DOI:10.19651/j.cnki.emt.2417494

基于改进 RT-DETR 的轻量航拍图像检测算法*

张淑卿 肖 凡 葛 超
(华北理工大学电气工程学院 唐山 063210)

摘 要:针对航拍遥感图像场景中目标体积小、背景复杂的问题,提出了一种基于 RT-DETR 改进的轻量化目标检测算法 ELS-RTDETR。该算法提出并使用一种基于 Vovnet 网络改进的新主干网络 LOB-Vovnet 对原主干网络进行替换。在 LOB-Vovnet 中,设计提出了一种新的特征增强模块 LRFF,提高检测模型对小目标的检测精度。同时为抑制复杂背景干扰,引入自适应通道提取的注意力机制 SE。最后为均衡模型精度与体积,LOB-Vovnet 将部分卷积替换为深度可分离卷积,并通过进行大量消融实验,对主干网络的深度和宽度重新调整。在 AIFI 中,引入级联群体注意力机制(CGA)有效减少多头注意力机制中的计算冗余。在数据集方面将 RSOD 数据集和 NWPU VHR-10 数据集进行融合,并通过添加仿射变换、相机底噪等效果对原始数据进行离线数据增强,使训练数据集更贴近真实应用场景。实验结果表明,改进模型 ELS-RTDETR 与原模型对比 mAP@50 提升 2.7%,模型参数量减少了 32.9%,面对困难检测目标实现了较好的检测效果,在 SIMD 数据集上进一步验证了改进方法的有效性。

关键词:航空遥感;目标检测;深度学习;RT-DETR;轻量化
中图分类号:TP391.41;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.6040

Lightweight aerial image detection algorithm based on improved RT-DETR

Zhang Shuqing Xiao Fan Ge Chao
(School of Electrical Engineering, North China University of Science and Technology, Tangshan 063210, China)

Abstract: In response to the challenges posed by small target volumes and complex backgrounds in aerial remote sensing images, a lightweight object detection algorithm named ELS-RTDETR, based on enhancements to RT-DETR, has been proposed. This algorithm introduces and utilizes a new backbone network called LOB-Vovnet, which is an improved version based on the Vovnet network, to replace the original backbone network. Within the LOB-Vovnet architecture, a novel feature enhancement module named LRFF (Lightweight receptive field focus) has been designed to enhance the detection accuracy of small targets. To address complex background interference, an attention mechanism called SE (Squeeze-and-Excitation) based on adaptive channel extraction has been introduced. To strike a balance between model accuracy and size, LOB-Vovnet replaces some convolutions with depthwise separable convolutions. Extensive ablation experiments have been conducted to readjust the depth and width of the backbone network. In the AIFI section, a Cascaded Group Attention (CGA) mechanism has been introduced to effectively reduce computational redundancy in multi-head attention mechanisms. Regarding datasets, the RSOD dataset and NWPU VHR-10 dataset have been merged. Additionally, offline data augmentation techniques such as affine transformations and camera noise have been applied to the original data to make the training dataset more closely aligned with real-world applications. Experimental results indicate that the improved ELS-RTDETR model has shown a 2.7% increase in mAP@50 compared to the original model, with a reduction in model parameters by 32.9%. It has demonstrated good detection performance for challenging targets. Further validation of the enhanced method has been conducted on the SIMD dataset to verify its effectiveness.

Keywords: aerial remote sensing; target detection; deep learning; RT-DETR; lightweight

0 引 言

近年来,无人机技术得到了迅速发展,在农业检测、航拍摄影和安全监控等领域发挥着愈发重要的作用。得益于无人机机动性强、活动空间大的优势,工作人员可以在较短时间内快速获得广泛地区中丰富的图像信息。然而与其他

收稿日期:2024-11-30
* 基金项目:河北省自然科学基金(F2021209006)项目资助

常规目标检测场景相比，航拍图像具有检测目标体积小、背景复杂和尺度跨越大等特点，严重影响了目标检测的精度。在实际应用，检测模型往往被部署在资源受限的端侧平台上，这对模型的体积提出了限制要求。如何在兼顾模型体积的同时提高检测精度是一个具有挑战性的研究问题。

自 2012 年 AlexNet 在 ImageNet 中使用卷积神经网络(convolutional neural networks，CNN)夺得冠军以来，基于 CNN 的深度学习模型迅速取代传统机器学习成为目标检测领域的主要方法。文献[1]提出的以 Faster R-CNN 为代表的两阶段检测算法表现出优秀的检测精度。白宗宝等[2]在 Faster R-CNN 主干网络输出端引入一种三叉戟注意力机制有效提升了检测精度。然而高昂的计算成本和繁琐的训练过程阻碍了两阶段模型在多个领域的进一步应用，尤其不适用于资源受限的端侧平台。随后，文献[3]提出的 SSD 算法和文献[4-7]提出的 YOLO 系列算法为代表的单阶段目标检测模型得到了广泛的应用，与双阶段检测算法相比，其以较低的计算开销获得了令人满意的检测精度，取得了在速度和精度之间的均衡。赵耘彻等[8]提出一种基于 YOLOv4 的航拍图片识别方法，其通过增加浅层检测层提高了检测精度。徐光达等[9]通过增加浅层网络的高分辨率特征图和加入对应尺寸检测头，提出了一种基于 YOLOv5 改进的无人机航拍图像检测算，新算法有效提高对小体积目标的检测精度。徐坚等[10]针对航拍图像中图像视角变化大检测困难的问题，使用可变形卷积和特征平衡金字塔对 YOLOv5 算法进行改进，严苛条件下的漏检和错检情况得到改善。CNN 结构在对长距离信息的捕捉上存在劣势，并且在检测过程中依赖以 NMS(non-maximum suppressio)为代表的后处理技术对冗余边界框进行消除，严重影响了检测效率。CenterNet 摆脱了对 NMS 的依赖，许延雷等[11]通过设计自适应阈值预测分支在 CenterNet 的基础上进行改进，新算法取得了良好的检测精度。

近些年，文献[12]提出的 Transformer 在计算视觉领域异军突起，得到了广泛的应用。与 CNN 网络相比，其能充分捕获长距离信息，有效减少采样过程中特征信息的丢失，使目标检测精度得到较高的性能提升。陈朋磊等[13]使用 Swin Transformer 作为 RetinaNet 主干网络，有效增强算法对全局信息的提取能力，提高了模型的检测精度。文献[14]提出的 DETR 是首个基于 Transformer 结构的端到端的稀疏检测模型。DETR 使用编码器直接输出目标边界框和类别，省去传统检测算法中以 NMS 为代表的后处理步骤，极大简化模型算法的检测流程。文献[15]提出的 Group DETR 针对 DETR 模型训练成本高、收敛慢的缺点，通过引入多个 Object queries 加快了模型的收敛速度。文献[16]提出的 DAB-DETR 算法提出了动态锚框概念，通过对 Transformer 解码器进行逐层动态更新，加快了 DETR 模型训练的收敛速度。文献[17]提出的 Deformable DETR 将可变形注意力机制引入模型结构中，使检测模型聚焦于图像中的关键区域，有效减少了 DETR 的计算开销。Zhang 等[18]在先前工作的基础上提出了 DINO，通过使用对比去噪训练和混合查询选择在有效降低模型参数量的同时提高了模型的训练效率。上述工作虽然针对 DETR 模型进行了一系列改进。由于 Transformer 结构的约束，模型在参数量和计算开销方面无法与 YOLO 系列进行比较。

针对上述问题，Zhao 等[19]提出了轻量化的实时检测模型 RT-DETR(Real-Time Detection Transformer)，该模型采用新的 IoU 感知查询机制，使模型能够更高效的对目标特征信息进行提取。RT-DETR 具有相较于其他 Transformer 检测模型更少的参数量和更快的检测速度，在实时性和检测精度上均优于同等大小的 YOLO 模型。

本研究以 RT-DETR 模型为基础，针对航拍图像中小体积和背景复杂目标检测困难的问题进行改进，在提高检测精度的同时，对模型参数量进行压缩使模型更适用于资源受限的端侧平台，主要工作如下：

1)原模型使用 Resnet 结构作为主干网络，针对 Resnet 存在大量冗余层结构的问题，在 Vovnet[20]的基础上进行改进，提出了新主干网络 LOB-Vovnet，使用其替换原主干网络 Resnet。

2)针对航拍图像中目标体积小检测困难的问题，在 LOB-Vonvnet 提出并使用了轻量级特征增强模块 LRFF。通过使感受野中新尽可能落在目标区域内，增加对于小目标的预测样本，从而提升模型对于小目标的检测识别能力。

3)针对航拍检测场景下目标背景复杂的问题，在 LOB-Vonvnet 中引入自适应通道提取的注意力机制 SE 模块，通过建模卷积特征通道之间的相互依赖性，抑制复杂背景对于检测目标的干扰。

4)在 AIFI 部分，引入级联群体注意力机制 CGA，通过增强图像特征之间的交互提高注意力图的多样性和计算效率。

5)仿照文献[21]提出的 Convnext，通过消融实验，使用合理的优化策略对网络结构进行重新调整，将部分卷积替换为深度可分离卷积，并遵循 MAC 最低设计思想对网络的深度和宽度进行调整。

最终实验结果表明改进模型在 RSOD 和 NWPU VHR-10 混合数据集、SIMD 数据上较原模型均表现出色。

1　RT-DETR 网络架构

RT-DETR 模型是基于 CNN 和 Transformer 混合结构的单阶段实时目标检测模型，并有 R18、R50、L 和 X 等版本。考虑到航拍图像检测场景对于模型参数量与实时性的限制要求，选择 RT-DETR-R18 作为基线算法。其网络结构主要由主干网络(Backbone)、AIFI(单尺度内的特征交互)、CCFM(跨尺度的特征融合)和改进解码器(Transformer decoder)四部分组成。模型结构图如图 1 所示。

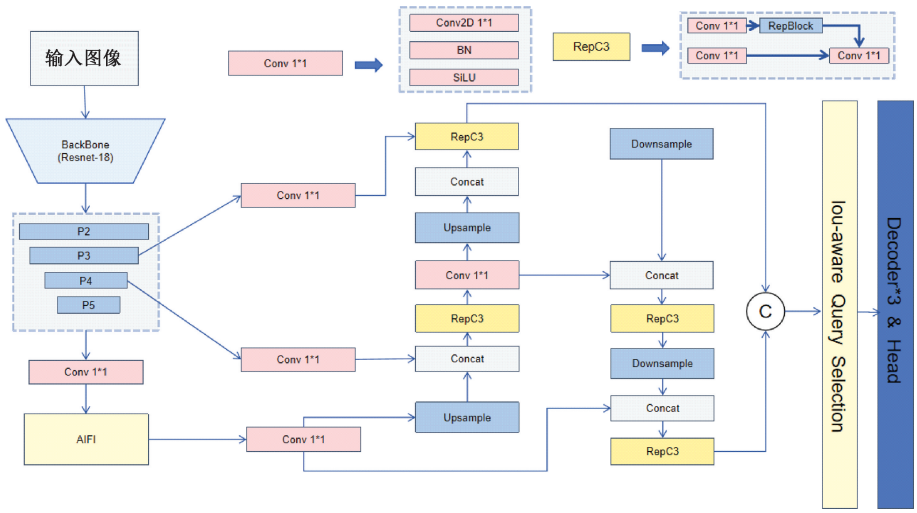

图 1 RT-DETR 网络结构图

Fig. 1 RT-DETR network structure

2 ELS-RTDETR 网络结构

RT-DETR 作为 DETR 系列首个轻量化模型,其计算效率有所提高,Attention 机制使模型获得全局感受野和样本间信息交换的能力。相较于纯 CNN 网络结构,RT-DETR 实现了图像特征的稀疏采样(sparse sampling)和端到端检测(end-to-end detection)。

然而 RT-DETR 在面对目标像素信息模糊、尺度变化大和背景复杂等困难检测目标时容易发生漏检和误判,严重影响模型检测精度。同时由于 Transform 结构的使用,RT-DETR 模型参数量较 CNN 模型结构仍稍显臃肿。由于边缘端硬件平台计算性能的限制,需要使用更高效的网络结构对 RT-DETR 进行改进以在兼顾检测精度的同时尽可能压缩模型体积。基于上述问题,本文提出一种改进模型 ELS-RTDETR,模型结构图如图 2 所示。

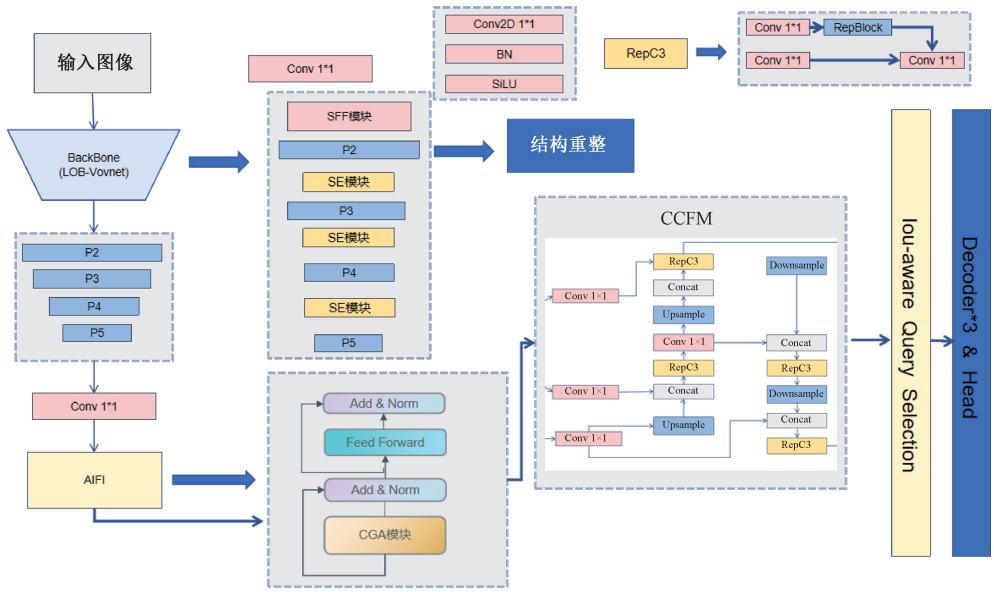

图 2 ELS-RTDETR 网络结构图

Fig. 2 Network structure of ELS-RTDETR

本文使用基于 Vovnet 改进的的新主干网络 LOB-Vovnet 替换原主干网络 Resnet。Vovnet 网络通过使用一次聚合(one-shot aggregation,OSA)模块在继承文献[22]提出的 DenseNet 多感受野表示多种特征优点的同时解决了密集连接效率地下的问题,在速度和效率上由于 ResNet 网络。

LOB-Vovnet 提出并使用了一种新的特征增强模块 LRFF 提高了对于小体积模板的检测精度。针对检测背景复杂的问题,LOB-Vovne 引入了自适应通道提取的注意力机制 SE。最后通过大量消融实验,对网络结构进行了优化

改进，将部分卷积替换为深度可分离卷积并对模型的深度和宽度进行了调整。

在 AIFI 部分，本文引入级联的分组注意力模块 CGA 通过提高注意力图的多样性，使模型能够学习到更加丰富的特征信息。

在航拍遥感图像检测中模型网络往往被部署在无人机等硬件资源受限的边缘平台上，与其他常规检测场景不同，航拍图像检测在要求检测精度的同时也对模型体积作出了限制。本文提出的 ELS-RTDETR 检测模型，通过设计改进，在有效提升模型对于困难目标检测精度的同时兼顾应用场景对模型体积的限制，压缩模型体积，实现了模型的轻量化。

2.1　LOB-Vovnet 主干网络

在目标检测任务中，主干网络是基本的特征提取器，主要负责对原始图像进行特征提取，直接影响着模型对于目标的检测效率与准确度。原 RT-DETR 模型使用的主干 Resnet 与轻量级主干模型相比参数量较大，存在冗余层结构，限制了其在计算资源有限的端侧平台上的应用使用。为更好的均衡检测精度与效率。本文提出了一种基于 Vovnet 网络改进的新主干网络 LOB-Vovnet。

1)LRFF 模块设计

小体积目标检测长期以来是目标检测中的一个难点。遵循使尽可能多的感受野中心落在目标区域内的设计准则，提出一种新的轻量级特征增强模块 LRFF 用于提高模型对于小体积目标的检测精度。

感受野理论上是指卷积神经网络输出的特征图上的像素点映射回输入图像区域上的大小，如图 3 所示为感受野示意图。

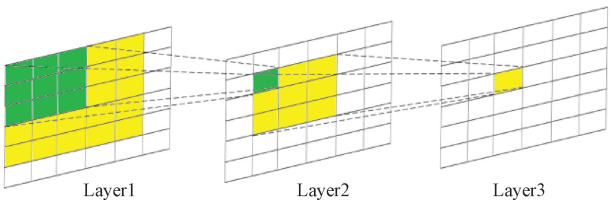

图 3　感受野示意图

Fig. 3　Schematic diagram of receptive field

Layer1 中方格每个方格为一个元素，左上角的 3×3 方格表征为一个 3×3 的卷积核。Layer2 由一个 3×3 的卷积核经过卷积运算输出，输出尺寸为 3×3(stride=1，padding=0)，layer2 中由虚线交叉得到的左上角单一方格是由 layer1 中 3×3 的方格经卷积运算得到的，layer2 中的左上角单一方格蕴含了 layer1 中左上角 3×3 区域的特征信息，即感受野为 layer1 中 3×3 区域的部分。

理论感受野大小的计算式如式(1)所示。

$$R_x = R_{x-1} + (K_x - 1) * \prod_{i=0}^{x-1} S_i \tag{1}$$

其中，R_{x-1} 为第 $X-1$ 层的感受野大小，K_x 为第 X 层的卷积核大小，S_i 为第 i 层的卷积步长。

不同感受野区域对于特征学习的贡献并不相同，由此有效感受野的概念被提出。有效感受野(effective receptive field，ERF)是指在卷积计算时，实际有效的感受野区域。有效感受野是一种超参数，无法像理论感受野那样被精确计算，但一般认为感受野中心的像素对输出有更大的影响。

较大的感受野可以更好的提取目标的上下文背景信息，但不可避免的会使原图像上的感受野中心变得稀疏，不利于小体积目标的检测。感受野中心越密集，目标被命中的概率也就越大。每个命中目标区域的 Point 可以被认为是检测模型学习的最小单元。通过增多小目标的学习样本数，进而增强对目标的特征表达能力，从而最终提高对于小体积目标的检测精度。

感受野中心命中边界框(bounding box)区域的数量可以由式(2)得出。

$$N_{hit} = \frac{W_{bbox}}{S_w} * \frac{H_{bbox}}{S_h} \tag{2}$$

其中，W_{bbox} 和 H_{bbox} 分别是 Bbox(目标框)的宽和高，S_w 和 S_h 分别是特征图在宽和高方向上的步长。从式(2)中可得 Srtide(步距)对感受野大小起决定性作用。一般认为在相同大小感受野的情况下，网络越深对于目标特征的学习表达能力越强。

感受野对于目标检测至关重要，通过显著增多小目标的学习样本数，能有效提高模型的检测精度。基于上述理论，本文提出了一种特征增强模块 LRFF，同时为有效避免下采样过程中的信息丢失，LRFF 模块被放置在 2 倍下采样后，用以替换原 Vovnet 中的 stem1 部分，stem1 结构如图 4 所示。

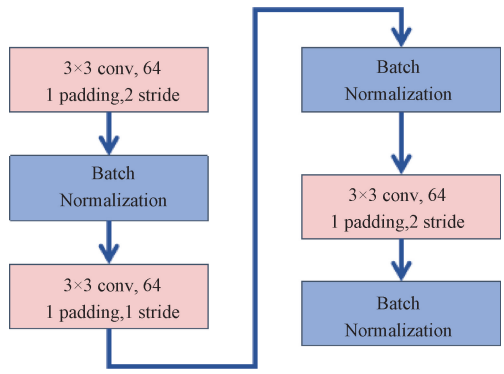

图 4　Steam1 结构示意图

Fig. 4　Steam1 structure diagram

LRFF 模块使用倒瓶颈结构、深度可分离卷积和残差连接，不引入复杂的网络结构、不额外增加模型参数量，模块结构如图 5 所示。

LRFF 模块在模型结构中处于靠前的位置，位于图像 2 倍下采样后，较低的下采样倍数在减少图像特征信息丢

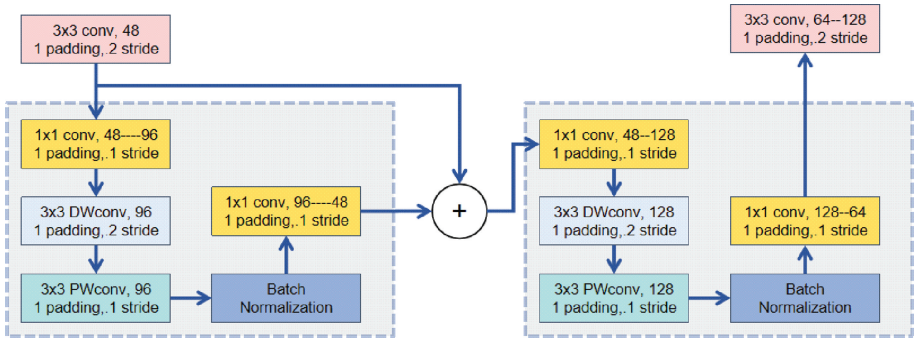

图 5 LRFF 结构示意图

Fig. 5 LRFF structure diagram

失的同时带来了很大计算开销。因此 LRFF 模块使用由 DW 卷积和 PW 卷积组成的深度可分离卷积替代普通卷积,解决了计算开销较大的弊端。普通卷积和深度可分离卷积计算量和参数量计算式如式(3)所示。

$$\begin{cases} Cp_c = D_K \times D_K \times M \times N \\ Pq_c = D_K \times D_K \times M \times N \times D_F \times D_F \\ Cp_{dsc} = D_K \times D_K \times M + M \times N \\ Pq_{dsc} = D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \end{cases} \tag{3}$$

其中,D_k 为卷积核大小,M 和 N 分别为输入和输出通道,D_F 为输出的特征图尺寸,Cp_c 和 Pq_c 分别标准普通卷积的计算量和参数量,Cp_{dsc} 和 Pq_{dsc} 分别为深度可分离卷积的计算量和参数量。由式(3)可得普通卷积的参数量和计算量是深度可分离卷积的 9 倍。

深度可分离卷积的引入在降低参数量和计算量的同时回带来一定的精度损失。针对此问题,LRFF 模块通过使用倒瓶颈结构、扩大通道数、残差连接弥补检测精度的损失。LRFF 模块通过恰当的结构设计在不增加参数量、不额外引入计算开销的同时,在相同感受野情况下加深了网络结构,提高模型对于小目标的检测精度。

2)SE 模块

注意力机制在目标检测网络中被广泛使用,在航拍图像检测场景下,图像的拍摄视角属于高空俯视角,具有视角大、信息多的特点,复杂多变的背景信息提高了目标检测的难度。

LOB-Vovnet 引入了自适应通道提取的注意力机制 SE。SE 是一种轻量高效的注意力模块,其结构如图 6 所示。

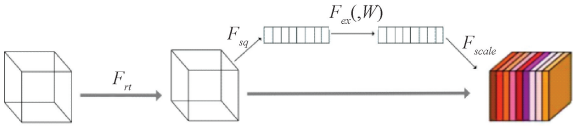

图 6 SE 结构示意图

Fig. 6 Schematic diagram of SE structure

图 6 中 F_{rr} 作为标准的卷积算子对特征图进行转换,随后通过 F_{sq} 和 F_{es} 分别进行全局信息嵌入和自适应重新校正对图像特征信息进行再次加工,最终使用 Scale 操作对信息进行加权。通过显式地建模卷积特征通道之间的相互依赖性,SE 模块改善了卷积神经网络中通道间信息的传递效率,有效提高了网络的表示能力,抑制了复杂背景对检测目标的干扰,其详细计算过程如式(4)所示。

$$X' = Scale(X) = X \cdot sigmoid(W_2 \cdot ReLU(W_1 \cdot Pool(X))) \tag{4}$$

SE 模块使用平均池化层将全局空间信息进行压缩,特征图被压缩为一个特征向量,解决了通道间依赖性的问题。随后被压缩的特征向量通过由两层全连接构成的门机制和 Relu 激活函数得到特征图中每个 Feature map 中权重。

SE 模块将不同通道之间的权重交由网络训练学习后自适应分配,相较于人工设计权重的方式,SE 模块能在训练过程中对通道特征反复校准大大提高了网络的表示能力。

2.2 CGA 模块

RT-DETR 的 AIFI(单尺度内的特征交互)部分使用 Transform Encoder 对特征进行再次编码。Transform 结构的使用在引入全局特征信息,提高全局上下文建模能力的同时带来了与 Cnn 结构相比更大的计算开销。庞大的计算开销不利于模型在资源有限的端侧硬件平台上的部署应用。因此如何降低 Transform Encoder 中的计算冗余,提高模型的计算效率和性能是一个至关重要的问题。

针对多头自注意力(MHSA)中的冗余问题,引入级联的分组注意力模块 CGA,其结构如图 7 所示。

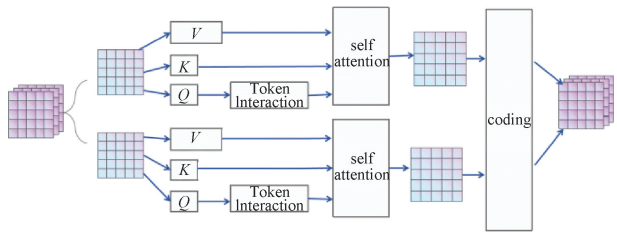

图 7 CGA 结构示意图

Fig. 7 Schematic diagram of CGA structure

CGA 模块将输入的图像特征进行分割，每一个被分割出的特征信息被输入进一个注意力头中。每个头完成注意力计算后，所有头的输出被级联汇总到一起，通过一个线性层投影回原输入维度形成最后输出。GCA 通过跨头级联计算输出特征信息的方式在减少多头注意力中计算冗余的同时，使模型网络深度增加进而提高了模型的泛化学习能力。CGA 计算式如式(5)～(7)所示。

$$\widetilde{X}_{ij} = Attn(X_{ij}W_{ij}^{Q}, X_{ij}W_{ij}^{K}, X_{ij}W_{ij}^{V}) \tag{5}$$

$$\widetilde{X}_{i+1} = Concat\left[\widetilde{X}_{ij}\right]_{j=1:h} W_{i}^{P} \tag{6}$$

$$X'_{ij} = X_{ij} + \widetilde{X}_{i(j-1)},\ 1 < j \leqslant h \tag{7}$$

其中，$\widetilde{X}_{ij}$ 表示输入特征 X_i 的第 j 次分割，W_{ij}^{Q}，W_{ij}^{K}，W_{ij}^{V} 是各不相同的投影层，W_{i}^{P} 是将输出特征投影回于输入相一致维度的投影层。X'_{ij} 是各个注意力头输出特征的和。

通过使用 CGA 注意力模块替代原模型的多头注意力机制，输入到注意力头图像特征的多样性得到了增强。与原先的自注意力机制不同，CGA 为每个头提供了不同的输入分割。多头级联的设计结构增加了图像特征之间的交互，提高了注意力图的多样性和 Transformer Encoder 的计算效率，在减少计算冗余的同时提高了模型的检测精度，使目标检测模型更适用于航拍图像检测场景。

2.3　模型结构重整

航拍图像检测对模型的识别能力与体积均提出要求。一方面航拍图像中检测目标具有体积小、背景复杂等特点，要求模型面对上述困难检测目标具有较好的识别能力。另一方面由于航拍图像检测场景具有端侧部署的特殊需要，受限于硬件平台有效的计算资源，要求对模型参数量与计算开销作出限制。因此如何在不引入额外计算开销的同时提高模型的检测精度成为了一个重要的问题。

ConvNeXt 中提出即使不在网络框架和搭建思路上做重大创新，仅在现有网络结构上进行恰当调整改进仍然能使检测精度有较大幅度的提高。

本文将模型结构重整的重点放在参数量与计算开销占比最大的主干网络部分，使用 Vovnet 网络替代原模型中使用的 Resnet 网络。原 Vovnet 网络结构如表 1 所示。

Vovnet 由 Stem stage1 和 4 个 OSA 模块组成，在 Stage 中，Vovnet 使用 3 * 3 的普通卷积。本文通过大量消融实验，遵循最低 MAC 准则，依次从宏观设计、深度可分离卷积和细节设计对原 Vovnet 网络的深度和宽度进行了调整，并将 OSA 模块中的部分卷积替换为由 DW 卷积和 PW 卷积组成的深度可分离卷积。

1)宏观设计

根据每个阶段的计算占比，重新调整每个网络块的重复次数进而改变网络模型深度。分别参考基线模型和 SOTA 模型的设计理念，在 ResNet-50 中网络块的重复次数为(3,4,6,3)，而在文献[23]中提出的 Swin-Transforme

表 1　Vovnet 网络

Table 1　Vovnet network

Type	Vovnet-19
Stem Stage1	3 * 3conv,64,stride=2 3 * 3conv,64,stride=1 3 * 3conv,64,stride=1
OSA Module Stage2	[3 * 3conv,128,*3 Concat * 1 conv,256] * 1
OSA Module Stage3	[3 * 3conv,160,*3 Concat * 1 conv,512] * 1
OSA Module Stage4	[3 * 3conv,192,*3 Concat * 1 conv,768] * 1
OSA Module Stage5	[3 * 3conv,224,*3 Concat * 1 conv,1024] * 1

中网络块的重复次数为(1,1,3,1)。本文通过消融实验分别比较了网络块重复次数为(1,1,1,1)、(1,1,2,2)、(2,1,1,2)、(1,2,1,2)4 种情况，实验结果如表 2 所示。

表 2　网络块消融实验

Table 2　Network block ablation experiment

网络块重复次数	mAP@50/%	参数量/M	计算量/G
(1,1,1,1)	80.79	10.96	51.5
(1,1,2,2)	80.90	12.72	54.6
(2,1,1,2)	80.19	12.17	61.0
(1,2,1,2)	79.74	12.41	57.6

从表 2 中可以看到检测精度和参数量、计算量并不具有因果关系，性能的提升不完全依托于 FLOPs 的增加，合理设计网络结构对于提高检测、减少计算冗余具有着重要作用。关于各个网络块的计算量分配，并没有理论上的参考，但在文献[24]提出的 RegNet 和文献[25]提出的 EfficinetNetV2 论文中均指出靠后的网络块应占用更多的计算量，本文实验结果符合这一次结论，最确定终将网络块结构调整为(1,1,2,2)。

2)深度可分离卷积的使用

文献[26]提出的 ResNeXt 提出通过采用 Group Conv(分组卷积)提升性能，ConvNeXt 在此基础上更进一步使用深度可分离卷积替代普通卷积在损失一定检测精度的同时大幅度压缩了模型体积。由于 Vovnet 网络相较于其他主干网络，其结构具有更多的通道数。ShufflenetV2[27]中指出较大的网络宽度有利于减少图像信息的损失。依托于这一特性，使用深度可分离卷积替代部分普通卷积并扩增通道数，最终消融实验结果如表 3 所示。

在 Stgae2 中使用深度可分离卷积替代普通卷积检测精度仅下降了 0.89%，而模型的计算量下降了 10.37%，大

表 3 深度可分离卷积消融实验

Table 3 Depthwise separable convolution ablation experiment

深度可分离卷积替换	mAP@50/%	参数量/M	计算量/G
不替换	81.45	13.89	64.6
Stgae2	80.72	13.39	57.9

大降低了计算开销,提高了检测效率。除替换 Stage2 外,本文另分 3 组进行实验,分别将不同 Stage 的普通卷积替换为深度的可分离卷积。第 1 组替换 Stgae1、Stage2 处卷积;第 2 组替换 Stage1、Stage2、Stage3 处卷积;第 3 组将所有卷积替换为深度可分离卷积。由于这 3 组实验其余变量未控制一致,故不以表格方式详细列出。最终替换 Stage1、Stage2 和 Stage3 中第 1 个 Block 处卷积的实验效果最佳,实现了检测精度和计算量之间的均衡。

3)细节调整

对 Vovnet 网络中各 Stgae 的通道数进行了重新调整。原 Vovnet 网络中各 Stgae 中 3 * 3 卷积的通道数分别为 128、160、192 和 224,1 * 1 卷积的通道数分别为 256、512、768 和 1 024。

经过调整后网络中各 Stgae 中 3 * 3 卷积的通道数分别为 102、160、96 和 112,1 * 1 卷积的通道数分别为 204、512、384 和 512。通过降低网络的通道数能大幅度减少模型的计算开销,但网络的宽度直接影响着模型能否学习到更高细粒度的特征信息,EfficientNetV2 中指出减少网络宽度会使得模型变得难以训练。为尽量避免通道数减少带来的负面影响,在调整网络时 Stage1 中 3 * 3 卷积的通道数只进行了小幅减少,Stage2 的通道数保持不变。Stgae3 和 Stgae4 获得的信息是经网络高度抽象后的图像特征信息可以由较低的通道数进行表征,因此通道数的降低主要在 Stgae3 和 Stage4 中。

最终经过结构重整后的主干网络模型如表 4 所示。

表 4 重整后 Vovnet 网络

Table 4 Reorganized vovnet network

Type	Vovnet-19
Stem Stage1	LRFF 模块
OSA Module Stage2	$\begin{bmatrix}3*3\text{PW、DW},102,*3\\ \text{Concat}*1\ \ \text{conv},204\end{bmatrix}*1$
OSA Module Stage3	$\begin{bmatrix}3*3\text{PW、DW},160,*3\\ \text{Concat}*1\ \ \text{conv},512\end{bmatrix}*1$
OSA Module Stage4	$\begin{bmatrix}3*3\text{PW、DW},96,*3\\ \text{Concat}*1\ \ \text{conv},384\end{bmatrix}+\begin{bmatrix}3*3\text{conv},96,*3\\ \text{Concat}*1\ \ \text{conv},384\end{bmatrix}$
OSA Module Stage5	$\begin{bmatrix}3*3\text{PW、DW},112,*3\\ \text{Concat}*1\ \ \text{conv},512\end{bmatrix}*2$

3 实验结果与分析

3.1 数据集

本文将 RSOD 数据集和 NWPU VHR-10 数据集进行融合,并在此基础上通过基于 python 下 Albumentations 库对原图像进行数据增强,通过随机引入仿射变换、相机底噪、雨天天气等效果使图像数据更加贴近实际环境,最后将图像增强后的数据最后作为实验数据集。

RSOD 数据集是一个用于遥感图像目标检测的开放数据集,其影像大小多为 1 044 pixel×915 pixel,包含四类检测对象共计 976 张图像。NWPU VHR-10 是一个公开可用地理空间对象检测数据集,其影像大小多为 800 pixel×600 pixel,包含 10 类对象共计 800 张图像。

将混合后的数据集按 1∶1.5 比例进行数据增强,仿射变换、相机底噪、雨天天气等效果按设定比例对原图像进行数据增强,在增强过程多种效果可能会叠加出现在同一张图片上,从而造成图像特征严重失真。为确保训练严谨性在初次训练开始前,人工筛选除去畸变严重、失去原本类别特征的图像数据得到包含 13 类检测对象的共 2 209 张图像,按 8∶1比例划分为训练集和验证集。数据集中每个类别分布如图 8 所示。混合后原数据的图像尺寸不变,主要分为两种:1 044 pixel×915 pixel 和 800 pixel×600 pixel,经过仿射变换后图像尺寸会减少 1/4 到 1/3 不等。数据集中大部分目标尺寸长宽比例在 0.1 倍以下。各个实例分布图如图 9 所示。

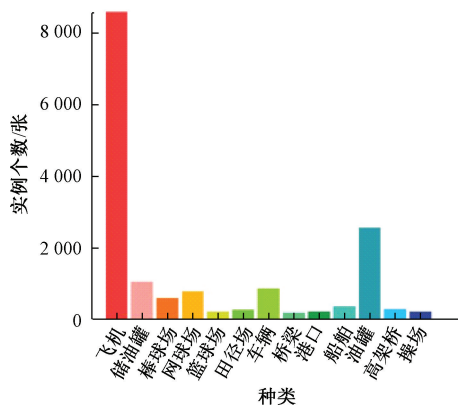

图 8 目标类别分布图

Fig. 8 Distribution of target categories

3.2 实验环境

模型开发语言为 Python,编译环境 Python 3.8.0,深度学习框架 Pytorch 2.0.0,CUDA 版本 11.7。实验在 Linux 操作系统中进行,CPU 选用 AMD EPYC 7551P,GPU 选用 NVIDIA RTX A5000,24 G 显存。

训练过程中根据数据集训练效果对参数进行微调,以

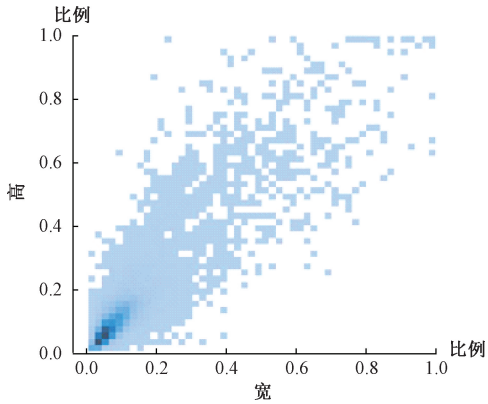

图 9　目标尺寸分布图

Fig. 9　Distribution of target sizes

达到整个网络最佳检测效果，优化器选择 AdamW，在前 2 000 个 epochs 进行热身训练，表 5 为训练具体参数。

表 5　训练参数

Table 5　Training parameters

参数	参数值
Epoch	120
Image input	640×640
Momentum	0.900
warmup_epochs	2 000
Learn rate	0.000 1～1.0
Batch size	12

通过观察模型训练过程中损失曲线，可发现训练轮次到达 120 时模型趋于收敛，损失曲线如图 10 所示。

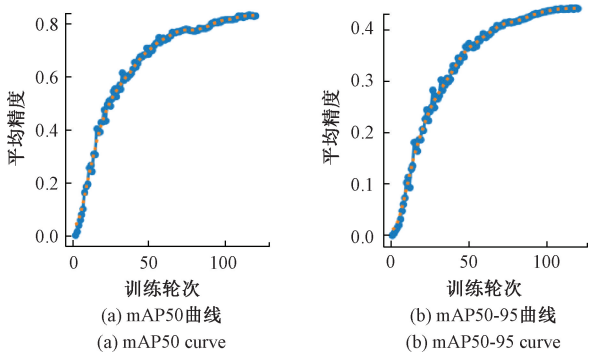

图 10　损失曲线图

Fig. 10　Loss curve graph

已达到理想性能水平，进一步增加训练次数不会显著提升模型性能；同时考虑数据集大小和显存限制，经过实验，保证训练稳定性的前提下 Batch size 为 12 可获得良好训练效果，使得设备性能同训练收敛速度及效果达到平衡。

3.3　评估指标

为评估 ELS-RTDETR 模型性能和航拍遥感图像检测的准确性，采用阈值为 0.5 的均值平均精度（mAP@50）、计算量（GFLOPs）、模型参数量（Params）、召回率（R）、准确率（P）、F1 分数曲线作为模型性能评估指标。在计算 mAP 前需要计算召回率（R）、准确率（P），P 和 R 计算式分别如（8）和（9）所示。

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

式中：TP 表示真阳性，FN 表示假阴性，FP 代表假阳性。目标检测中 mAP 是最常见评价指标之一，用于衡量模型在多个类别上的平均性能指标，计算公式如（10）所示。

$$mAP = \frac{\sum_{i=1}^{n}\int_{0}^{1} P(R)\mathrm{d}R}{n} \tag{10}$$

其中，n 为数据集图像总张数。

F1 分数表征模型在不同阈值下假阳性和假阴性之间的平衡。计算量（GFLOPs）为每秒浮点运算次数，常用于衡量模型的时间复杂度，其与模型的计算时间直接相关。模型参数量（Params）即网络模型需要训练参数总数，单位一般为 Mb，常被用于衡量模型的时间复杂度，其与内存占用直接相关。

3.4　消融实验

以 RT-DETR 作为基准模型，在其基础上进行了一系列改进，为系统分析各改进点对于模型性能的影响共设计 7 组对照实验。其中 A、B、C、D、E、F、H 分别代表不同阶段的实验模型，A 为基准模型，B 为用 Vovnet 提高 Resnet 后的模型，C 为在 B 的基础上添加 SE 模块后，D 为在 C 的基础上对网络深度宽度进行调整后，E 为在 D 的基础上添加 LRFF 模块，F 为在 E 的基础上添加 CGA 模块，H 为在 F 的基础上再次调整网络结构。

结构重整详细改动如 2.3 节所示，仿照 ConvNeXt 中对深度学习模型的优化改进，在不引入复杂结构的同时通过大量消融实验将部分卷积替换为深度可分离卷积，并结合利用 SE 模块和 CGA 模块的优势特点对网络的深度和宽度重新进行优化调整，在压缩模型体积、减少计算量的同时，减少模型的精度损失。消融试验结果如表 6 所示。

实验结果表明，ELS-RTDETR 改进模型检测精度 mAP@50 提高了 2.7%，模型参数量减少 32.9%，较原模型大幅度降低。Vovnet 主干网络的引入显著提高模型检测精度的同时为后续优化改进留下了空间；SE 模块的使用有效降低了卷积网络中的冗余信息，通道间的信息传递变得更加高效；通过设计并使用 LRFF 模块，在不引入复杂结构、不增加额外计算开销的同时，提高了模型对于小体积目标的检测能力；CGA 模块的使用减少了原模型中多头注意力机制里的冗余信息提高了图像特征信息的丰富度；最后通过优化重整网络结构，较好的平衡了检测精度与模型体积之间的矛盾，使其更适用于航拍图像检测场景。

表 6 消融实验

Table 6 Ablation experiments

编号	模型	mAP@50/%	召回率/%	准确率/%	F1	计算量/G	参数量/M
A	RT-DETR	80.8	73.31	84.84	0.78	57.3	19.04
B	A+Vovnet 主干替换	84.1	80.12	85.36	0.82	92.3	16.46
C	B+SE 模块	85.4	78.50	86.99	0.81	92.3	18.34
D	C+结构重整_1	81.4	77.95	79.26	0.77	64.6	13.89
E	D+LRFF 模块	81.6	75.26	85.24	0.78	64.2	13.42
F	E+CGA 模块	82.6	78.36	84.28	0.80	64.3	13.26
H	F+结构重整_2	83.6	79.16	83.32	0.80	57.5	12.78

3.5 主流目标检测模型对比试验

为进一步验证模型改进有效性,挑选现阶段主流目标检测模型与改进后模型进行对比试验。其中实验环境相同、训练参数按原默认项配置,其中 SSD 模型未使用预训练权重,训练效果较差为达到最终收敛,训练轮次为 240 轮(其他模型均为 120 轮),如表 7 所示分别对 mAP50、GFLOPs、参数量和 mAP50:95 进行比较以证明改进后模型有效性。

表 7 主流目标检测模型对比

Table 7 Comparison of state-of-the-art object detection models

模型	mAP50/%	mAP50:95/%	计算量	参数量
SSD	35.77	15.59	267.5	42.80
YOLOv5m	75.78	39.74	64.4	23.91
YOLOv8n	68.36	32.78	8.2	2.87
YOLOv8m	79.88	41.80	79.1	24.66
YOLOv10m	80.27	44.62	64.0	15.73
RT-DETR	80.80	46.56	57.3	19.04
Ours	83.60	44.35	57.5	12.78

3.6 泛化性分析

表 8 为模型在 SIMD 遥感图像数据上的泛化性实验结果,SIMD 数据集是一个多类别、开源、高分辨率的遥感对象检测数据集,共包含 15 类检测对象、5 000 张图片。

表 8 SIMD 数据集实验结果

Table 8 Experimental results on the simd dataset

模型	mAP50/%	mAP50:90/%	召回率/%	准确率/%
RT-DETR	74.86	61.06	75.13	74.17
ELS-RTDETR	77.40	61.69	76.99	76.89

实验结果证明改进后模型在 SIMD 数据集上 mAP@50 较原模型显著提高了 2.54%,进一步证明了改进模型的泛用性。

3.7 可视化实验

为对改进后模型的检测性能进行直观展示,使用 Grad-CAM 热力图显示技术实现模型网络在推理过程中的可视化,原模型和改进后模型展示结果分别如图 11、12 所示。Grad-CAM 通过对选定卷积层中的梯度进行计算分析从而获取不同特征图对于检测结果的重要性。Grad-CAM 为目标检测模型的结果提供了可解释性。在航拍遥感图像检测场景下,改进模型表现出了显著优势,其能准确识别原模型未能发现的小体积、特征模糊目标。热力图可视化结果进一步证明了改进方法对于提高模型在航拍遥感场景下检测精度有效性。

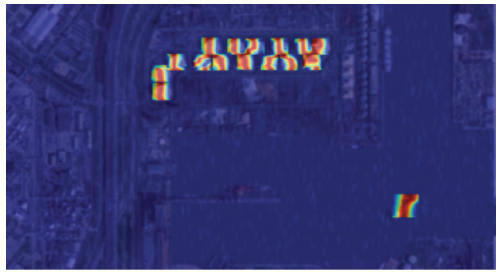

图 11 原模型热力图

Fig. 11 Original model heatmap

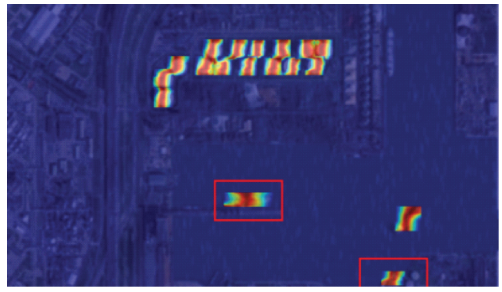

图 12 改进模型热力图

Fig. 12 Improved model heatmap

4 结 论

ELS-RTDETR 模型首先通过引入 Vovnet 主干网络获得了较原模型更高的检测精度;其次设计基于感受野中

心匹配的特征增强模块 LRFF，在不引入复杂结构，不带来额外计算开销的同时，提高模型对于小体积目标的检测精度；再次通过引入 CGA 模块和 SE 模块分别降低了原模型中多头注意力机制和卷积通道中的冗余信息，使模型获得了更丰富的语义特征和图像细粒度信息。最终对网络模型进行结构重整，实现了检测精度与体积、计算开销之间的均衡，使模型更适用于航拍图像检测场景。

在 RSOD 数据集和 NWPU VHR-10 数据集混合并进行数据增强后的自建数据上对模型性能进行测试和比较，证实各部分改进的有效性，在 SIMD 数据集上进一步验证了模型的泛用性。面对航拍图像场景下目标体积小、背景复杂和特征模糊的检测挑战，改进后的 ELS-RTDETR 模型在计算量不变，模型参数量大幅度降低的同时，检测精度得到了较大提高。解决了遥感航拍场景对于检测模型高性能的要求，同时较小的参数量和简单的网络结构给端侧平台部署带来了便捷。

参考文献

[1] REN SH Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[2] 白宗宝，张俊举，高原，等. 基于注意力机制的航拍图像目标检测算法[J]. 激光与光电子学进展，2023，60(12)：322-332.
BAI Z B, ZHANG J J, GAO Y, et al. Aerial image objectdetection algorithm based on attention mechanism[J]. Progress in Laser and Optoelectronics, 2023, 60(12): 322-332.

[3] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]. European Conference on Computer Vision (ECCV). Cham: Springer, 2016: 21-37.

[4] FARHAD A, REDMON J. YOLOv3: An incremental improvement [C]. Computer Vision and Pattern Recognition, Berlin/Heidelberg: Springer, 2018, 1804: 1-6.

[5] WU W, LIU H, LI L, et al. Application of local fullyconvolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image[J]. PloS One, 2021, 16(10): e0259283.

[6] LI C, LI L, JIANG H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. ArXiv preprint arXiv: 2209.02976, 2020.

[7] VARGHESE R, SAMBATH M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness[C]. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE, 2024: 1-6.

[8] 赵耘彻，张文胜，刘世伟. 基于改进 YOLOv4 的无人机航拍目标检测算法[J]. 电子测量技术，2023，46(8)：169-175.
ZHAO Y CH, ZHANG W SH, LIU SH W. Unmanned aerial vehicle aerial object detection algorithm based on improved YOLOv4[J]. Electronic Measurement Technology, 2023, 46(8): 169-175.

[9] 徐光达，毛国君. 多层级特征融合的无人机航拍图像目标检测[J]. 计算机科学与探索，2023，17(3)：635-645.
XU G D, MAO G J. Unmanned aerial vehicle aerial image object detection with multi-level feature fusion[J]. Computer Science and Exploration, 2023, 17(3): 635-645.

[10] 徐坚，谢正光，李洪均. 特征平衡的无人机航拍图像目标检测算法[J]. 计算机工程与应用，2023，59(6)：196-203.
XU J, XIE ZH G, LI H J, et al. Feature-balanced object detection algorithm for unmanned aerial vehicle aerial images [J]. Computer Engineering and Applications, 2023, 59(6): 196-203.

[11] 许延雷，梁继然，董国军，等. 基于改进 CenterNet 的航拍图像目标检测算法[J]. 激光与光电子学进展，2021，58(20)：192-201.
XU Y L, LIANG J R, DONG G J, et al. Aerial image object detection algorithm based on improved CenterNet[J]. Progress in Laser and Optoelectronics, 2021, 58(20): 192-201.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach: Curran Associates, 2017: 6000-6010.

[13] 陈朋磊，王江涛，张志伟，等. 基于特征聚合与多元协同特征交互的航拍图像小目标检测[J]. 电子测量与仪器学报，2023，37(10)：183-192.
CHEN P L, WANG J T, ZHANG ZH W, et al. Small object detection in aerial images based on feature aggregation and multi-element collaborative feature interaction[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(10): 183-192.

[14] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]. European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 213-229.

[15] CHEN Q, CHEN X K, WANG J, et al. Group DETR: Fast DETRtraining with group-wise one-to-

many assignment [C]. IEEE/CVF International Conferenceon Computer Vision(ICCV). Washington D.C.: IEEEPress,2023: 6610-6619.

[16] LIU SH L, LI F, ZHANG H, et al. DAB-DETR: Dynamic anchor boxes are better queriesfor DETR[C]. ICLR 2022 Conference, 2022.

[17] ZHU X ZH, SU W J, LU L W, et al. Deformable DETR: Deformable Transformersforend-to-endobjectdetection[C]. ICLR 2021 Conference, 2021.

[18] ZHANG H, LI F, LIU SH L, et al. DINO: DETR withimproved denoising anchor boxes for end-to-end object detection[C]. ICLR 2023 Conference, 2023.

[19] ZHAO Y AN, LYU W Y, XU SH L, et al. DETRs beat YOLOs on real-time object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.

[20] LEE Y, HWANG J, LEE S, et al. An energy and GPU-computation efficient backbone network for real-time object detection[C]. IEEE/CVF Conferenceon Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA: IEEE, 2019.

[21] LIU Z, MAO H, WU C Y, et al. A convnet for the2020s[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022:11976-11986.

[22] HUANG G, LIU Z, VAN DER M L, et al. Densely connected convolutional networks [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[23] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision, 2021:9992-10002.

[24] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2020.

[25] TAN M, LE Q V. EfficientNetV2: Smaller models and faster training[C]. International Conference on Machine Learning. New York: ACM, 2021: 7102-7110.

[26] XIE S N, GIRSHICK R, DOLLAR P, et al. Aggregated residual transformations for deep neural networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE Press, 2017: 5987-5995.

[27] MA N N, ZhANG X Y, ZHENG H T, et al. ShuffleNet v2: Practical guidelines for efficient CNN architecture design [C]. Computer Vision-ECCV 2018:15th European Conference,Munich:ACM,2018: 122-138.

作者简介

张淑卿(通信作者),硕士,副教授,硕士生导师,主要研究方向为深度学习、复杂系统建模与控制。

E-mail:88638696@qq.com

肖凡,硕士研究生,主要研究方向为深度学习、目标检测。

E-mail:3171289016@qq.com

葛超,博士,教授,硕士生导师,主要研究方向为网络控制系统稳定性分析、多智能体系统建模与控制。