北京師範大學

人工智能学院

硕士研究生中期考核报告

研 究 生：	高兴林
学　　号：	202121081041
导　　师：	余先川教授
专　　业：	计算机应用技术
研究方向：	机器感知与视觉
考核时间：	2023 年 10 月 29 日
考核地点：	电子楼 306 会议室

论文题目：	基于注意力机制和高斯分布表示的遥感图像旋转框检测研究

中期考核一般不少于 6000 字（英文不少于 3500 词）

一、学位论文的研究意义及国内外研究现状分析

1.研究背景与意义

目标检测任务是找出图像或视频中人们感兴趣的物体，并同时检测出它们的位置和大小。不同于图像分类任务，目标检测不仅要解决分类问题，还要解决定位问题，是属于 Multi-Task 的问题，倍受大家的关注。

广义上的遥感技术是指利用间接的手段来获取目标状态信息的方法，这里的间接手段往往指利用人造卫星或无人机来对地面观测，以感知目标的某些特性并加以分析[1]。遥感技术的最大优点在于能够短时间内获取大范围的数据和替代人类前往难以抵达或危险的地方进行观测，其在气象、航海、农业、环境等各个领域有着广泛应用。同时随着国内高分系列卫星[2,3]和国外 Landsat 系列卫星的逐渐投入使用，使大量具有高分辨率的遥感图像的获取变得更为容易。

目前，具有高分辨率的遥感图像包含着丰富的空间和地物纹理等信息，其在农业作物生产预测、大气成分分析、天气预测、土地利用调查等与人类生产生活息息相关的活动中能够发挥重要作用。大量遥感图像若只利用人类先验知识来分析，其存在着耗时且费力的特点，难以满足实际需求。将深度学习相关方法应用至大量遥感图像数据的分析，目前已成为学术界和工业界的共识。

深度学习尤其是计算机视觉领域在最近 10 年内得到快速发展。自 2012 年，AlexNet[4]用卷积神经网络构建较深的网络模型并以高第二名近 10%优势获得 ILSVRC 竞赛冠军，使得卷积神经网络成为大家关注的重点对象。随即 VGG Net[5]尝试使用小卷积核堆叠的形式来达到与直接使用较大卷积核达到相同的感受野，来构建层数更深的网络模型；同时在模型的宽度方面，Inception[6]系列模型尝试在一层模型中并行使用不同尺寸的卷积核来增强网络对尺度的适应性；ResNet[7]则采用残差连接的结构来使模型层数达到 152 层，并且获得良好效果，现已成为图像分类任务进行比较的主流方法；DenseNet[8]采用紧密连接的形式，即每个层都会接受前面所有层作为其额外的输入，使得网络更加容易训练。2017 年以自注意力机制为特点的 Transformer[9]模型被提出，其首先在自然语言处理领域得到广泛应用，典型代表为 BERT[10]和 GPT[11]系列模型， 其逐渐成为自然语言处理的机器翻译、文本分类、命名体识别任务的主流选择。此外 2020 年 VIT[12]首度将 Transformer 模型首次应用到视觉领域并获得了良好效果，随之也涌现出相关工作。2022 年生成模型的发展,其代表有 Stable diffusion 和 ChatGPT，其引人注目的效果获得社会各界对 AIGC 领域的关注。同时，近年来用于深度学习的硬件发展较为迅速，新型号的 GPU 已允许研究者训练规模更大的模型，同时极大加快训练速度。

遥感图像的大部分场景存在物体倾斜的情况，通常会在水平框标注的基础上添加旋转角度维度来完成对目标的精确标注，其相比水平框检测任务更具挑战性。

遥感图像目标检测任务是遥感图像应用和分析的重要应用之一，其是指在遥感图像中自动检测出感兴趣的目标，比如：建筑物、轮船和飞机等，同时该任务也为计算机视觉领域中的研究热点之一。遥感图像由于拍摄方式和地物分布的原因而存在如下特点：(1)复杂背景，相比自然图像中背景较为单一，遥感图像的背景则存在着多样化的特点，比如城市、森林、草原和沙漠均可能有车辆这一待检测的物体；（2）物体分布密集：比如机场某一地方存在多架分机时，在检测时也会遇到密集目标检测的物体；（3）目标差异大，遥感图像中包含的目标种类丰富，同时这些目标的尺度、形状、纹理和颜色等特征各不相同；（4）小目标，单张遥感图像一般包含较大的视野，部分目标相对于整张图像，尺寸较小；（5）类别不平衡，大部分数据集对每种目标所包含的实例数有着较大差距，以上这些特点给遥感图像的目标检测任务带来了巨大挑战。

使用深度学习方法来实现对遥感图像中目标的自动化检测，能够避免人工目视解译存在的耗时缺陷，同时其相比使用滑动窗口等传统机器学习方法来获取候取区域以及使用 SVM 来对候选区域进行分类，进一步提高检测的效率和准确性。对于使用旋转框检测的任务，在移植自然图像检测(水平框标注检测)方法时需要额外考虑角度带来的影响。因此，根据遥感图像自身的的特点,尝试应用计算机视觉领域内较新的方法来获得识别准确率和检测效率较高的旋转框检测方法，来推动目标检测技术和遥感图像分析技术的发展。同时遥感图像中的部分特点，比如密集场景和小目标检测也存在水平框检测任务当中，因此旋转框检测技术的发展也能为水平框检测技术以及 OCR 领域中的场景文本检测和等其他领域的图像处理和分析技术提供借鉴和启示。

本研究选择相比水平框检测更具挑战性的旋转框检测领域，并选用具有代表性的遥感图像数据集，来评测旋转框检测器的检测能力。

2.国内外研究现状分析

遥感图像目标检测（旋转框检测）与自然图像目标检测（水平框检测）的发展息息相关，国内外研究现状部分则从水平框检测器和旋转框检测器两方面来进行。

2.1 水平框检测器

水平框检测器通常根据是否有获取候选区域的步骤而将检测方法分为一阶段方法和两阶段方法。两阶段检测器通常具有更高的精度，但其检测速度较慢，一阶段检测器具有较快的检测速度，但其检测精度相比两阶段检测器较低。两阶段检测器将检测任务分为两个阶段，第一阶段通常利用区域提议网络来生成一些候选区域；第二阶段则对候选区域进行分类和位置回归。两阶段检测器的典型代表为 R-CNN 系列检测器，Faster R-CNN[13]将区域提议、特征提取、分类和边界框回归等全部融入到一个网络中，相比采用传统方法能极大提升检测的效率和准确率。2017 年 Mask

RCNN[14] 则引用双线性插值的方式来对候选区域的特征图进行处理，以避免边界导致的特征不准确问题，同时该方法通过加入 mask 分支使检测结果达到像素级别，对每一个目标物体，能够对边界框内的每个像素进行是否属于该物体的标记。一阶段检测器则直接从输入图像预测目标的位置和类别，相比两阶段检测器，其具有计算效率较高的特点，但精度略低。一阶段检测器中的典型代表有 YOLO 系列、SSD 和 RetinaNet。YOLO-V1 首次将两阶段进行合并，在各种网络规模和计算资源之间提供了很好的平衡，同时能够在具有实时要求的场景中进行应用，该系列模型的后续改进使 YOLO 模型的精度不断提高，现已成为工业界目标检测任务的主流方法；SSD 将多尺度特征的思想引入到单价段检测器中，其在不同尺度上的特征进行预测，然后将所有尺度上的预测进行合并，首次达到与两阶段检测器相当的精度，同时能够满足实时推理的速度要求；RetinaNet 则使用 Focal loss 来解决前背景类别不平衡的问题，通过重构交叉熵损失来增加困难样本的权重和减轻简单样本的损失贡献来进行在线困难样本挖掘，同时此外使用 FPN 结构来进行多尺度特征融合，具有训练简单、易于实现的特点，同时在精度和运行时间方面达到比较好的权衡。

水平框检测器也可以根据是否采用锚框分为 Anchor-based 和 Anchor-free 两类方法。Anchor-based 方法会预设先验框的尺寸、长宽比和位置，然后对每个 Anchor 进行分类和回归，最终得到目标的位置和类别，前面提到的一阶段和两阶段检测器（除了 YOLOv1）均属于 Anchor-based 检测框架。Anchor-free 方法则不依赖于锚框，通过密集检测的方式来进行预测,其典型代表 FCOS[18]将 FCN 的思想用到检测任务当中，为了避免产生低质量的定位框，使用 center-ness 分支来预测像素位置相比边界框中心的偏差值；CornerNet[19]使用单一卷积模型生成热点图和连接矢量来检测边界框的左上角和右下角顶点来以识别目标，同时提出 corner pooling 来更好地定位目标角点；RepPoints[20]通过一些点的集合来表明物体的空间范围和重要的局部语义区域，其训练由目标定位和识别共同驱动以引导检测器正确分类，同时使用可变性卷积提供了适合于指导自适应采样的识别反馈。

近年来，随着 Transformer 被应用到计算机视觉中并获得良好效果，作为计算机视觉的基础任务之一，目标检测领域中也涌现出一系列相关工作。DETR[21]首次将 Transformer 作为主干架构引入到目标检测框架当中，其将目标检测任务视为集合预测问题，先用 CNN 提取特征然后送入 Transformer 做关系建模，得到的输出通过匈牙利匹配算法来与图片上的 ground-truth 做匹配,有效地消除了部分手工组件的设计，比如对 NMS 和 Anchor 的需求。Deformable DETR[22]针对 DETR 中存在的训练周期长、小目标性能差的特点，将 Deformable Conv 的稀疏空间采样与 Transformer 相关性建模相融合并进行多尺度检测。Dynamic DETR[23]将动态注意力引入 DETR 的编码和解码阶段，以打破训练收敛慢和小特征分辨率低的两个限制。

2.2 旋转框检测器

旋转框检测器目前已有的工作可以分为两个方向，一个方向将水平框检测器增加关于角度维度的参数从而将角度看作为一个回归问题，这类方法虽能够取得良好效果但因为角度的边界周期性特点和旋转框的表示方法而存损失函数不连续和回归不一致的问题；另一个方向则将旋转框检测中的角度看作一个分类问题，CSL[24]则通过设计圆形平滑标签来解决角度的周期性并增加相邻角度之间的误差容限来预测角度来解决角度回归造成的边界问题；DCL[25]针对CSL采用稀疏编码需要较长的位数来进行编码的缺点，采用二值编码和格雷编码等密集编码方法来替代稀疏编码。

基于角度回归的检测模型也可以和水平框检测器一样划分为一阶段模型和两阶段模型。两阶段模型在定义先验框时，需要将角度、长宽比、放缩比三方面的设置进行组合，同时根据需要设计 Rotated RPN、Rotated RoI Align 和 Rotated RoI Pooling 等部分组件。两阶段的水平框检测器遵循 Faster R-CNN 的范式，其典型代表工作如：ROI Transformer[26]提出学习模块，其第一部分为 RRoI Learner 学习从水平 RoIs 到旋转 RoIs 的转换，第二部分为 RRoI warping 从旋转 RRoI 中提取 旋转不变的特征，以用于后续的分类和回归子任务；SCRDet[27]则针对场景中的小型目标，其从特征融合和采样角度设计特征融合网络、针对杂乱、密集的场景则设计有监督的注意力网络减少背景噪声的不利影响、针对任意旋转方向通过添加 IoU 常数因子来设计改进的平滑 L1 损失；ReDet[28]明确地对旋转同变形和旋转不变性特征进行提取，通过旋转等变网络来提取旋转同变特征，该研究同时设计 RiRoIAlign，其通过循环地切换朝向通道与特征插值，来对齐朝向维度的特征以提取完整的旋转不变特征。一阶段检测器模型遵循 RetinaNet 的范式，其代表有：R3Det[29]同时使用水平锚框和旋转锚框来解决紧密分布的问题，针对现有的检测器中存在的特征未对齐问题，设计了一个特征修正模块，使用特征插值来获取具体的位置信息并重建特征图实现特征对齐，针对 SkewIoU loss 不可微的问题，则将 SkewIOU 以权重的形式添加到边界框的回归过程中；RSDet[30]采用八参数回归的方法来避免回归不一致的问题，通过排序算法来对角点进行排序以解决预测框与真实框角点顺序不一致所带来的问题；S2A-Net[31]在特征修正模块中使用可变形卷积来将卷积特征和任意朝向对象进行对齐，然后在朝向检测模块中分别提取方向敏感和方向不变特征，随即分别进行定位和分类等回归子任务。

随着 anchor-free 方法在水平框检测任务中被提出，旋转框见检测领域中也有相关工作。Oriented RepPoints[32]利用自适应点表示去捕捉任意方向实例的几何信息，并利用点集表示的目标结构直接回归带有方向的检测框，针对点集学习提出了一种有效的自适应点评估和分配样本方案。同时，[33]运用凸包表示来学习不规

则的形状和布局从而来避免特征不一致的问题。KLD Loss[34]将旋转框转换为高斯分布，然后使用KL散度来计算两个分布之间的差异，通过这种方式来计算两个旋转框的匹配程度以获得比较好的学习效果。

二、学位论文的研究目标、研究内容及拟解决的关键问题

1.研究目标

本研究选取遥感图像旋转框检测任务，首先完成从水平框检测向旋转框检测任务的迁移；其次将Swin Transformer以Backbone的形式融入到模型中来提升检测网络的特征提取能力;针对遥感图像检测类别中的物体分布紧凑、目标尺度差异较大，小目标检测的问题，使用可变形卷积和注意力机制融合的方式来提升模型在该场景下的检测精度；针对旋转框检测任务中回归子任务使用的Smooth L1 loss与评价指标不一致的问题，将旋转框转化为高斯分布，利用高斯分布间距离去衡量预测旋转框与实际旋转框的定位偏差。包含内容具体如下：

1) 对RetinaNet模型进行改动，以符合旋转框检测任务的要求。
2) 使用Swin-Transfromer等视觉Transformer骨架网络，提升模型特征提取能力，进而提升检测精度。
3) 结合注意力机制和可变形卷积来改进网络结构，提升模型在小目标和密集场景的检测精度。
4) 将旋转框去转换为高斯分布，利用概率分布间的距离来衡量预测框与真实框的偏差，基于此改进回归子任务的损失函数，使回归子任务损失与评价指标一致，以进一步提升模型的检测精度。

2.研究内容

研究内容从提升模型特征提取能力、引进注意力机制来提升小目标和密集场景检测精度，使用高斯分布距离衡量旋转框偏差出发，其详细内容如下：

1) 探究网络特征提取能力对检测精度的影响，验证检测模型是否存在因特征提

取能力而引起检测精度较低的问题。

2) 针对网络在尺度较小和紧密排列目标场景下检测效果较差的问题，探究注意力机制和可变形卷积方法对场景的提升效果。

3) 针对回归子任务损失与评价指标不一致的问题，研究高斯分布旋转框表示方法对该问题的解决能力。

3.拟解决的关键问题

本研究从模型特征提取能力、小目标及密集场景检测和回归子任务损失与评价指标存在不一致问题出发，拟解决的关键问题如下：

1) 增强旋转框检测模型的特征提取能力，减少因特征提取能力不足带来的不利影响。

2) 提升旋转框检测器在小目标和密集场景下的检测精度。

3) 使用高斯分布之间的距离，解决基于角度回归存在的回归子任务损失与评价指标不一致、角度边界不一致、长宽边交换等问题。

三、学位论文的研究计划、研究方法、创新之处和预期成果

1 研究计划

2023.04-2023.12：收集整理旋转框检测评测数据集、查阅相关文献并进行初步分析、算法设计及实验环节，完成修改 RetinaNet 水平框检测流程以适应旋转框检测任务，替换 Backbone 为 Swin-Transformer 和在模型中引入注意力机制和可变形卷积来提升网络检测精度部分的工作。

2024.01-2024.04：完成基于高斯分布的旋转框损失函数设计及实验部分，完善实验过程，制作图表，撰写毕业论文。

2 研究方法：

目标检测领域的检测流程已非常成熟，典型的目标检测网络通常包含：Backbone、Neck 和 Head 三部分结构。Backbone 网络主要从输入的图像中提取特征，目前其可用的结构为卷积神经网络和视觉 Transformer 两类；Neck 部分主要将从不同尺度提取的特征进行融合，常用的代表结构为 FPN；Head 部分负责分类和定位子任务以检测出对应目标。首先选择 RetinaNet 网络进，在锚框生成和回归子网络进行改造以适应旋转框的检测任务；针对 RetinaNet 采用 ResNet50 作为 Backbone 进行图像特征提取，近年来视觉 Trasformer 发展迅速并已经被广泛应用至自然图像的分类、检测和分割任务当中，则用 Swin-Transforme 替换 ResNet50 网络来增强模型的特征提取能力，并把改进后的网络作为全文的 Baseline。

针对旋转框检测在小目标和紧密排列场景中检测效果较差的问题，在 BackBone

部分分别在不同阶段的特征上使用空间和通道注意力来以使模型更容易关注到目标区域，同时在 Neck 部分使用引入特征校正模块,特征校正模块通过此位置预测框与真实框的偏差来学习可变形卷积感受野位置偏差，从特征图的校对来适应不同旋转角度的目标，希望通过这种方式来提升检测精度；目前回归子网络的损失存在与评价指标不完全一致的问题，选择将旋转框转换为高斯分布，用巴氏距离来衡量两个高斯分布的近似程度来评估两个框的匹配程度，进而提出针对定位子任务的损失函数，以解决基于角度回归存在的问题，期待能够进一步提升检测精度。本研究的研究内容路线如图 1 所示。

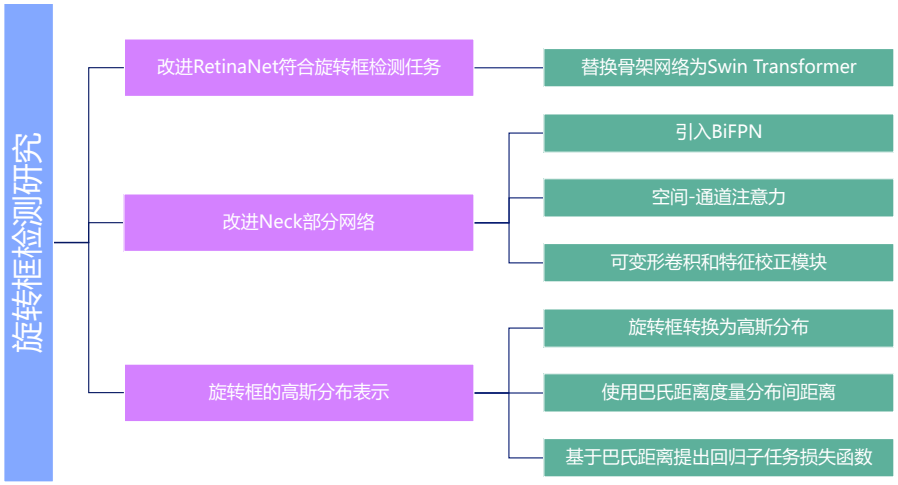

图 1 研究内容路线图

为了评估上述改进方法的有效性，选择在 DOTA-V1.0 和 HRSC2016 两个旋转框检测遥感图像领域的基准数据集上进行实验，采用的评价指标为 mAP。

3 创新之处

本研究主要包含如下创新点：

1) 将 Swin-Transformer 引入到遥感图像旋转框检测任务中，提升检测模型的特征提取能力。
2) 将注意力机制引入到网络模型中，使用可变形卷积来修正卷积网络的位置偏差来提升小目标和密集场景下的检测精度。
3) 将旋转框转换为高斯分布，使用巴氏距离衡量两个高斯分布的相似性来评估旋转框的偏差，并基于此定义回归子任务的损失函数。

4 预期成果：

1) 完成注意力机制和可变形卷积融合于旋转框检测器的研究
2) 完成基于高斯分布表示的旋转框度回归子任务损失研究
3) 完成一篇学术论文

四、学位论文研究工作情况和已完成的研究内容

1. 学位论文研究工作情况

1.1 改造水平框检测器 RetnaNet 并融入 Swin Transformer

本研究选取 RetinaNet 作为基础检测器，其能够有效解决正负样本不平衡问题，具有结构简单、具有较强的泛化能力、易于训练和实现的特点。但其最初在水平框检测任务中被提出，不能直接应用至遥感图像的旋转框检测任务当中。因此，首先对锚框生成方式添加角度维度，并修改模型回归子任务的输出通道数由 4 维变为 5 维来满足旋转框检测任务的要求，修改后的模型为 RRetinaNet(Rotated RetinaNet)进行。

RetinaNet 为保持较快的推理速度 Backbone 采用的网络为 ResNet50。但近年来 Vision Transformer 迅速发展，并被广泛应用至自然图像的分类、目标检测和分割任务等各个领域中，其代表为 ViT 和 Swin Transformer,其在图像的特征提取能力要优于 ResNet50. 为了获得较高的旋转框检测精度，尝试使用 Swin-Transformer 来作为 RRtinaNet 的 Backbone 网络，以验证原有检测器是否有因图像特征提取能力不足而引起的检测精度较低的问题。修改后的检测器结构如图 2 所示。

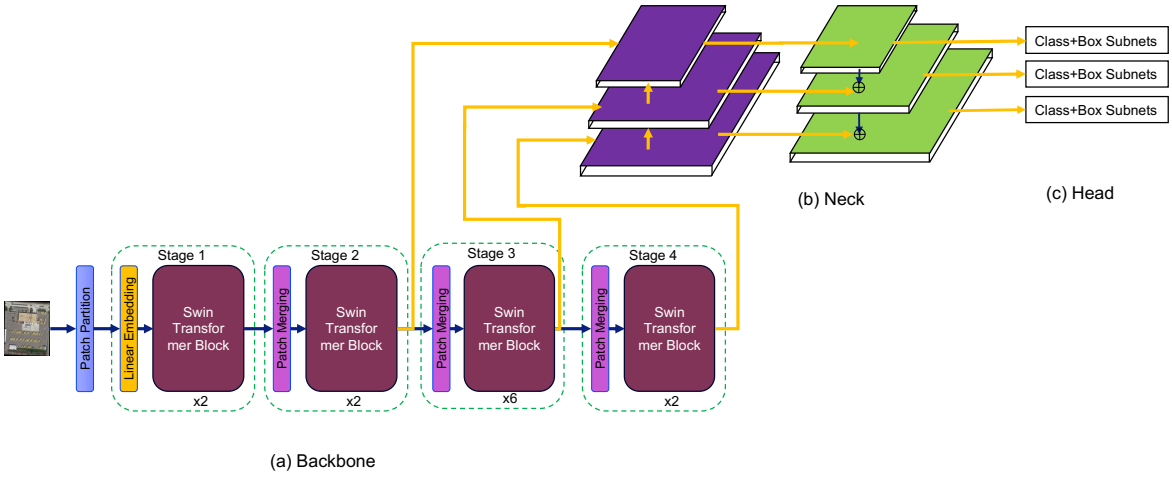

图 2 引入 Swin Transformer 后的 RRetinaNet 结构图

在替换 RRetinaNet 的 ResNet50 为 Swin Transformer 后在 DOTA-V1.0 和 HRSC2016 两个数据集上进行实验验证。为了更好验证改进方法的有效性，避免其它因素的干扰，该部分实验没有使用复杂数据增广以及多尺度训练技巧。实验结果如表 1 和表 2 所示,此处使用的评价指标为阈值为 0.5 时的 mAP。

表 1 RRtinaNet 在 DOTA-V1.0 实验结果

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
RRetinaNet	Res50	89.30	78.00	40.24	69.50	77.60	62.60	77.80	90.90	81.33	81.46	55.66	63.34	53.47	63.35	33.85	67.89
RRetinaNet	SwinT	88.59	69.70	38.34	50.05	76.13	73.93	86.12	90.76	79.36	84.27	50.66	57.57	61.64	65.12	53.61	**68.40**

表 2 RRetinaNet 在 HRSC2016 实验结果

Method	Backbone	mAP
RRteinaNet	Res50	83.0
RRetinaNet	Swin-T	**85.5**

当我们替换 RRetinaNet 中的 ResNet50 为 Swin Transformer 后，在 DOTA-V1.0 和 HRSC2016 两个数据集上均有较为显著的提升效果，说明原有的检测器存在部分因特征提取能力而引起检测效果较差的情况，更换新的 BackBone 能够弥补这种不足。此外通过对比表 1 中各个类别的检测精度发现，当进行替换后模型在部分类别出现检测精度下降的情况，对于直升机等尺度较大的类别，检测效果有非常显著的提升效果，对于较小物体类别的精度并没有太多改善。针对该问题，则尝试后面使用注意力机制和可变形卷积来提升模型在该场景下的检测效果。同时为了更公平地进行比较，选用改进后的 RRetinaNet 作为该研究的 Baseline.

1.2 引入可变形卷积和注意力机制提升密集场景和小目标检测性能

在 RRetinaNet 换用 Swin Transformer 作为新的主干网络后，其在小目标和密集检测的场景的精度并没有太大的提升，甚至还出现下降现象，这一现象证明了小目标和密集场景下的检测需要通过其它方法来解决。

RRetinaNet 的 Neck 部分网络采用特征金字塔结构来进行不同尺度特征的融合，但其只有自顶向下一条路径，高层特征不能很好地与低层特征进行跨层连接。基于此考虑我们选用在[36]中提出的 BiFPN 结构，其引入双向的特征传播机制且特征融合的权重为动态可学习，来进行更好的多尺度特征融合以增强模型的泛化能力，此外融合过程支持重复多次，具有更灵活的网络结构，希望通过这种方式能够提升检测器在目标尺度差异较大的场景的检测精度。

针对小目标和密集场景下的检测，目前的主流做法为引入注意力机制。注意力机制这几年在计算机视觉和自然语言处理等各个领域得到迅速发展。视觉领域目前采用的注意力机制可以分为：空间注意力、通道注意力和空间-通道注意力三种。为了尽量减少引入新的参数，选择分别使用 CBAM 的空间注意力和通道注意力模块，并把他们分别添加至不同层级的特征中。高层次的特征一般具有较多的通道数和较低的分辨率，因此我们选择在主干网络输出的最高层次特征使用通道注意力机制，同时在主干网络输出的最低层次特征使用空间注意力机制。

另外在旋转框检测领域还存在着卷积层的感受野和目标类别形状不一致的问题，卷积核大小为3×3的示意图如图 3(a)所示。

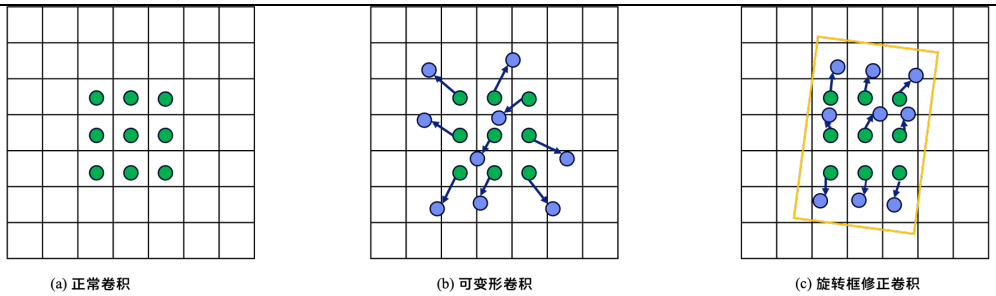

图 3 不同类型卷积感受野示意图

从上图可知，正常卷积的感受野为一个标准的正方形框，可变形卷积的感受野可能会发生不规则变化，其通过在卷积核中引入偏移量以允许感受野的范围进行自适应变化位置和形状，以这种方式来适应不规则的目标形状，从而提升模型对于目标形变的鲁棒性。可变形卷积此外通过采样操作减轻了对网络结构的假设，提高对检测目标的描述能力。希望通过向模型中引入可变形卷积来提升检测器在小目标和紧密排列场景下的检测性能。在引入可变形卷积后，希望在进行回归和分类子任务前，按照图 3(c)的示意来对模型特征进行校正，基于此提出一个特征校正模块。同理为了尽量减少新引入的参数，放弃在 Head 网络中的每层特征上引入特征校正模块，而是选择在 Neck 网络中引入该模块。特征校正模块的结构如图 4 所示。

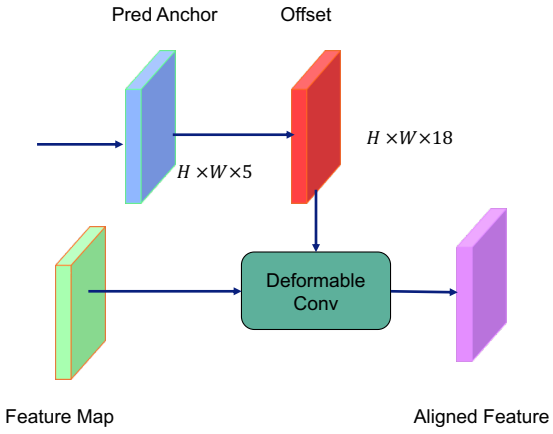

图 4 特征校正模块网络结构图

特征校正模块主包含卷积感受野位置偏差学习，在带偏差的感受野上进行卷积两部分。卷积感受野偏差通过在输入特征上进行锚框预测，并根据预测锚框和实际锚框来的偏差来确定感受野位移偏差，而后可变形卷积根据该偏差确定感受野范围,进而在适应后的感受野上进行卷积。上述使用 BiFPN 结构、引入空间注意力机制和通道注意力机制和引入可变形卷积的特征校正模块均在Neck 网络部分。新的Neck 网络结构则如图 5 所示。

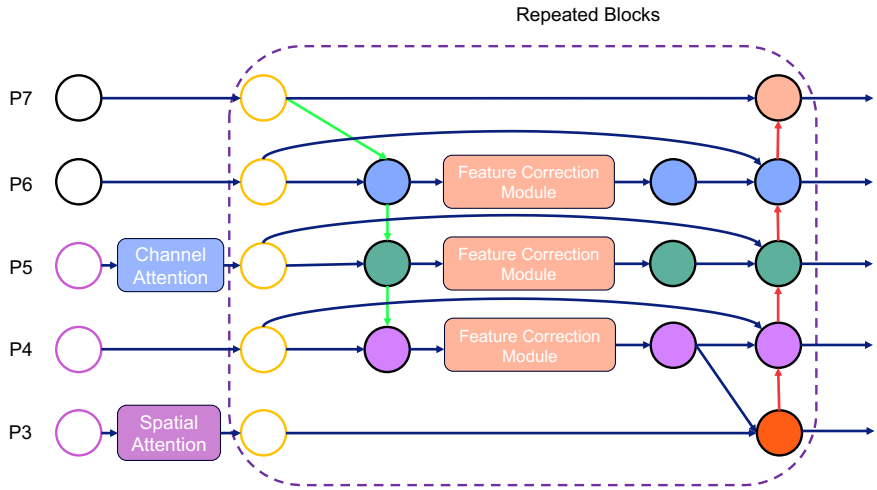

图 5 修改后的 Neck 网络结构图

下面就图中的符号和主要模块的结构和功能进行介绍：P3~P6 为主干网络输出的不同尺度特征，当主干网络中输出的尺度数不满足要求时，则在最上层特征基础上进行下采样来产生不同尺度的特征满足输入要求。对于给定特征F,其空间注意力$M_s(F)$和通道注意力$M_c(F)$的公式表述如下所示。

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$

上式中σ为 Sigmoid 激活函数，$f^{7\times7}$则为卷积核大小为7×7的卷积操作。图 3 中还有另外一个重要模块即特征校正模块,其利用可变形卷积使模型修改卷积感受野的位移和范围以灵活适应不同旋转角度的检测目标。修改后的 Neck 部分包含多种元素，将修后的模型命名为：ADBiFPN(Attettion Module with Deformable convolution BiFPN)。

我们在 DOTA1.0 和 HRSC2016 两个数据集上进行实验验证，实验结果如表 3 和表 4 所示。将 ADBiFPN 模块添加至 RRetinaNet 后，在不使用数据增广和多尺寸训练等方法，模型在 DOTA-V1.0 和 HRSC2016 两个数据集上相较 Baseline 有 6.86%和 4.6%的提升。通过观察各个类别的精度，可发现该改进能够有效提升小目标和紧密排列场景下的检测精度。

表 3 ADBiFPN 在 DOTA-V1.0 实验结果

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
FR-O	Res101	79.42	77.13	17.70	64.05	35.30	38.02	37.16	89.41	69.64	59.28	50.30	52.91	47.89	47.40	46.30	54.13
RoI Trans*	Res101	88.64	78.52	43.44	**75.92**	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
Center Map-Net	Res50	88,88	81.24	53.15	60.65	**78.62**	66.55	78.10	88.83	77.80	83.61	49.36	66.19	72.10	72.36	58.70	71.74
SCRDet	Res101	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	**86.86**	**65.02**	66.68	66.25	68.24	65.21	72.61
P-IOU	DLA-34	80.90	69.70	24.10	60.20	38.30	64.40	64.80	90.90	77.20	70.40	46.50	37.10	57.10	61.90	64.00	60.50
RRetinaNet	SwinT	88.59	69.70	38.34	50.05	76.13	73.93	86.12	90.76	79.36	84.27	50.66	57.57	61.64	65.12	53.61	68.40
DRN	H-104	88.91	80.22	43.52	63.65	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58,41	67.62	68.60	52.50	70.70
R3Det	Res101	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	68.93	**67.17**	73.74
RSDet	Res50	**90.10**	**82.00**	53.80	68.50	70.20	**78.70**	73.60	**91.20**	87.10	84.70	64.30	**68.20**	66.10	69.30	63.70	74.10
Ours	SwinT	89.39	75.35	**55.20**	73.54	74.01	77.59	**87.53**	90.91	86.74	84.92	59.20	63.25	**74.09**	**70.47**	66.67	**75.26**

表 4 ADBiFPN 在 HRSC2016 实验结果

Method	Backbone	mAP
RRteinaNet	SwinT	85.50
RoI Transformer	Res50	86.20
Gliding Vertex	Res50	88.52
R3Det	Res101	89.26
Our	SwinT	**90.10**

此外为了评估新引入各个部分的必要性，在 HRSC2016 数据集上进行了各个改动的消融实验。实验结果如表 5 所示。实验数据表明 ADBiFPN 的每步改进均能提升检测精度，由此说明给模块的改进是必要的。

表 5 ADBiFPN 在 HRSC2016 消融实验结果

Method	BiFPN	Spatial & Channel Attention	Feature Correction Module	mAP
RRetinaNet				85.5
	√			87.5
	√	√		88.4
	√	√	√	**90.1**

此外我们也对 ADBiFPN 的重复次数，进行了消融实验，实验数据如表 6 所示。从模型检测精度和模型推理方面综合权衡，将 NUM_STAGES 超参数设置为 3。

表 6 NUM_STAGES 消融实验结果

Method	NUM_STAGES	mAP
ADBiFPN	1	87.7
ADBiFPN	2	88.7
ADBiFPN	**3**	90.1
ADBiFPN	4	90.3

1.3 基于旋转框的高斯分布表示，改进回归子任务损失函数

目前旋转框常用的表示方法整体可分为 OpenCV 和长边表示法两种。为了方便进行表述做如下符号约定，在给定旋转框元组:(x, y, w, h, θ)其分别表示旋转框的中心点坐标，长度、宽度和旋转角度，此外规定逆时针的旋转角度为负。长边表示法根

据角度范围又分为 le135 和 le90 两种角度，旋转框的表示如图 6（该图参考 MMRotate 中的示意图）所示。

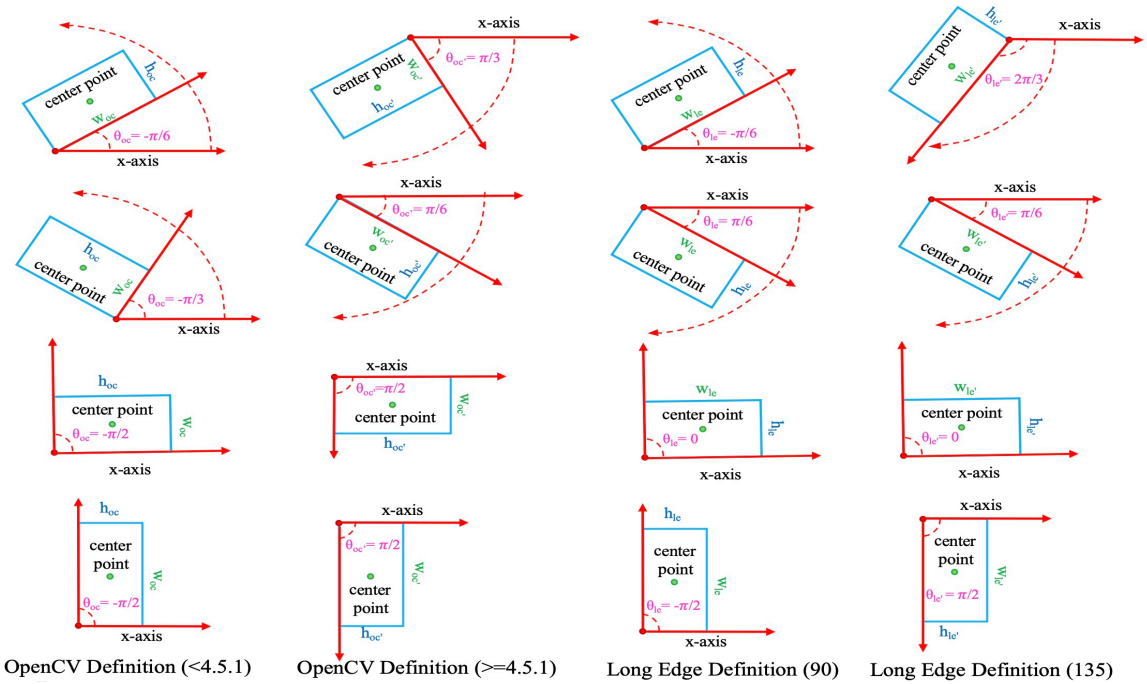

图 6 旋转框表示示意图

OpenCV 的新旧表示法的角度范围分为为：$(0^o, 90^o]$和$[-90^o, 0^o)$，le135 和 le90 的角度范围分为为：$(-45^o, 135^o]$和$[-90^o, 90^o)$. 且不同的表示方法能够进行相互转换。

旋转框检测任务的损失函数包含回归子任务和分类子任务两部分的损失，分类子任务的损失一般使用 Focal loss，其为改进版的交叉熵损失，对于困难样本的损失赋予更大的权重。由于计算两个旋转框的 IOU 过程不满足可导的要求，目前主流的做法为使用 Smooth-L1 loss 作为回归子任务的损失函数。目前[35]指出该种方式存在损失函数与评价指标不一致的问题，即较小的 loss 值并不能确保有一个较高的 IOU；Smooth-L1 loss 对于旋转框的长宽比变化不变，但 IOU 对旋转框的长宽比有着较大的敏感性。此外由于旋转框表示法存在的角度周期性和长宽边交换问题而存在着边界不连续和 Square-Like 问题。OpenCV 表示法和长边表示法只能分别解决其中的一个问题。因此本研究参考[34, 35]选择将旋转框转换为高斯分布，然后利用概率分布距离来衡量预测框和真实框的偏差，进而提出基于此的损失函数。将旋转框转换为二维高斯分布$\mathcal{N}(\mu, \Sigma)$}的过程如下所示。

$$
\begin{aligned}
\Sigma^{\frac{1}{2}} &= \mathcal{R}S\mathcal{R}^T \\
&= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \frac{w}{2} & 0 \\ 0 & \frac{h}{2} \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\
&= \begin{pmatrix} \frac{w}{2}\cos^2\theta + \frac{h}{2}\sin^2\theta & \frac{w-h}{2}\cos\theta\sin\theta \\ \frac{w-h}{2}\cos\theta\sin\theta & \frac{w}{2}\sin^2\theta + \frac{h}{2}\cos^2\theta \end{pmatrix} \\
m &= (x, y)
\end{aligned}
$$

$\mathcal{R}$表示旋转矩阵，S表示对角矩阵的特征值。在完成旋转框检测转换后，需要使用一种度量方式来评估两个高斯分布的相似程度来间接评估两个旋转框的偏差。在[34, 35]中作者分别使用KL散度和W(Wasserstein) distance来评估两个高斯分布的相似程度.计算两个高斯分布$\mathcal{N}_1(\boldsymbol{m_1}, \boldsymbol{\Sigma_1})$和$\mathcal{N}_2(\boldsymbol{m_2}, \boldsymbol{\Sigma_2})$的 W 距离$d$其结果如下所示。

$$
d^2 = \|\boldsymbol{m_1} - \boldsymbol{m_2}\|_2^2 + \boldsymbol{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})
$$

W距离表示能够解决前面所提到的定位子任务的损失与评价指标不一致的问题，也能很好解决角度周期性和长宽边交换带来的损失边界不连续和 Square Lik 等问题。该种方法的缺点为，不具有尺度敏感性，在检测类别尺度差异较大时，效果仍较差。W 距离整体上分为两部分，其分别为计算旋转框中心点的偏差和另外一部分，旋转框中心点偏差仅与旋转框的中心点坐标有关，并不能够体现出旋转框的高度和宽度对计算中心偏移的引导作用，容易导致模型预测的中心点偏差较大。

为了解决 W 距离不具有尺度不变性的特点，[34]从不同尺度的目标对旋转框的五个参数敏感程度并不相同角度出发，使用 KL 散度进行改进，使用 KL 散度计算两个高斯分布$\mathcal{N}_1(\boldsymbol{m_1}, \boldsymbol{\Sigma_1})$和$\mathcal{N}_2(\boldsymbol{m_2}, \boldsymbol{\Sigma_2})$的距离$d$,其结果如下所示。

$$
d(\mathcal{N}_1\|\mathcal{N}_2) = \frac{1}{2}(\boldsymbol{m_1} - \boldsymbol{m_2})^T\Sigma_1^{-1}(\boldsymbol{m_1} - \boldsymbol{m_2}) + \frac{1}{2}\boldsymbol{Tr}(\Sigma_2^{-1}\Sigma_1) + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - 1
$$

使用 KL 散度衡量的距离大部分也可整体分为两部分，但在前半部分计算中心点偏差时会和一个概率分布的协方差矩阵耦合，该点使得 KL 距离相比 W 距离有着更好的中心点优化机制。在[34]中作者进一步通过对梯度进行分析，KL 距离能够在面对不同尺度的目标时关注点会不同。当目标的长宽对比较大时，模型会比较关注角度的优化，而对于长宽比较低的目标，模型会比较关注长度和宽度的优化。

虽然使用KL散度计算高斯分布距离相比W距离和Smooth L1 Loss解决许多问题，但由于KL散度自身具有不对称性的特点，是与距离定义的对称性相违背的，这点可以通过使用 JS 散度来进行改进即可。KL 散度自身的取值范围可能会导致惩罚和梯度差距极大，另外当两个概率分布相距较远时，此时KL散度的值并没有意义，JS 散度值则为一个常数，此时梯度为 0，不利于模型收敛。为了综合以上两种距离的优点，本研究则尝试使用巴氏距离来衡量两个高斯分布的距离，其能够很好地结合 W 距离和 KLD 散度的优点。使用巴氏距离计算给定的两个高斯分布$\mathcal{N}_1(\boldsymbol{m_1}, \boldsymbol{\Sigma_1})$和

$\mathcal{N}_2(\boldsymbol{m_2}, \boldsymbol{\Sigma_2})$的距离$d$,其结果如下所示。

$$d(\mathcal{N}_1 \| \mathcal{N}_2) = \frac{1}{8}(\boldsymbol{m_1} - \boldsymbol{m_2})^T \Sigma^{(}\boldsymbol{m_1} - \boldsymbol{m_2}) + \frac{1}{2}\log\frac{\det(\Sigma)}{\sqrt{\det(\Sigma_1\Sigma_2)}}$$

$$\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$$

从上式可知，巴氏距离满足对称性，且和 KL 散度一样在计算中心点偏移时会有两个分布的协方差矩阵参与，同样具有较好的中心点优化机制，也能够对不同尺度的目标的旋转框参数的敏感性进行一定的调控。巴氏距离能够很好地结合 W 距离和 KL 散度的优点。基于此距离，改进回归子任务的损失函数。新的损失函数应该满足与 IOU 变化趋势一致、在反向传播时支持可微分和在角度的边界区域平滑等特点。只针对巴氏距离可能存在值域范围变化较大的情况，则可以通过log等函数做适当的变换。

该部分研究将按照该思路对巴氏距离的值域和变化趋势进行分析，确定损失函数形式，并在 DOTA-V1.0 和 HRSC2016 两个数据集进行有效性验证。

2. 已完成的研究内容

已完成的研究内容已经在研究工作情况中进行体现，该部分按照研究目标概括描述如下：

1) 对 RetinaNet 模型进行改造，以符合旋转框检测任务的要求，该部分目前已经完成，实验结果在第二点中体现。
2) 使用 Swin-Transfromer 等视觉 Transformer Backbone 网络，提升模型特征提取能力，进而提升检测精度。该部分工作已经完成，实验数据见表 1，2。
3) 结合注意力机制和可变形卷积来改进网络结构，提升模型在小目标和密集场景的检测精度。该部分实验设计和工程实现均已完成，实验数据见表 3，4，5，6。
4) 将旋转框去转换为高斯分布，利用概率分布间的距离来衡量预测框与真实框的偏差，基于此提出回归子任务的损失函数，使回归子任务损失与评价指标一致，以再进一步提升检测精度。该部分确定选用巴氏距离作为度量标准，并从理论上初步验证可行性，等待后面在基准数据集上验证实验效果和进一步分析。

五、所取得的阶段性成果

该研究取得的阶段性成果概括如下：

1) 完成 RetinaNet 水平框检测器的改造以适应旋转框检测的要求，并使用 Swin Transformer 作为 BackBone 分别在 DOTAV1.0 和 HRSC2016 上提升 0.51%和 2.0%精度，该研究以改进后的精度作为 Baseline。

2) 在模型的 Neck 部分引入 BiFPN、空间注意力和通道注意力、特征校正模块，以完成完成注意力机制和可变形卷积融合于旋转框检测器的研究，其相比 Baseline 在 DOTAV1.0 和 HRSC2016 数据集上分别有 6.86%和 4.6%的提升，并通过在 HRSC2016 数上进行消融实验，验证了改进的每部分均能提升检测器精度。观察 DOTAV1.0 数据集上每个检测类别的精度可发现，模型在改进后对紧密排列和小目标类别的检测精度有所提升。在不使用其它复杂数据增广和多尺度训练等技巧，我们的方法也能达到较高的精度。在添加这些训练技巧后，理论上模型精度能够进一步提升。

3) 基于旋转框的高斯分布表示来改进回归子任务损失函数，该部分确定使用巴氏距离来结合使用 W 距离和 KL 散度的优点。从理论上进行分析了该实验的可行性，预期通过实验验证该方法的有效性，并从梯度和实验结果等多个维度对该方法的有效性进行分析。

主要参考文献目录

[1] 翟敏.人工智能时代测绘遥感技术的发展机遇与挑战[J].工程与建设,2022,36(03):633-634.

[2] 付鸿鹏.高分七号卫星数据质量评价及处理方法研究[J].测绘与空间地理信息,2022,45(12):143-145+149.

[3] 范立佳,于龙江,姜洋等.高分多模卫星工作模式设计与在轨验证[J].航天器工程,2021,30(03):36-42.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

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

[6] 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.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceed

ings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[8] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 201 7: 4700-4708.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neura l information processing systems, 2017, 30.

[10] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transform ers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by g enerative pre-training[J]. 2018.

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

[13] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with r egion proposal networks[J]. Advances in neural information processing systems, 2015, 28.

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

[15] 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 recogni tion. 2016: 779-788.

[16] 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 1 1–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[17] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedi ngs of the IEEE international conference on computer vision. 2017: 2980-2988.

[18] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 96 27-9636.

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

[20] Yang Z, Liu S, Hu H, et al. Reppoints: Point set representation for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9657-9666.

[21] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020: 213-229.

[22] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.

[23] Dai X, Chen Y, Yang J, et al. Dynamic detr: End-to-end object detection with dynamic attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2988-2997.

[24] Yang X, Yan J. Arbitrary-oriented object detection with circular smooth label[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. Springer International Publishing, 2020: 677-694.

[25] Yang X, Hou L, Zhou Y, et al. Dense label encoding for boundary discontinuity free rotation detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15819-15829.

[26] 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.

[27] Yang X, Yang J, Yan J, et al. Scrdet: Towards more robust detection for small, cluttered and rotated objects[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8232-8241.

[28] Han J, Ding J, Xue N, et al. Redet: A rotation-equivariant detector for aerial object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2786-2795.

[29] Yang X, Yan J, Feng Z, et al. R3det: Refined single-stage detector with feature refinement for rotating object[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(4): 3163-3171.

[30] Qian W, Yang X, Peng S, et al. Learning modulated loss for rotated object detection [C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(3): 2458-246

6.

[31] Han J, Ding J, Li J, et al. Align deep features for oriented object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-11.

[32] Li W, Chen Y, Hu K, et al. Oriented reppoints for aerial object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1829-1838.

[33] Guo Z, Liu C, Zhang X, et al. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8792-8801.

[34] Yang X, Yang X, Yang J, et al. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence[J]. Advances in Neural Information Processing Systems, 2021, 34: 18381-18394.

[35] Yang X, Yan J, Ming Q, et al. Rethinking rotated object detection with gaussian wasserstein distance loss[C]//International conference on machine learning. PMLR, 2021: 11830-11841.

[36] 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.

六、导师意见

该论文选题有意义，将注意力机制和 DCN 融合应用到旋转框检测领域，并基于高斯分布表示来设计回归任务损失函数来提取结果，论文选题符合专业培养目标。同意中期。

导师签字： 李芝川

年　　月　　日