Computer Science and Application 计算机科学与应用, 2024, 14(4), 105-114
Published Online April 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.144081

Hans汉斯

基于Transformer的遥感图像目标检测算法研究

魏玉梅，江　涛*，白金燕

云南民族大学数学与计算机科学学院，云南　昆明

收稿日期：2024年3月5日；录用日期：2024年4月5日；发布日期：2024年4月16日

摘　要

针对遥感图像中目标特征不明显等导致的精度低、性能差问题。我们给出基于改进Transformer的遥感图像目标检测模型。首先，运用迁移学习加载模型，并且用ResNet101替换原始主干；其次在特征提取阶段，在主干网的bottlenet层中引入SE注意力机制；最后，将原有损失函数优化为L1损失和CIOU损失的结合。实验结果证实，改进之后的模型相对于基准而言，在精度和性能上都有一定的提高。

关键词

遥感图像，目标检测，Transformer，SE注意力机制

Research on Remote Sensing Image Target Detection Algorithm Based on Transformer

Yumei Wei, Tao Jiang*, Jinyan Bai

College of Mathematics and Computer Science, Yunnan Minzu University, Kunming Yunnan

Received: Mar. 5th, 2024; accepted: Apr. 5th, 2024; published: Apr. 16th, 2024

Abstract

Aiming at the problem of low accuracy and poor performance caused by unobvious target features in remote sensing images, we give a remote sensing image target detection model based on improved Transformer. Firstly, transfer learning is used to load the model, and ResNet101 is used to replace the original trunk. Secondly, in the feature extraction stage, the SE attention mechanism is

*通讯作者。

文章引用: 魏玉梅, 江涛, 白金燕. 基于Transformer的遥感图像目标检测算法研究[J]. 计算机科学与应用, 2024, 14(4): 105-114. DOI: 10.12677/csa.2024.144081

introduced into the bottlenet layer of the backbone network; finally, the original loss function is optimized to a combination of L1 loss and CIOU loss. The experimental results show that the improved model has a certain improvement in accuracy and performance compared with the benchmark.

Keywords

Remote Sensing Image, Target Detection, Transformer, SE Attention Mechanism

1. 引言

遥感在民生建设、城市规划、国防安全等方面都有重要的用途[1]。由于拍摄方式特殊，采用俯瞰形式，因此遥感目标存在密集排列、尺寸差大等特点[2] [3] [4]。近年来，随着卫星航拍技术的快速发展，遥感图像的处理与应用成为研究的热点[5] [6] [7]。

目标检测[8]作为计算机视觉的一项基本任务，旨在识别图片或者视频中人们感兴趣的物体。当前，基于遥感图像的目标检测算法大致可以分为三类：一是基于传统的检测，主要包含有两个阶段：1) 训练，2) 预测。较为经典的算法有尺度不变特征变换(SIFT) [9]、方向梯度直方图(HOG) [10]，形变部件模型(DPM) [11]等。该方法提取特征存在着不充分、模型泛化能力差等问题，因此导致检测的精度不高。二是基于深度学习的检测，按实现方式差异可分为两类：一是二阶段法，通过候选区域实现遥感图像物体检测，较为经典的方法为 R-CNN [12] [13] [14]系列，优点在于检测速度快；二是一阶段算法，通过回归分析来实现检测，较为经典的方法有 YOLO 系列[15] [16] [17] [18]，SSD 系列[19]，其优势在于检测精度高。三是基于 Transformer [20]的端到端的检测，Transformer 最初运用在自然语言处理中与卷积神经网络带给模型复杂的结构不同，因此在图像处理领域也取得了不错的成果。

当前，基于 Transformer 的遥感图像目标检测已经成为研究热点。近年来，许多学者提出了使用 Transformer 进行遥感图像目标检测的方法。使得模型在准确率、稳定性等方面都有所提升，对于实际问题具有重要的意义。Ding J [21]等提出的 ROI Transformer，主要是运用旋转感兴趣区域来提取特征，并将提取到的特征和 Transformer 结合，以此来实现旋转目标检测任务。Yang X [22]等提出了 R3Det 通过使用从粗粒度到细粒度的逐步回归的方法来快速和准确的检测目标。O2DETR [23]是 DETR [24]在遥感检测领域的首次尝试，通过引入深度可分离卷积，以此来替代注意力机制实现检测任务。

同时查找相关资料表明基于 Transformer 的目标检测还存在着一些挑战，其中较为经典的有模型复杂度高、计算资源要求大等。针对这些问题，本文选择 DETR 模型，在此基础上进行了改进，主要工作是将主干网进行替换，并在其 bottlenet 层引入 SENet [25]，随后将 *GIOU* 损失更换为更加稳定的 CIOU (Complete-IOU) [26]，通过上述方法来达到提高遥感目标检测准确度的目的，使得其在该领域内具有更为实际的意义。

2. DETR 模型

DETR [24]是将 Transformer [20]初次引入到目标检测领域，它是 Transformer 进入目标检测领域的开山之作。模型结构由三部分组成：backbone、Transformer、检测头。

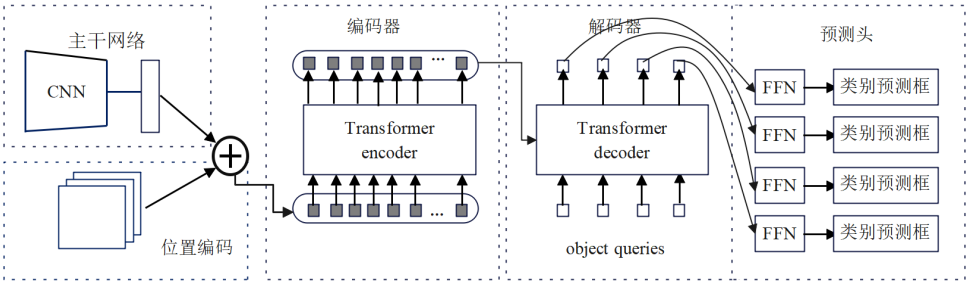


Figure 1. DETR structure
图 1. DETR 结构

观察图 1 可知：

首先，输入图像，由 Resnet-50 实现特征提取，并且融入位置编码信息，通过上述操作得到模型的输入；

其次，提取的特征图送入 Transformer 结构，在这里，值得一提的是解码器位置的输入是由两个部分组成，一个其有前面的编码器得到的结果，另外一个则是 object queries；

最后，经过上述两个步骤之后，得到包含图像特征信息的特征向量，之后经过 FFN 获得最终的预测结果。FFN 本质上是一个感知机，由 3 层网络结构构成，其中使用的激活函数为 ReLU。在检测的最后阶段，每个对象查询都需要使用该方式来得出物体的预测框和类别，这里的预测框由三个值构成，即目标的中心点以及宽和高。

DETR 模型的损失在预测对象和真实对象之间产生了最佳的二分匹配，然后优化特定对象(边界框)的损失。其主要分为以下几个步骤：

1) 检测框和真实框的匹配代价

$$\hat{\sigma}=\underset{\sigma\in S_N}{\arg\min}\sum_{i}^{N}L_{match}\left(y_i,\hat{y}_{\sigma(i)}\right) \tag{1}$$

其中 y 表示对象的真实集，$\hat{y}=\left\{\hat{y}_i\right\}_{i=1}^{N}$ 表示 N 个预测的集合，$L_{match}\left(y_i,\hat{y}_{\sigma(i)}\right)$ 表示的是真实值 y_i 和索引为 σ_i 之间的预测成本。其定义如下所示：

$$L_{match}=-1_{\{c_i\neq\varnothing\}}\hat{p}_{\sigma(i)}\left(c_i\right)+1_{\{c_i\neq\varnothing\}}L_{box}\left(b_i,\hat{b}_{\sigma(i)}\right) \tag{2}$$

公式(2)中 1 是一个符号，后面括号内容为真时取值为 1，否则取值为 0；i、c_i 和 b_i 分别表示真实框中的第 i 个元素、类别和边界框，后面两个分别记为 $bboc_i$；$\sigma(i)$是某个组合中真实框第 i 个元素对应预测框中的索引；$\hat{p}_{\hat{\sigma}(i)}$ 和 $\hat{b}_{\sigma(i)}$ 分别表示预测框中第 $\sigma(i)$个 *probs* 和 *bbox*，分别记为 $probs_{\sigma(i)}$和 $bbox_{\sigma(i)}$。其中 L_{box} 的计算表达式如下：

$$L_{box}\left(b_i,\hat{b}_{\sigma(i)}\right)=\lambda_{iou}L_{iou}\left(b_i,\hat{b}_{\sigma(i)}\right)+\lambda_{L1}\left\|\left(b_i,\hat{b}_{\sigma(i)}\right)\right\|_1 \tag{3}$$

公式(3)中 λ_{iou} 和 λ_{L1} 分别为 *GIOU* 损失和 L1 损失的权重系数，L_{iou} 是 *GIOU* 损失函数，L_{L1} 是 *L*1 损失函数。

2) 计算我们的损失函数，表示为：

$$L_{Hungarian}\left(y,\hat{y}\right)=\sum_{i=1}^{N}\left[-\log\left(\hat{p}_{\hat{\sigma}(i)}\left(c_i\right)\right)+1_{\{c_i\neq\varnothing\}}L_{box}\left(b_i,\hat{b}_{\hat{\sigma}(i)}\right)\right] \tag{4}$$

其中 $\hat{\sigma}$ 为公式(1)中计算所得到的最优分配。

3. 改进 DETR 算法

本文将 DETR 应用到遥感检测领域，在原始模型的基础上进行了改进，首先将主干网用 resnet101 替

换原始的 resnet50，其次在主干网中的 bottlenet 层中引入 SENet (Squeeze-and-Excitation Networks) [25]，随后将 *GIOU* 损失更换为更加稳定的 CIOU (Complete-IOU, CIOU) [26]，使其更加适用于本文的目标任务。

3.1. 引入 SENet

SENet (Squeeze-and-Excitation Networks) [25]是由 Hu 等人所提出的注意力模块。它将通道和空间注意力有机结合，形成一个复杂的网络结果，其具体的结构如图 2 所示[25]。

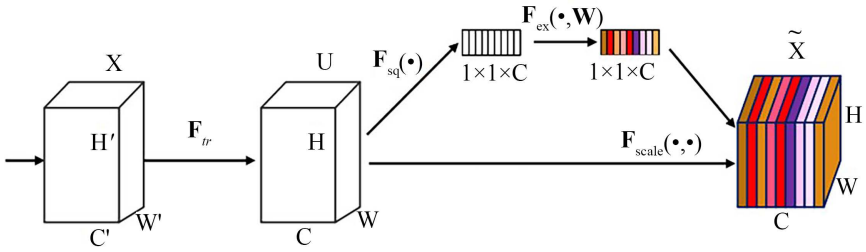


Figure 2. Squeeze-and-Excitation block
图 2. SE 模块

SENet 注意力机制处理全局信息是通过挤压和激励两种方式来进行的，具体操作所运用公式如下：

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{5}$$

$$s = F_{ex}(z,W) = \sigma\left(g(z,W)\right) = \sigma\left(W_{2\delta}(W_1 z)\right) \tag{6}$$

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c u_c \tag{7}$$

式子(5~7)比较详细的解释了 SENet 的具体过程，换言之，即输入图像，通过挤压、激励输出图像尺度的过程，式子中的 H,W 代表了图像的高宽，$u_c(i,j)$ 代表 u 的第 c 个通道的第 (i,j) 个元素。

SENet 很通用，可以非常简单、容易的和现有的网络进行结合，它利用二维全局池化计算出通道信息特征图，以较低的计算成本实现了性能的提升；并且它不改变原始模型的结构。将其引入到比较典型的主干网络—残差网络，使得模型最终的精确性得到较好的优化。

下图 3 就是将 SENet 注意力机制引入到残差网络中的结构示意图。

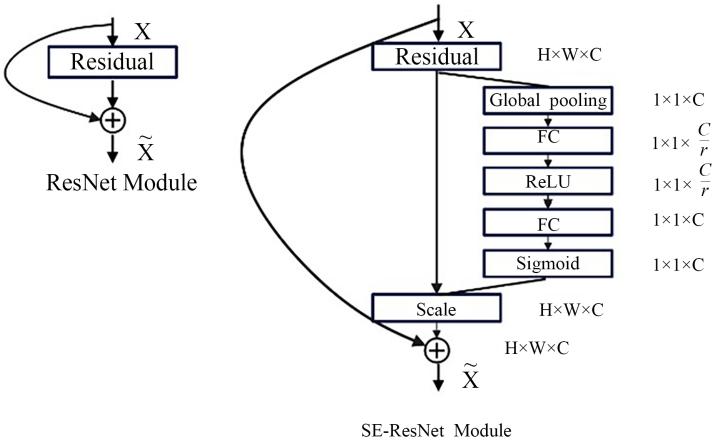


Figure 3. SENet in residual network
图 3. 残差网络中的 SENet

3.2. 损失函数设计

针对 *GIOU* 损失存在的稳定性较差的问题，我们引入了 *CLOU* Loss [26]，将 DETR 模型中的损失优化为 *CIOU* loss 与 *L*1 loss 的结合，它考虑了目标框之间的完整交叉，并通过修正因子的方式，使得当目标框有重叠甚至是有包含时，我们的回归更加快速和准确。

本文引入的 *CIOU* 损失能从一个好的边界框损失的三要素进行考虑，即分别从重叠面积、中心点距离和纵横比。通过对其充分考虑，不仅在一定程度上解决了 *GIOU* 损失所存在的问题－当预测框和真实标签包含是，我们的损失会退化为 *IOU* 的情况；而且它相较于 *GIOU* 有着很好的收敛性。

CIOU 损失的定义如下：

$$L_{ciou}\left(b_i,\hat{b}_{\sigma(i)}\right)=1-IOU+\frac{\rho^2\left(b_i,\hat{b}_{\sigma(i)}\right)}{q^2}+\alpha v \tag{8}$$

$$v=\frac{4}{\pi^2}\left(\arctan\frac{\omega}{h}-\arctan\frac{\hat{\omega}}{h}\right)^2 \tag{9}$$

$$\alpha=\frac{v}{\left(1-IOU\right)+v} \tag{10}$$

上述式中的 ρ 指预测框和真实框中心点的距离，q 表二者之间的最小闭包区域的对角线的长度，其中的 v 表纵横比参数。

4. 实验与结果分析

4.1. 数据集

本次实验选用 DOTA 数据集[27]，DOTA_V1.0 的图像来自于 Google Earth，其中一些由 JL – 1 卫星拍摄，其余由中国资源卫星数据和应用中心的 GF – 2 卫星拍摄。它包含 2806 幅图像，包含 15 个类别，共计 188,282 个实例。其中 15 个类别分别为飞机(Plane, PL)，直升机(Helicopter, HC)，游泳池(Swimming Pool, SP)，环形车道(Roundabout, RA)，港口(Harbor, HA)，篮球场(Basketball Court, BC)，足球场(Soccer Ball Field, SBF)，网球场(Tennis Court, TC)，田径场(Ground Track Field, GTF)，棒球场(Baseball Diamond, BD)，储油罐(Storage Tank, ST)，桥梁(Bridge, BR)，船舶(Ship, SH)，小型车辆(Small Vehicle, SV)，大型车辆(Large Vehicle, LV)，然后将图片按照 3:1:2 划分为训练集、验证集、测试集。

4.1.1. 数据预处理

因所采用的数据集是公开的，图片尺寸比较大，若直接把它输入到模型中进行训练，存在显存问题，使得计算资源不能承担。所以我们需对图片数据进行处理，采用官方给出的 DOTA 配套数据处理工具 DOTA devkit_master 进行处理，首先利用 DOTA.py 来加载原始图片并绘制出目标边框；其次使用 ImgSplit.py 将数据切分为 1024 × 1024 的图像块。

4.1.2. 数据转换

由于 DETR 所支持的数据集格式为 COCO、VOC 格式，因此我们需要将 DOTA 数据标签的 TXT 文本转换为 COCO 格式，DOTA 的数据集格式如图 4 所示。

观察图 4 可知，该数据包含五类，其中 image source 表图片来源 Google Earth，GF-2 and JL-1 satellite，第二个值 *gsd* 相当于比例尺，第三个值 8 个坐标值 *x*1, *y*1, *x*2, *y*2, *x*3, *y*3, *x*4, *y*4，任意四边形的四个顶点的坐标顶点按顺时针顺序排列，第一个点为起点，第四个值表示的是实例类别，最后一个值表示该实例是否难以检测(1 表示困难，0 表示不困难)。

```
imagesource:GoogleEarth
gsd:0.257268700806
641 570 672 603 658 620 624 595 swimming-pool 0
1351 298 1402 298 1402 324 1351 324 swimming-pool 0
1996 1734 2044 1739 2038 1779 1991 1772 swimming-pool 0
1966 2106 1992 2106 1985 2150 1959 2153 swimming-pool 0
1964 2031 2002 2036 1996 2094 1961 2094 swimming-pool 0
1646 2089 1671 2097 1671 2140 1643 2137 swimming-pool 0
```

Figure 4. DOTA data format
图 4. DOTA 数据格式图

4.2. 实验评价指标和条件

本实验运用的评价指标为召回率(Re*call*)、精确率(*Precision*)、均值平均精确率(*mAP*, *mean Average Precision*)。相关计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{12}$$

$$AP = \int_0^1 P(r)dr \tag{13}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{14}$$

上述公式中，*TP* 和 *TN* 分别表示为识别正确的正、负样本的个数；*FN* 和 *FP* 分别表示为识别错误的正、负样本的个数。

本文使用在 coco2017 数据集上提前训练好的模型，首先基于迁移学习加载预训练，然后在此基础上进行训练。实验所需环境和参数如表 1 所示。

Table 1. Experimental conditions
表 1. 实验条件

参数	取值
输入图像批处理大小	2
初始学习率	10^{-4}
权重衰减	10^{-4}
训练轮次	200
参数	实验配置
系统	ubuntu18.04
CPU	Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60GHz
GPU	NVIDIA GeForce RTX 3090 24G
深度学习框架	Pytorch
开发环境	torch1.9.0，cuda11.1，Python3.8

4.3. 结果分析

本文将改进之后的 DETR 算法命名为 S-C-DETR，和 DETR 在相同的训练环境下分别训练 200 轮次，检测的结果如表 2 所示。

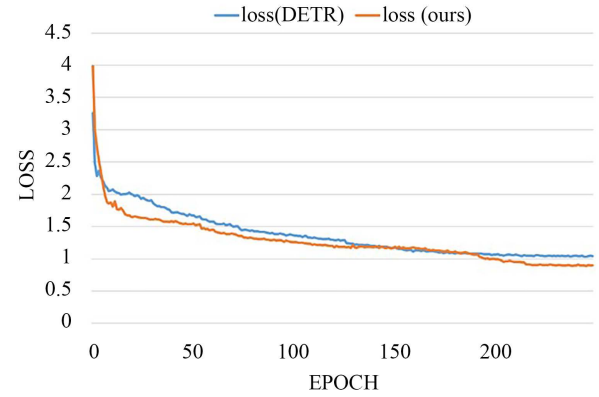
Table 2. Comparison of target detection algorithms
表 2. 目标检测算法对比

Algorithm	mAP@0.5%
DETR	58.23
DETR + CIOU	58.6
DETR + SE	59.89
S-C-DETR (Ours)	60.24

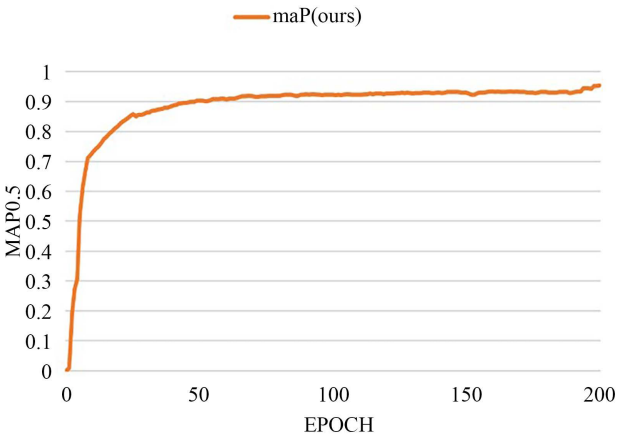
观察表 2 可知，改进之后的算法相较于原始模型而言，其检测得到的平均精度提高了 2.01%，通过实验表明改进方式有效。

4.4. 性能对比实验

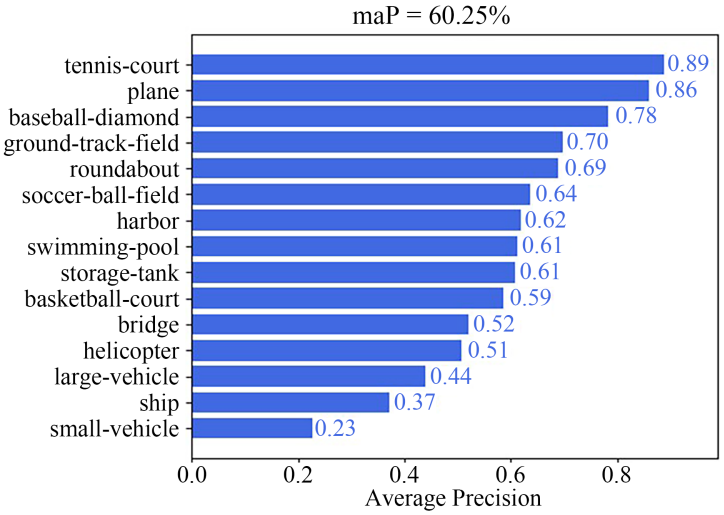
为更加直观的显示出改进之后算法的优越性，将在相同环境中训练得到的改进前后的损失和 MAP 曲线，以及各个类别的准确率曲线。如图 5 所示。



(a)



(b)



(c)

Figure 5. Training loss, mAP change curve and AP of each category. (a) Loss variation curve; (b) mAP change curve; (c) AP by category
图 5. 训练损失、mAP 变化曲线及各类别的 AP。(a) 损失变化曲线；(b) mAP 变化曲线；(c) 各类别 AP

观察图 5(a)知道，在训练过程，模型的损失在前 200 轮迅速收敛，在 200 轮之后趋于稳定，没有出现过拟合和欠拟合现象，因此改进之后的训练效果较理想。在改进算法下，训练 batch 为 2，初始学习为 10-4，迭代次数为 200 轮，在 200 轮左右的时候，模型处于收敛状态，模型损失稳定在 0.8 左右，训练过程得到正常的收敛效果，且改进之后模型的收敛效果更优。

观察图 5(b)知道，改进前后模型的 mAP 的变化趋势，在训练轮次到达 200 轮次时，我们改进之后的算法 S-C-DETR 的均值平均精度在阈值为 0.5 时到达了 0.623，相较于改进之后的模型的平均精度 0.582 提升了 0.2，证实了我们模型的改进有效。

根据图 5(c)可以很直观的观察到模型 S-C-DETR 在 DOTA 数据集各个类别的准确率。

4.5. 目标检测算法对比实验

为了验证算法针对于遥感图像目标检测的性能，将本文提出的方法与目标的主流算法进行比较，所有算法都在相同环境下进行，检测结果表 3 所示。

Table 3. Comparison of mainstream algorithms for target detection
表 3. 目标检测主流算法对比

类别	SSD [19]	Faster R-CNN [14]	DETR [24]	S-C-DETR
PL	80.9	74.7	85.3	85.8
BD	70.3	66.4	77.4	78.2
BR	18.2	14.0	46.1	51.9
GTF	68.7	63.7	68.7	69.8
SV	22.0	8.8	22.1	22.8
LV	58.4	38.0	41.2	44.2
SH	34.6	13.2	35.2	37.0
TC	88.0	84.6	90.3	88.7

续表

BC	61.2	53.2	59.8	58.6
ST	23.5	17.4	59.2	60.6
SBF	65.3	57.3	64.9	63.8
RA	32.5	28.2	65.6	68.9
HA	70.8	56.3	60.7	61.7
SP	38.4	25.7	57.4	60.9
HC	53.5	27.8	39.7	50.8
mAP	52.4	42.0	58.2	60.2

根据表 3 可以知道，我们的模型 S-C-DETR 相较于比较经典的一阶段模型 SSD 和二阶段模型 Faster R-CNN 的平均均值精度有很大的提升。S-C-DETR 模型相较于 DETR 模型精度提升了 2%；相较于 SSD 提升个 10%；除此之外和 Faster R-CNN 而言提升了 20%。因此根据实验证明，模型 S-C-DETR 针对 DOTA 的各个类别以及模型的 mAP 值都有较大的提升，改进有效。

5. 总结

本文针对于遥感图像的目标检测问题，因遥感图像存在的检测难度大等方面的问题，所以本文提出了基于 transformer 的遥感图像目标检测算法，首先使用迁移学习的方式加载预训练模型，然后再将主干网用 Resnet101 进行替换，并且再替换后的主干网中引入 SENet 注意力机制，最后优化损失函数。经过实验表明，改进之后算法再精度方面有一定的提升，因此本文的研究方法在一定程度上为以后的遥感图像检测的研究奠定基础。

参考文献

[1] Yang, F., Fan, H., Chu, P., *et al*. (2019) Clustered Object Detection in Aerial Images. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 8310-8319. https://doi.org/10.1109/ICCV.2019.00840

[2] 张意, 阚子文, 邵志敏, 等. 基于注意力机制和感知损失的遥感图像去噪[J]. 四川大学学报(自然科学版), 2021, 58(4): 45-55.

[3] 李国祥, 马文斌, 王继军. 稠密特征编码的遥感场景分类算法[J]. 小型微型计算机系统, 2021, 42(4): 766-772.

[4] 刘通, 胡亮, 王永军, 等. 基于卷积神经网络的卫星遥感图像拼接[J]. 吉林大学学报(理学版), 2022, 60(1): 99-108.

[5] Cheng, G. and Han, J. (2016) A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS Journal of Photogrammetry and Remote Sensing*, **117**, 11-28. https://doi.org/10.1016/j.isprsjprs.2016.03.014

[6] Maktav, D. and Berberoglu, S. (2018) Different Digital Image Processing Methods for Remote Sensing Applications. *Journal of the Indian Society of Remote Sensing*, **46**, 1201-1202. https://doi.org/10.1007/s12524-018-0829-4

[7] Wei, W., Zhang, J., Zhang, L., *et al*. (2018) Deep Cube-Pair Network for Hyperspectral Imagery Classification. *Remote Sensing*, **10**, Article 783. https://doi.org/10.3390/rs10050783

[8] 李章维, 胡安顺, 王晓飞. 基于视觉的目标检测方法综述[J]. 计算机工程与应用, 2020, 56(8): 1-9.

[9] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, **60**, 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[10] Kuang, H.-L., Chan, L.L.H. and Yan, H. (2015) Multi-Class Fruit Detection Based on Multiple Color Channels. 2015 *International Conference on Wavelet Analysis and Pattern Recognition* (*ICWAPR*), Guangzhou, 12-15 July 2015, 1-7. https://doi.org/10.1109/ICWAPR.2015.7295917

[11] Felzenszwalb, P., McAllester, D. and Ramanan, D. (2008) A Discriminatively Trained, Multiscale, Deformable Part model. 2008 *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 23-28 June 2008, 1-8.

https://doi.org/10.1109/CVPR.2008.4587597

[12] Girshick, R., Donahue, J., Darrell, T., *et al*. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[13] Girshick, R. (2015) Fast R-CNN. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[14] Ren, S., He, K., Girshick, R.B., *et al*. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[15] Redmon, J., Santosh, K.D., Ross, B.G., *et al*. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[16] Redmon, J. and Farhadi, A. (2017) Yolo9000: Better, Faster, Stronger. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[17] Redmon, J. and Farhadi, J. (2018) YOLOv3: An Incremental Improvement. arXiv:1804.02767.

[18] Ge, Z., Liu, S.T., Wang, F., *et al*. (2021) YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430.

[19] 张馨月, 降爱莲. 融合特征增强和自注意力的 SSD 小目标检测算法[J]. 计算机工程与应用, 2022, 58(5): 247-255.

[20] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[21] Ding, J., Xue, D., Long, Y., *et al*. (2019) Learning RoI Transformer for Oriented Object Detection in Aerial Images. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, 15-20 June 2019, 2849-2858. https://doi.org/10.1109/CVPR.2019.00296

[22] Yang, X., Yan, J.C., Feng, Z.M., *et al*. (2021) R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 3163-3171. https://doi.org/10.1609/aaai.v35i4.16426

[23] Ma, T., Mao, M., Zheng, H., *et al*. (2021) Oriented Object Detection with Transformer. arXiv:2106.03146.

[24] Carion, N., Massa, F., Synnaeve, G., *et al*. (2020) End-to-End Object Detection with Transformers. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, JM., Eds., *Computer Vision—ECCV* 2020. *ECCV* 2020. *Lecture Notes in Computer Science*, Springer, Cham, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[25] Hu, J., Shen, L., Albanie, S., *et al*. (2020) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

[26] Zheng, Z.H., Wang, P., Liu, W., *et al*. (2020) Distance-Iou Loss: Faster and Better Learning for Bounding Box Regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12993-13000. https://doi.org/10.1609/aaai.v34i07.6999

[27] Xia, G.S., Bai, X., Ding, J., *et al*. (2018) DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3974-3983. https://doi.org/10.1109/CVPR.2018.00418