计了本课题的方法，实现了优异性能，甚至超过了大部分二阶段的方法。

表 4-1 DOTA 数据集检测结果

Table.4-1 Detection results on DOTA database.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage																
FR-O	79.4	77.1	17.7	64.1	35.3	38.0	37.2	89.4	69.6	59.3	50.3	52.9	47.9	47.4	46.3	54.1
ICN	81.4	74.3	47.7	70.3	64.9	67.8	70.0	90.8	79.1	78.2	53.6	62.9	67.0	64.2	50.2	68.2
RoI Trans.	88.6	78.5	43.4	75.9	68.8	73.7	83.6	90.7	77.3	81.5	58.4	53.5	62.8	58.9	47.7	69.6
SCRDet	90.0	80.7	52.1	68.4	68.4	60.3	72.4	90.9	**87.9**	**86.9**	65.0	66.7	66.3	68.2	65.2	72.6
Glid. Ver.	89.6	**85.0**	52.3	**77.3**	73.0	73.1	86.8	90.8	79.0	86.8	59.6	**70.9**	**72.9**	**70.9**	57.3	**75.0**
One-stage																
IENet	80.2	64.5	39.8	32.1	49.7	65.0	52.6	81.6	44.7	78.5	46.5	56.7	64.4	64.2	36.8	57.1
RetinaNet	88.9	67.7	33.6	56.8	66.1	73.3	75.2	90.9	74.0	75.1	43.8	56.7	51.1	55.9	21.5	62.0
DRN	88.9	80.2	43.5	63.6	73.5	70.7	84.9	90.1	83.9	84.1	50.1	58.4	67.6	68.6	52.5	70.7
O2-Det	89.3	82.1	47.3	61.2	71.3	74.0	78.6	90.8	82.2	81.4	60.9	60.2	58.2	67.0	61.0	71.0
R^3Det	89.2	80.8	51.1	65.6	70.7	76.0	78.3	90.8	84.9	84.4	**65.1**	57.2	68.1	69.0	60.9	71.7
RSDet	**90.0**	82.0	**53.8**	68.5	70.2	78.7	73.6	**91.0**	87.1	84.7	64.3	68.2	66.1	69.3	63.7	74.1
Ours	89.9	75.3	50.0	68.4	**78.5**	**82.4**	**87.0**	90.8	82.5	85.6	61.6	61.4	70.9	67.5	**71.0**	74.9

在 VisDrone 数据集检测中，如表 4-2 所示，本课题的方法在验证集上取得了 33.81% mAP 的性能，本次实验对比的方法包括有单阶段方法 Slim-YOLOv3[48] 和 YOLOv3-SSP3[48]，双阶段方法 DREN[49]，ClusDet[50]和 SAMFR[51]，甚至多阶段方法 DS-Cascade[52]。本文实验将 VisDrone 数据集标签水平边框视为特殊的旋转四边形（旋转角均为 0°），实验结果也充分说明了本文方法在通用的航拍影像中也能实现极佳的检测效果。另外对 AP_{50} 和 AP_{75} 对比分析发现，AP_{50} 即为在目标与预测边框交并比在 0.5 以上则预测为对，而 AP_{75} 则为目标与预测边框交并比在 0.75 以上预测为对，本方法在 AP_{75} 则劣于其他方法，也说明了在预测的边框回归的准确度上还不足，原因也是优于边框尺寸的回归不能对每个目标产生一致性的损失，但在检测目标覆盖率上，即交并比阈值没有特别严格情况下（0.5），本课题的方法能实现最佳效果 AP 达到了 66.50，这也充分体现了基于关键点检测方法能很好地应对密集小目标检测问题，极大程度避免了因为锚框覆盖率不足导致的目标漏检问题，证明了基于关键点检查的方法对小目标检查具有显著优势，因此如果能进一步提高尺寸回归的精度，本方法还能有一个非常可观的性能提升。

表 4-2 VisDrone 数据集检测结果

Table.4-2 Detection results on VisDrone database.

Method	AP_{50}	AP_{75}	mAP
Multi-stage			
DS-Cascade	58.02	27.53	30.12
Two-stage			
DREN	-	-	30.30
ClusDet	56.2	31.6	32.40
SAMFR	58.62	**33.88**	33.72
One-stage			
Slim-YOLOv3	-	-	25.80
YOLOv3-SSP3	-	-	26.40
Ours	**66.50**	30.01	**33.81**

图 4-8 展示了在拥挤、旋转等复杂场景下的检测效果，其中第一行图片来自 DOTA 数据集的图像裁剪图片，可以看到无论是大尺度的码头目标，还需小尺度密集的船只目标，本课题的方法能有效的检测出目标，在飞机类别检测中可以看出，目标的旋转边框没有严格满足矩形框约束，是由于本课题将目标表示为姿态图而非旋转的矩形框。第二行图片来自与 VisDrone 数据集的图像，可以看到对道路车辆的检测效果，不同颜色框代表不同的目标类别，由于 VisDrone 数据集标签只提供水平边框，因此本课题将其视为特殊的旋转边框，即旋转角为 90° 的边框。保证准确定位出目标位置的同时，能正确地检测出目标的类别，得益于回归与定位分支进行多任务预测。

图 4-8 检测结果可视化

Fig. 4-8 Visualization of detection results.

4.3.2 消融实验分析

为了验证本文所提的每个技术策略，本节将讨论在 DOTA 验证集上做的一系列消融实验，如表 4-3 所示，实验将对本课题提出的自适应特征融合特征金字塔网络以及选择性采样策略的有效性进行验证，同时也对检查方法使用不同特征提取的骨架网络进行了对比验证。

自适应融合特征金字塔网络。本文通过增加可学习的权重，将传统的特征金字塔网络改进为一般化形式，利用可学习权重，使融合网络能够动态地学习不同尺度特征的重要性。同时，融合特征在进行 1×1 卷积之前根据公式(3.11)进行归一化。为了对比分析，本文先使用传统金字塔特征网络作为基准模型进行使用，然后再使用本文所改进的自适应特征图融合网络进行实验，如表 4-3 所示，检测性能从 71.57% mAP 提升到 72.87% mAP，有 1.30%的提升。为了进一步验证，本文将自适应特征融合网络的学习权重重置为 0.5，实验发现性能有严重下降，这说明了不同尺度的特征对融合特征具有不同的贡献度，学习到的权重能自动地引导融合网络选择更具判别性的尺度特征，本次实验也同样引发对特征提取网络改进的思考，在 HRNet 多路并行的高分辨率分支网络对图像进行多尺度特征融合时候，是否也能加入可学习的权重进行特征选择，如图 3-3 的 HRNet 网络所示，在每次多尺度特征交互融合阶段，加入特征自适应选择，使在对图像进行特征提取过程中，能够从低层信息到高层信息过程进行特征适应性选择，从而为最后的特征融合阶段提供更高维度的选择空间。

选择性采样。本文所提出的选择性采样策略是为了提高网络的训练效率，同时进一步缓解训练正负样本不均衡问题。同样，本文选用滑动窗口看均匀采样策略作为对比基准，如表 4-3 所示，在使用选择性采样策略之后，网络的整体检测性能有 2.07% mAP 的提升，结合自适应特征融合网络后，最终在验证集上实现了 75.17% mAP。同时，为了进一步验证方法的有效性，本文通过开源代码复现了 R3Det 的方法，在仅增加选择性采样策略之后，网络也实现了 0.80% mAP 的性能提升，这样充分说明选择性采样策略能无代价提升网络模型性能，采样选择性采样能够去除大量冗余的训练样本，留下更有训练价值的样本，从而提升网络的训练效率的同时，还能保证网络模型的性能。本课题的选择性采样是在图像的像素空间即长和宽进行采样，同样的设想，在整个数据集空间亦可引入选择性采样，因为在整个数据集空间，并非每张图像对网络训练都

有很大的训练价值，例如当一张图像如果已经能被已训练的网络准确预测，那么这张图像对网络优化无法产生有价值的反传梯度，因此，可以利用预先训练的网络对训练数据集提前做一遍预测，然后根据预测分数来进行选择性加入训练，同样能提供网络训练效率。

特征提取网络的选择。本课题在特征提取网络的选择上进行了对比分析，其中对比的特征提取网络包括有常用的 ResNet，基于关键点检查方法所常用的 Hourglass 网络和本课题使用的 HRNet，如表 4-3 所示，打勾表示使用了该方法，在相同的检查方法下，采用了特征的高分辨率表示的 HRNet 的效果明显优于低分辨率特征表示的 ResNet，虽然 Hourglass 网络在特征压缩之后有重新同个上采样和转制卷积将特征映射到高分辨率表示，但如上文 3.1.2 所述，HRNet 能够保留图像的高分辨率特征表示，采用递进方式，从高分辨率的子网络逐步增加高分辨率到低分辨率的自网络，并在多个不同分辨率特征表示中通过上下采样进行特征融合，使得特征网络能够获取更丰富多尺度特征，避免了反复上下采样容易造成的特征空间信息的损失，同时在中层特征进行高层与底层语义特征的融合，从而保证了特征的丰富性，另外结合近期计算机视觉技术研究热点，未来也将尝试利用基于自注意力的体系结构视觉 Transformer[53]结构的特征提取网络，利用图像块和特征的位置编码来提取图像的空间特征，避免因图像高度下采样导致图像信息丢失。

表 4-3 消融分析实验结果

Table.4-3 Results of ablation studies.

Method	Backbone	Selective Sample	AFPN	mAP
R^3Det[13]	ResNet101			71.37
	ResNet101	✓		72.17
Ours	HRNet32			71.57
	HRNet32	✓		73.64
	HRNet32		✓	72.87
	ResNet101	✓	✓	73.89
	Hourglass104	✓	✓	74.27
	HRNet32	✓	✓	**75.17**

4.4 本章小结

本章对本课题提出的定向目标检测方法进行充分的实验验证，通过实验发现本课题的方法在遥感数据集 DOTA 和无人机航拍数据集 VisDrone 均取得不错的结果，相比于现有的单阶段方法和双阶段方法都有相应优势。在实验中针对航空影像的高分辨率特点，提出选择性采样策略，有效提升网络训练效率，同时使模型的性能取得显著提升。通过与前人方法的对比分析，充分证明的了本课题在方法创新上带来的性能提升，同时也实验结果引发相应的思考，为未来的工作提供了改进的思路。

结论与展望

定向目标检测是传统目标检测在特定场景下一项延伸的研究课题，是一个极具挑战且有极大的实际应用价值的任务。近些年得益于深度学习的发展，计算机视觉领域的的目标检测方法得以多样化发展，也解决了许多实际应用问题，如人脸检测，行人检测，工业元器件检测等。在航空影像检测领域，例如遥感影像目标检测，无人机航拍物体检测，由于复杂多变的视角带来复杂背景和和任意方向的目标，传统目标检测的水平边框检测难以完美地解决这个问题，因此，引入定向目标检测来解耦这个复杂问题。虽然定向目标检测可以基于传统目标检测增加额外的角度回归来实现，但是由于角度的周期特性，会带来旋转敏感度误差问题，因此本课题针对定向目标检测的方法的设计展开研究，旨在以高效的单阶段方法实现旋转目标检测，通过关键点姿态表示的方法解决角度周期性带来的转敏感度误差问题，本课题的主要研究内容与工作总结如下：

（1）本课题首先阐述了基于深度学习的目标检测方法的理论基础与设计要点，对传统目标检测方法的水平边框表示和定向边框表示展开对比分析，针对转敏感度误差问题提出了基于关键点的旋转目标姿态表示，将不同旋转角目标表示成不同姿态，通过检测目标的中心的位置及回归顶点相对坐标来实现旋转目标的检测，最后提出了单阶段基于姿态表示的旋转目标检测网络。

（2）为了解决目标检测中目标的多尺度问题，本课题基于传统的特征金字塔网络基础上进行改进，利用数据驱动的方式进行特征动态选择，并提出了自适应特征金字塔网络，在不同尺度特征融合之前增加可学习权重变量，利用已学习的权重变量使网络自动地从多尺度特征中选择更具判别性的特征，从而使得检测网络能够根据目标尺度来自动调整特征融合策略，提升对多尺度目标检测的性能。

（3）结合实际的航空影像检测场景，本课题对定向目标检测方法进行了充分的实验，包括有遥感影像和无人机航拍影像的实验，同时在实验中对目标稀疏的数据提出了选择性采样策略，特别是针对超过分辨率的航空影像上，利用目标的位置标签提供采样先验信息，使最终的训练样本具有更高的目标占有率，从而有效地提高网络训练效率和缓解网络正负样本不平衡问题，系列的对比实验分析和消融实验分析充分证明

了本课题的设计方法的有效性。

本课题提出的基于关键点的定向目标检测方法在航拍的遥感影像 DOTA 数据集和无人机航拍 VisDrone 数据集上取得优异是性能表现，虽然实验结果表现不错，但若要将算法结合到实际应用场景依然有两点需要进一步进行探索，也是后续的研究将要重点关注的两点。

（1）对检测方法泛化性进行研究分析，所提的方法是否能在不同数据域之间具有很好的迁移学习能力，即在实验数据集训练的实验结果能否在实际应用场景取得同样的性能表现。因此未来的研究将在小样本学习上进行探索，使得模型在非常有限的训练数据上训练出泛化性较好的检测模型，这将极大减小对大量标注数据的依赖程度，提高设计方法实际应用落地的可能性。

（2）提升定向目标检测方法运行效率，虽然本课题所提的检测方法是基于单阶段无锚框的方法，理论上算法的复杂度会低于二阶段有锚框的方法，但本课题尚未对检测效率展开详细的实验分析，因此，未来会针对算法效率问题进行研究，同时可以利用现有网络加速方法进行模型优化，例如进行网络剪枝或网络蒸馏，在保证检测性能可靠的同时加速网络的运行，使得检测方法能在算力有限的平台得以运行，例如在无人机平台上进行实时检测，保证无人机在飞行中能实时获得检测目标的信息，这无论时在军用领域还是在民用场景都十分具有应用价值。

参考文献

[1] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[2] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. arXiv preprint arXiv:1506.01497, 2015.

[3] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[4] 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.

[5] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[6] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[7] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[8] 钟映春, 孙思语, 吕帅, 等. 铁塔航拍图像中鸟巢的 YOLOv3 识别研究[J]. 广东工业大学学报, 2020, 37(03): 42-48.

[9] Lin T Y, Goyal P, Girshick R , et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2999-3007.

[10] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.

[11] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[12] Liao M, Shi B, Bai X, et al. Textboxes: A fast text detector with a single deep neural network[C]. Proceedings of the AAAI conference on artificial intelligence. 2017,

31(1).

[13] Liao M, Shi B, Bai X. Textboxes++: A single-shot oriented scene text detector[J]. IEEE transactions on image processing, 2018, 27(8): 3676-3690.

[14] Liu Y, Jin L. Deep matching prior network: Toward tighter multi-oriented text detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1962-1969.

[15] Ma J, Shao W, Ye H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

[16] He W, Zhang X Y, Yin F, et al. Deep direct regression for multi-oriented scene text detection[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 745-753.

[17] Zhou X, Yao C, Wen H, et al. East: an efficient and accurate scene text detector[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 5551-5560.

[18] Zhong Z, Sun L, Huo Q. An anchor-free region proposal network for Faster R-CNN-based text detection approaches[J]. International Journal on Document Analysis and Recognition (IJDAR), 2019, 22(3): 315-327.

[19] Ding J, Xue N, Long Y, et al. Learning roi transformer for detecting oriented objects in aerial images[J]. arXiv preprint arXiv:1812.00155, 2018.

[20] Yang X, Yang J, Yan J, et al. Scrdet: towards more robust detection for small, cluttered and rotated objects[C]. IEEE International Conference on Computer Vision. Seoul, Korea: IEEE, 2019: 8232-8241.

[21] Xu Y, Fu M, Wang Q, et al. Gliding vertex on the horizontal bounding box for multi-oriented object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[22] Qian W, Yang X, Peng S, et al. Learning modulated loss for rotated object detection[J]. arXiv preprint arXiv:1911.08299, 2019.

[23] Yang X, Liu Q, Yan J, et al. R3det: refined single-stage detector with feature

refinement for rotating object[J]. arXiv preprint arXiv:1908.05612, 2019.

[24] Pan X, Ren Y, Sheng K, et al. Dynamic refinement network for oriented and densely packed object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA：IEEE, 2020: 11207-11216.

[25] Lecun Y, Bottou l, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[26] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

[27] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[28] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

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

[30] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[31] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. European conference on computer vision. Springer, Cham, 2016: 483-499.

[32] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5693-5703.

[33] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]. International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[34] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C].

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[35] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of machine learning research, 2011, 12(7).

[36] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[37] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.

[38] Liu W, Anguelov D , Erhan D , et al. SSD: Single Shot MultiBox Detector[J]. Springer, Cham, 2016.

[39] Tian Z, Shen C, H Chen, et al. FCOS: Fully Convolutional One-Stage Object Detection[J]. 2019.

[40] Everingham M, Van Gooll, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[41] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context[C]. European Conference on Computer Vision. Zurich, Switzerland: Springer, Cham, 2014: 740-755.

[42] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[43] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]. IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UTAH, USA: IEEE, 2018: 3974-3983.

[44] Zhu P, Wen L, Bian X, et al. Vision meets drones: a challenge[J]. arXiv preprint arXiv:1804.07437, 2018.

[45] Azimi S M, Vig E, BAHMANYAR R, et al. Towards multi-class object detection in unconstrained remote sensing imagery[C]. Asian Conference on Computer Vision.

Perth Australia: Springer, Cham, 2018: 150-165.

[46] Lin Y, Feng P, Guan J. Ienet: interacting embranchment one stage anchor free detector for orientation aerial object detection[J]. arXiv preprint arXiv:1912.00969, 2019.

[47] Wei H, Zhang Y, Chang Z, et al. Oriented objects as pairs of middle lines[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 169: 268-279.

[48] Zhang P, Zhong Y, Li X. SlimYOLOv3: Narrower, faster and better for real-time UAV applications[C]. IEEE International Conference on Computer Vision Workshops. Seoul, Korea: IEEE, 2019: 0-0.

[49] Zhang J, Huang J, Chen X, et al. How to fully exploit the abilities of aerial image detectors[C]. IEEE International Conference on Computer Vision Workshops. Seoul, Korea: IEEE, 2019: 0-0.

[50] Yang F, Fan H, Chu P, et al. Clustered object detection in aerial images[C]. IEEE International Conference on Computer Vision. Seoul, Korea: IEEE, 2019: 8311-8320.

[51] WANG H, WANG Z, JIA M, et al. Spatial attention for multi-Scale feature refinement for object detection[C]. IEEE International Conference on Computer Vision Workshops. Seoul, Korea: IEEE, 2019: 0-0.

[52] Zhang X, Lzquierdo E, Chandramouli K. Dense and small object detection in uav vision based on cascade network[C]. IEEE International Conference on Computer Vision Workshops. Seoul, Korea: IEEE, 2019: 0-0.

[53] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020

攻读学位期间取得与学位论文相关的成果

发表和投稿与学位论文相关学术论文

1. 张国生，冯广，李东. 基于姿态表示的航空影像旋转目标检测网络[J]. 广东工业大学学报，2021.（自然版）（对应于本文的第三章，科技核心，已录用）

申请发明专利

1. 李东，张国生等. 基于视频的行为分析方法、装置、设备、系统及存储介质. 发明专利申请号：201810790994.X.

2. 张国生，李东等. 实现眼球三维视线跟踪的方法、装置、设备及存储介质. 发明专利申请号：201811375929.7.

学位论文独创性声明

本人郑重声明：所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，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。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，并表示了谢意。本人依法享有和承担由此论文所产生的权利和责任。

论文作者签名：张国生 日期：2021年6月19日

学位论文版权使用授权声明

本学位论文作者完全了解学校有关保存、使用学位论文的规定：“研究生在广东工业大学学习和工作期间参与广东工业大学研究项目或承担广东工业大学安排的任务所完成的发明创造及其他技术成果，除另有协议外，归广东工业大学享有或特有”。同意授权广东工业大学保留并向国家有关部门或机构送交该论文的印刷本和电子版本，允许该论文被查阅和借阅。同意授权广东工业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、扫描或数字化等其他复制手段保存和汇编本学位论文。保密论文在解密后遵守此规定。

论文作者签名：张国生 日期：2021年6月19日

指导教师签名：李东 日期：2021年6月19日

致 谢

行文至此，意味着我即将结束我的三年研究生学习生涯，三年的研究生学习除了让我掌握了丰富的专业知识，还让我掌握了科研方法，提高了我的处事能力。在此我要为我所得进行诚挚地致谢。

首先我要感谢我的导师李东，在他的悉心教导和鼓励之下，我顺利完成了本次关于定向目标检测的研究。李东老师时刻秉承着一丝不苟的学术精神，在科研道路上给予了我极大的帮助，我深刻明白授人以鱼不如授人以渔，李东老师的三年教导让我掌握了一套高效且科学的学习方法，这无论对我未来的科研道路还是事业道路都带来举足轻重的影响。我非常庆幸能够师从李东老师，在此表达对李东老师最诚挚的敬意和感谢。其次，我要感谢课题组的同学们在学习上乐于分享的精神，在遇到科研问题时他们都能为我提供宝贵的意见和建议，帮助我解决了许多科研难题。同时我还要感谢我的家人，感谢他们为我提供了稳定生活后盾，让我能全身心的投入到科研生活中。最后，我要感谢能在百忙之中抽出时间阅读我论文的各位评审老师，谢谢。