文章编号　1004-924X(2024)24-3616-16

跨层注意力交互下的多特征交叉无人机图像检测

张志豪，杜丽霞，侯　越*，郝紫微，尹　杰
(兰州交通大学 电子与信息工程学院，甘肃 兰州 730000)

摘要：针对无人机交通巡检存在的航拍图像背景复杂、目标密集、目标尺度分布不均匀等问题，提出了一种跨层注意力交互下的多特征交叉无人机目标检测算法(Multi-feature Crossover under cross-layer Attentional Interaction，MCAI)。首先，在主干网络部分设计自适应跨层注意力交互模块(Adaptive Cross-layer Attentional Interaction，ACAI)，使模型聚焦于关键特征区域以实现对全局关键特征信息的有效筛选，从而淡化复杂背景的影响。其次，设计了一种基于可变形自注意力的编码器(Deformable Encoder，DeEncoder)，该编码器通过扩大特征层感受野，弥补丢失的目标特征。最后，为有效识别区域内不同尺度的微小目标，提出多尺度交叉融合模块(Multi-scale cross fusion module，MSCF)，该模块通过小波变换和特征表示结合的方式，融合浅层空间信息和深层语义信息，以有效捕捉不同尺度目标的细粒化特征。通过在VisDrone2019-DET、BDD-100K数据集和LZTraffic Video数据集上的实验结果表明，MCAI相较于RT-DETR模型，*mAP*0.5分别提高了3%，2.2%和4.5%，显著提高了无人机巡检的检测精度。此外，在阴雨场景中，MCAI的*mAP*0.5相较于RT-DETR模型提升了2.1%，极端天气鲁棒性表现更优。

关　键　词：无人机巡检；目标检测；注意力交互；编码器；小波变换
中图分类号：TP391.4　　**文献标识码**：A　　**doi**：10.37188/OPE.20243224.3616

Multi-feature cross UAV image detection algorithm under cross-layer attentional interaction

ZHANG Zhihao，DU Lixia，HOU Yue*，HAO Ziwei，YIN Jie

(*College of Electronic and Information Engineering*，*Lanzhou Jiaotong University*，*Lanzhou* 730000，*China*)
* *Corresponding author*，*E-mail*：*houyue@mail.lzjtu.cn*

Abstract：Aiming at the problems of complex background of aerial images，dense targets，and uneven target scale distribution in UAV traffic inspection，a multi-feature crossover under cross-layer attentional interaction (Multi-feature crossover under cross-layer attentional interaction，MCAI) UAV target detection algorithm was proposed. Firstly，an Adaptive Cross-layer Attentional Interaction (Adaptive Cross-layer Attentional Interaction，ACAI) module was designed in the backbone network part so that the model focused on the key feature regions to achieve effective screening of global key feature information，thus fading the influence of the complex background. Secondly，a deformable self-attentive encoder (Deformable Encoder，DeEncoder) was designed，which compensated for the lost target features by expanding the fea-

收稿日期：2024-06-05；**修订日期**：2024-08-16.
基金项目：国家自然科学基金(No. 62063014，No. 62363020)；甘肃省自然科学基金(No. 22JR5RA365)

ture layer receptive field. Finally, in order to effectively identify tiny targets at different scales in the region, the multi-scale cross-fusion module (Multi-scale cross fusion module, MSCF) was proposed, which fused shallow spatial information and deep semantic information by combining the wavelet transform and feature representation in order to efficiently capture the fine-grained features of targets at different scales. The experimental results on the VisDrone 2019-DET, BDD-100K dataset, and LZTraffic Video dataset show that MCAI improves *mAP*0.5 by 3%, 2.2%, and 4.5%, respectively, compared to the RT-DETR model, which significantly improves the detection accuracy of the UAV inspection. In addition, in the cloudy and rainy scenario, the *mAP*0.5 of MCAI improves by 2.1% compared to the RT-DETR model, with better extreme weather robustness performance.

Key words: UAV inspection; target detection; attentional interaction; encoder; wavelet transform

1　引　言

无人机的高机动性和优异的性价比使其在智能交通、城市管理、交通巡检等领域展现出巨大的应用前景[1]。然而，受无人机高空巡检、飞行高度不定等作业方式影响，其所捕获的目标通常被复杂的背景环境包围，同时排列密集且尺度不均匀的目标特征增加了无人机精准巡检的难度，因此无人机图像目标检测仍是一项具有挑战性的任务。

传统目标检测技术通过区域选择器遍历候选区域，并结合HOG，Haar等特征提取器及AdaBoost，SVM分类器进行特征分类[2-4]。但此类方法在处理候选框时存在复杂度高、窗口冗余等问题，致使模型在复杂场景和多类检测任务中，检测精度不高。随着深度学习技术的不断发展，基于卷积神经网络(Convolutional Neural Network, CNN)的目标检测算法已成为无人机航拍图像处理的主流方法[4]。目前，基于CNN的目标检测模型主要分为两类：两阶段检测器和单阶段检测器。两阶段检测算法如R-CNN[5]，Faster-RCNN[7]和Mask-RCNN[8]等，该类算法需要在对目标进行分类和定位之前生成区域建议，并根据区域建议进行候选边界框提取，以实现目标的识别和分类。然而，此类方法需生成大量边界框以提高检测的准确性，因此导致了检测速度的降低，难以满足实时处理的要求。相较于两阶段检测算法，单阶段目标检测具有较高的实时性，包括YOLO系列[9-11]、SSD系列[11]、Efficient系列[13]和Refinedet模型[14]等，其主要通过提取特征直接对物体进行识别和分类，网络结构简单、检测速度快，更能满足无人机航拍图像检测的需求。Wang[15]等基于YOLOv5检测框架提出STC-YOLOv5，通过增加全局依赖关系改善特征提取网络结构，丰富空间位置信息，提高了单类目标检测的准确性，但仍存在小目标漏检和误检的问题。Zhang[16]等提出了一种改进的单阶段检测模型YOLOv5，该模型采用自适应上采样和空间相关性增强策略，通过优化局部与全局特征的相关性，使模型在特定类别的检测精度上取得了提升，但整体性能有所下降。Gao[17]等利用多特征交叉融合注意力和跨层级联多尺度特征融合金字塔来提升不同尺度目标的检测准确率，由于多尺度金字塔时间复杂度较高，无疑增加了模型的计算负荷，影响模型实时性。Li[18]等提出无人机航拍图像目标检测算法DA-YOLO，通过多尺度动态特征加权融合网络和并行选择性注意力机制，提升不同尺度目标的特征表达，从而在提高网络对相似目标区分能力的基础上，降低了微小目标误检和漏检的风险，但此方法忽略了实时性的要求。然而，上述算法从不同角度入手对微小目标检测任务进行优化，但并未考虑到复杂背景对无人机航拍图像检测任务的影响，以及微小目标特征随尺度变化而丢失的问题。深度学习建模理论中的特征提取与融合方法，为实时、准确地提取微小目标特征，提供了更为有效的解决方案。Wu[19]等提出了自上而下的特征聚合注意力，通过结合浅层特征优化目标空间位置信息的表示。Xie[20]等设计了特征交互和自注意力融合网络，增强了区域上下文信息特征，但对局部交互信息的提取表现较差。Chen[21]等利用深度可分离卷积减少模型参数以提高特征提取效率。

尽管上述方法在提取微小目标特征方面取得了一定进展，但在融合语义不一致特征时的冲突问题上仍有改进空间。当前的研究大多集中于单一尺度特征，对不同尺度特征的交互和融合研究不足，限制了模型在应对尺度变化时的适应性和鲁棒性。

针对无人机航拍图像背景复杂，全局上下文信息难以有效利用，造成不均匀尺度微小目标难以提取的问题，本文提出了一种跨层注意力交互下的多特征交叉无人机图像检测算法。该网络由自适应跨层注意力交互模块、可变形编码器和多尺度交叉融合模块组成。其中，自适应跨层注意力交互模块主要包括外部残差连接和跨层注意力交互模块。外部残差连接主要对局部信息进行恒等映射，跨层注意力交互模块利用稀疏注意力机制过滤掉区域内不相关的键值对，以实现背景关键特征信息的筛选，提升网络对全局上下文信息的捕获能力。在此基础上，本文提出基于可变形自注意力的编码器，该编码器通过可变形偏移网络的方式以扩大特征层感受野，有效提取目标全局上下文信息，进而增强模型的空间位置信息和特征表达能力，并缓解了小目标信息丢失问题。最后，针对区域内尺度不均目标检测精度低的问题，本文设计了多尺度交叉融合模块，该模块由小波变换和特征表示模块组成。首先，利用小波变换提取目标空间位置信息，然后将包含细粒度特征的浅层空间信息引入到特征表示模块中，以增强模型对目标深层特征信息的提取和融合，从而提升小目标检测的准确性。此外，验证模型在极端天气条件下的鲁棒性，本研究引入了阴雨极端天气的干扰测试，实验证明了 MCAI 网络在极端天气条件下的稳健性和适用性。为解释改进思路在模型决策中的过程，本文对模型的全局感受野、整体模型对小目标的关注进行了解释性分析，以突出本文模型改进点对决策结果的影响。

2　本文算法

本文提出基于 RT-DETR[22] 模型的跨层注意力交互下的多特征交叉无人机航拍图像检测网络，其网络结构如图 1 所示，其中，本文利用

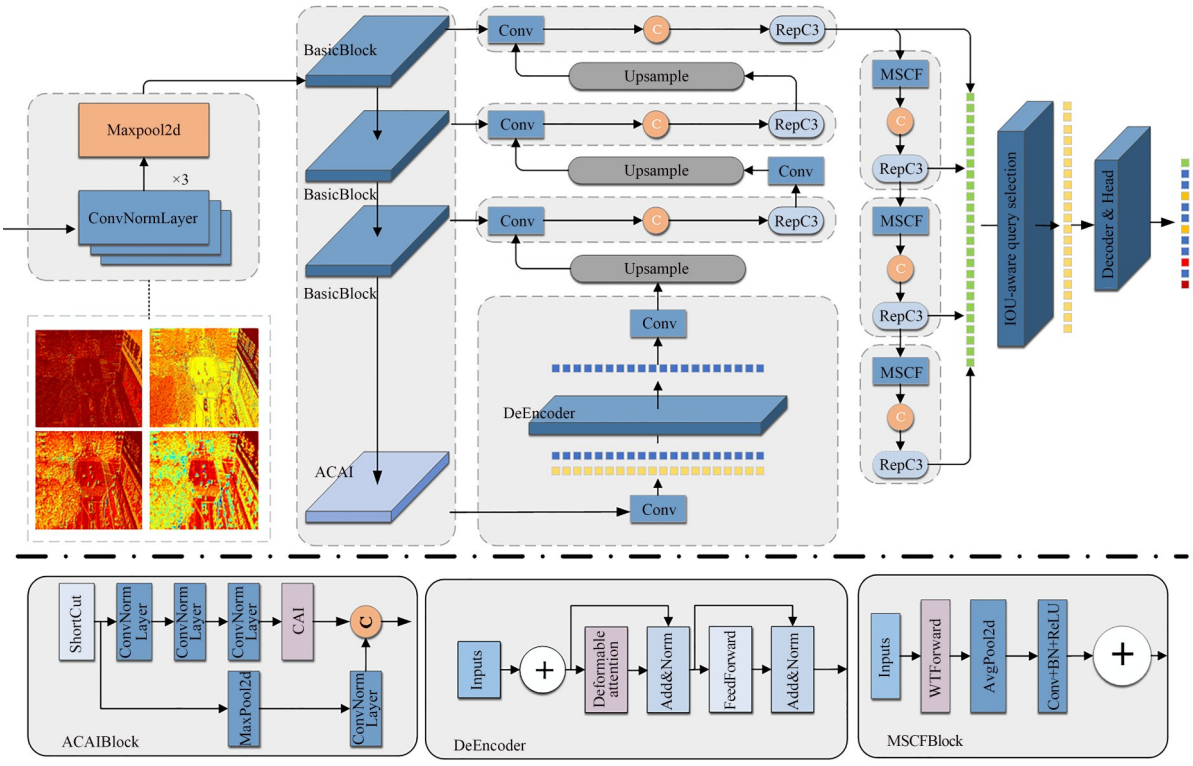


图 1　跨层注意力交互下的多特征交叉无人机目标检测网络

Fig. 1　Multi-feature crossover UAV target detection network under cross-layer attentional interaction

ResNet18[23]作为主干网络提取图像特征，在主干网络末端设计自适应跨层注意力交互模块，以充分筛选出与检测任务密切相关的关键特征区域，使模型聚焦于待检测区域。其次，设计可变形编码器以丰富目标特征信息，该编码器通过特征层感受野的扩张，在弥补了目标丢失特征的同时，显著提高了特征提取效果。最后，设计多尺度交叉融合模块，将小波变换所提取的目标信息映射至特征表示模块中，以达到对区域内小目标显著化提取的目的。

2.1　自适应跨层注意力交互模块

RT-DETR[22]的主干网络（ResNet18[23]）主要通过卷积神经网络提取目标特征，但受限于卷积运算的固定感受野，图像特征的全局依赖关系难以有效表示，进而导致目标关键特征无法进行有效提取。鉴于此，本文提出自适应跨层注意力交互模块以有效筛选复杂背景的关键特征，自适应跨层注意力交互模块结构如图 2 所示，主要以残差结构为基础，构建以卷积层，池化层和跨层注意力为核心的自适应跨层注意力交互模块。其主要通过跨层注意力（Cross-layer Attentional Interaction，CAI）和残差结构对全局目标关键特征进行关注。其中，外层残差连接对局部信息进行恒等映射，内层结构中使用跨层注意力交互模块，该模块利用动态稀疏注意力能够在粗糙区域背景级别过滤掉大部分不相关的键值对，只保留一小部分特征区域。然后，根据当前查询（query）的内容动态调整所关注的特征区域。最后，通过计算查询（value）和键（key）之间的相似度矩阵，可以识别和过滤掉不相关的背景区域，只保留与目标相关的关键特征区域。与传统卷积神经网络相比，这种动态调整使得跨层注意力交互模块能够自适应地聚焦于不同图像中与检测目标相关的区域，从而提升特征提取的有效性。此外，自适应跨层注意力交互模块可以有效结合不同层次的位置特征信息，选择性地关注和整合对目标识别最为关键的特征。这种跨层特征的交互使得模型能够自适应地提取复杂背景中的关键目标，增强对目标全局和局部特征的综合理解。

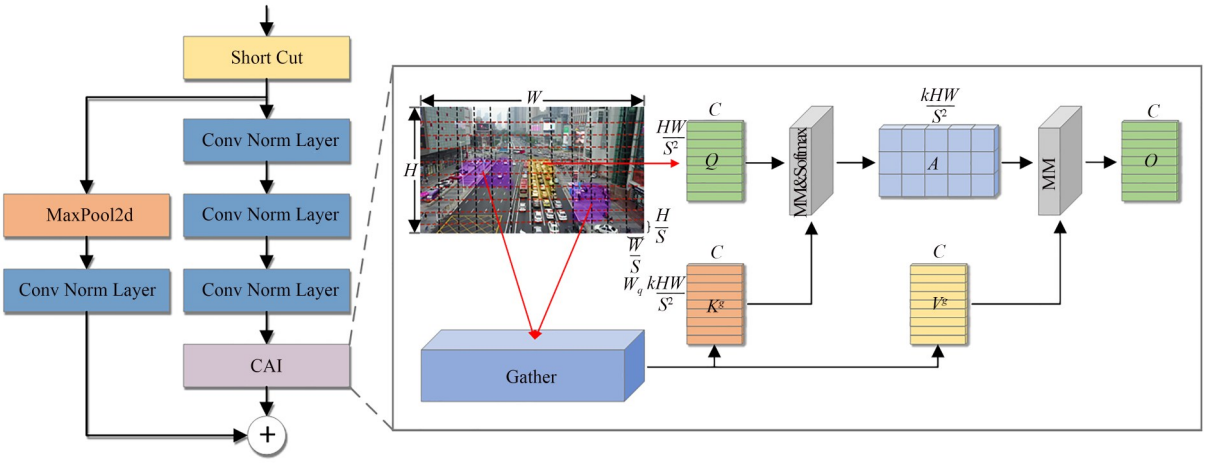


图 2　自适应跨层注意力交互模块结构图

Fig. 2　Sructure of the adaptive cross-layer attention interaction module

首先对输入图像 $\boldsymbol{X}\in\boldsymbol{R}^{H\times W\times C}$ 进行区域划分和输入投影，将输入图像划分为 $S\times S$ 的不重叠子区域，然后将 $\boldsymbol{X}$ 重塑为 $\boldsymbol{X}_r\in\boldsymbol{R}^{S^2\times HW/S^2\times C}$，使其每个子区域都含有 HW/S^2 个特征向量，以此获得特征向量组成的特征图。最后，使用线性投影导出查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$、值 $\boldsymbol{V}$ 的张量，具体过程如公式（1）所示：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{X}_r\boldsymbol{W}_q\\ \boldsymbol{K}=\boldsymbol{X}_r\boldsymbol{W}_k\\ \boldsymbol{V}=\boldsymbol{X}_r\boldsymbol{W}_v\end{cases}\tag{1}$$

其中：$\boldsymbol{X}\in\boldsymbol{R}^{H\times W\times C}$ 是输入图像（H，W 是输入图像的高度和宽度，C 输入图像的通道数），$\boldsymbol{X}_r\in\boldsymbol{R}^{S^2\times HW/S^2\times C}$ 是特征向量组成的特征图，$\boldsymbol{W}_q$，

W_k, W_v分别是查询,键,值的投影权重。

其次,利用获得的三个投影向量进行有向图的特征信息交互,并计算投影向量之间的相似度矩阵,在此基础上,根据相似度矩阵选择相邻区域,具有过程如式(2)~式(5)所示:

$$\boldsymbol{Q}^r = avgpool(\boldsymbol{Q}), \tag{2}$$

$$\boldsymbol{K}^r = avgpool(\boldsymbol{K}), \tag{3}$$

$$\boldsymbol{A}^r = \boldsymbol{Q}^r(\boldsymbol{K}^r)^{\mathrm{T}}, \tag{4}$$

$$\boldsymbol{I}_r = topIndex(\boldsymbol{A}^r), \tag{5}$$

其中:查询$\boldsymbol{Q}^r$和键$\boldsymbol{K}^r$分别是$\boldsymbol{Q}$,$\boldsymbol{K}$的每个区域的平均池化输出且$\boldsymbol{Q}^r$,$\boldsymbol{K}^r \in \boldsymbol{R}^{S^2\times C}$,Avgpool是平均池化操作,$\boldsymbol{A}^r$是矩阵相乘(Matrix Multiplication, MM)所得的相关性矩阵且$\boldsymbol{A}^r \in \boldsymbol{R}^{S^2\times S^2}$,$\boldsymbol{I}_r$是索引矩阵,该矩阵中包含与该矩阵最相关的$k$个区域的索引。

在去除不相关区域获得$\boldsymbol{I}_r$后,不同区域的关键特征信息会散布在整个特征图上,因此,需要对键和值张量进行收集,运用Gather函数收集键和值张量的具体公式如(6)~式(8)所示:

$$\boldsymbol{K}^g = gather(\boldsymbol{K}, \boldsymbol{I}_r), \tag{6}$$

$$\boldsymbol{V}^g = gather(\boldsymbol{K}, \boldsymbol{I}_r), \tag{7}$$

$$\boldsymbol{O} = Attention(\boldsymbol{Q}, \boldsymbol{K}^g, \boldsymbol{V}^g, \boldsymbol{I}_r), \tag{8}$$

其中:$\boldsymbol{K}^g$, $\boldsymbol{V}^g$是相关区域内所有的$\boldsymbol{K}$, $\boldsymbol{V}$信息,$Gather(\cdot)$是收集操作,$\boldsymbol{O}$是关键特征,$Attention(\cdot)$是跨层注意力。

2.2 可变形编码器

在自适应跨层注意力交互模块利用动态稀疏注意力进行背景特征信息筛选时,由于图像尺度变化,易导致远处目标的特征信息丢失,从而影响无人机巡检的检测性能。鉴于此,本文设计了一种基于可变形自注意力的编码器,以获取目标的空间位置信息和特征信息。可变形编码器结构如图3所示,该编码器主要是由归一化层,反馈层和可变形自注意力机制组成。其中,图3将编码器架构下将自注意力机制替换为可变形自注意力,该编码器能够在关键特征区域的指导下,利用卷积层学习offset偏移网络,使卷积核集中在关键区域的采样点发生偏移,以此实现在当前区域附近随意采样而不局限于常规卷积的规则采样。

首先,基于输入的特征图,通过卷积层对其进行线性投影,生成查询(query)、键(key)、值(value)三个初始采样区域。然后,利用偏移网络(Offset Network),对初始采样点进行学习,生成与之对应的偏移量。这个偏移量用于调整初始采样点的位置,使其偏离常规的规则采样网格。然后,原本规则的采样点位置会被移动到新的位置,从而形成变形采样点。利用变形采样点可以有效地扩大检测网络感受野,而不再局限于常规卷积操作中固定的感受野。由于远距离目标尺度特征微小,利用变形采样点能够根据图像中的实际目标特征动态调整采样位置,从而能够更好地捕捉微小尺度目标特征。其次,利用双线性插值的方法,将变形采样点位置上的目标特征输入到后续的键和值投影中,生成变形后的键和值。最后,应用多头自注意力关注参加查询的采样键和聚合功能的变形值。由此,可变形编码器通过查询内容自适应地选择键和值点,并动态调整键和值点的数量和分布,实现局部增强标记之间的全局关系建模,从而有效解决了远处目标尺寸变化导致的尺度缩小问题。

传统的自注意力头首先对输入的特征图进行线性投影,得到查询$\boldsymbol{Q}$,键$\boldsymbol{K}$和值$\boldsymbol{V}$三个子特征图$\boldsymbol{q}$,$\boldsymbol{k}$,$\boldsymbol{v}$:

$$\begin{cases} \boldsymbol{q} = \boldsymbol{x}\boldsymbol{W}_q \\ \boldsymbol{k} = \boldsymbol{x}\boldsymbol{W}_k \\ \boldsymbol{v} = \boldsymbol{x}\boldsymbol{W}_v \end{cases}, \tag{9}$$

$$\boldsymbol{z}^{(m)} = \sigma\left(\boldsymbol{q}^{(m)}\boldsymbol{k}^{(m)\mathrm{T}}\Big/\sqrt{d}\ \right)\boldsymbol{v}^{(m)}, \tag{10}$$

$$\boldsymbol{z} = Concat(\boldsymbol{z}^{(1)}\cdots\boldsymbol{z}^{(M)})\boldsymbol{W}_o, \tag{11}$$

其中:$\boldsymbol{x}$是输入的特征图且$\boldsymbol{x} \in \boldsymbol{R}^{H\times W}$,$\boldsymbol{q}$,$\boldsymbol{k}$,$\boldsymbol{v}$分别表示查询,键和值,$W_q$, W_k, W_v是所生成的投影矩阵,$\boldsymbol{z}^{(m)}$是第m个自注意力输出,σ表示$soft\max$函数,$\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}$是计算查询和键的点积,$d$表示每个头的尺寸,每个头的特征连接在一起并通过$W_o$投影以获得最终输出$\boldsymbol{z}$。

将输入特征图$\boldsymbol{x}$进行线性投影到$\boldsymbol{Q}$上,然后将$\boldsymbol{Q}$输入到偏移网络,得到每个相关点的偏移量,输出每个相关点变换后的位置,最终对变化

后的点进行特征采样，作为可变形键 K 和值 V 的具体计算公式如(12)～式(14)所示：

$$\begin{cases} \boldsymbol{q}=\boldsymbol{x}W_q \\ \tilde{\boldsymbol{k}}=\tilde{\boldsymbol{x}}W_k \\ \tilde{\boldsymbol{v}}=\tilde{\boldsymbol{x}}W_v \end{cases} \tag{12}$$

$$\Delta\boldsymbol{p}=\theta_{\text{offset}}(\boldsymbol{q}), \tag{13}$$

$$\tilde{\boldsymbol{x}}=\phi(\boldsymbol{x};\boldsymbol{p}+\Delta\boldsymbol{p}), \tag{14}$$

其中：$\tilde{x}$ 是变形的特征区域，$\tilde{k}$ 和 $\tilde{v}$ 分别是变形的键和值嵌入，Δp 是偏移量输出，θ_{offset} 是偏移网络，$\phi(\bullet)$ 是双线性插值。

同时，双线性插值 ϕ 使其可微的具体过程如式(15)～式(17)所示：

$$g(\boldsymbol{p},\boldsymbol{r})=g(\boldsymbol{p}_x,\boldsymbol{r}_x)g(\boldsymbol{p}_y,\boldsymbol{r}_y), \tag{15}$$

$$\boldsymbol{z}(\boldsymbol{r})=\boldsymbol{z}\left[\boldsymbol{r}_y,\boldsymbol{r}_x,:\right], \tag{16}$$

$$\phi(z;(\boldsymbol{p}_x,\boldsymbol{p}_y))=\sum_{(r_x,r_y)}\boldsymbol{G}(\boldsymbol{p},\boldsymbol{r})\times\boldsymbol{z}(\boldsymbol{r}), \tag{17}$$

其中：$(\boldsymbol{p}_x,\boldsymbol{p}_y)$ 指的是图像中像素的位置信息，$g(a,b)=\max(0,1-|a-b|)$，(r_x,r_y) 索引 $\boldsymbol{R}^{H\times W}$ 上的所有位置信息。

最后，对 q,k,v 进行可变形自注意力计算，并采用相对位置偏移 R，所得式(18)如下：

$$z=\sigma(q\tilde{k}/\sqrt{d}+\phi(\tilde{B};\boldsymbol{R}))\tilde{v}, \tag{18}$$

其中：z 为可变形自注意力的输出，$(\tilde{B};\boldsymbol{R})$ 对应位置嵌入。

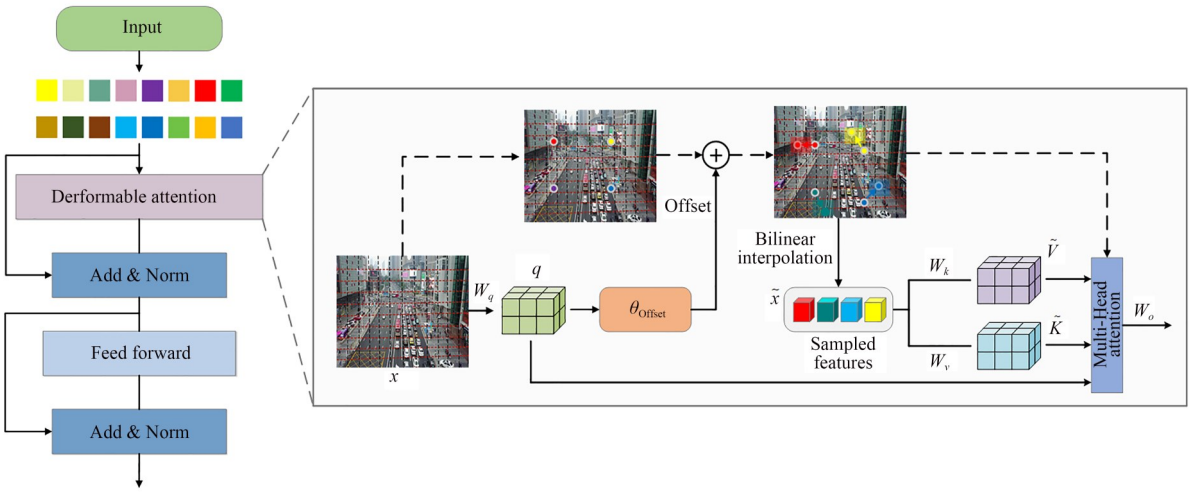


图 3　可变形编码器结构图

Fig. 3　Deformable encoder structure

2.3　多尺度特征融合网络

针对无人机巡检过程中局部区域小目标特征的识别问题，本文提出多尺度交叉融合模块以实现微小目标多尺度特征提取，该模块由小波变换模块和特征表示模块组成，其具体结构如图 4 所示。小波变换模块利用时频域的转换机制，在保留空间信息的同时有效地降低了特征图的分辨率。特征表示模块由池化层、标准卷积层、批量归一化和 ReLU 激活层组成，主要用于提取目标特征。其中，小波变换模块首先使用离散小波变换(DWTForward)进行时频域的转换，利用分解滤波器，将分辨率为 H 的图像，分解为低频滤波器 H_0 和高频滤波器 H_1，这些滤波器分别用于从图像中提取低频和高频信息，最后生成四个分量分别是低频分量、水平、垂直和高频分量，每个分量的大小为 $H/2\times W/2$，而特征图的通道数多了两倍，因此小波变换可以将空间维度的部分信息编码到通道维度，不会丢失任何信息。特征表示模块由平均池化层、标准卷积层、批量归一化层和 ReLU 激活函数组成，在该模块中，首先利用平均池化层提取小波变换模块的特征图详细信息，其次，采用标准卷积来调整特征图的通道数，以及尽可能过滤冗余信息，使后续层能够更有效地学习目标代表性特征。

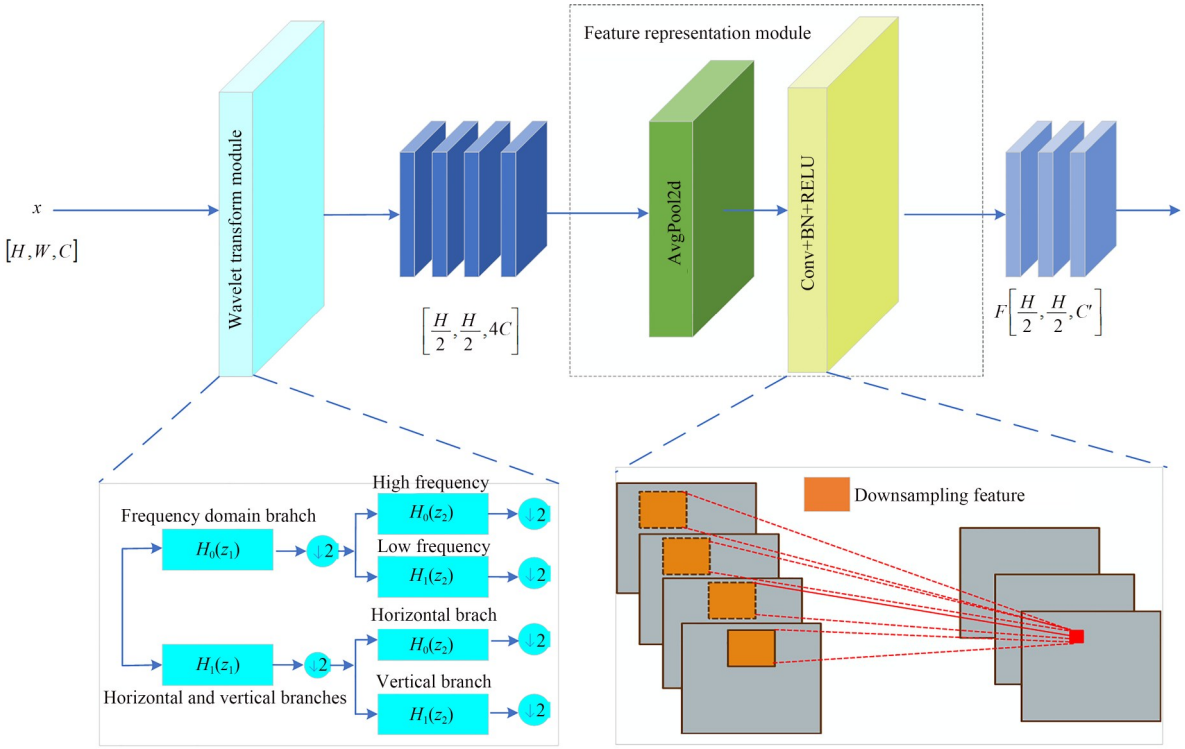


图 4　多尺度交叉融合模块结构图

Fig. 4　Structure diagram of the multi-scale feature fusion module

3　实验结果和分析

3.1　数据集

VisDrone2019-DET：VisDrone2019-DET 数据集是由 6 471 张训练图像，548 张验证图像和 3 190 张测试图像组成，选择预定义的类别：行人、人、汽车、货车、公共汽车、卡车、机动车、自行车、遮阳篷三轮车和三轮车。

BDD-100K：BDD-100K 数据集包含多个场景，如城市街道，加油站和高速公路。本文选择了五个预定义的类别：公共汽车、自行车、摩托车、汽车、卡车和行人。本文随机选择 20% 的图像并将其文件组织为 YOLO 格式，其中训练集 13 961 张图像，验证集为 6 100 张图像。

LZ Traffic Video 数据集：在某市交通监控捕获的实际数据集上验证了本文方法，该实际数据集共包含 8 851 张训练集，2 213 张验证集，像素维持在 1 920×1 920。预定类别为汽车，行人，摩托车，面包车，公交车，卡车，三轮车，自行车，摩托车，送货车，施工车。

3.2　实验环境

实验采用的硬件环境：GPU 为 NVIDIA GeForce RTX 3060-12GB，CPU 为 Intel 7.5 GHz；软件环境为 Windows11 操作系统，选用 Pytorch1.8 为深度学习框架。采用改进 RT-DETR 模型进行实验，分别在训练集和测试集上进行训练和测试，输入尺寸大小为 640×640，批次为 4，Epoch 为 100，初始学习率为 0.01，终止学习率为 0.2，采用随机梯度下降策略。

3.3　评价方法

目标检测的评价方法主要基于模型在识别和定位目标上的准确性。因此，为了比较所有最先进的方法，本文使用 $mAP0.5$，$mAP0.5{:}0.95$，*Recall*。*Precision* 和 *mAP* 指标作为模型性能评估指标。本文中 *AP* 为单一类别检测精度的相关指标。对各类别的 *AP* 值进行相加再除以类别数为 *mAP*，$mAP0.5$ 代表 $IOU=0.5$ 时的 *mAP*，$mAP0.5{:}0.95$ 表示设置 10 个 IOU 阈值，从 0.5 到 0.95，步长为 0.05 时的 *mAP*。*AP* 和 *mAP* 的计算公式如式(19)～式(22)：

$$P=\frac{TP}{TP+FP}, \tag{19}$$

$$R=\frac{TP}{TP+FN}, \tag{20}$$

$$AP=\int_0^1 P(R)d(R), \tag{21}$$

$$mAP=\frac{\sum_{i=1}^{N}AP_i}{N}, \tag{22}$$

其中：N为物体类别个数；FP为假正样本，即误检；FN为假负样本，即漏报；TP为真正样本，即正确检测；TN为真负样本。

同时，为了验证本文算法的稳健性和有效性，本文设计消融实验和类别错误实验以验证每个改进模型的有效性，设计可解释性实验来了解改进模型的内部决策过程，设计鲁棒性实验来验证本文算法的稳健性和鲁棒性，最后，设计三组不同数据集的总体模型对比实验来验证本文算法的先进性和准确性。

3.4　对比模型

为了验证MCAI模型的有效性，本文比较了几种最先进的检测框架，Faster-RCNN，RetinaNet，CenterNet，SSD，ATSS-FPN-DyHead，TOOD，DINO，YOLOX-Tiny，GFL，RTMSet，YOLOv3，YOLOv4，YOLOv5，YOLOv7和RT-DETR。所有检测框架都是在相同的VisDrone2019-DET数据集和BDD-100K数据集进行测试。

在实际生活中，由于单阶段检测框架实时性表现较好，因此，在LZTraffic Video数据集中，特别与单阶段YOLO算法进行比较，以YOLOv3，YOLOv5，YOLOv6，YOLOv8，YOLOv8-Decoder以及RT-DETR为基准线。

3.5　消融实验

为验证所提出的自适应跨层注意力交互模块、可变形编码器和多尺度交叉融合模块的有效性，选用RT-DETR[22]模型为消融实验的基准模型，依次加入各个模块进行消融实验，所有消融实验均在VisDrone2019-DET数据集上进行。首先在基准模型RT-DETR中加入自适应跨层注意力交互模块，并将其命名为G-ACAI，在基础上，加入可变形编码器，将其命名为G-DMHSA，最后加入多尺度交叉融合模块，将其命名为G-MSCF。同时，设置Epoch=100，Batch_Size=4，输入图像为640×640，Momentum=0.9，IOU=0.7，Scale=0.5，控制各个消融实验的实验参数保持一致，依次验证G-ACAI模型，G-DMHSA模型，G-MSCF模型的准确性和有效性。

由此进行消融实验分析，具体结果如表1所示。

表1　改进模块消融实验

Tab.1　Improved module ablation experiments（%）

Methods	*R*	*P*	*mAP*0.5	*mAP*0.5:0.95	*mAP*
RT-DETR	40.8	56.5	42.5	25.7	34.2
G-ACAI	41.9	57.6	44.4	27.5	35.6
G-DMHSA	43	58.6	45.5	28.2	36.5
G-MSCF	44.2	58.8	46.1	28.5	36.8

表1给出了不同改进模块的微小目标识别性能。由表1可知，基础RT-DETR模型的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95，*mAP*分别为40.8%，56.5%，42.5%，25.7%，34.2%，其目标识别精度不高。G-ACAI模型的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95，*mAP*评价指标分别为41.9%，57.6%，44.4%，27.5%，35.6%，相较于RT-DETR模型，其性能分别提升了1.1%，1.1%，1.9%，1.8%，1.4%，这表明自适应跨层注意力交互模块能够有效利用动态稀疏注意力去掉不相关的背景特征并筛选出关键目标特征，从而减少背景噪声的影响。G-DMHSA模型在G-ACAI模型的基础上，增加了可变形编码器，其*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95，*mAP*评价指标分别是43%，58.6%，45.5%，28.2%，36.5%，相较于G-ACAI模型，识别精度进一步提升，其指标分别提升了1.1%，1%，1.1%，0.7%，0.9%，证明可变形编码器有效利用偏移网络扩张了特征层感受野，解决了目标特征信息丢失问题。G-MSCF模型在G-DMHSA模型基础上，增加了多尺度交叉融合模块，其*Recall*、*Precision*、*mAP*0.5，*mAP*0.5:0.95，*mAP*评价指标分别是44.2%，58.8%，46.1%，28.5%，36.8%，其特征提取和融合能力进一步提升，其性能分别提升了1.2%，0.2%，0.6%，0.3%，0.3%，意味着多尺度交叉融合模块可以很好地处理局部目标多尺度特征的提取和融合问题。

3.6 类别错误消融实验

传统消融实验主要选用平均精度值(mean Average Precision, *mAP*)作为模型性能评估指标。然而,平均精度值(*mAP*)通常不透明,无法具体指明模型改进的细节,如性能的具体提升点或优化的错误类型。因此,需要更细致的分析方法来解释模型的改进,在本文类别实验中,实现了准确的类别错误检测,具体错误类型包括分类错误(Cls)、定位错误(Loc)、重复检测框错误(Dupe)、背景误差(Bkg)、漏检(Miss)、错误预测为负样本的数量(*FN*),以及错误预测为正样本的数量(*FP*)。这种细致的错误分析方法不仅揭示了模型在哪些方面需要改进,而且展示了改进模块在提高无人机巡检图像分析准确性方面的具体贡献。

因此,为验证跨层注意力交互下的多特征交叉检测网络的具体类别错误,选用RT-DETR模型为类别实验的基准模型,依次加入各个模块进行类别错误实验,所有类别错误实验均在VisDrone2019-DET数据集上进行。首先在基准模型RT-DETR中加入自适应跨层注意力交互模块,并将其命名为G-ACAI,在此基础上,加入可变形编码器,将其命名为G-DMHSA,最后加入多尺度交叉融合模块,将其命名为G-MSCF。同时,设置Epoch=100,Batch_Size=4,输入图像为640×640,Momentum=0.9,IOU=0.7,Scale=0.5。由此进行类别错误实验分析,具体结果如表2所示。

表2 改进模块的类别错误实验

Tab. 2 Category error experiments of the improved module

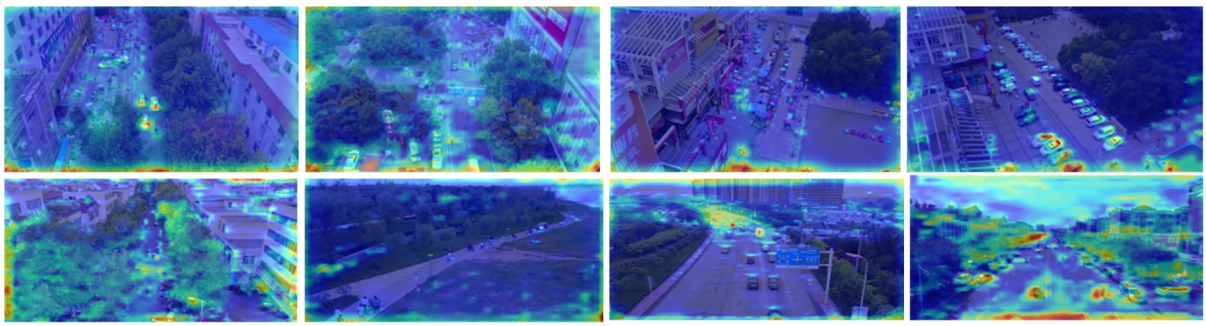
Methods	Cls	Loc	Dupe	Bkg	Miss	*FP*	*FN*
RT-DETR	20.98	3.82	0.21	2.12	13.31	10.92	37.54
G-ACAI	**20.79**	**3.30**	0.43	2.24	14.07	11.26	**37.08**
G-DMHSA	20.81	**3.26**	**0.38**	**2.17**	**13.86**	**11.05**	**36.65**
G-MSCF	**20.25**	3.29	**0.37**	2.27	**13.76**	11.12	36.70

从表2可以看出,RT-DETR模型的Cls,Loc,Dupe,Bkg,Miss,*FP*和*FN*分别是20.98,3.82,0.21,2.12,13.31,10.92和37.54,其模型检测性能较差。在此基础上,加入自适应跨层注意力交互模块后,G-ACAI模型的Cls,Loc和*FN*分别是20.79,3.30和37.08,相较于RT-DETR模型,其筛选区域目标关键特征信息的性能表现较优,Cls,Loc和*FN*分别降低了0.19,0.52和0.46,表明本文设计的自适应跨层注意力交互模块有效捕捉到全局目标关键信息,使模型聚焦于目标关键区域。在此基础上,添加可变形编码器后,G-DMHSA的Loc,Dupe,Bkg,Miss,*FP*和*FN*分别是3.26,0.38,2.17,13.86,11.05和36.05,相较于G-ACAI模型,G-DMHSA的Loc,Dupe,Bkg,Miss,*FP*和*FN*分别降低了0.04,0.05,0.07,0.21,0.21和0.43,由此看出,可变形编码器有效提取了关键区域内的目标特征信息,从而降低模型漏检、误检错误。最后,添加多尺度特征融合模块,G-MSCF的Cls,Dupe,Miss分别是20.25,0.37,13.76,相较于G-DMHSA模型,其Cls,Dupe,Miss分别降低了0.56,0.01和0.1,表明多尺度交叉融合模块有效融合了不均匀目标的尺度特征。综上所述,在无人机巡检领域,本文所提出的跨层注意力交互下的多特征交叉检测网络明显优于RT-DETR模型。

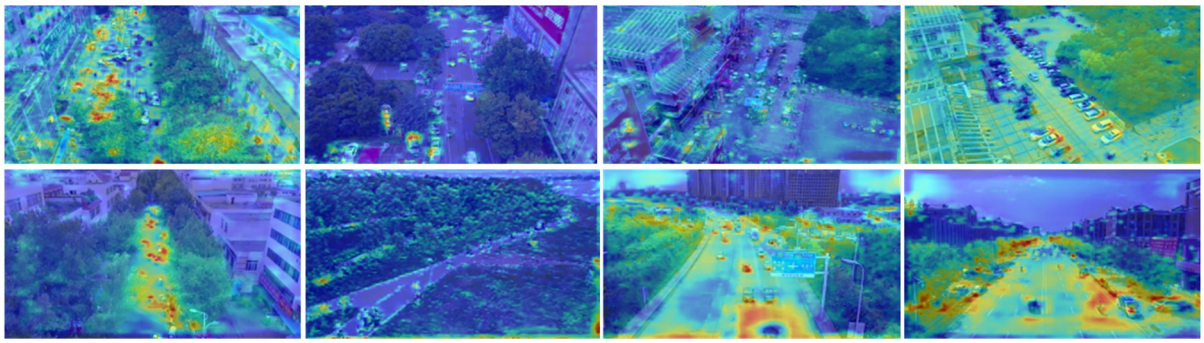
3.7 可解释性实验

为使研究者更加直观了解模型内部运行机制,本实验对跨层注意力交互下的多特征交叉网络进行可视化解释。其中,红色的区域为模型关注区域,颜色越深代表关注度越高。

由图5可知,MCAI的红色关注特征区域相较于RT-DETR更加广泛和准确,这意味着MCAI能够关注到全局特征信息,并利用全局和局部信息的依赖性来突出显示目标关键特征(彩图见期刊电子版)。相比之下,RT-DETR模型的特征关注点较为有限,RT-DETR模型未能观察到遮挡和密集区域的细节,从而导致对小目标特征识别不足。而MCAI模型的优势不仅体现在其对关键目标特征信息的理解上,还在于其能够有效利用可变形自注意力机制增



RT-DETR



MCAI

图 5　MCAI网络的模型可视化过程(红色部分是关注特征)

Fig. 5　Model visualisation process for the MCAI network (features of interest in red)

大特征层感受野，并有效识别遮挡区域和密集区域的小目标，对于在复杂场景下的无人机交通巡检网络至关重要。

3.8　感受野可视化实验

在评估不同模型的有效感受野(ERF)时，如图 6 所示，图像中的暗区分布范围可以作为 ERF 大小的指标。根据下述图示，将 YOLOv5-Decoder，YOLOv8-Decoder，RT-DETR 和 MCAI 模型的 ERF 进行了比较。结果表明，MCAI 模型的全局感受野相较于其他模型更为广泛，这意味着 MCAI 在捕获图像中的全局信息方面具有优势。通过比较 YOLOv5-Decoder，YOLOv8-Decoder，RT-DETR 和 MCAI 网络的有效感受野(ERF)来展示它们在处理小目标时的差异性。YOLOv5-Decoder 和 YOLOv8-Decoder 在扩展 ERF 方面表现不够明显，而 RT-DETR 模型虽然具有比前两者更大的感受野，但主要集中在局部信息上，对全局信息的提取能力有限。相比之下，MCAI 模型通过自适应跨层注意力交互模块筛选出关键特征信息，有效建立了全局结构的依赖性，增大

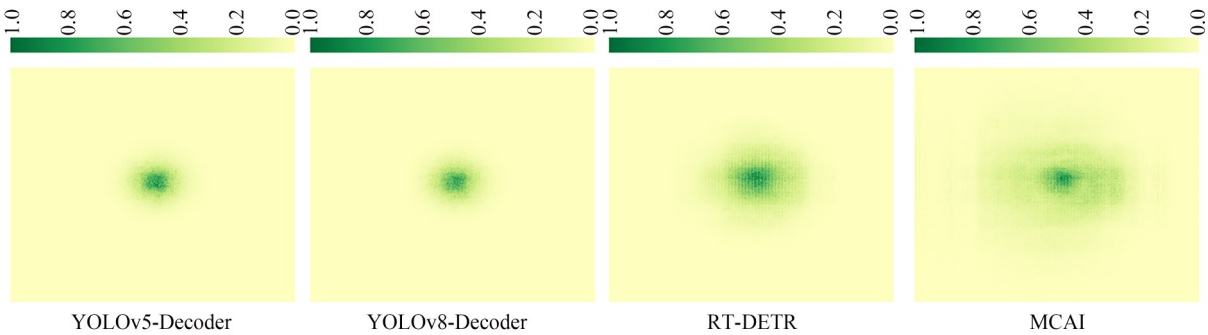


图 6　感受野可视化实验

Fig. 6　Receptive field visualization experiment

了感受野。相较于传统CNN模型，MCAI模型展现出更好的检测效果。本研究的成果有助于深入理解跨层注意力交互下的多特征交叉网络在目标检测中的作用机制，以及它们如何提升模型的整体性能。

3.9　鲁棒性实验

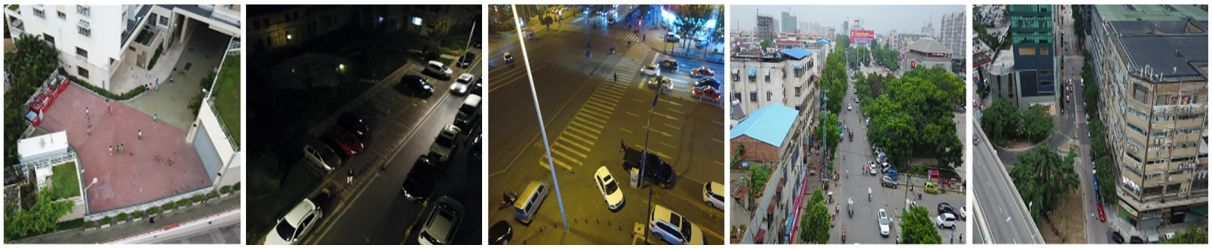
近年来，极端天气的频繁出现对无人机交通巡检的准确性构成了挑战。为了探索无人机交通巡检网络在复杂天气条件下的性能，本文实验在VisDrone2019-DET数据集中增加阴雨天作为干扰因素来测试MCAI模型在阴雨天气中的稳健性。下表为不同检测模型在阴雨天气中的无人机航拍图像检测结果。其中，Epoch=50，Batch_Size=10，输入图像为640×640，Momentum=0.9，IOU=0.7，Scale=0.5，数据增强策略为设置阴雨天气，数据集按照5∶5的比例分为阴雨天气数据和正常天气数据以此验证检测模型的鲁棒性。本文模型对于鲁棒性的评价指标为*mAP*0.5，*mAP*0.5∶0.95。其中，*mAP*0.5@*rain*和*mAP*0.5∶0.95@*rain*表示阴雨天气的评价指标，*mAP*0.5@*clean*和*mAP*0.5∶0.95@*clean*表示正常天气的评价指标。

表3　VisDrone2019-DET数据集的阴雨天气鲁棒性实验

Tab. 3　Robustness experiments on rainy weather on the VisDrone2019-DET dataset　(%)

Methods	*mAP*0.5@*clean*	*mAP*0.5@*rain*	*mAP*0.5∶0.95@*clean*	*mAP*0.5∶0.95@*rain*
YOLOv5-Decoder	33.9	21.5	19.9	11.7
YOLOv8-Deocder	32.0	20.9	18.3	11.5
RT-DETR	42.5	32.6	25.7	19
MCAI	46.1	34.7	28.5	20.7

表3为阴雨天气下的模型检测实验结果，图7为阴雨天气的可视化。由表3数据可知，在阴雨天气情况下，MCAI的*mAP*0.5@*rain*是34.7%，*mAP*0.5∶0.95@*rain*是20.7%，而YOLOv5-Decoder，YOLOv8-Decoder和RT-DETR模型的*mAP*0.5@*rain*是21.5%，20.9%，32.6%，*mAP*0.5∶0.95@*rain*分别是11.7%，11.5%和19%，相较于MCAI模型，YOLOv5-



(a) 原始图像
(a) Original image

(b) 阴雨天气图像
(b) Rain weather

图7　阴雨天气的图像可视化

Fig. 7　Image visualisation of a rainy day

Decoder，YOLOv8-Decoder和RT-DETR模型的*mAP*0.5@*rain*分别降低了13.2%，13.8%和2.1%，*mAP*0.5∶0.95@*rain*分别降低了9%，9.2%和1.7%，虽然阴雨天气的评价指标低于正常天气的评价指标，但MCAI模型相较于其他检测模型仍具有较高的检测精度，证明MCAI模型在阴雨天气中的鲁棒性较好。因此，在极端天气的影响下，天气变化虽然对无人机巡检领域的识别结果具有一定的影响性，但本文所提出的跨层注意力交互下的多特征交叉检测算法通过自适应跨层注意力交互模块有效识别无人机目标关键特征信息，并利用可变形编码器中的偏移网络有效扩张了感受野，增强了检测网络特征提取的能力，减少了目标特征信息丢失。最后，在局部密集区域，多尺度交叉融合模块可以有效融合和提取局部不同尺度的目标特征信息，使其具有较高的检测准确度。因此，跨层注意力交互下的多特征交叉检测算法在阴雨天气中仍具有较高的准确性。

3.10　总体算法分析

为了验证所提方法的有效性，在VisDrone2019-DET数据集上将所提方法与目前主流的检测模型对比，即Faster-RCNN，RetinaNet，CenterNet，SSD，ATSS-FPN-DyHead，TOOD，DINO，YOLOX-Tiny，GFL，RTMSet和YOLO系列。如表4所示，MCAI模型相较于其他几种经典的通用目标检测算法中具有良好的检测性能，其中，MCAI模型的*mAP*和*mAP*0.5分别是36.8%和46.1%，相较于Faster-RCNN算法和SSD算法，*mAP*和*mAP*0.5分别提升了15.3%，10.4%和28.2%，29.5%；相较于ATSS-FPN-DyHead和TOOD，*mAP*和*mAP*0.5分别提升了16.4%，12.3%和16.4%，12.2%，而且MCAI的参数量远小于ATSS-FPN-DyHead和TOOD；相较于DINO，GFL和RTMDet，*mAP*和*mAP*0.5分别提升了11.5%，1.6%和17.5%，14%和18.4%，14.9%；相较于RetinaNet算法和CenterNet模型，*mAP*和*mAP*0.5分别提升了20.7%，18.8%和24.4%，23.4%；相较于YOLOX-Tiny，*mAP*和*mAP*0.5分别提升22%和18.3%；与YOLOv3模型相比，*mAP*提升了28.7%，*mAP*0.5提升了28.4%；与YOLOv4模型相比，*mAP*提升了18.2%，*mAP*0.5提升了20.72%，相较于RT-DETR模型，*mAP*提升了4.2%，*mAP*0.5提升了3.6%。与上述模型相比，本文模型依然具有良好的检测能力。

表4　不同算法模型在VisDrone2019-DET上的结果对比

Tab.4　Comparison of results of different algorithmic models on VisDrone2019-DET

Method	Backbone	*mAP*	*mAP*0.5	*Param*
Faster-RCNN	ResNet50	21.5	35.7	41.7
RetinaNet	ResNet50	16.1	27.3	21.37
CenterNet	DLA-34	12.4	22.70	25.6
SSD	VGG-16	8.6	16.6	4.23
YOLOv3	MobileNet	8.1	17.7	3.75
YOLOv4	ResNet50	18.6	25.38	6.06
YOLOV5s	CSPDarkNet	19.1	29.8	7.05
YOLOv7	CSPDarkNet	22.9	34.5	37.26
RT-DETR	ResNet18	32.6	42.5	21.3
ATSS-FPN-DyHead	ResNet50	20.4	33.8	38.91
TOOD	ResNet50	20.4	33.9	32.4
DINO	DETR	25.3	44.5	47.56
YOLOX-Tiny	CSPDarkNet	14.8	27.8	5.035
GFL	CSPDarkNet	19.3	32.1	32.279
RTMDet	CSPDarkNet	18.4	31.2	4.876
MCAI	ResNet18	36.8	46.1	22.4

图 8　VisDrone2019-DET 数据集的实际应用场景可视化

Fig. 8　Visualization of the actual application scenario of the VisDrone2019-DET dataset

图 8 显示了 MCAI 检测网络在 VisDrone2019-DET 数据集的实际应用场景可视化(彩图见期刊电子版)。可以看出,MCAI检测模型能够高度准确地检测对象,对于应用场景中的实时性能,本文使用 *TP*、*FP* 和 *FN* 作为指标,其中绿色框(*TP*)表示正确的检测和真实的标签。红色框(*FN*)为漏检,未检测到。蓝色框(*FP*)表示误检测,误检测。

为进一步验证本文算法的有效性,本文使用 BDD-100K 数据集验证 MCAI 模型的泛化性。如表 5 所示,MCAI 模型相较于其他几种经典的通用目标检测算法中具有良好的检测性能,其中,MCAI 模型的 *mAP*0.5 是 72.4%,相较于 Faster-RCNN 算法,SSD 算法,*mAP*0.5 分别提升了 37.95% 和 43.98%;与 RetinaNet 模型相比,*mAP*0.5 提升了 34.75%;相较于 ATSS-FPN-DyHead 和 TOOD,*mAP*0.5 分别提升了 5.5% 和 3.3%;相较于 DINO,GFL 和 RTMDet,*mAP*0.5 分别提升了 2%,5.6% 和 4%;相较于 CenterNet 模型,*mAP*0.5 提升了 32%;相较于 YOLOX-Tiny,*mAP*0.5 提升了 10.1%;与 YOLOv3 模型、YOLOv4 模型相比,*mAP*0.5 分别提升了 30% 和 27.1%,相较于 RT-DETR 模型,*mAP*0.5 提升了 2.2%。与上述检测模型相比,改进的 MCAI 网络实现了比其他框架更高的精度。

表 5　BDD-100K 数据集的整体性能比较

Tab. 5　Overall performance comparison of BDD-100K dataset

Method	Year	BackBone	*mAP*0.5/%
Faster-RCNN	2016	ResNet50	34.45
SSD	2017	VGG-16	28.42
RetinaNet	2018	ResNet50	37.65
CenterNet	2020	DLA-34	40.40
YOLOv3	2018	MobileNet	42.4
YOLOv4	2020	CSPDarkNet	45.3
YOLOv5s	2020	CSPDarkNet	61
YOLOv7	2022	CSPDarkNet	48.7
RTDETR	2023	ResNet18	70.2
ATSS-FPN-DyHead	2024	ResNet50	66.9
TOOD	2021	ResNet50	69.1
DINO	2023	ResNet50	70.4
YOLOX-Tiny	2021	CSPDarkNet	62.3
GFL	2020	CSPDarkNet	66.8
RTMDet	2023	CSPDarkNet	68.4
MCAI	2024	ResNet18	72.4

在某市实际无人机航拍图像数据集中测试

了MCAI网络，以评估在实际场景中的泛化能力，并与YOLOv3，YOLOv5，YOLOv6，YOLOv8，YOLOv8-Decoder和RT-DETR进行对比，实验结果如表6所示，本研究所提网络的*Rcall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95分别是75.8%，82.8%，80.6%，59.9%，而YOLOV3的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95分别是41.4%，82.6%，52.1%，30.9%，相较于YOLOv3，MCAI的*Recall*，*Precision*分别提升了34.4%和0.2%，*mAP*0.5和*mAP*0.5:0.95分别提升了28.5%和29%；而YOLOv5的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95分别是48.5%，88.1%，56.7%，38.6%，相较于YOLOv5，MCAI的*Recall*提升了27.3，*mAP*0.5和*mAP*0.5:0.95分别提高了23.9%和21.3%；YOLOv6的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95分别是40.4%，72%，45.2%，31.7%，相较于YOLOv6模型，MCAI的*Recall*和*Precision*分别提升了35.4%和10.8%，*mAP*0.5和*mAP*0.5:0.95分别提升了35.4%和28.2%。YOLOv8的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95分别是47.4%，61.3%，52.1%，36.7%，相较于YOLOv8模型，MCAI的*Recall*和*Precision*分别提升了28.4%和21.5%，*mAP*0.5和*mAP*0.5:0.95分别提升了28.5%和23.2%。YOLOv8-Decoder的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95分别是49.5%，76.7%，55.1%和37.2%，相较于YOLOv8-Decoder模型，MCAI的*Recall*和*Precision*分别提升了26.3%和6.1%，*mAP*0.5和*mAP*0.5:0.95分别提升了25.5%和22.7%。RT-DETR的*Recall*，*Precision*，*mAP*0.5，*mAP*0.5:0.95分别是64.9%，87.9%，76.1%和51.7%，相较于RT-DETR模型，MCAI的*Recall*提升了10.9%，*mAP*0.5和*mAP*0.5:0.95分别提升了4.5%和8.2%。在表6中，与YOLO系列相比，改进的跨层注意力交互下的多特征交叉无人机检测网络实现了比其他框架更高的*mAP*。综上所述，MCAI网络通过利用自适应跨层注意力交互模块，有效地捕捉全局和局部的上下文信息，使模型在处理复杂场景时可以有效筛选出关键目标特征。相比之下，YOLO系列模型主要依赖于标准的卷积神经网络结构，局限于局部区域特征，无法有效捕捉长距离依赖和全局特征信息。因此，MCAI网络能够更好地处理背景复杂、目标分布不均匀的实际无人机交通巡检场景。

图9显示了MCAI检测网络在真实数据集LZ Traffic Video的实际应用场景可视化（彩图

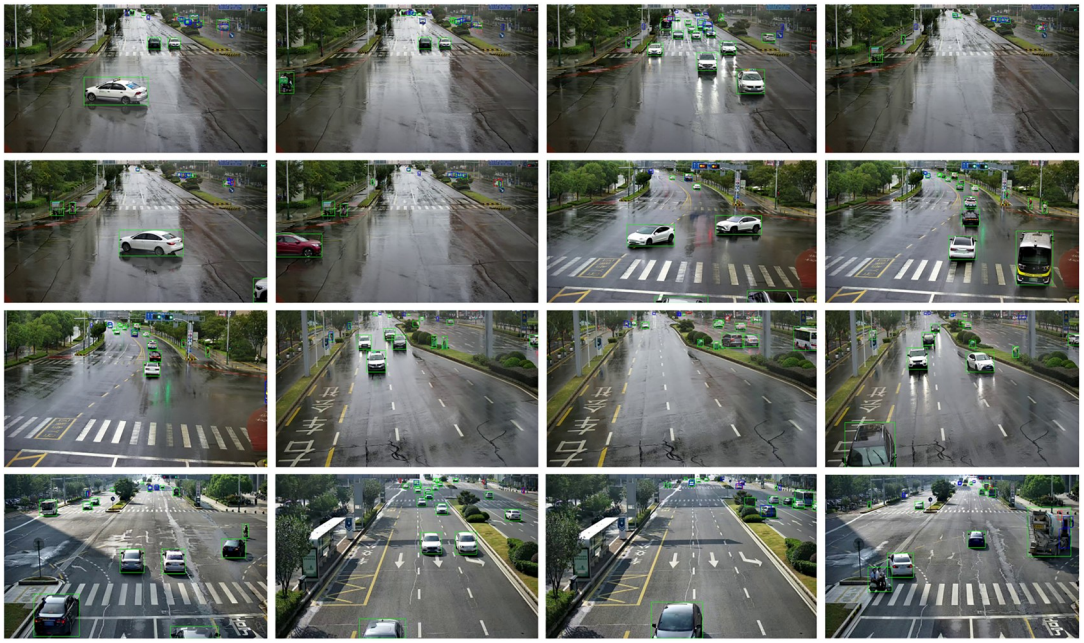


图9　LZ Traffic Video的实际应用场景可视化

Fig. 9　Visualization of the actual application scenario of LZ Traffic Video

表 6　LZ Traffic Video Detection Dataset 的总体性能比较

Tab. 6　Overall performance comparison of LZ traffic video detection dataset

Baseline	*Recall*	*Precision*	*mAP*0.5	*mAP*0.5:0.95
YOLOv3	41.4	82.6	52.1	30.9
YOLOv5	48.5	88.1	56.7	38.6
YOLOv6	40.4	72	45.2	31.7
YOLOv8	47.4	61.3	52.1	36.7
YOLOv8-Decoder	49.5	76.7	55.1	37.2
RT-DETR	64.9	87.9	76.1	51.7
MCAI	75.8	82.8	80.6	59.9

见期刊电子版)。可以看出,MCAI检测模型能够高度准确地检测对象,对于应用场景中的实时性能,本文使用 *TP*、*FP* 和 *FN* 作为指标,其中绿色框(*TP*)表示正确的检测和真实的标签。红色框(*FN*)为漏检,未检测到。蓝色框(*FP*)表示误检测,误检测。

4　结　论

无人机航拍图像中存在背景信息复杂、尺度分布不均、微小目标密集等问题,这对无人机交通巡检任务带来了一定的困难,针对上述问题,本文提出跨层注意力交互下的多特征交叉检测网络。首先,该方法在ResNet18主干网络的末端构建自适应跨层注意力交互模块,通过残差连接和动态稀疏注意力对全局上下文关键特征进行筛选和提取,从而缓解背景噪声问题。在此基础上,构建可变形编码器,通过可变形自注意力机制以扩大特征层感受野,弥补目标丢失特征。最后,为有效识别关注区域内的目标特征,构建多尺度交叉融合模块,将小波变换处理的浅层空间信息引入到特征表示模块中,以获取小目标的显著特征。此外,为验证检测模型的鲁棒性,增加阴雨天气干扰因素,证明了本文算法的鲁棒性更优。实验结果表明,本文所提算法在不同场景中均表现出良好的检测性。

参考文献:

[1] HONG T, LIANG H M, YANG Q Y, *et al*. A real-time tracking algorithm for multi-target UAV based on deep learning[J]. *Remote Sens*, 2022, 15: 2.

[2] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]. *In Proceedings of the* 2001 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR* 2001. IEEE, 2001.

[3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]. 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*'05). *San Diego*, *CA*, *USA*. IEEE, 2005: 886-893.

[4] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, *et al*. Object detection with discriminatively trained part-based models [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(9): 1627-1645.

[5] XIE S N, GIRSHICK R, DOLLÁR P, *et al*. Aggregated residual transformations for deep neural networks[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Honolulu*, *HI*, *USA*. IEEE, 2017: 5987-5995.

[6] GIRSHICK R, DONAHUE J, DARRELL T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014 *IEEE Conference on Computer Vision and Pattern Recognition. Columbus*, *OH*, *USA*. IEEE, 2014: 580-587.

[7] REN S, HE K, GIRSHICK R, *et al*. "Faster R-CNN: Towards real-time object detection with region proposal networks," *in Proc*. Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 1-9.

[8] HE K, G G., D P, *et al*. Mask R-CNN[C]. *Proc. IEEE Int. Conf. Comput. Vis.* (*ICCV*), *Oct*. 2017, 2961-2969.

[9] REDMON J, DIVVALA S, GIRSHICK R, *et al*. You only look once: unified, real-time object detection[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Las Vegas*, *NV*, *USA*. IEEE, 2016: 779-788.

[10] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Honolulu*, *HI*, *USA*. IEEE, 2017: 6517-6525.

[11] REDMON J, FARHADI A. YOLOv3: An Incre-

mental Improvement [EB/OL]. 2018: *arXiv*: 1804.02767. http://arxiv.org/abs/1804.02767

[12] LIU W, ANGUELOV D, ERHAN D, *et al*. SSD: Single Shot Multibox Detector [EB/OL]. 2015: *arXiv*: 1512.02325. http://arxiv.org/abs/1512.02325

[13] TAN M X, PANG R M, LE Q V. EfficientDet: scalable and efficient object detection [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Seattle*, *WA*, *USA*. IEEE, 2020: 10778-10787.

[14] ZHANG S F, WEN L Y, BIAN X, *et al*. Single-shot refinement neural network for object detection [C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 4203-4212.

[15] 王尔申,张宏轩,徐嵩,等. 基于注意力特征融合的无人机检测算法研究[J]. 北京航空航天大学学报,2024: 1-11.
WANG E S, ZHANG H X, XU S, *et al*. Research on UAV detection algorithm based on Attention feature fusion [J]. *Journal of Beijing University of Aeronautics and Astronautics*, 2024: 1-11. (in Chinese)

[16] 张会娟, 李坤鹏, 姬淼鑫, 等. 基于空间相关性增强的无人机检测算法[J]. 浙江大学学报(工学版), 2024, 58(3): 468-479.
ZHANG H J, LI K P, JI M X, *et al*. UAV detection algorithm based on spatial correlation enhancement [J]. *Journal of Zhejiang University* (*Engineering Science*), 2024, 58(3): 468-479. (in Chinese)

[17] 高武奇, 杨婷, 李亮亮. 基于多特征交叉融合及跨层级联的航拍目标检测算法[J]. 西北工业大学学报, 2023, 41(6): 1179-1189.
GAO W Q, YANG T, LI L L. Aerial military target detection algorithm based on multi-feature cross fusion and cross-layer concatenation[J]. *Journal of Northwestern Polytechnical University*, 2023, 41(6): 1179-1189. (in Chinese)

[18] 李校林,刘大东,刘鑫满,等. 改进YOLOv5的无人机航拍图像目标检测算法[J]. 计算机工程与应用, 2024, 60(11): 204-214.
LI X L, LIU D D, LIU X M, *et al*. Improved YOLOv5 UAV Aerial image target Detection Algorithm [J]. *Computer Engineering and Applications*, 2024, 60(11): 204-214. (in Chinese)

[19] WU Z H, LIU C L, WEN J, *et al*. Selecting high-quality proposals for weakly supervised object detection with bottom-up aggregated attention and phase-aware loss[J]. *IEEE Transactions on Image Processing*, 2023, 32: 682-693.

[20] XIE Y M, ZHANG L W, YU X Y, *et al*. YOLO-MS: multispectral object detection via feature interaction and self-attention guided fusion [J]. *IEEE Transactions on Cognitive and Developmental Systems*, 2023, 15(4): 2132-2143.

[21] 陈健松, 蔡艺军. 面向垃圾分类场景的轻量化目标检测方案[J]. 浙江大学学报(工学版), 2024, 58(1): 71-77.
CHEN J S, CAI Y J. Lightweight object detection scheme for garbage classification scenario[J]. *Journal of Zhejiang University* (*Engineering Science*), 2024, 58(1): 71-77. (in Chinese)

[22] ZHAO Y A, LV W Y, XU S L, *et al*. DETRs Beat YOLOs on real-time object detection [C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Seattle*, *WA*, *USA*. IEEE, 2024: 16965-16974.

[23] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition [C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Las Vegas*, *NV*, *USA*. IEEE, 2016: 770-778.