FVIT-YOLO v8：基于多尺度融合注意机制的改进 YOLO v8 小目标检测

刘富宽，罗素云，何 佳，查超能
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘要：本文研究了遥感与无人机航拍图像中的小目标检测问题。由于这类图像存在目标尺度小、目标分布密集、背景复杂等特点，使得特征提取困难。目前针对小目标检测的算法，为了提升精度，大多忽略了参数量与推理速度的影响，这使得算法缺乏实用性。针对上述问题，本文提出了一种基于轻量化的多尺度融合注意机制的改进 YOLO v8 小目标检测算法。算法首先在 YOLO v8 的 FPN 结构中加入 F 算子，设计了多尺度特征的加权融合；然后在网络预测层剔除了 P4、P5 预测层，加入 P2 层用于小目标的预测；最后对轻量化自注意力机制进行图像输入网格化分割整合改进，并用它替换了 FPN 中的 C2f 模块，使得算法具有更好的全局感知能力，并大幅降低了参数量。与 YOLO v8s 相比，本文算法在 DOTA 数据集上的 mAP 提升了 4.4%，网络参数量下降了 52%，FPS 达到了 46 帧/s。在 VisDrone 数据集中，本算法在精度上提升了 8.2%。

关键词：YOLO v8；小目标检测；Transformer；轻量化实时性
中图分类号：TP391 **文献标识码：**A **文章编号：**1001-8891(2024)08-0912-11

FVIT-YOLO v8: Improved YOLO v8 Small Object Detection Based on Multi-scale Fusion Attention Mechanism

LIU Fukuan，LUO Suyun，HE Jia，ZHA Chaoneng
(*School of Mechanical and Automotive Engineering*, *Shanghai University of Engineering Science*, *Shanghai* 201620, *China*)

Abstract: This study investigates the problem of small-target detection in remote sensing and drone aerial images. These images have the characteristics of a small target scale, dense target distribution, and complex background, which makes feature extraction difficult. Most current algorithms for small-target detection ignore the impact of parameter quantity and inference speed on the practicality of the algorithm to improve accuracy. Therefore, this algorithm is impractical. To address these problems, this study proposes an improved YOLO v8 small target detection algorithm based on a lightweight multiscale fusion attention mechanism. The algorithm first adds the F operator to the FPN structure of YOLO v8, designs the weighted fusion of multiscale features, removes the P4 and P5 prediction layers in the network prediction layer, adds a P2 layer for small target prediction, improves the image input grid segmentation integration of the lightweight attention mechanism, and replaces the C2f module in the improved FPN with it, thereby improving the algorithm have better global perception ability and greatly reducing the parameter quantity. Compared to YOLO v8s, the mAP of this algorithm on the DOTA dataset increased by 4.4%, the network parameter quantity was reduced by 52%, and the FPS reached 46 frames. For the VisDrone dataset, this algorithm improved the accuracy by 8.3%.
Key words: YOLO v8, small target detection, Transformer, lightweight real-time

0 引言

高分辨率遥感图像与无人机航拍图像中的目标检测在智慧交通、城市建设、军事应用等领域具有举

收稿日期：2023-04-26；**修订日期：**2024-08-02.
作者简介：刘富宽（2000-），男，硕士研究生，主要从事无人驾驶车辆环境感知方向与计算机视觉的研究。E-mail：liufukuan927@163.com。
通信作者：罗素云（1975-），女，副教授，主要从事无人驾驶汽车环境感知及控制的研究。E-mail：lsyluo@163.com。
基金项目：国家自然科学基金（62101314）。

足轻重的作用。而近年来，随着深度学习在视觉领域的应用，虽然目标检测得以快速发展，但基于无人机和遥感图像的小目标检测仍是一大难点。

目前基于卷积神经网络（convolution neural network，CNN）的深度学习特征提取方案，已经广泛应用于目标检测的各个领域。Pascal VOC[1]与 MS COCO[2]这些基线数据集的出现对于目标检测的发展与应用起到了至关重要的作用，并随之出现了以 RCNN[3]、Faster-RCNN[4]、Mask-RCNN[5]、以及 SPP-Net[6]为代表的 Two-Stage 检测算法，此类算法精度较高但是速度慢，为此 Redmon 等人提出了以 YOLO（You Only Look Once[7-10]）与 SSD[11]为代表的 One-Stage 算法。Yolo 算法的诞生极大地促进了目标检测的应用与发展，目前已经更新到 YOLO v8。以上算法均采用通用框架因此对于常规数据集中目标检测的泛化效果较好，但对于小目标检测，因其目标尺度较小，会在特征提取过程中出现特征丢失的情况。

特征金字塔结构在近年来的小目标检测中被广泛应用。为了提升小目标检测能力，Zhu[12]和 Dong[13-14]等主要关注浅层特征提取，构建了自下而上的特征提取网络，与常规特征提取网络不同，此类网络同时融合了浅层特征与高层特征的语义信息。Zhang[15-16]等采用自上而下的特征提取方式，通过构建特征提取网络结构分别负责上采样与下采样，并在两个网络结构之间构建连接以实现不同层信息流的交叉融合，构成特征金字塔网络（feature pyramid network，FPN），有效提升了对小目标的表征能力。Cheng[17]等在 FPN 的基础上在特征融合过程中采用 1×1 卷积操作降维，实现跨通道的特征整合，并且可以降低参数，以实现多尺度目标检测。

在注意力领域，受处理序列信息的 Transformer 网络启发，Dosovitskiy[18]等人提出了一种用于图片处理的 Transformer 模型，Vaswani[19]等人在此基础上进一步提升，在小目标检测领域 Zhu[20]等人利用改进的 Transformer 检测头在小目标检测数据集 VisDrone 上取得了最先进（state of the art, SOTA）的表现，VIT（vision transformer）的参数量过大，因此 Liu[21]等人设计了 Swin-Transformer 采用滑动窗口的方式对每一个窗口进行局部化计算。

常规场景的目标检测已取得了很好的效果，但相较于常规场景的目标检测，基于高分辨率遥感图像与无人机航拍图像的小目标检测因为目标尺度小、目标分布密集、背景复杂等特点，导致常规的检测模型难以达到理想的检测效果。并且目前小目标检测算法几乎只关注于精度提升而并不考虑模型的参数量增加相对于精度提升的边际效应，缺乏实用性。

因此，本文针对高分辨率遥感图像与无人机航拍图像的小目标的尺度特性，与基线算法的结构特点，提出了一种小目标检测算法：加速视觉变换器改进 YOLOv8 算法（fast vision transformer-YOLO v8，FVIT-YOLO v8），该算法主要特性有：

改进双向多尺度特征融合网络，在常规提取的过程中，通过增加连结结构实现跨特征层融合，会加强浅层神经网络的语义特征提取与不同层特征的融合。

对 YOLO v8s 预测层进行改进，去除原网络中用于预测大目标的层，同时增加用于小目标检测的 P2 层，以增强对小目标检测识别的效果。

添加基于 Transformer 的全局注意力机制，在 Transformer Encoder 特征图输入阶段采用网格化处理并 reshape 为序列输入，大幅降低了参数量并在 Encoder 层加入平均池化层弱化复杂背景影响。

1 FVIT-YOLO v8 算法

1.1 YOLO v8s 概述

YOLO v8s 作为目前最新的 YOLO 系列网络，其网络简化结构如图 1 所示，在主干网络部分依旧沿袭 YOLO v5 系列的跨阶段部分连接（cross stage partial，CSP）思想，采用 C2f 模块实现了进一步的轻量化；在特征金字塔（feature pyramid network，FPN）结构上使用了 PAN-FPN 结构；在检测头部分采用 Decoupled-Head 检测头，并且使用了 Anchor-Free 的思想；在损失函数方面 YOLO v8 使用 VFL Loss 作为分类损失，使用 DFL Loss＋CIOU Loss 作为回归损失；并抛弃了以往的 IOU 匹配或者单边比例的分配方式，而是使用了任务对齐分配（task-aligned assigner）匹配方式。

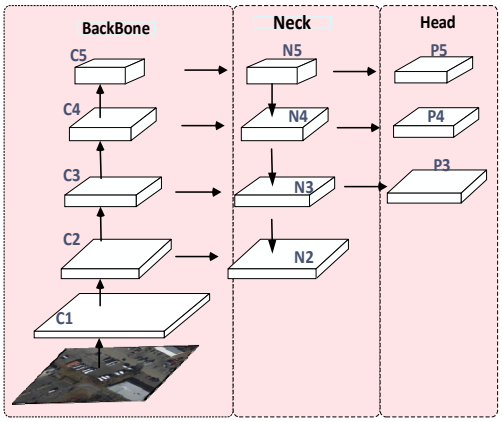


图 1 YOLO v8 模型结构简图

Fig. 1 Schematic diagram of YOLO v8 model structure

1.2 FVIT-YOLO v8 介绍

本文改进的 FVIT-YOLOv8 的网络结构如图 2 所示，主要由 YOLO v8s 主干网络（BackBone），双向

多尺度融合交叉 FPN（Neck），全局动态自注意力机制轻量化视觉变化器（light vision transformer, LVIT）以及轻量化检测头（Head）组成。在主干特征提取网络中我们沿用了 YOLO v8s 的 BackBone，在 Neck 部分我们重新设计了一种双向多尺度融合交叉 FPN 用于小目标特征提取，并且在 FPN 内部加入本文改进的轻量化 Transformer Encoder 结构提升网络对于全局特征的注意力，最后在预测层加入专用于小目标检测的 P2 层并移除原网络中的 P4、P5 层，使 FVIT-YOLO v8 适用于高分辨率遥感图像与无人机航拍图像的小目标检测。

1.3 多尺度加权融合交叉 FPN——CROSS-FPN

多尺度特征融合是小目标检测的重要手段，本文在 YOLO v8 原 FPN（如图 1 所示）基础上重新设计了多尺度加权融合交叉 FPN 网络结构如图 3，本文在改进 FPN 中加入 F 算子，算子结构如图 3 所示，其中红色箭头表示特征输入，蓝色箭头表示特征输出，在特征输入下采样过程中使用步长为 2 的 3×3 卷积对输入特征进行卷积，其他方向采样则使用 1×1 卷积进行降维输入，最终通过 F 算子实现跨特征层融合，再将融合特征通过 1×1 卷积进行通道整合并二次降维输出。相对于 YOLO v8s 直接从 N5 层到 N4、N3、N2 的上采样，我们将 N5 层的特征信息传入 F 算子，通过 1×1 卷积降维，在降低参数的同时，也使得 N4 层获取了更为丰富的通道特征信息。

在预测层，本文首先基于 DOTA 数据集[22]对高分辨率遥感图像与无人机航拍图像的小目标的特点进行了分析，再参考小目标定义与数据集分布的阈值范围，将 YOLO v8 的预测层进行重新设计，剔除了原始模型中关注大目标的预测层，并且增加了 P2 层获取融合特征进行小目标的预测，在提升数据集精度的同时，还降低了一部分的参数量，最终的改进 FPN 在 DOTA 数据集上实现了 3%的精度提升，相对于原网络（YOLO v8s）降低了 69.1%的参数量。

具体步骤如下：

输入图像为 640×640 时，通过多次特征提取经历 C_1、C_2、C_3、C_4、C_5 下采样，再经过 N_5、N_4、N_3、N_2 的上采样，形成双向特征融合网络，下采样过程特征图从 320×320 像素减到 20×20 像素，每经历一次特征提取，特征图像变为原图的 1/4。在进行上采样时以 N_4 层为例，首先通过 1×1 的卷积将 N_5 与 C_4 的特征输出作降维处理，对于 C_3 的输出采用 3×3 的卷积进行处理，通过 F 算子接收来自这三层的跨特征层信息做交叉融合，将融合特征通过 1×1 卷积操作进行降维输出，N_3 与 N_2 层同理。在预测层同样使用 F 算子进行通道整合输出操作，相较于原网络融合了更多的跨通道特征，并且根据数据集样本分布情况，针对小目标检测剔除了原网络中的 P_4、P_5 层并增加了针对小目标检测所调优的 P_3 检测层，在降低原始网络参数量的同时较大幅度增加了网络对于小目标的

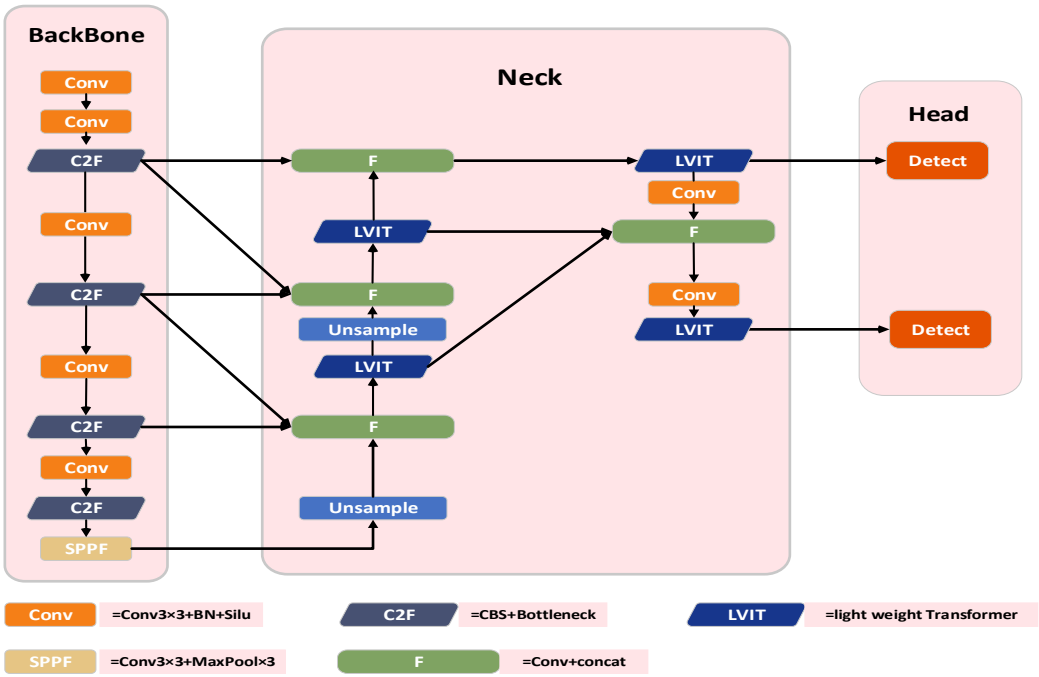


图 2 FVIT-YOLO v8 小目标检测算法网络结构

Fig.2 Network structure of the improved YOLO v8s small target detection algorithm

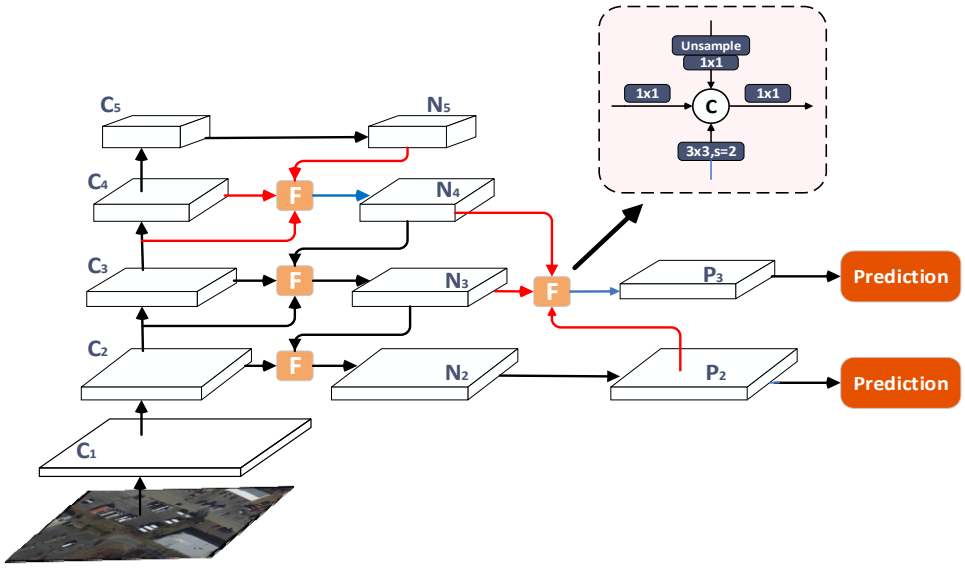


图 3 多尺度加权融合交叉 FPN 网络结构

Fig.3 The network architecture of small object multi-scale fusion cross FPN with multilevel feature integration

表征与检测能力。本文在做跨通道信息交互时多次采用 1×1 的卷积操作，在整合通道的同时也起到了降维与减少计算与参数量的作用。

C_1、C_2、C_3、C_4、C_5 对应主干网络的不同层次的特征，对应下采样倍率 2、4、8、16、32，使用 F 算子进行多尺度的融合交叉连接，融合 3 种不同层次的特征，使用 stride 为 2 和上采样加卷积的操作实现特征尺度对齐，1×1 卷积实现不同层次的特征融合，并引入了 C_2 特征进一步融合浅层特征加强空间建模能力，将 F 算子引入自下而上的特征融合，在最后的预测层阶段，本文舍弃了原始的 P_4、P_5 层，并增加了 P_2 层负责检测更大输出特征层尺度上的小目标，并保持原有的 P_3 层进行 bbox 的预测。图 3 中的特征图以 P_3 与 N_4 为例，可以使用公式(1)与公式(2)来表示融合计算:

$$P_3 = \mathrm{Conv}\left(\frac{\mathrm{Conv}\left(\omega_1 \cdot N_4^{\mathrm{out}}\right)+\mathrm{Conv}\left(\omega_2 \cdot N_3^{\mathrm{out}}\right)+\mathrm{Conv}\left(\omega_3 \cdot P_2^{\mathrm{out}}\right)}{\omega_1+\omega_2+\omega_3+\varepsilon}\right) \tag{1}$$

$$N_4 = \mathrm{Conv}\left(\frac{\mathrm{Conv}(\omega_1' \cdot C_4^{\mathrm{out}})+\mathrm{Conv}(\omega_1' \cdot N_5^{\mathrm{out}})+\mathrm{Conv}(\omega_1' \cdot C_3^{\mathrm{out}})}{\omega_1+\omega_2+\omega_3+\varepsilon}\right) \tag{2}$$

式中：ω_1 与 ω_1' 代表不同权值，采用上角标用以区分不同特征处理的通道，使用下角标数值区分不同层的输入，因采用跨特征加权融合操作使得同一层可能含有多个输出，因此上式涉及输出通道均已用红色箭头于图 2 表示。Conv 表示卷积操作；ε为融合加权超参数，此处设为 0.001。

因为本文在设计 F 算子进行跨通道特征融合时，出现了混叠效应，所以在进行特征层下采样的过程中并未使用 1×1 的卷积进行降维操作，而是采用 3×3 的卷积处理下采样的输出特征，这种方法可以有效消除横向连接与下采样特征融合过程中产生的混叠效应。

1.4 C2f 模块

对于原网络中的 C2f 结构如图 4 所示。首先输入的信息流经过 CBS 模块，该模块由 Conv、BatchNorm 和后面的 SiLU 组成。相对于 YOLOv5，YOLOv8 将 Conv 操作换成了 3×3 的卷积，Bottleneck 与 YOLOv5 相同，但第一个 conv 的卷积核大小从 1×1 更改为 3×3。从这些信息中，我们可以看到 YOLO v8 开始回归到 2015 年提出的 ResNet[23]网络，在 C2f 中所有 Bottleneck 的输出都通过残差结构在 Concat 层进行整合，可以通过调节 N（Bottleneck 堆叠个数），来增加网络的深度，类似 YOLO v7 中的 ELAN 结构融合了丰富的梯度流信息。

1.5 Transformer encoder 模块的轻量化及 C2f 模块改进

本文将 YOLO v8s 的 C2f 模块采用经过轻量化改进的 Transformer Encoder 模块进行替换。近两年 Vision Transformer（VIT）发展迅速，并且为图像处理领域带来了新的思路，受其启发本文在设计网络时将 Transformer 结构与原网络进行融合，并且对于 Transformer 中的一些结构进行调整，在提升精度的同时，很大程度减少了网络参数量。VIT[24]中的多头注

意力机制结构相较于 YOLOv8 Neck 结构中的 C2f 结构，其对全局特征的感知要优于 C2f 中的残差结构。因考虑本网络输入窗口较小，因此本文并未进行 position embedding 操作，同时 Transformer 在计算注意力矩阵的过程中会增加网络的参数量并且与序列长度相关，而原始 Transformer 中的 patch 操作会产生大量序列，因此本文采用网格化划分替代 patch 操作以降低参数量。因遥感图像背景复杂，本文在 LN 层后加入了平均池化操作可以一定程度上弱化背景影响并降低参数，经过上述改进在提升精度 1.4%的同时模型整体参数量减小到 5.41 M，参数量仅为原模型的 48%。

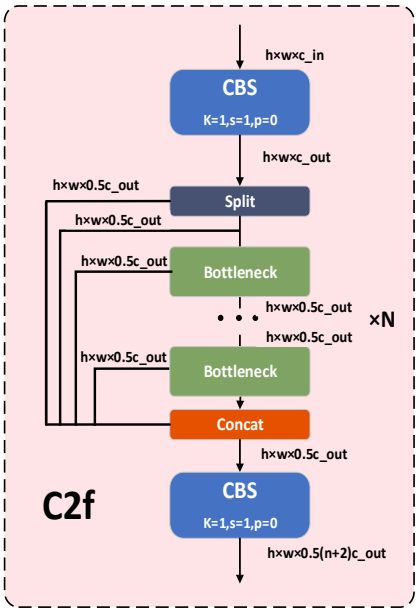


图 4 C2f 模块结构

Fig.4 The structure of the C2f module

Transformer 架构基于一种自注意力机制，该机制可以有效学习序列中各个元素之间的关系，与递归的处理序列元素并且只关注于短期的上下序列之间关系的网络相反，Transformer 可以关注到完整的序列信息，与其他注意力机制相比，如硬注意力，需要蒙特卡洛采样来确定位置，其本质上是随机的，ViT 模型在计算效率和精度方面比目前最先进的 CNN 几乎高出 4 倍。

Transformer Encoder 中有多个块，每个块由 3 个主要处理元素组成：Layer Norm、Multi-head Attention Network（MHSA）、Multi-Layer Perceptrons（MLP），其中 Layer Norm 使训练过程保持稳定，并让模型适应训练图像之间的变化。

在多头注意力机制中，每个头都有自己的投影矩阵 $\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$，它们分别计算使用这些矩阵投影的特征值的注意力权重。多头注意力网络（MHSA）是负责从给定的嵌入式视觉标记生成注意力图的网络。这些注意力图有助于网络将注意力集中在图像中的目标区域。MLP（多层感知机）是一个两层的分类网络，最后是 GELU（Gaussian Error Linear Unit）。MLP 模块也称为 MLP 头，用作 ViT 的输出。在此输出上应用 softmax 可以提供图像分类标签。

自注意力机制动态建模能捕获图像的全局特征，能大幅度提升目标检测器的小目标的检测能力。对于用做图像分类的 Transformer，其工作流程如下：首先通过 Embedding 将图片进行变换，将输入原图处理成诸多固定大小的 N 个 Patches， 如果输入图片大小为 640×640，按 16×16 大小的 Patch 进行划分，可以得到，$N=(H\times W)/p^2=6420/162=1600$，$N$ 同时也代表了序列的长度，进而将处理好的序列输入 encoder 中进行处理最后通过 MLP 输出特征。但在特征图分辨率较大的情况下，硬件设备难以承受自注意力巨大的计算量成本。在这个问题上，本文提出了网格化（Grid）特征图的方法，将计算特征图以 Grid size 为 10 的大小依次划分开，进而对网格特征图进行 Reshape 处理为 1×64 的序列，特征图网格化输出网格个数所对应的通道即为序列的个数，并对处理后的序列进行 Encoder。对于 Encoder 操作，本文在 K、V 获取时使用平均池化的方法，将 10 的窗口大小池化到 5，进一步减少计算量，并且加强了对于背景的处理能力，之后送入 MLP 学习注意力机制后的高维空间关系，最后通过反 Reshape 和反 Grid 操作还原为原特征图大小，其原理示意如图 5 所示。

1.6 Loss 函数

本文算法使用 YOLOv8 损失函数进行损失计算，主要由 2 部分组成，分别是分类分支和回归分支。分类分支使用二分类交叉熵损失函数（BCE Loss）进行计算；回归分支损失采用 Distribution Focal Loss（DFL Reg_max 默认为 16）＋CIoU Loss 来计算，且只计算正样本的回归损失，3 个 Loss 使用加权平均进行计算，其权值分别为λ_1（dfl loss）、λ_2（cls loss）、λ_3（box loss）。

计算公式为：

$$L_{\text{loss}}=\lambda_1 l_{\text{dfl}}+\lambda_2 l_{\text{cls}}+\lambda_3 l_{\text{box}}$$

$$l_{\text{dfl}}=-[(p_{i+1}-p)\log(S_i)+(p-p_i)\log(S_{i+1})]$$

$$l_{\text{cls}}=-\sum_{i=1}^{N}\hat{y}_i\ln(y_i)+(1-\hat{y}_i)\ln(1-y_i)$$

$$l_{\text{box}}=1-I_{\text{oU}}+\frac{\rho^2}{c^2}+\alpha v$$

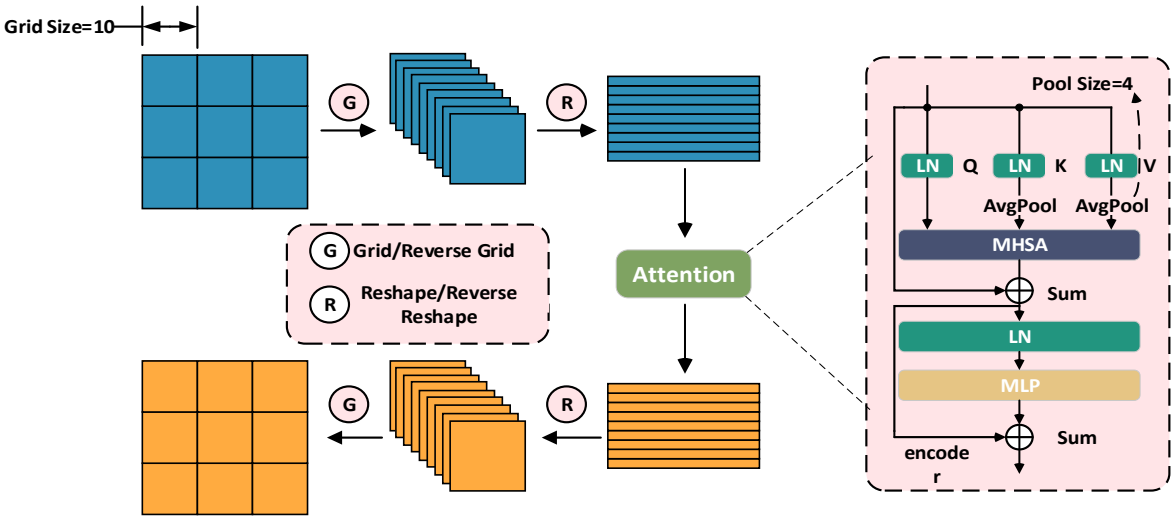


图 5　轻量化 Transformer 网络结构

Fig.5　Lightweight transformer network architecture

其权值分别为λ_1=1.5、λ_2=0.5、λ_3=7.5。δ_i 与 y_i 表示置信度对某个类别的真实概率；p 则表示预测概率值；I_{OU} 为预测框与真实框的交并比；ρ为预测框与真实框中心点的距离；c 为预测框与真实框交并矩形对角线长度；α为影响因子，它决定 v 预测框与真实框宽高比相似度的权重。

现阶段的 bbox 表示大多是通过对 bbox 方框狄拉克分布结果进行建模而形成的单一分布。狄拉克分布可以认为在一个点概率密度为无穷大，其他点概率密度为 0，这是一种极端地认为离散标签是绝对正确的。然而，在复杂场景下特别是 DOTA 遥感数据集中，很多检测目标的边界因背景的复杂性致使目标边界并非十分明确。DFL loss[25]通过采用任意分布来建模边界框，而后使用 softmax 函数对离散变量进行回归，通过将狄拉克分布的积分形式转化为一般积分形式来实现对于 bbox 边界框的表示，可以解决边界识别问题。

2　实验设置

2.1　实验环境与实验参数

本文实验使用 Ubuntu20.04 操作系统，采用 Pytorch1.10 深度学习框架，CUDA 版本为 11.3，显卡 GPU 采用 NVIDIA GeForce RTX 2080Ti 显存 11G。训练过程中采用随机梯度下降算法（stochastic gradient descent，SGD）训练 300 epoch。初始 learning rate 设为 0.01，batchsize 设为 8。在训练初期的 3 个 epoch 采用 warm up 进行训练，并在最后 10 个 epoch 关闭 Mosaic 操作。

2.2　数据集及预处理

本文使用的是 DOTA1.0 数据集。DOTA1.0 数据集由不同传感器所采集的 2806 幅航拍图像所构成。由于传感器的差异导致图像的画幅有所区别，图像的大小为 800×800 像素～4000×4000 像素之间，而一般数据集如 PASCAL-VOC 和 MSCOCO 图像尺寸都在 1000×1000 像素之间。DOTA 数据集，包含的图像在尺度、方向与形状上都存在差距，图像经由航天图像判读专家选取 15 个重要类别使用旋转框进行标注，包括：飞机、船舶、储罐、棒球场、网球场、游泳池、地面田径场、港口、桥梁、大型车辆、小型车辆、直升机、环形交叉口、足球场和篮球场。DOTA 数据集是遥感图像小目标检测的重要基线模型。数据集的类别分布如图 6 所示。

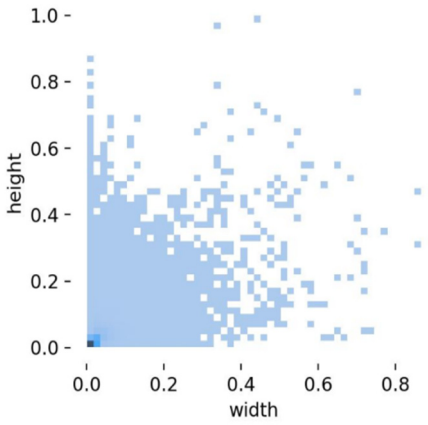


图 6　DOTA1.0 数据集不同类别标签数

Fig.6　The number of different category tags in the DOTA1.0 dataset

DOTA 数据集中小目标物体占据绝大多数，如图 7 所示，可以根据水平边界框的高度将数据集中的所有实例分为 3 部分：小范围为 10～50 像素占总数据集的 57%，中范围为 50～300 像素占总数据集的 41%，大范围为 300 像素以上仅占数据集比例的 2%。对于小型交通工具像素一般在 20 像素，而一座大桥可能所占像素点可以达到 1200 像素，但此类目标在数据集中所占比例比较小，因此模型必须足够灵活，才能够很好地同时处理微小目标与大目标。

2.3 数据处理

DOTA 数据集共包含训练集 1411 张、验证集 459 张、测试集 910 张。标注的图像实例为 188282 个。对于小目标的定义，COCO 数据集中采用绝对定义，将 32×32 像素及其以下目标定义为小目标，相对定义概念是将目标所占比例小于图像 0.1 的目标定义为小目标，如图 7 所示，显然 DOTA 数据集中小目标在以上定义标准下均符合小目标定义。

目前很多算法在进行数据预处理时将数据集通过设置一定的 Gap，将原图切分成 1024×1024 像素的图像，从而扩展数据集，但这种方式破坏了原始数据集的特性，使得原有图像在输入的时候实际上被放大了，降低了网络提取特征的难度，并且不同的 Gap 同样会对网络检测效果产生影响，这种方式虽然会大幅提升检测精度但同时也破坏了数据集的原始特性，因此本文对原数据集仅进行 Resize 处理，将其缩放为 640×640 像素图像进行网络输入，因比例问题造成的空白采用灰条补全，相较于多数 1000×1000 像素及其以上的网络输入大幅降低了推理时间。

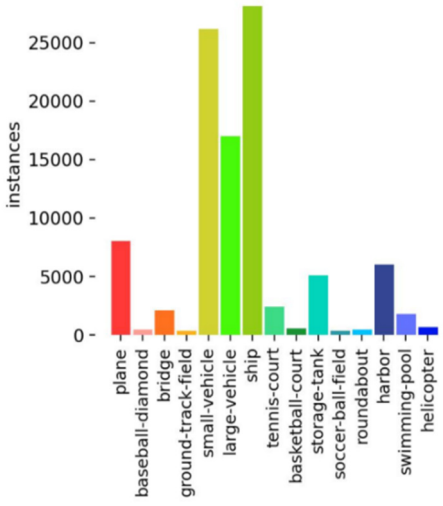


图 7 DOTA1.0 数据集目标标注框尺度分布

Fig.7 DOTA1.0 data set target label box scale distribution

3 实验结果与分析

本文所改进的基准模型为 YOLO v8s 模型，模型参数 11.2Params(M)、28.6FLOPS(B)。为证明本文所提出改进算法的有效性和轻量化效果，本文在 DOTA 数据集上进行了消融实验与对比试验，并在 VisDrone 数据集上进行迁移实验，验证算法对于不同小目标数据集的泛化性。

3.1 消融实验

实验结果如表 1 所示从上至下依次为 YOLO v8s 原算法、YOLO v8s 加改进 Transformer 算法（LVIT）、YOLO v8s 加改进多尺度加权融合交叉 FPN 算法（CROSS-FPN）、YOLO v8s 同时加入 LVIT 与 CROSS-FPN（FVIT-YOLO v8）的实验结果。

由表 1 可知与 YOLO v8 相比，CROSS-FPN 在基本不增加运算成本的条件下，大幅提升了检测速度，并且将 FPS 提升至 60.2 帧，同时还将 mAP 提升了 3.0%，同时加入 LVIT 后 mAP 提升了 1.4%，总体网络精度同比增长 4.4%，网络参数量降低为原始的 48%，虽然运算成本有了一定的增加，但模型总量依然较小，并且检测速度也能满足实时性的要求。

表 1 消融实验结果

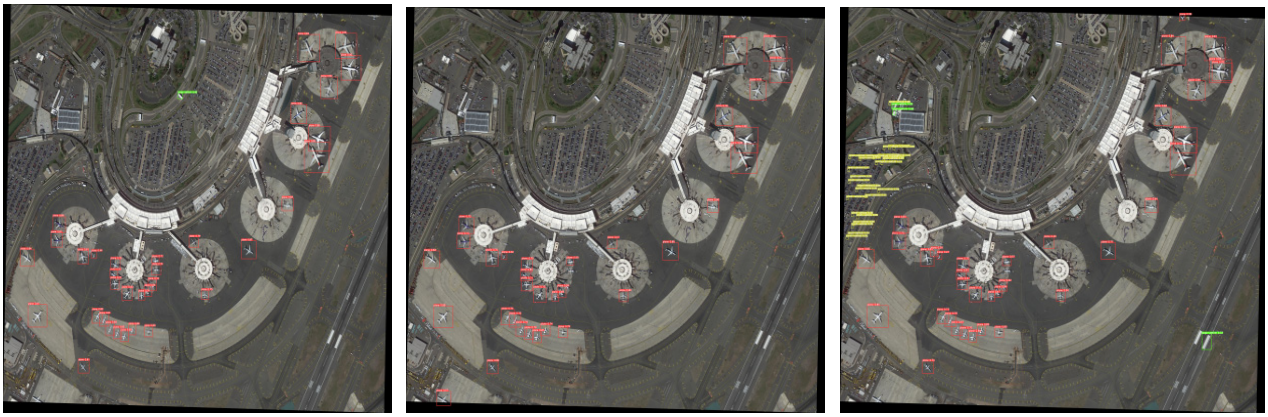
Table 1 Ablation results

Methods	mAP/(%)	Parameters/(M)	FPS
YOLO v8s	45.6	11.2	42.3
YOLO v8s+LVIT	46.0	16.8	48.7
YOLO v8s+ CROSS-FPN	48.6	3.46	60.2
YOLO v8s+LVIT+ CROSS-FPN	50.0	5.41	46.0

图 8 为本文改进算法的对比检测效果，可以看出改进的多尺度加权融合交叉 FPN 网络（CROSS-FPN）对于精度的影响更多，而添加全局注意力之后的网络检测出了 YOLO v8s 无法检测的小型交通工具，验证了本文改进的有效性。

3.2 对比实验

由表 2 可知，本文提出的 FVIT-YOLOv8 算法在 DOTA 数据集的每一个类别上都实现了精度的提升，值得注意的是在棒球场、桥梁、储油罐目标检测中 mAP 分别提升了 10.7%、10.4%、10.8%。对于桥梁目标因其横纵比原因导致检测难度大大提升，并且也同时存在背景模糊的问题。本文算法在实现相对于原网络轻量化的同时也相对提升了对各类目标的检测精度，相对于其他算法本文提出算法也表现出了良好的效果，基本持平 TPH-YOLO v5，但后者的整体参数量为本算法的 9 倍。



(a) YOLO v8s (b) YOLO v8s+CROSS-FPN (c) YOLO v8s+CROSS-FPN+LVIT

图 8 改进算法图像检测效果对比

Fig. 8 Comparison of detection effects of the improved algorithm

表 2 不同算法识别准确率对比

Table 2 Comparison of recognition accuracy of different algorithms

%

	SSD	R-FCN[26]	Faster R-CNN	YOLO v2[10]	YOLO v4	YOLO v5s	TPH-YOLOv5	YOLO v8s	(Ours)FVIT-YOLO v8
Plane	57.8	39.6	74.7	76.9	69.2	68.3	72.6	71.2	73.3
Baseball field	32.7	46.1	66.4	33.9	49	48.9	56.4	42.6	53.3
Bridge	16.1	3	14	22.7	16.2	15.9	25	9.3	19.7
Athletic field	18.7	38.5	63.7	34.9	29.3	28.4	34.7	33.3	37.6
Small vehicle	0.1	9.1	8.8	38.7	49.2	48	53.7	55.2	58.9
Large vehicle	36.9	3.7	38	32	71.2	70	77.6	78.7	79.6
Ship	24.7	7.5	13.2	52.4	48.2	46.4	61.3	58.3	63.9
Tennis court	81.1	42	84.6	61.7	88.7	88.1	90.2	90.2	91.7
Basketball court	25.1	50.4	53.2	48.5	35.7	34.6	41.9	36.4	38.4
Oil tank	47.4	67	17.4	33..9	23.5	22.5	38.6	28.9	39.7
Soccer field	11.2	40.3	57.3	29.3	33.6	32	36.3	33.6	36.1
Roundabout	31.5	51.3	28.2	36.8	14.6	14.2	16.9	12.6	16.4
Port	14.1	11.1	56.3	36.4	65.2	64.4	72.7	69.5	72.5
Swimming pool	9.1	35.6	25.7	38.3	42.7	40.8	49.4	37.8	43.1
Helicopter	0	17.5	27.8	11.6	28.8	27.8	32.8	26.1	26.7
mAP/(%)	29.9	30.8	42	39.2	44.2	43.2	50.7	45.6	50

3.3 检测结果可视化

如图 9 本文选取了不同场景下的检测效果图，可见本文改进算法 FVIT-YOLO v8 在检测精度与检出数量上均超过了 YOLO v8s 原网络并且对比(e)与(f)图，FVIT-YOLO v8 还检出了 YOLO v8s 未检出的游泳池。

为验证算法改进的有效性，本文通过对比 GradCAMPlusPlus，GradCAM，XGradCAM 三种工具对本网络模型的 Heat MAP 可视化效果，最终采用 GradCAM[27]对网络的第 9 层输出进行可视化操作，其中 conf_threshold 为 0.6，按置信度排序取前 2%的数据进行热力图计算，在反向传播中本文将 score＋box 同时进行反向传播，进而进行梯度求和。如图 10，其中热力图可反映网络的感兴趣区域，其颜色越深表示其注意程度越强，可以看出在进行网络改进之后本文算法对检测目标的注意力更为集中与准确，对比(a)与(b)可以看出，FVIT-YOLO v8 相较 YOLO v8s 对环

岛的注意力更为准确，同时也大幅降低了对其他无关目标的关注。对比(c)与(d)图可以看出，对于飞机的关注 FVIT-YOLOv8 要优于 YOLO v8s。

3.4 基于 VisDrone 数据集泛化性实验

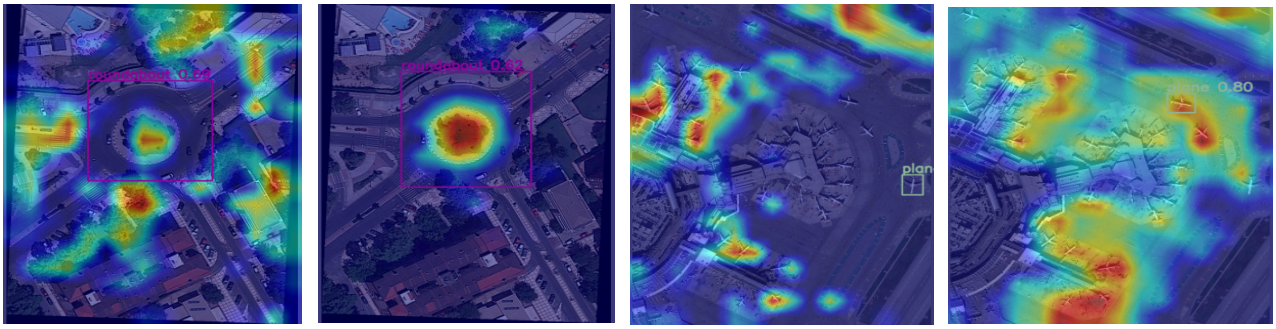
本文基于 VisDrone[28]数据集进行迁移实验，用于评估 FVIT-YOLO v8 对小目标数据集的泛化性能，VisDrone 数据集为无人机航拍数据集，主要类别为车辆与行人，属于小目标数据集。如表 3 所示 FVIT-YOLO v8 在 VisDrone 数据集上实现了 8.2%的精度提升，这表示 FVIT-YOLO v8 在无人机航拍小目标检测方面也实现了在网络参数量大幅降低的情况下，精度有较大提升，表现出对不同数据较强的泛化性。

(a) YOLO v8s results 1 (b) FVIT-YOLO v8 results 1 (c) YOLO v8s results 2

(d) FVIT-YOLO v8 results 2 (e) YOLO v8s results 3 (f) FVIT-YOLO v8 results 3

图 9 YOLO v8s 与 FVIT-YOLO v8 在不同场景的目标检测结果对比

Fig. 9 Comparison of object detection results of YOLO v8s and FVIT-YOLO v8 in different scenarios



(a) Roundabout(raw) (b) Roundabout(ours) (c) Plane(raw) (d) Plane(ours)

图 10 YOLO v8s 与 FVIT-YOLO v8 在不同场景热力图对比

Fig.10 Heatmap comparison of YOLO v8s and FVIT-YOLO v8 in different scenarios

表 3 基于 VisDrone 数据集的消融实验

Table 3 Ablation experiment of VisDrone dataset

Methods	mAP/(%)
YOLOv8s(baseline)	39.5
YOLOv8s+CROSS-FPN	46.2
YOLOv8s+CROSS-FPN+LVIT	47.7

本文同时也在 Visdrone 数据集中进行了检测效果对比试验，对于 Tph-YOLO v5 本文使用了其包含 P2 层与 Transformer 的算法进行对比试验，结果如表 4 所示。通过表 4 可以看出本文算法在众多算法中表现良好，并且很好地兼顾了轻量化与精确度。验证了本算法对不同数据良好的泛化性。

表 4 基于 VisDrone 数据集的不同算法识别准确率

Table 4 Algorithm comparison experiment on VisDrone dataset

Methods	mAP/(%)
RetinaNet[29]	21.37
RefineDet[30]	28.76
DetNet59[31]	29.23
Cascade-RCNN[32]	31.91
FPN[33]	32.20
Light-RCNN[34]	32.78
CornetNet[35]	34.12
Faster-RCNN	38.20
YOLO v5s	34.70
Tph-YOLO v5	42.10
Ours(FVIT-YOLO v8)	47.70

4 结论

针对遥感图像及无人机拍摄图像的目标检测，本文在改进 YOLO v8s 模型基础上进行改进，形成了 FVIT-YOLOv8 算法。该算法提出了一种双向多尺度融合交叉 FPN 网络，在检测层根据数据集小目标分布情况进行了针对化调整，集成了基于 Transformer 的自注意力机制，同时对其进行轻量化处理。FVIT-YOLO v8 相比于 YOLO v8s，参数量下降了 52%；精度在 DOTA 数据集上提升了 4.4%，在 VisDrone 数据集上提升了 8.2%，可促进遥感图像及无人机拍摄图像的小目标检测算法的工程化应用。

参考文献：

[1] Everingham M, Van Gool L, Williams C K I, et al. The pascal vision object classes (voc) challenge[J]. *International Journal of Computer Vision*, 2009, **88**: 303-308.

[2] LIN T Y, Maire M, Belongie S, et al. Microsoft coco: lofxol common objects in context[C]//13*th European Conference*, 2014: 740-755.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580-587.

[4] Girshick R. Fast r-cnn[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2015: 1440-1448.

[5] HE K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2961-2969.

[6] Purkait P, Zhao C, Zach C. SPP-Net: Deep absolute pose regression with synthetic views[J]. arXiv preprint arXiv:1712.03452, 2017.

[7] 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.

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

[9] WANG C Y, Bochkovskiy A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.

[10] HAN X, CHANG J, WANG K. Real-time object detection based on YOLO-v2 for tiny vehicle object[J]. *Procedia Computer Science*, 2021, **183**: 61-72.

[11] LIU W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//*Computer Vision–ECCV* 2016: 14*th European Conference, Amsterdam*, 2016: 21-37.

[12] ZHU M, XU Y, MA S, et al. Effective airplane detection in remote sensing images based on multilayer feature fusion and improved nonmaximal suppression algorithm[J]. *Remote Sensing*, 2019, **11**(9): 1062.

[13] DONG Z, LIN B. BMF-CNN: an object detection method based on multi-scale feature fusion in VHR remote sensing images[J]. *Remote Sensing Letters*, 2020, **11**(3): 215-224.

[14] ZHU H, ZHANG P, WANG L, et al. A multiscale object detection approach for remote sensing images based on MSE-DenseNet and the dynamic anchor assignment[J]. *Remote Sensing Letters*, 2019, **10**(10): 959-967.

[15] ZHANG X, ZHU K, CHEN G, et al. Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network[J]. *Remote Sensing*, 2019, **11**(7): 755.

[16] ZHUANG S, WANG P, JIANG B, et al. A single shot framework with multi-scale feature fusion for geospatial object detection[J]. *Remote Sensing*, 2019, **11**(5): 594.

[17] CHENG G, SI Y, HONG H, et al. Cross-scale feature fusion for object detection in optical remote sensing images[J]. *IEEE Geoscience and Remote Sensing Letters*, 2020, **18**(3): 431-435.

[18] 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.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in Neural Information Processing Systems*, 2017, **30***:* 6000-6010.

[20] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 2778-2788.

[21] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 10012-10022.

[22] XIA G S, BAI X, DING J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 3974-3983.

[23] Targ S, Almeida D, Lyman K. Resnet in resnet: generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.

[24] HAN K, XIAO A, WU E, et al. Transformer in transformer[J]. *Advances in Neural Information Processing Systems*, 2021, **34**: 15908-15919.

[25] LI X, WANG W, WU L, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection[J]. *Advances in Neural Information Processing Systems*, 2020, **33**: 21002-21012.

[26] DAI J, LI Y, HE K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. *Advances in Neural Information Processing Systems*, 2016, **29***:* 379-387.

[27] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Vision explanations from deep networks via gradient-based localization[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2017: 618-626.

[28] Visdrone Team. Visdrone2020leaderboard [EB/OL][2020-07-10]. http://aiskyeye.com/ visdrone-2020-leaderboard/.

[29] CHENG X, YU J. RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection[J]. *IEEE Transactions on Instrumentation and Measurement,* 2020, **70**: 1-11.

[30] ZHANG S, WEN L, BIAN X, et al. Single-shot refinement neural network for object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 4203-4212.

[31] LI Z, PENG C, YU G, et al. Detnet: a backbone network for object detection[J]. arXiv preprint arXiv:1804.06215, 2018.

[32] CAI Z, Vasconcelos N. Cascade r-cnn: delving into high quality object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio*n, 2018: 6154-6162.

[33] 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.

[34] LI Z, PENG C, YU G, et al. Light-head r-cnn: In defense of two-stage object detector[J]. arXiv preprint arXiv:1711.07264, 2017.

[35] Law H, DENG J. Cornernet: detecting objects as paired keypoints[C]//*Proceedings of the European Conference on Computer Vision* (ECCV), 2018: 734-750.