中图法分类号:TP391　　文献标识码: A　　文章编号: 1006-8961(2025)11-3547-17

论文引用格式:Feng Q H, Wang Z X, Sun C C and Shao Z W. 2025. Small object detection in drone images via foreground refinement and multidimensional inductive bias self-attention. Journal of Image and Graphics, 30(11):3547-3563(冯琪涵, 王志晓, 孙成成, 邵志文. 2025. 融合前景细化和多维归纳偏置自注意力的无人机图像小目标检测. 中国图象图形学报, 30(11):3547-3563)[DOI:10.11834/jig.250017]

融合前景细化和多维归纳偏置自注意力的无人机图像小目标检测

冯琪涵[1],王志晓[1*],孙成成[1],邵志文[1,2]

1. 中国矿业大学计算机科学与技术学院,徐州　221116; 2. 上海交通大学计算机科学与工程学院,上海　200240

摘　要: 目的　无人机小目标检测受运动模糊等因素影响,导致目标细节在成像过程中丢失。尽管超分辨技术能够还原目标细节,但全图处理会造成大量计算资源浪费。此外,现有模型通过顺序堆叠卷积和Transformer,因固有的范式冲突导致局部空间特征和全局特征存在分离性。**方法**　本文提出融合前景细化和多维归纳偏置自注意力的无人机小目标检测。前景细化通过多层协同显著性映射方法筛选前景,并利用扩散模型仅对前景进行超分辨处理,从而在减少背景计算负担的同时提高检测精度。多维归纳偏置自注意力网络包括多维自注意力模块、混合增强前馈模块、尺度耦合模块和邻域特征交互模块。多维自注意力将自注意力分解到水平和垂直两个维度,强化对空间信息的感知,同时引入并行的归纳偏置感知路径,实现局部与全局特征的协同表征,避免特征分离。混合增强前馈模块和尺度耦合模块分别通过动态卷积和多重卷积与自注意力交互,能够最大限度保留局部与全局特征。最后,邻域特征交互模块通过逐层聚合邻域特征,确保预测特征图中包含充分的小目标信息。**结果**　在3个数据集上与先进检测方法进行对比实验,精确率、召回率、平均检测精度和交并比(intersection of union,IoU)阈值为0.5的平均准确率均有显著提高,IoU阈值为0.5的平均准确率在3个数据集上分别达到53.2%、38.7%和93.9%,相较于基线方法分别提高9.7%、9.5%和1.2%。**结论**　实验结果表明,所提方法在无人机场景下具有处理复杂背景和小目标检测的强大能力。代码开源链接https://github.com/CUMT-GMSC/MIBSN。

关键词: 小目标检测;Transformer;扩散模型;归纳偏置;无人机图像

Small object detection in drone images via foreground refinement and multidimensional inductive bias self-attention

Feng Qihan[1], Wang Zhixiao[1*], Sun Chengcheng[1], Shao Zhiwen[1,2]

1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou　221116, China;
2. School of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai　200240, China

Abstract: Objective Drones equipped with 4K cameras have emerged as a transformative technology across diverse domains, which range from traffic surveillance and criminal investigation to military reconnaissance and disaster response. The capability to capture aerial images over vast areas enables unprecedented visual analysis capabilities. In drone-based applications, object detection serves as a fundamental technology for scene perception, which facilitates precise object rec-

收稿日期:2025-01-15;**修回日期:**2025-04-05;**预印本日期:**2025-04-12

*通信作者:王志晓　zhxwang@cumt.edu.cn

基金项目:国家自然科学基金项目(62472424,62402496);江苏省自然科学基金项目(BK20242084)

Supported by: National Natural Science Foundation of China (62472424, 62402496); Natural Science Foundation of Jiangsu Province, China (BK20242084)

ognition and tracking to enhance operational efficiency and safety in complex scenes. Despite the success of deep learning-based object detection models on natural scenes, drone-based detection presents unique challenges. A notable challenge is the motion blur introduced by the high-speed flight of the drone, which degrades fine details of small objects in the captured images. High-speed motion often leads to blurred images that are difficult for traditional models to process, especially when detecting small or distant objects that require high spatial resolution. Although recent advances in diffusion models have shown promise in addressing super-resolution tasks (i. e., providing finer details in blurry images), their use in drone images remains limited due to the sparse distribution of objects in the scenes. Full-image super-resolution is computationally expensive and inefficient when the objects of interest are few and scattered across large areas. Another major challenge arises from the high-altitude perspective of drone images, where ground objects appear smaller and are easily overwhelmed by complex background. This inherent scale reduction makes small objects particularly difficult to detect and classify, as their features become less distinctive against varying ground patterns and environmental conditions. The accurate detection of such small objects demands fine-grained local spatial information and global contextual features. Local features are essential for capturing distinctive small object characteristics, while global features provide crucial contextual cues for distinguishing objects from similar background patterns. Although hybrid architectures that combine convolutional neural networks and Transformers attempt to leverage their complementary advantages in local feature extraction and global dependency modeling, these approaches still present inherent limitations. The sequence-based processing of the Transformer, even with positional encoding, fundamentally constrains local spatial perception due to the conversion of 2D spatial structures into 1D sequences. Furthermore, the independent stacking of convolution and self-attention layers, rather than facilitating feature integration, tends to intensifies the disconnect between local and global representations. This architectural limitation results in feature separation rather than meaningful fusion, which particularly impacts the capability of the model to detect and classify small objects in complex aerial scenes. **Method** This study proposes a novel small object detection model based on foreground refinement and a multidimensional inductive bias self-attention to address the aforementioned challenges. We introduce a foreground refinement module (FRM) that employs a class-agnostic multi-layer aggregated activation map to identify regions of interest for overcoming the computational inefficiency of processing high-resolution drone images. This module enables targeted refinement that preserves computational resources and enhances critical small object details. At the core of our method, we propose a multidimensional inductive bias self-attention network (MIBSN), which consists of multidimensional self-attention, hybrid enhanced feed-forward, scale coupling module, and progressive interaction module. The multidimensional self-attention module decomposes attention operations across horizontal and vertical dimensions, and it implements adaptive window-based attention to significantly enhance spatial position awareness. Within the multidimensional self-attention, we innovatively incorporate an inductive bias perception path that operates in parallel with the attention mechanism, which is specifically designed to capture fine-grained local features. The parallel computation generates global contextual information and local spatial details, which are jointly processed by a hybrid enhanced feed-forward module utilizing dynamic convolution kernels. This design ensures the Transformer inherits strong inductive bias properties while maintaining comprehensive feature representation capabilities in complex backgrounds. To effectively handle the multi-scale nature of objects in drone images, the scale coupling module employs adaptive kernel configurations while dynamically interacting with multidimensional self-attention. This way maximizes the preservation of local and global features across different scales. Finally, the progressive interaction module progressively aggregates neighborhood features in the network decoder, which enables sufficient restoration of small object details in the final prediction maps. By integrating FRM and MIBSN, our method provides a comprehensive solution for small object detection. The FRM enhances fine-grained details through targeted refinement of regions of interest, while MIBSN ensures robust feature extraction through its multi-component architecture. Specifically, the combination of local spatial information and global context in MIBSN, along with the enhanced detail preservation from FRM, enables accurate detection and classification of small objects even in complex drone images. **Result** To validate the effectiveness of our proposed method, we conduct extensive experiments on three challenging datasets: vision meets drones 2019 (VisDrone2019), unmanned aerial vehicle detection and tracking (UAVDT), and Northwestern Polytechnical University Very High Resolution 10-Class Dataset (NWPU VHR-10). Our method demonstrates remarkable performance on VisDrone2019, which is a large-scale drone-captured dataset. Specifi-

cally, it achieves 53.2% average precision at the IoU threshold of 0.5 (AP@50) and 32.8% mean average precision across all IoU thresholds from 0.5 to 0.95 (mAP), which substantially surpass those of existing state-of-the-art object detection methods. Notably, compared with the current leading Transformer-based detector DINO (detection Transformer with improved denoising anchor boxes), our approach achieves a significant improvement of 7.9%. On UAVDT benchmark, our approach achieves 38.7% AP@50 and 23.4% mAP. This substantial gain can be attributed to our innovative multidimensional inductive bias self-attention network, which effectively enhances local feature learning and global dependency modeling. To further demonstrate the generalization capability of our method in handling small objects from aerial perspectives, we evaluate our approach on the NWPU VHR-10 remote sensing dataset. Despite the different imaging characteristics between drone and satellite imagery, both scenarios share common challenges in small object detection, including scale variation and complex backgrounds. Our method achieves a state-of-the-art mAP of 64.5% and AP@50 of 93.9% on this dataset, which validates its robust feature extraction capabilities across different aerial imaging domains. We conduct comprehensive ablation studies on FRM and MIBSN to systematically analyze the contribution of each proposed component to the overall performance improvement for providing deeper insights into the effectiveness of our method. **Conclusion** In this study, we propose a novel small object detection method for drone images. This method integrates a foreground refinement module for targeted refinement processing and a multidimensional inductive bias self-attention network for enhanced feature learning. Our approach effectively addresses the challenges of small object detection. Comprehensive evaluations on VisDrone2019, UAVDT, and NWPU VHR-10 datasets demonstrate the superior performance of our method, which greatly advances the detection accuracy of small objects in drone images. The code is available at https://github.com/CUMT-GMSC/MIBSN.

Key words: small object detection; Transformer; diffusion model; inductive bias; drone image

0 引　言

随着无人机技术的迅速普及，其应用场景广泛涉及交通监测（于静茹 等，2024）、农业生产、军事侦察（田昊东 等，2023）、环境监测和灾害救援等（王胜科等，2024）。无人机凭借其高机动性和灵活性，能够快速获取大范围高分辨率视觉数据，为场景理解提供重要支持。目标检测作为无人机场景理解的核心技术，起着至关重要的作用。精确的目标检测（Glenn等，2023）可以帮助无人机在复杂环境中识别和跟踪目标，提高任务执行的效率和安全性。

尽管基于深度学习的目标检测模型（Carion等，2020）在 MS COCO（Microsoft common objects in context）（Lin等，2014）等自然场景下的数据集上表现出色，但针对无人机场景仍存在挑战。由于无人机在高速飞行过程中进行数据采集，图像容易出现运动模糊现象。运动模糊会导致图像中的小目标细节丢失，使目标的边缘变得模糊不清，增加了检测难度。其次，由于无人机的飞行高度较高，地面目标在图像中的尺寸通常较小。小目标在图像中往往会被复杂的背景所淹没，导致其特征难以被准确提取和识别。

针对无人机图像运动模糊现象，深度生成模型取得了显著进展。早期工作（Bai 等，2018；Lei 等，2020；Yang 等，2022）通过构建对抗生成框架有效缓解了图像模糊对检测模型的影响，但受限于模型训练的稳定性问题。随着 Ho 等人（2020）提出扩散模型理论，该领域迎来突破性进展（Croitoru 等，2023）：Saharia 等人（2023）通过引入条件约束增强了模型的定向生成能力，Fei 等人（2023）提出的 GDP（generative diffusion prior）框架突破了传统方法在图像尺寸上的限制，Xiao 等人（2024）则在计算效率与重建精度间实现了更优的平衡。尽管如此，扩散模型的高计算成本在处理无人机图像时面临严峻挑战。如图 1 所示，由于无人机图像中前景目标往往仅占据较小的图像区域，对整幅图像执行细化操作不仅会导致计算资源的过度消耗，而且导致对关键区域的关注度不足，削弱模型对前景目标细节信息的还原能力。

针对无人机视觉场景中因小目标与背景高相似性导致的特征淹没问题，早期工作（Lin 等，2017）构建多尺度特征融合，有效增强了小目标的特征表示。Du 等人（2023）和 Liang 等人（2022）分别引入稀疏卷积、剪枝等技术，在保证特征质量的同时显著提升了

计算效率。尽管基于卷积神经网络(convolutional neural network, CNN)的方法取得了一定成功,但仍需要非极大值抑制等后处理操作。Transformer的引入为目标检测带来了新的范式。DETR(detection Transformer)(Carion等,2020)首次实现了端到端的目标检测框架,但在小目标检测和训练收敛性方面表现不佳。Zhu等人(2021b)和Wang等人(2021)致力于构建多尺度特征表示,DNTR(denoising feature pyramid network with trans region-based convolutional neural network)(Liu等,2024)则利用局部区域特征融合全局信息,实现高检测精度,但此类方法仍需要大量数据来稳定收敛过程。DETR中的对象查询是固定数量的可学习向量,一些研究已经注意到初始对象查询的设计显著影响收敛速度,Zhang等人(2022)提出两阶段初始化策略,虽然提升了收敛速度,但固定数量的查询难以适应无人机场景中目标数量的动态变化。为此,Huang等人(2025)设计了类别计数模块实现动态查询选择,但此类方法在小目标检测精度和推理速度间的权衡仍待优化。RTDETR(real time DETR)(Zhao等,2024)在保持检测精度的同时实现了实时推理。

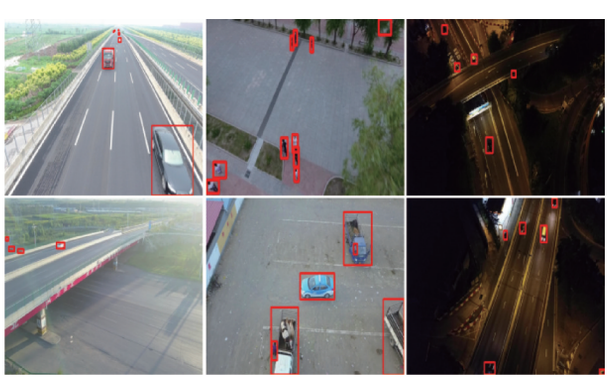


图1　无人机中前景目标占比示意

Fig. 1　Illustration of foreground object proportion in drone

另外,基于Transformer的目标检测模型将图像转换为一维序列,并在全局范围内计算特征之间的关系(Dong等,2022),这种全局建模能力虽然有助于捕捉长距离依赖关系,但在建模局部空间位置信息缺乏卷积操作的强归纳偏置,因此不利于从复杂的背景中定位只占极少像素的小目标。现有研究(Zhu等,2021a;Zhang等,2021)通过顺序堆叠Transformer和卷积层平衡全局与局部特征,但这种简单的叠加策略往往导致特征表示的分离性,即全局特征和局部特征独立处理,缺乏有效的交互融合机制,使得模型难以同时利用两种特征的优势。其次,该类模型仅使用单尺度特征图进行预测或直接聚合信息差异较大的特征,这种聚合策略造成信息不足或引入噪声,削弱了对小目标特征的可分性。

本文提出融合前景细化和多维归纳偏置自注意力的无人机图像小目标检测模型,以解决上述问题。具体而言,设计了前景细化模块(foreground refinement module, FRM)。该模块通过类别无关的多层协同显著性映射方法(class-agnostic multi-layer aggregated activation map, CAM-AM)提取包含前景区域的图像块,过滤掉不影响小目标检测性能的背景区域,随后,利用条件扩散模型仅生成前景区域的细化图像,从而减少条件扩散模型的输入数据量,降低资源消耗的同时复原小目标的关键细节,进而提升小目标检测性能。其次,提出多维归纳偏置自注意力网络(multidimensional inductive bias self-attention network, MIBSN)。该网络包含多维自注意力模块(multidimensional self-attention, MSA),将自注意力在水平和垂直两个空间维度上并行分解,从而显著提高自注意力对空间位置的感知能力。同时,在MSA内部引入归纳偏置感知路径(inductive bias perception path, IBPP),通过在模型训练过程中注入归纳偏置,增强模型对局部特征的编码能力。该路径与多维自注意力并行,使得自注意力模块能够协同学习局部特征和全局依赖关系,避免特征分离问题。MIBSN的混合增强前馈模块(enhanced hybrid feed-forward, EHFF)通过引入动态卷积核,允许模型在应对复杂背景时动态调节特征表达,进一步提升对局部特征的敏感度。此外,MIBSN的尺度耦合模块(scale coupling module, SCM)通过多尺度卷积核的灵活选择与自注意力的动态交互,适应无人机场景中目标尺度的多样性。最后,MIBSN的邻域特征交互模块(localized neighborhood interaction, LNI)通过级联交互逐层聚合邻域特征,以渐进的方式逐层恢复目标信息,确保最终的预测特征图保留充分的小目标信息。

本文主要贡献如下:1)设计了前景细化模块。首先,提出类别无关的多层协同显著性映射方法以有效筛选包含前景区域的图像块,随后利用条件扩散模型仅对选中区域进行细节重建,降低资源浪费

的同时提高小目标检测性能。2)设计了多维归纳偏置自注意力网络。首先多维自注意力模块通过将自注意力在水平和垂直维度上分解，能够有效捕捉空间位置信息，补偿图像转为一维序列后空间结构的损失，并引入归纳偏置路径实现局部与全局特征的协同学习，避免了特征分离问题。融合后的特征经过混合增强前馈模块，该模块通过动态卷积核自适应地关注不同目标区域，进一步加深对局部特征的感知能力。接下来，尺度耦合模块通过多层次的卷积与自注意力的交互操作，确保局部与全局信息的平衡保留。最后，邻域特征交互模块通过汇聚多层次的邻域信息，确保编码的小目标关键信息能够充分保存在预测特征图中。3)在无人机图像数据集 VisDrone2019（vision meets drones 2019）（Du 等，2019）、UAVDT（unmanned aerial vehicle detection and tracking）（Du 等，2018）和遥感图像数据集 NWPU VHR-10（Northwestern Polytechnical University Very High Resolution 10-Class Dataset）（Cheng 等，2016）上进行了实验验证。所提方法在无人机图像小目标检测任务中取得较高的检测精度，在遥感图像数据集上的良好表现进一步验证了方法的泛化能力。

1　方　法

本文提出融合前景细化和多维归纳偏置自注意力的无人机图像小目标检测模型，模型总体结构如图 2 所示。前景细化模块通过类别无关的多层协同显著性映射方法提取前景区域，将其输入条件扩散模型进行图像细化。多维归纳偏置自注意力网络首先通过多维自注意力模块在水平和垂直两个空间维度上并行分解自注意力操作，同时引入归纳偏置感知路径，在训练过程中注入归纳偏置加强对局部特征的捕捉。该路径与多头自注意力并行处理，最终将融合后的特征输入到混合增强前馈模块。该网络采用了双重结构设计，包含与自注意力交互的尺度耦合模块，有效平衡了局部与全局特征的表达。在解码过程中，该网络进一步集成了邻域特征交互模块，逐层聚合邻域特征，能够确保最终的预测特征图包含充分的小目标信息。

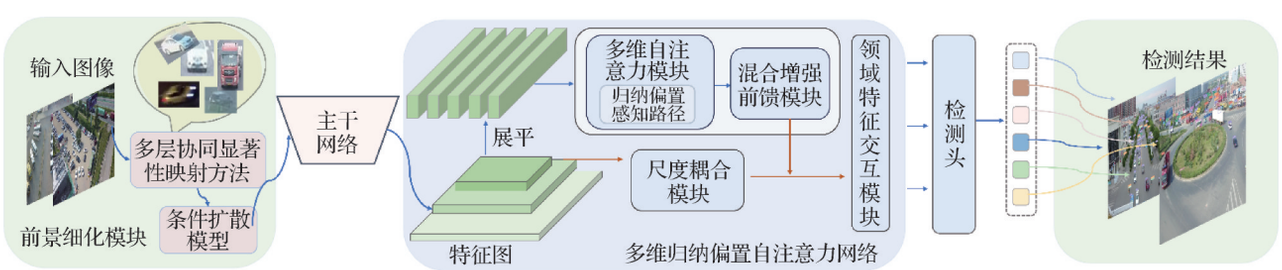


图 2　总体框架

Fig. 2　The overall architecture of our model

1.1　前景细化模块

针对无人机视觉数据中的运动模糊现象，本文设计了前景细化模块。结构如图 3 所示。该模块通过类别无关的多层协同显著性映射方法提取前景区域。其中，主特征向量表征了特征空间中前景—背景的最佳判别方向，其方向编码了特征空间中前景与背景的本质差异；而激活图则量化了图像各区域沿该方向的投影强度，高激活值区域对应潜在前景目标。通过这种基于特征空间统计特性的映射关系，实现了前景区域的精确提取以及过滤无关背景，从而减少了条件扩散模型的输入数据量。

类别无关的多层协同显著性映射方法中，前景与背景的区分基于特征空间中统计分布的显著性差异。具体而言，首先从主干网络提取多尺度特征图 $\boldsymbol{A}^{(l)} \in \mathbf{R}^{\mathcal{H}_l \times \mathcal{W}_l \times \mathcal{C}_l}$，$l = 1, 2, \cdots, L$，实现从低层次边缘纹理到高层次语义信息的全面捕捉。特别在目标密集分布的场景下，多尺度特征的协同作用能够有效捕捉不同尺度目标的特征表达，显著增强模型对密集小目标的感知能力。对每层特征图，通过变换 $\mathcal{F}$: $\boldsymbol{A}^{(l)} \rightarrow \boldsymbol{A}^{'(l)} \in \mathbf{R}^{(\mathcal{H}_l \times \mathcal{W}_l) \times \mathcal{C}_l}$ 将其展平，并计算每层特征图的协方差矩阵 $\boldsymbol{C}^{(l)}$，量化特征通道之间的线性相关性，具体为

$$\boldsymbol{C}^{(l)} = \frac{1}{\mathcal{H}_l \times \mathcal{W}_l} \left(\boldsymbol{A}^{'(l)} \right)^{\mathrm{T}} \boldsymbol{A}^{'(l)} \tag{1}$$

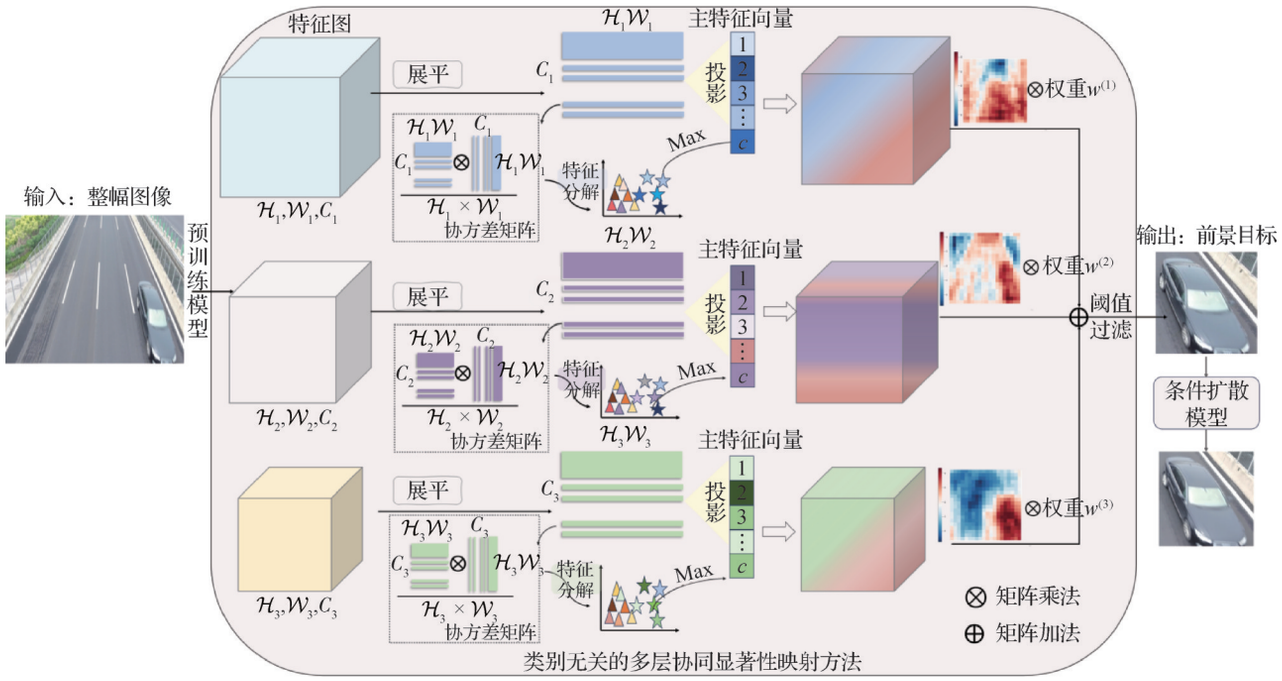


图3　前景细化模块结构

Fig. 3　The architecture of foreground refinement module

式中，$\mathcal{H}_l$和$\mathcal{W}_l$分别代表第l层特征图$\boldsymbol{A}^{(l)}$的高度和宽度。

随后，对每层特征图的协方差矩阵进行特征分解，求解其特征对$\{(\boldsymbol{\lambda}_i, \boldsymbol{v}_i)\}_{i=1}^{\mathcal{C}_l}$。选择对应最大特征值$\boldsymbol{\lambda}_{\max}^{(l)}$的特征向量作为主特征向量，对应的特征值称为主特征值。具体计算为

$$\boldsymbol{C}^{(l)}\boldsymbol{v}_i^{(l)} = \boldsymbol{\lambda}_i^{(l)}\boldsymbol{v}_i^{(l)} \tag{2}$$

式中，$\boldsymbol{v}_i^{(l)}$和$\boldsymbol{\lambda}_i^{(l)}$分别为协方差矩阵$\boldsymbol{C}^{(l)}$的特征向量和特征值。主特征向量计算为

$$\boldsymbol{v}_{\max}^{(l)} = \arg\max_{\boldsymbol{v}_i^{(l)}} \boldsymbol{\lambda}_i^{(l)} \tag{3}$$

主特征向量$\boldsymbol{v}_{\max}^{(l)}$表征特征空间中数据分布方差最大的正交基方向，其指向本质反映了前景与背景特征响应在统计分布上的最大可分性方向。将特征图线性投影到主特征向量上生成显著性激活图，具体为

$$\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}\,(l)} = \boldsymbol{A}^{(l)}\boldsymbol{v}_{\max}^{(l)} \tag{4}$$

式中，$\boldsymbol{v}_{\max}^{(l)}$代表主特征向量，$\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}\,(l)}$为$\boldsymbol{A}^{(l)}$经投影得到的显著性激活图，其强度反映像素与主方向的对齐程度。前景区域因在$\boldsymbol{v}_{\max}^{(l)}$方向上具有显著高维响应，投影后呈现高激活值，而背景区域则表现为低激活值。基于此，前景区域定义为在主特征向量方向上投影值显著偏离背景分布的像素集合。

随后，根据每层特征图的主特征值大小，为每层特征图动态分配权重$w^{(l)}$，具体为

$$w^{(l)} = \frac{\boldsymbol{v}_{\max}^{(l)}}{\sum_{l=1}^{\mathcal{L}} \boldsymbol{v}_{\max}^{(l)}} \tag{5}$$

式中，主特征值$\boldsymbol{v}_{\max}^{(l)}$反映了该层对模型最终决策的贡献度，权重分配机制确保了更具判别性的特征层获得更高的权重，当目标密集分布时，具有强判别性的特征层会自适应获得更高权重，有效提升了模型区分相邻目标的能力。最后，将多个层的激活图按权重加权平均，生成最终的激活图，并将最终的激活图归一化。归一化处理能进一步增强激活图的对比度，使得密集分布的目标区域能够被清晰区分。计算式为

$$\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}} = \sum_{l=1}^{\mathcal{L}} w^{(l)} \boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}\,(l)} \tag{6}$$

$$\boldsymbol{f}_{\mathrm{CAM\ norm}}^{\mathrm{AM}} = \frac{\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}} - \min\left(\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}}\right)}{\max\left(\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}}\right) - \min\left(\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}}\right)} \tag{7}$$

式中，$\min\left(\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}}\right)$和$\max\left(\boldsymbol{f}_{\mathrm{CAM}}^{\mathrm{AM}}\right)$分别表示未归一化激活图中的最小值和最大值。最后，设定动态阈值$\tau = \mu + 2\sigma$（μ和σ为最终激活图均值与标准差），通过二值化操作（$\boldsymbol{f}_{\mathrm{CAM\ norm}}^{\mathrm{AM}} > \tau$）生成前景掩码$M_{\mathrm{fg}}(i,j)$，此过程等价于在特征空间中沿主特征向量方向划分超平面，实现前景—背景线性可分，为条件扩散模型提供

关键的空间引导信息。这种基于主特征投影的前景提取方法，有效利用了特征空间中的显著性信息，并提升了计算效率，即使是在停车场等密集场景，条件扩散模型也只需要处理前景数据，计算复杂度从 $O(HWC)$ 下降到 $O\left(M_{\mathrm{fg}}C\right)$，其中，$M_{\mathrm{fg}}$ 表示前景区域的像素数量。

针对前景区域，本文采用条件扩散模型重构目标细节，对背景区域则使用双线性插值进行上采样，以提升效率。在条件扩散模型中，噪声数据的重建基于原始数据的条件信号，从而确保生成结果满足特定需求。本文将低分辨率图像作为条件嵌入去噪过程，使得生成的高分辨图像与输入保持整体结构的一致性。模型包括前向扩散过程和反向生成过程，前向扩散过程通过逐步添加噪声将真实数据 $\boldsymbol{x}_0$ 转换为噪声数据 $\boldsymbol{x}_T$，具体为

$$q\left(\boldsymbol{x}_t\left|\boldsymbol{x}_{t-1}\right.\right)=N\left(\boldsymbol{x}_t;\sqrt{1-\beta_t}\,\boldsymbol{x}_0,\beta_t\boldsymbol{I}\right) \tag{8}$$

式中，t 为时间步，$t=1,2,\cdots,T$，$\boldsymbol{x}_t$ 表示当前时间步 t 的噪声图像，β_t 是时间步 t 的噪声系数，$N(\cdot)$ 为高斯分布，$\boldsymbol{I}$ 为单位矩阵。通过多次迭代，最终生成的 $\boldsymbol{x}_T$ 近似纯高斯噪声。

反向生成过程通过将噪声数据 $\boldsymbol{x}_T$ 逐步去噪，还原为数据样本 $\boldsymbol{x}_0$。在每个时间步 t 内，模型首先需要预测噪声，然后基于该预测值估计均值，进而利用估计的均值生成下一时间步的数据 $\boldsymbol{x}_{t-1}$。具体步骤如下：

1）噪声预测。条件扩散模型的神经网络接收当前时间步 t 的噪声样本 $\boldsymbol{x}_t$、时间步 t 以及条件信息 z，即低分辨率图像，并输出预测噪声 $\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t,t,z\right)$。

2）估计均值。基于预测噪声，模型计算出生成下一时间步所需的均值 $\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t,z\right)$，其表达式为

$$\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t,z\right)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t,t,z\right)\right) \tag{9}$$

式中，$\alpha_t=1-\beta_t$，代表去噪强度，$\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$，代表噪声累积值。

3）生成数据。根据均值 $\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t,z\right)$，模型从高斯分布中采样生成下一时间步的图像 $\boldsymbol{x}_{t-1}$，逐步去噪直至生成最终的清晰图像，计算过程为

$$p_\theta\left(\boldsymbol{x}_{t-1}\left|\boldsymbol{x}_t,z\right.\right)=N\left(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t,z\right),\sigma_t^2\boldsymbol{I}\right) \tag{10}$$

式中，z 为条件信息（低分辨率图像），$\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,t,z\right)$ 和 σ_t^2 分别为条件扩散模型通过神经网络估计的均值和方差，本文将方差 σ_t^2 设置为 β_t。

1.2 多维归纳偏置自注意力网络

自注意力机制依赖固定大小的补丁进行序列化处理，这不仅降低了对空间位置信息的敏感度，还忽视了局部细节的捕捉。而卷积与自注意力模块的顺序堆叠虽然分别擅长捕捉局部和全局信息，但在架构上未能实现两者的深度融合，导致全局与局部特征分离，无法协同作用。为解决这一问题，本文提出多维归纳偏置自注意力网络，网络结构如图 4 所示。

网络首先引入多维自注意力模块。该模块将自注意力操作分解在水平和垂直两个空间维度上并行执行。给定输入特征图 $\boldsymbol{S}\in\mathbf{R}^{(H\times W)\times C}$，将其分为两个部分 $\left\{\boldsymbol{S}_{\mathrm{H}},\boldsymbol{S}_{\mathrm{V}}\in\mathbf{R}^{(H\times W)\times C'},C'=C/2\right\}$。对于水平方向，特征图 $\boldsymbol{S}_{\mathrm{H}}$ 划分为均等的水平条带 $\boldsymbol{X}^i$，其中，$i=1,2,\cdots,M$。每个条带 $\boldsymbol{X}^i$ 的高度为 h，即 $\boldsymbol{X}^i\in\mathbf{R}^{h\times W\times C/2}$。如果高度 H 不是 h 的整数倍，最后一个水平条带的大小通过下式得到，具体为

$$\boldsymbol{X}^M=pad\left(\boldsymbol{X}^M,p_s=\left(h\times M-H,0\right)\right) \tag{11}$$

式中，$pad(\cdot)$ 代表填充操作，填充大小为 p_s，高度填充为 $h\times M-H$，宽度方向不进行填充。通过上述步骤，特征图被划分为 M 个不重叠的水平条带，每个条带的高度为 h，宽度为 w，通道维度为 $C/2$。对于每个水平条带 $\boldsymbol{X}^i$，通过 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$ 计算自注意力，具体为

$$\tilde{\boldsymbol{X}}^i=f_{\mathrm{Attention}}\left(\boldsymbol{W}^Q\boldsymbol{X}^i,\boldsymbol{W}^K\boldsymbol{X}^i,\boldsymbol{W}^V\boldsymbol{X}^i\right) \tag{12}$$

式中，$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$ 分别是用于生成查询、键和值的参数矩阵。为了减少参数开销，模型采用值共享策略，直接使用值表示作为对应的键，即 $K=V$。最后，将所有水平条带的自注意力结果 $\tilde{\boldsymbol{X}}^i$ 通过拼接操作得到整体的水平自注意力输出，具体为

$$H_{\text{-Attention}}=f_{\mathrm{Concat}}\left(\tilde{\boldsymbol{X}}^1,\tilde{\boldsymbol{X}}^2,\cdots,\tilde{\boldsymbol{X}}^M\right) \tag{13}$$

式中，$f_{\mathrm{Concat}}(\cdot)$ 代表张量拼接操作。垂直方向类似，具体为

$$V_{\text{-Attention}}=f_{\mathrm{Concat}}\left(\tilde{\boldsymbol{Y}}^1,\tilde{\boldsymbol{Y}}^2,\cdots,\tilde{\boldsymbol{Y}}^M\right) \tag{14}$$

式中，$\tilde{\boldsymbol{Y}}^i$ 是第 i 个垂直条带自注意力的计算结果。除此之外，在多维自注意力模块内部嵌入平行的归纳偏置感知路径。该路径通过特征空间分解和自适应投影，可以有效地注入归纳偏置并利用硬件加速提升计算效率。给定 $\boldsymbol{S}_{\mathrm{H}}\in\mathbf{R}^{(H\times W)\times C'}$，首先通过结构化分解将特征维度转化为 $\boldsymbol{S}_{\mathrm{H}}\in\mathbf{R}^{(H\times W)\times p\times q}$，$p$ 和 q 满足 $C'\approx p\times q=2^{\log_2(C')/2}\times 2^{\log_2(C')-\log_2(C')/2}$。基于分解后的特征空

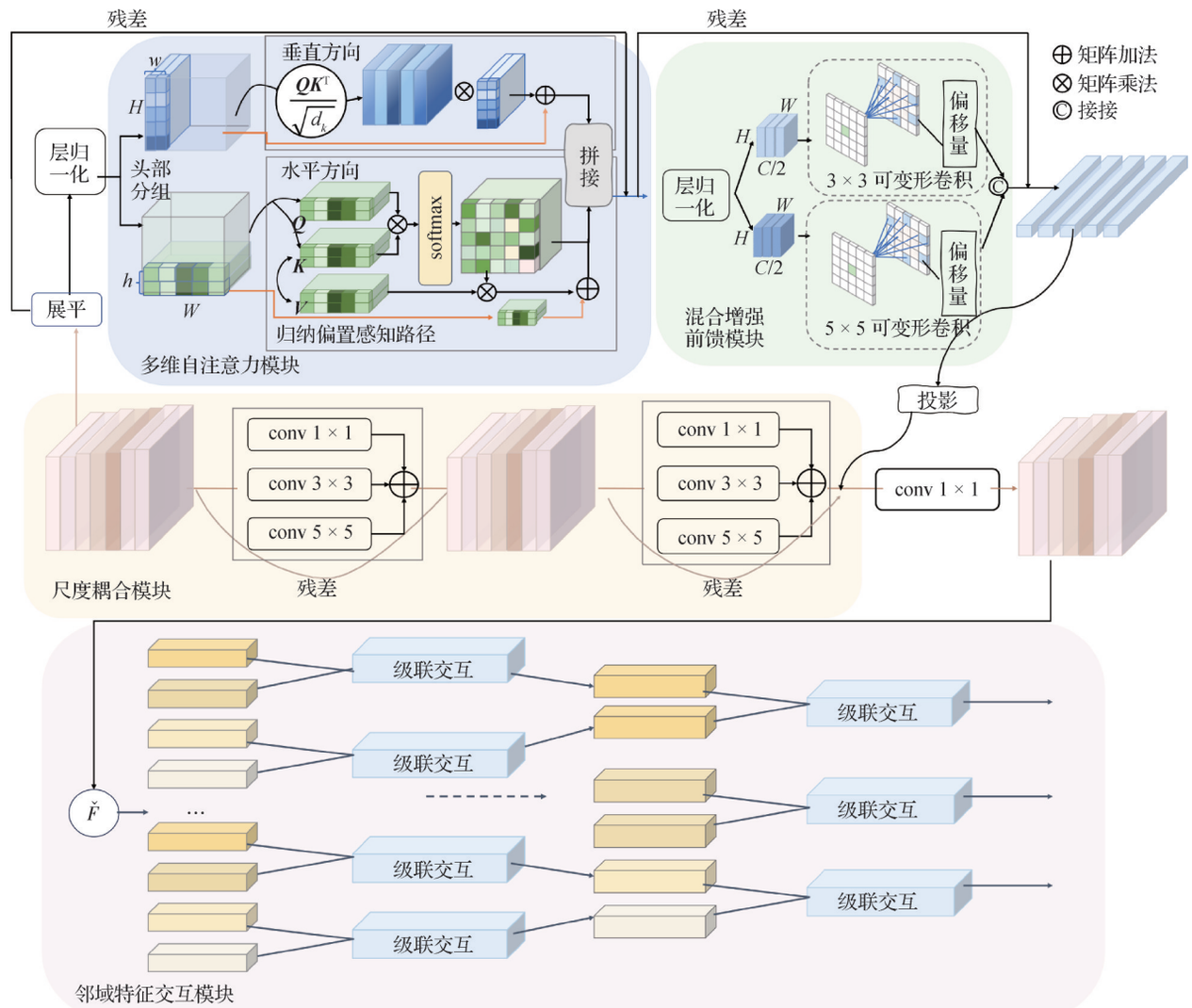


图4　多维归纳偏置自注意力网络结构

Fig. 4　The architecture of multidimensional inductive bias self-attention network

间，采用双重投影策略进行特征变化，最终输出为

$$IBPP(\boldsymbol{S}_{\mathrm{H}}) = f_{\mathrm{Flatten}}\left(\sigma(\boldsymbol{S}_{\mathrm{H}}\boldsymbol{W}^{p})\boldsymbol{W}^{q}\right) \tag{15}$$

式中，σ 表示 SiLU（sigmoid gated linear unit）激活函数，$\boldsymbol{W}^{p} \in \mathbf{R}^{(p\times p)}$ 为内部投影矩阵，$\boldsymbol{W}^{q} \in \mathbf{R}^{(q\times q)}$ 为外部投影矩阵。内部投影矩阵实现投影维度内的特征重组，外部投影矩阵完成维度映射。

为了使模型在不同子空间中学习多样化的特征表示，引入了多头自注意力机制。多头机制中，头部分为两组，分别处理水平和垂直两个维度，从而实现两个维度自注意力的并行计算。计算过程为

$$z_k = \begin{cases} H_{\text{-Attention}_k} + IBPP(\boldsymbol{S}_{\mathrm{H}}) & k = 1,\cdots,\dfrac{K}{2} \\ V_{\text{-Attention}_k} + IBPP(\boldsymbol{S}_{\mathrm{V}}) & k = \dfrac{K}{2}+1,\cdots,K \end{cases} \tag{16}$$

式中，K 为总头数。

最后，两个维度并行的自注意力结果拼接起来，得到最终的输出，具体为

$$MSA = \boldsymbol{W}^{O} f_{\mathrm{Concat}}(z_1, z_2, \cdots, z_K) \tag{17}$$

式中，$\boldsymbol{W}^{O}$ 是输出层参数。多维自注意力实现了在两个正交方向上捕捉特征，既可以聚合局部信息，增强模型对空间位置信息的感知能力，又能通过累积水平和垂直两个维度的编码特征实现对全局信息的捕捉。

多维自注意力模块虽然增强了对局部信息的编码，但在无人机视觉数据中，小目标往往存在复杂的形状变化。为了增强网络对局部几何形状的敏感度，本文又设计了混合增强前馈模块，将自注意力编码得到的特征输入到具有 $c \times r$ 卷积核的卷积层，沿

通道方向将特征图分割为 r 组独立的特征图 $\boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_r$。为节省计算量，本文只设置了两组特征图，每个特征图通道数为 $c/2$。接下来，将 $\boldsymbol{U}_1$ 和 $\boldsymbol{U}_2$ 输入到可变形卷积，以适应不同目标的形状和尺寸，提高网络对复杂几何形状目标的捕捉能力。计算过程为

$$EHFF(\boldsymbol{X}') = LN\left(F_2\left[\phi_3(\boldsymbol{U}_1) \| \phi_5(\boldsymbol{U}_2)\right]\right) \quad (18)$$

式中，F_2 代表全连接层，LN(layer normalization)代表层归一化，$\|$ 代表向量拼接操作，$\phi_d(\cdot)$ 代表偏移量为 d 的可变形卷积。

多维归纳偏置自注意力网络采用双重结构，纳入另外一条与自注意力进行交互的尺度耦合模块 SCM，使用 1 × 1、3 × 3 和 5 × 5 这 3 个具有不同膨胀率的多重卷积，将输入图像下采样至多尺度特征空间。通过这种双重结构，能够有效保证模型最大程度保留局部特征和全局特征。虽然基于 Transformer 的目标检测模型利用多尺度特征的稀疏采样更新了编码特征，显著提高多尺度特征的使用效率。但在解码过程中，没有充分利用编码后的多尺度特征。多尺度特征有助于分离目标和背景，特别是在复杂背景下。利用多尺度特征，可以更有效地抑制噪声，增强对小目标的敏感性，从而提高检测的精度和鲁棒性。因此，本文在多维归纳偏置自注意力网络中设计了邻域特征交互模块，其核心是逐层聚合特征图。首先将从混合增强前馈模块和尺度耦合模块输出的特征通过 1 × 1 卷积融合后获得高维序列特征，然后通过变换 $\check{F}$ 转换为 $\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_n$，以适应后续的处理。级联交互通过聚合相邻的特征进行空洞卷积和 ReLU(rectified linear unit)非线性变化，能够有效地增强特征图之间的联系，增强局部特征的聚合效果，同时扩大感受野，逐步提取小目标的细微特征，计算过程为

$$\boldsymbol{F}_i' = f_{\text{ReLU}}\left(f_{\text{Conv dil}}\left(\boldsymbol{F}_i \| \boldsymbol{F}_{i+1}\right)\right) \quad (19)$$

式中，$f_{\text{Conv dil}}(\cdot)$ 表示空洞卷积。

2 实　验

2.1 数据集及评价指标

VisDrone2019 是由中国计算机学会为无人机视觉挑战赛组织的数据集，旨在推动无人机视觉算法的发展。该数据集涵盖多种城市和乡村环境的视觉任务，如目标检测、目标跟踪和视频目标检测，包含 10 209 幅图像，分辨率从 960 × 540 像素到 2 000 × 1 500 像素不等，提供行人、车辆、自行车等 11 类目标的边界框标注，详细标注目标的类别、位置、尺寸以及可见部分等信息。该数据集是无人机小目标检测任务常用的一个数据集。

UAVDT 数据集是一个专门针对无人机视角目标检测与跟踪的大规模基准数据集。该数据集包含约 80 000 帧高质量视频图像，分辨率在 1 080 像素左右。这些图像来自不同高度、不同视角和不同位置的无人机航拍场景。数据集中标注了超过 841 488 个边界框，涵盖了车辆、行人等多个目标类别。UAVDT 数据集在无人机视觉感知研究中具有重要的参考价值。

NWPU VHR-10 数据集则是一个高分辨率遥感图像数据集，图像来源于 Google Earth，分辨率在 0.5 m～2 m 之间。包含 800 幅分辨率为 1 000 × 1 000 像素的图像，标注了包括飞机、船只、车辆等 10 类目标，目标在图像中具有不同的尺度和旋转角度，适用于在复杂场景下进行小目标检测和识别的研究。本文随机采样 650 幅图像作为训练集，剩下的 150 幅图像用于验证。

本文采用 4 个标准指标进行性能评估：精确率(precision，P)、召回率(recall，R)、多 IoU 阈值(0.5∶0.95)下的平均精度均值(mean average precision，mAP)以及 IoU 阈值为 0.5 时的平均检测精度(average precision，AP@50)。

2.2 实验设置

本文所有实验均在 Ubuntu 20.04.4 环境下进行，CPU 为 AMD EPYC 7402P 24-Core Processor，内存 86 GB，GPU 为 2 块 Tesla T4，显存 12 GB。算法采用 Meta AI 开源的 PyTorch 深度学习框架实现，基线模型为 RTDETR。模型采用 Ultralytics 开发的 Auto 优化器，根据训练数据集和模型架构自动选择最佳的优化器，简化训练过程。默认权重衰减为 5×10^{-4}，动量为 9×10^{-1}。初始学习率设置为 1×10^{-2}，在 VisDrone2019 训练集上训练 80 个 epoch。输入图像长边设置为 1 536，批处理大小为 4。在 UAVDT 数据集上同样训练 80 个 epoch，输入图像长边设置为 1 080，批处理大小为 10。在 NWPU VHR-10 数据集上训练 50 个 epoch，输入图像边长设置为 1 024 像素。

2.3 对比实验

VisDrone2019的实验结果如表1所示，本文方法在测试集上的AP@50和mAP分别达到53.2%和32.8%，显著优于大多数先进的目标检测方法。对比算法主要分为两类：基于CNN的方法包括通用目标检测YOLOv8（you only look once version 8）（Glenn等，2023）和YOLOv9（Wang等，2025），以及面向无人机场景优化的CEASC（context-enhanced adaptive sparse convolutional network）（Du等，2023）和EdgeYOLO（Liang等，2022）；基于Transformer的方法包括DETR（detection Transformer）（Carion等，2020）、Deformable DETR（Zhu等，2021b）、DINO（DETR with improved denoising anchor boxes）（Zhang等，2022），以及面向无人机场景的DNTR（denoising transformer）（Liu等，2024）和RTDETR（real time DETR）（Zhao等，2024）。与现有较先进的基于Transformer的目标检测方法相比，本文方法同样展现出明显的优势，尤其是相较于DINO，mAP实现了7.9%的提升。

表1 VisDrone2019测试集上实验结果对比

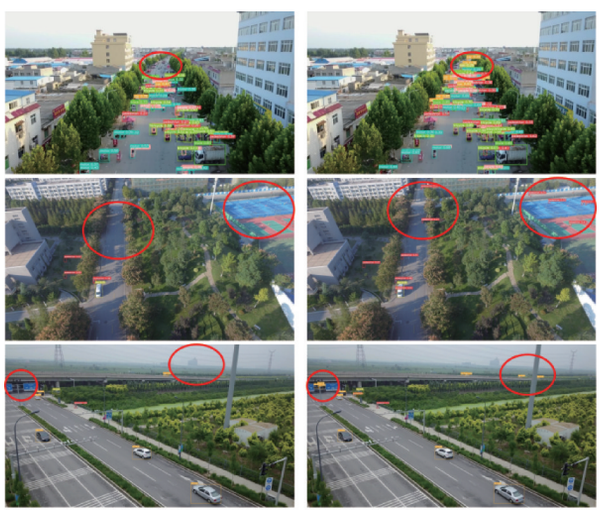
Table 1 The comparative experimental results on VisDrone2019 testset

/%

模型	主干网络	精确率	召回率	AP@50	mAP
YOLOv8	CSPDarkNet	57.6	45.8	46.1	27.7
YOLOv9	GELAN	56.9	47.0	46.7	28.3
CEASC	ResNet50	61.0	44.2	48.7	26.6
EdgeYOLO	ELAN	59.2	45.6	43.0	24.4
DETR	ResNet50	52.8	39.0	40.0	21.9
Def-DETR	ResNet50	56.6	42.9	41.0	24.4
DINO	ResNet50	59.5	45.8	43.1	24.9
DNTR	ResNet50	61.2	47.2	48.1	28.3
RTDETR	HGNetv2	60.4	44.1	43.5	25.0
本文	HGNetv2	**61.6**	**52.7**	**53.2**	**32.8**

注：加粗字体表示各列最优结果。

图5展示了本文方法与RT-DETR方法在VisDrone2019数据集上的检测效果对比。实验结果表明，在复杂场景下，本文方法对远距离小目标和遮挡目标的检测能力显著优于RTDETR。特别是在处理树木遮挡、远距离低分辨率等挑战性场景时，本文方法均能保持较高的检测精度和置信度。如图中红色标注区域所示，对于树木遮挡区域内的车辆目标、远处运动场和道路区域内的多个遮挡行人目标，以及较远距离高架桥上低分辨率车辆目标，本文方法都展现出了优秀的检测性能。这表明本文提出的多维归纳偏置机制不仅保持了Transformer的全局建模优势，还有效捕获了对小目标检测至关重要的局部空间细节信息。



RTDETR　　本文

图5 本文方法与RTDETR的检测效果可视化对比

Fig. 5 Visual comparison of the detection results between RTDETR and ours

表2展示了不同模型在UAVDT数据集的性能对比。实验结果表明，本文方法综合性能表现出显著优势。本文方法的AP@50和mAP达到35.8%和22.4%，证实了多维归纳偏置自注意力网络在增强小目标空间位置信息感知方面的有效性。引入FRM后，模型的mAP为23.4%，超越DNTR模型2%，同时AP@50提升至38.7%。RTDETR通过Transformer的全局建模能力以及匈牙利匹配避免了传统卷积检测器中非极大值抑制引发的误检问题，准确率达到45.6%，但是，其基于高置信度筛选的保守预测策略导致召回率显著降低，仅有38.1%。相比之下，本文提出的多维自注意力策略展现出更均衡的检测性能。特别是在召回率方面，本文方法达到45.2%，显著优于其他对比方法，表明多维归纳偏置自注意力网络对局部细节信息的增强以及FRM中前景—背景解耦设计显著降低了复杂背景导致的小目标漏检。

NWPU VHR-10遥感数据集上的实验结果如表3所示，本文方法在复杂遥感场景中也表现出了显著的优势，最优mAP达到64.5%，明显优于当前主流的目标检测方法。相比其他检测方法，本文方法不仅在整体精度上取得了显著提升，更重要的是，在处理小目标和复杂背景的情况下，展现了更高的鲁棒性和适应性。

表2　UAVDT上实验结果对比

Table 2　The comparative experimental results on UAVDT

/%

模型	主干网络	精确率	召回率	AP@50	mAP
EdgeYOLO	ELAN	36.2	31.8	25.0	13.7
YOLOv9	GELAN	34.8	39.5	32.4	21.2
DNTR	ResNet50	40.5	37.9	33.0	21.4
RTDETR	HGNetv2	**45.6**	38.1	29.2	16.8
本文	HGNetv2	41.6	41.4	35.8	22.4
本文-FRM	HGNetv2	42.2	**45.2**	**38.7**	**23.4**

注：加粗字体表示各列最优结果。

进一步分析单个类别的mAP结果，如表4所示，本文方法在大多数类别上均优于其他现有方法。尽管TPHYOLO（Transformer prediction head YOLO）以及YOLO系列模型采用了聚焦和有效的数据增强策略，在如储罐等背景相对简单、目标尺寸较大的类别上表现出色，但在港口和地面轨道等背景复杂的类别上表现欠佳。这主要是因为这些模型在复杂背景下难以有效区分目标和背景，容易受到背景噪声的干扰，导致检测精度下降。相比之下，本文方法通过引入多维归纳偏置自注意力网络，显著增强了模型在复杂背景下提取局部和全局特征的能力，从而在港口和地面轨道等背景复杂的类别上取得了显著的性能提升。同时，在桥梁和交通工具等小尺寸目标的检测中也展现出显著优势，这些目标通常在图像中仅占据极少像素且易受遮挡干扰。这一优势表明，本文方法不仅适用于无人机图像小目标，也能够很好地处理遥感图像中的小目标检测任务。

表3　NWPU VHR-10上实验结果对比

Table 3　The comparative experimental results on NWPU VHR-10

/%

模型	主干网络	精确率	召回率	AP@50	mAP
TPHYOLO	CSPDarkNet	91.0	88.9	92.8	58.0
YOLOv7	ELAN	91.8	89.6	92.9	61.3
YOLOv8	CSPDarkNet	90.3	88.5	91.1	63.4
YOLOv9	GELAN	91.4	89.4	92.0	63.6
RTDETR	HGNetv2	91.8	89.5	92.7	63.6
本文	HGNetv2	91.9	91.0	93.0	63.7
本文-FRM	HGNetv2	**93.6**	**92.1**	**93.9**	**64.5**

注：加粗字体表示各列最优结果。

表4　NWPU VHR-10上各类别实验结果对比

Table 4　The comparative experimental results on NWPU VHR-10 for each class

/%

模型	飞机	船舶	储罐	棒球场	网球场	篮球场	地面轨道	港口	桥梁	交通工具	mAP
TPHYOLO	66.3	55.6	**38.7**	75.2	53.8	70.6	84.9	51.4	25.5	57.9	58.0
YOLOv7	71.3	70.1	30.1	75.9	57.7	75.2	86.5	**52.7**	24.9	68.2	61.3
YOLOv8	**73.7**	71.3	32.8	79.6	**64.6**	76.3	87.7	49.0	27.0	71.6	63.4
YOLOv9	73.3	**75.1**	31.6	**80.8**	63.2	75.1	**89.3**	49.6	27.0	70.9	63.6
RTDETR	72.9	72.9	31.5	79.2	62.8	**78.6**	88.2	49.7	27.9	**71.8**	63.6
本文	72.6	74.6	37.1	80.2	63.3	76.1	89.1	50.7	**31.1**	70.0	**64.5**

注：加粗字体表示各列最优结果。

2.4　消融实验

为了系统验证所提方法的有效性，在VisDrone2019数据集上进行详细的消融实验，包括前景细化模块FRM和多维归纳偏置自注意力网络MIBSN。

2.4.1　前景细化模块性能评估与分析

实验对前景细化模块进行全面的性能评估，从

前景区域选择、图像重建质量、细化策略的效果以及对检测器的普适性等多方面验证其在小目标检测中的有效性。

首先，对前景细化模块中提出的CAM-AM方法与当前主流的类激活映射方法进行比较。比较的方法包括Grad-CAM（gradient-weighted class activation map）（Selvaraju 等，2017）、Grad-CAM++（Chattopadhay 等，2018）、CR-CAM（contrast-ranking CAM）（Li 等，2024）。实验结果如图6所示，CAM-AM在前景区域选择上显著优于这些主流方法。这得益于CAM-AM通过自适应权重融合主干网络中不同层次的特征图，能够全面捕捉从低层次的边缘特征到高层次的抽象语义特征，进而有效减少噪声影响，筛选出更加精确的前景区域。

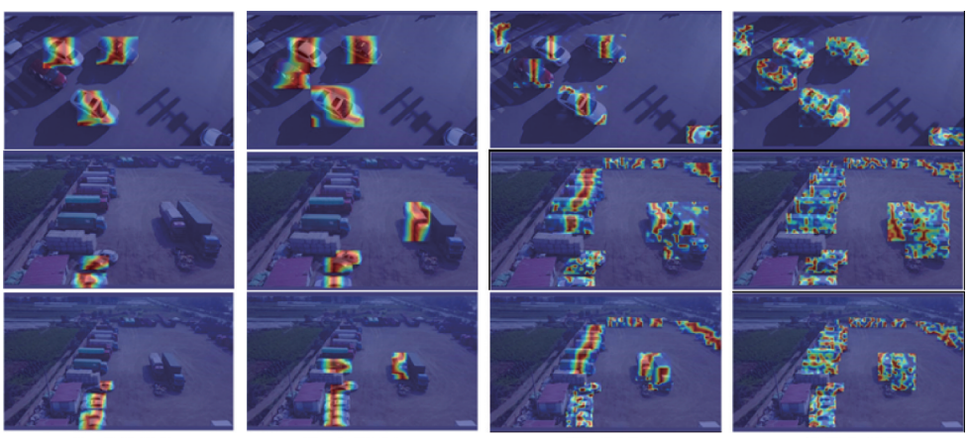


图6　CAM-AM对比实验结果

Fig. 6　The comparative experimental results of CAM-AM

其次，本文通过可视化对比实验评估了前景细化模块的图像重建质量。如图7所示，与传统的双线性插值（bilinear interpolation，BI）方法相比，SwinIR（image restoration using Swin-Transformer）（Liang等，2021）和GDP（Fei等，2023）在纹理上有所改善，但在汽车轮廓和前灯等细节处理上比较粗糙。相比之下，FRM在保持高分辨率重建的同时，显著提升了小目标的边缘锐度与细节保真度。定量评估结果进一步支持了这一结论，如表5所示，FRM在峰值信噪比（peak signal-to-noise ratio，PSNR）和结构相似性分数（structural similarity，SSIM）这两个图像质量评估指标上均优于对比方法，相比于BI方法，FRM的SSIM和PSNR分别提升了0.14和4.19 dB，表明前景掩码为扩散模型提供显式空间约束，避免生成过程中的结构失真。

图7　超分辨可视化结果

Fig. 7　The visualization results of super-resolution

另外，本文开展细化策略的消融实验，探究不同细化比例对小目标检测性能的影响。结果如图8所示，横坐标表示经过前景细化模块处理的图像块比例，纵坐标表示检测性能和相对性能增长率。实验结果表明，仅对20%图像中的关键区域应用条件扩散模型进行细化，检测性能就获得显著提升；当处理60%的关键区域时，其检测性能与使用100%双线性插值细化的基准水平相当。这验证了本文提出的CAM-AM选择性细化策略的高效性，即通过识别和

表 5　前景细化模块 SSIM 和 PSNR 实验结果对比

Table 5　The comparative experimental results of SSIM and PSNR of FRM

模型	SSIM	PSNR/dB
BI	0.767	21.463
SwinIR	0.879	24.252
GDP	0.881	24.651
FRM	**0.907**	**25.653**

注：加粗字体表示各列最优结果。

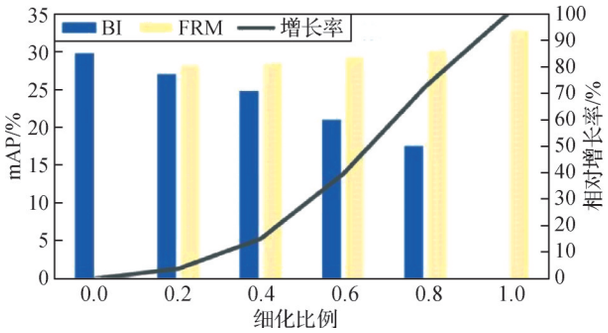


图 8　不同细化比例下的小目标检测性能对比

Fig. 8　Performance comparison of small object detection under different refinement ratios

细化包含小目标的关键区域，只需处理少量图像的关键区域就能实现理想的检测效果。

最后，本文进一步验证了前景细化模块的普适性，尝试将其应用于当前主流目标检测器的图像处理流程中。实验结果如表 6 所示，经过前景细化模块处理后的图像显著提升了各检测器的检测精度。具体而言，YOLOv9-FRM 的整体性能提高 1.0%，而 RTDETR 的性能提升 2.3%。表明前景细化模块的普适性，改善了模型在复杂场景下的表现，能够在不同架构的检测器上稳定发挥性能提升作用。综上所述，实验结果充分证明 FRM 的有效性。该模块通过精确的前景区域选择和高质量的图像重建，为小目标检测提供了一种兼具性能和效率的解决方案。

表 6　前景细化模块处理后的 VisDrone2019 测试集检测结果对比

Table 6　The comparison of detection results on VisDrone2019 test set processed by FRM

模型	主干网络	输入尺寸/像素	精确率/%	召回率/%	AP@50/%	mAP/%
YOLOv8	CSPDarkNet	1 536 × 1 536	**57.6**	45.8	46.1	27.7
YOLOv8-FRM	CSPDarkNet	1 536 × 1 536	56.8	**46.7**	**46.5**	**28.2**
YOLOv9	GELAN	1 536 × 1 536	56.9	47.0	46.7	28.3
YOLOv9-FRM	GELAN	1 536 × 1 536	**61.5**	**50.1**	**47.9**	**29.3**
RTDETR	HGNetv2	1 536 × 1 536	60.4	44.1	43.5	25.0
RTDETR-FRM	HGNetv2	1 536 × 1 536	**59.3**	44.1	**44.7**	**27.3**
本文	HGNetv2	1 536 × 1 536	60.7	50.5	51.0	31.1
本文-FRM	HGNetv2	1 536 × 1 536	**61.6**	**52.7**	**53.2**	**32.8**

注：加粗字体为每组对比的最优值，FRM 代表经前景细化模块处理。

2.4.2　多维归纳偏置自注意力网络性能评估与分析

为了验证多维归纳偏置自注意力网络有效性，分别对网络中的多维自注意力模块、混合增强前馈模块、尺度耦合模块和邻域特征交互模块进行消融实验。

1）多维自注意力模块。在基线模型 RTDETR 的基础上引入 MSA 模块，该模块剔除了 IBPP，以验证多维自适应自注意力的有效性。实验结果如表 7 和表 8 所示。在 VisDrone2019 数据集上的结果显示，AP@50 和 mAP 分别提升至 44.7% 和 27.3%。MSA 在大多数类别上都能有所提高，特别是行人、人和小型车辆等尺寸较小的无人机目标。在卡车和公交车类别上性能有所下降，原因可能在于样本数量有限，且复杂的自注意力操作导致模型出现过拟合现象。总体来看，多维自注意力模块在增强空间位置感知能力方面对无人机小目标检测性能起到了积极作用。

2）归纳偏置感知路径。在 MSA 模块中增加 IBPP，实验结果如表 7 所示。AP@50 和 mAP 分别进一步提高至 47.0% 和 28.5%。IBPP 在实例数量较少且较难区分的类别上有较大改善。其中，在三轮

车和公交车分别提升了3.0%和6.8%，表明通过引入归纳偏置感知路径，模型在编码局部特征时得到了显著增强，同时也表明该模块在稳定小规模数据集上的训练过程方面起到了关键作用。

3）混合增强前馈模块。增加EHFF模块后，mAP从基线的25.0%提升至29.3%，表现出显著提升。EHFF模块在行人、摩托车等目标较小、形状复杂的类别上表现尤为突出，说明可变形卷积有助于更准确地聚焦于这些无人机图像小目标的关键特征。此外，卡车和三轮车等类别的准确率分别达到41.2%和20.7%，也验证了EHFF对不同形状目标的适应能力。总体来看，EHFF通过可变形卷积对不同角度拍摄目标的灵活感知，对不规则形状的小目标检测效果显著。

4）尺度耦合模块。增加SCM模块后，AP@50和mAP分别提高至48.5%和29.4%。SCM模块的引入使得模型在几乎所有的类别上都有明显的性能提升，特别是在小型车辆类别上，性能提升了5.3%。尺度耦合模块的加入有效增强了多尺度局部特征，与多维自注意力模块的交互进一步改善了全局与局部特征的平衡，使得模型能够更有效地处理复杂背景下的无人机图像小目标检测任务。

5）邻域特征交互模块。在加入LNI后，AP@50大幅提升至51.0%，mAP提升至31.1%。与基准相比，LNI的引入在中型车辆、三轮车和公交车等类别上有显著的提升。实验结果表明，LNI模块在解码小目标的细微特征以及抑制背景噪声方面具有显著优势。

2.4.3　前景细化模块和多维归纳偏置自注意力网络整体性能评估与分析

表7　多维归纳偏置自注意力网络在VisDrone2019测试集消融实验

Table 7　The ablation study of MIBSN on VisDrone2019 test set

/%

模型	精确率	召回率	AP@50	mAP
基线模型	60.4	44.1	43.5	25.0
+ MSA	59.3	44.1	44.7	27.3
+ IBPP	57.4	47.4	47.0	28.5
+ EHFF	**61.4**	47.9	47.5	29.3
+ SCM	62.0	49.7	48.5	29.4
+ LNI	60.7	**50.5**	**51.0**	**31.1**

注：加粗字体表示各列最优结果。

为了系统评估所提方法的有效性，在VisDrone2019数据集上进行了完整的消融实验。实验从基线模型出发，依次引入前景细化模块和多维归纳偏置自注意力网络，实验结果如表9所示。

多维归纳偏置自注意力网络对行人、人和自行车等无人机图像小型目标的检测性能提升显著，行人的AP值从基线模型的13.8%提升至22.5%。这主要是由于多维归纳偏置自注意力网络通过邻域特征的有效聚合，增强了模型对小目标细节特征的捕捉能力。前景细化模块的引入进一步优化了检测效果，行人的AP值从22.5%提升至26.0%，自行车的AP值从13.4%提升至15.1%，表明前景细化模块通过细化前景区域，复原小目标的边缘、纹理等其他关键特征，进一步提高小目标的检测准确率。在公交车和卡车等较大目标的检测中，多维归纳偏置自注

表8　多维归纳偏置自注意力网络在VisDrone2019测试集各类别消融实验

Table 8　The ablation study of MIBSN on VisDrone2019 test set for each class

/%

模型	行人	人	自行车	小型车辆	中型车辆	卡车	三轮车	遮阳三轮车	公交车	摩托车	mAP
基线模型	13.8	8.93	9.69	49.3	34.8	38.8	16.8	15.4	45.0	17.1	25.0
+ MSA	**23.8**	**16.5**	11.3	**58.7**	34.3	30.3	18.6	11.9	43.9	23.2	27.3
+ IBPP	19.5	11.3	10.6	55.4	36.4	40.7	19.8	17.9	51.8	21.8	28.5
+ EHFF	23.7	13.6	11.8	57.4	36.4	41.2	20.7	16.1	46.7	24.9	29.3
+ SCM	23.7	15.7	12.9	54.6	34.0	39.6	22.5	18.0	48.5	**25.0**	29.4
+ LNI	22.5	13.3	**13.4**	57.2	**38.2**	**44.4**	**23.0**	**19.9**	**54.3**	24.7	**31.1**

注：加粗字体表示各列最优结果。

表 9　在 VisDrone2019 测试集上消融实验

Table 9　The ablation study on VisDrone2019 testset

/%

模型	行人	人	自行车	小型车辆	中型车辆	卡车	三轮车	遮阳三轮车	公交车	摩托车	mAP
基线模型	13.8	8.93	9.69	49.3	34.8	38.8	16.8	15.4	45.0	17.1	25.0
+ MIBSN	22.5	13.3	13.4	57.2	38.2	**44.4**	23.0	19.9	54.3	24.7	31.1
+ FRM	**26.0**	**15.9**	**15.1**	**60.0**	**40.5**	43.9	**23.1**	**21.0**	**55.1**	**27.3**	**32.8**

注：加粗字体表示各列最优结果。

意力网络和前景细化模块同样展现了显著效果。加入多维归纳偏置自注意力网络，公交车的 AP 值从基线模型的 45.0% 逐步提升至 54.3%，加入前景细化模块后，AP 值上升至 55.1%。消融实验结果表明，多维归纳偏置自注意力网络通过多维自注意力和邻域交互，增强了模型对无人机图像小目标局部和全局特征的提取能力；而前景细化模块仅使用少量计算资源进行图像细化，重建前景区域的细节，从而进一步提升无人机图像小目标检测精度。

3　结　论

本文提出一种融合前景细化和多维归纳偏置自注意力的无人机图像小目标检测模型，旨在提升小目标检测的精度和鲁棒性。前景细化模块通过筛选前景区域，并采用条件扩散模型对该区域进行细化处理，仅利用少量计算资源有效恢复了小目标的关键细节信息。此外，本文提出多维归纳偏置自注意力网络，摒弃了现有卷积和 Transformer 顺序堆叠架构，使编码器能够并行学习全局和局部特征的依赖关系，从而提升了编码器的特征表示能力；在网络的解码过程中，邻域特征交互模块通过逐层聚合编码器所编码的局部和全局特征，并通过多尺度预测层进行特征交互，确保最终的预测特征图保留足够的小目标信息。在 VisDrone2019、UAVDT 和 NWPU VHR-10 上的实验表明，提出的方法可以显著提高小目标的检测精度。

但本文算法也存在局限性，仍有改进空间。本文方法主要着眼于静态图像的小目标检测，未能充分利用视频序列中的时序相关性信息，且在前景细化方面过于依赖激活图的质量。未来研究将聚焦于时空特征的有效融合以增强时序建模能力，并通过整合视频序列的时序信息增强激活图的生成质量，从而进一步提升模型的性能。

参考文献（References）

Bai Y C, Zhang Y Q, Ding M L and Ghanem B. 2018. SOD-MTGAN: small object detection via multi-task generative adversarial network//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 210-226 [DOI: 10.1007/978-3-030-01261-8_13]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chattopadhay A, Sarkar A, Howlader P and Balasubramanian V N. 2018. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, USA: IEEE: 839-847 [DOI: 10.1109/WACV.2018.00097]

Cheng G, Zhou P C and Han J W. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 54(12): 7405-7415 [DOI: 10.1109/TGRS.2016.2601622]

Croitoru F A, Hondru V, Ionescu R T and Shah M. 2023. Diffusion models in vision: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850-10869 [DOI: 10.1109/TPAMI.2023.3261988]

Dong X Y, Bao J M, Chen D D, Zhang W M, Yu N H, Yuan L, Chen D and Guo B N. 2022. CSWin transformer: a general vision transformer backbone with cross-shaped windows//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12114-12124 [DOI: 10.1109/CVPR52688.2022.01181]

Du B W, Huang Y C, Chen J X and Huang D. 2023. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images//Proceedings of 2023 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 13435-13444 [DOI: 10.1109/CVPR52729.2023.01291]

Du D W, Qi Y K, Yu H Y, Yang Y F, Duan K W, Li G R, Zhang W G, Huang Q M and Tian Q. 2018. The unmanned aerial vehicle benchmark: object detection and tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 375-391 [DOI: 10.1007/978-3-030-01249-6_23]

Du D W, Zhu P F, Wen L Y, Bian X, Lin H B, Hu Q H, Peng T, Zheng J Y, Wang X Y, Zhang Y, Bo L F, Shi H L, Zhu R, Kumar A, Li A J, Zinollayev A, Askergaliyev A, Schumann A, Mao B J, Lee B, Liu C, Chen C R, Pan C H, Huo C L, Yu D, Cong D C, Zeng D N, Pailla D R, Li D, Wang D, Cho D, Zhang D Y, Bai F R, Jose G, Gao G Y, Liu G Z, Xiong H T, Qi H, Wang H R, Qiu H Q, Li H L, Lu H C, Kim I, Kim J, Shen J, Lee J, Ge J, Xu J J, Zhou J K, Meier J, Choi J W, Hu J H, Zhang J Y, Huang J Y, Huang K Q, Wang K Y, Sommer L, Jin L, Zhang L, Huang L H, Sun L, Steinmann L, Jia M X, Xu N, Zhang P Y, Chen Q, Lv Q X, Liu Q, Cheng Q S, Chennamsetty S S, Chen S H, Wei S, Kruthiventi S S S, Hong S, Kang S, Wu T, Feng T, Kollerathu V A, Li W Q, Dai W, Qin W D, Wang W Y, Wang X R, Chen X Y, Chen X, Sun X, Zhang X, Zhao X, Zhang X D, Zhang X Y, Chen X K, Wei X D, Zhang X Z, Li Y C, Chen Y F, Toh Y H, Zhang Y, Zhu Y, Zhong Y X, Wang Z X, Wang Z K, Song Z C and Liu Z M. 2019. VisDrone-DET2019: the vision meets drone object detection in image challenge results//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 213-226 [DOI: 10.1109/ICCVW.2019.00030]

Fei B, Lyu Z Y, Pan L, Zhang J Z, Yang W D, Luo T Y, Zhang B and Dai B. 2023. Generative diffusion prior for unified image restoration and enhancement//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9935-9946 [DOI: 10.1109/CVPR52729.2023.00958]

Glenn J, Ayush C and Jing Q. 2023. Ultralytics YOLOv8 [EB/OL]. [2025-01-09]. https://github.com/ultralytics/ultralytics

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851

Huang Y X, Liu H I, Shuai H H and Cheng W H. 2025. DQ-DETR: DETR with dynamic query for tiny object detection//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 290-305 [DOI: 10.1007/978-3-031-73116-7_17]

Lei S, Shi Z W and Zou Z X. 2020. Coupled adversarial training for remote sensing image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 58(5): 3633-3643 [DOI: 10.1109/TGRS.2019.2959020]

Li Y S, Liang H J, Zheng H F and Yu R. 2024. CR-CAM: generating explanations for deep neural networks by contrasting and ranking features. Pattern Recognition, 149: #110251 [DOI: 10.1016/j.patcog.2024.110251]

Liang J Y, Cao J Z, Sun G L, Zhang K, Van Gool L and Timofte R. 2021. SwinIR: image restoration using swin transformer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 1833-1844 [DOI: 10.1109/ICCVW54120.2021.00210]

Liang S Y, Wu H, Zhen L, Hua Q Z, Garg S, Kaddoum G, Hassan M M and Yu K P. 2022. Edge YOLO: real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 23(12): 25345-25360 [DOI: 10.1109/TITS.2022.3158253]

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2117-2125 [DOI: 10.1109/CVPR.2017.106]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu H I, Tseng Y W, Chang K C, Wang P J, Shuai H H and Cheng W H. 2024. A DeNoising FPN with transformer R-CNN for tiny object detection. IEEE Transactions on Geoscience and Remote Sensing, 62: #4704415 [DOI: 10.1109/TGRS.2024.3396489]

Saharia C, Ho J, Chan W, Salimans T, Fleet D J and Norouzi M. 2023. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4713-4726 [DOI: 10.1109/TPAMI.2022.3204461]

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]

Tian H D, Zhang J P and Wang Y H. 2023. Sparse constraint and spatial-temporal regularized correlation filter for UAV tracking. Journal of Image and Graphics, 28(2): 458-470 (田昊东, 张津浦, 王岳环. 2023. 稀疏约束的时空正则相关滤波无人机视觉跟踪. 中国图象图形学报, 28(2): 458-470) [DOI: 10.11834/jig.210611]

Wang C Y, Yeh I H and Liao H Y M. 2025. YOLOv9: learning what you want to learn using programmable gradient information//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 1-21 [DOI: 10.1007/978-3-031-72751-1_1]

Wang S K, Wang X D, Qu L, Yao F Q, Liu Y Y, Li C H, Wang Y T and Zhong G Q. 2024. Semantic segmentation benchmark dataset for coastal ecosystem monitoring based on unmanned aerial vehicle

(UAV). Journal of Image and Graphics, 29(8): 2162-2174 (王胜科, 王贤栋, 曲亮, 姚凤芹, 刘莹莹, 李聪慧, 王玉瑱, 仲国强. 2024. 面向无人机海岸带生态系统监测的语义分割基准数据集. 中国图象图形学报, 29(8): 2162-2174) [DOI: 10.11834/jig.230670]

Wang W H, Xie E Z, Li X, Fan D P, Song K T, Liang D, Lu T, Luo P and Shao L. 2021. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 548-558 [DOI: 10.1109/ICCV48922.2021.00061]

Xiao Y, Yuan Q Q, Jiang K, He J, Jin X Y and Zhang L P. 2024. EDiffSR: an efficient diffusion probabilistic model for remote sensing image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 62: #5601514 [DOI: 10.1109/TGRS.2023.3341437]

Yang J Z, Fu K, Wu Y M, Diao W H, Dai W and Sun X. 2022. Mutual-feed learning for super-resolution and object detection in degraded aerial imagery. IEEE Transactions on Geoscience and Remote Sensing, 60: #5628016 [DOI: 10.1109/TGRS.2022.3198083]

Yu J R, Yao S Y, Chen X Q, Lin Y L and Wang F Y. 2024. Cross-domain collaborative technology among vehicles, infrastructure, and drones for connected and autonomous driving deployment. Journal of Image and Graphics, 29(11): 3293-3304 (于静茹, 姚升悦, 陈喜群, 林懿伦, 王飞跃. 2024. 面向网联自动驾驶部署的车—路—无人机跨域协同技术. 中国图象图形学报, 29(11): 3293-3304) [DOI: 10.11834/jig.230786]

Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, Ni L M and Shum H Y. 2022. DINO: DETR with improved denoising anchor boxes for end-to-end object detection [EB/OL]. [2025-01-09]. https://arxiv.org/pdf/2203.03605.pdf

Zhang Z X, Lu X Q, Cao G J, Yang Y T, Jiao L C and Liu F. 2021. ViT-YOLO: transformer-based YOLO for object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 2799-2808 [DOI: 10.1109/ICCVW54120.2021.00314]

Zhao Y A, Lv W Y, Xu S L, Wei J M, Wang G Z, Dang Q Q, Liu Y and Chen J. 2024. DETRs beat YOLOs on real-time object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 16965-16974 [DOI: 10.1109/CVPR52733.2024.01605]

Zhu X K, Lyu S, Wang X and Zhao Q. 2021a. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 2778-2788 [DOI: 10.1109/ICCVW54120.2021.00312]

Zhu X Z, Su W J, Lu L W, Li B, Wang X G and Dai J F. 2021b. Deformable DETR: deformable transformers for end-to-end object detection [EB/OL]. [2025-01-09]. https://arxiv.org/pdf/2010.04159.pdf

作者简介

冯琪涵，女，博士研究生，主要研究方向为计算机视觉、小目标检测和航空图像处理。E-mail: fengqh@cumt.edu.cn

王志晓，通信作者，男，教授，主要研究方向为数据挖掘、社交网络分析和影响力最大化。E-mail: zhxwang@cumt.edu.cn

孙成成，男，博士后，主要研究方向为深度学习、图数据挖掘和图表示学习。E-mail: scc@cumt.edu.cn

邵志文，男，副教授，主要研究方向为计算机视觉、人脸表情识别、人脸表情合成和场景文本检测及识别。

E-mail: zhiwen_shao@cumt.edu.cn