中图法分类号:V279; TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)04-1125-17

论文引用格式:Liu X, Song P B, Bao F X and Du H W. 2026. Small-object detection model integrating wavelet convolution and frequency-domain attention. Journal of Image and Graphics, 31(4):1125-1141(刘旭, 宋佩博, 包芳勋, 杜宏伟. 2026. 融合小波卷积与频域注意力的小目标检测. 中国图象图形学报, 31(4):1125-1141)[DOI:10.11834/jig.250293]

融合小波卷积与频域注意力的小目标检测

刘旭[1],宋佩博[1],包芳勋[1*],杜宏伟[2]

1. 山东大学数学学院,济南　250100; 2. 浪潮通用软件有限公司,济南　250101

摘　要:**目的**　无人机拍摄图像存在小目标数量多、易受恶劣天气等噪声污染的特点,针对无人机拍摄图像的小目标检测技术在军用领域和商用领域都发挥着重要作用。然而,现有的目标检测方法在定位小目标方面仍然存在检测精度低的问题。针对这些问题,提出基于YOLOv8(you only look once)的融合小波卷积与频域注意力的改进模型(an enhanced YOLO model integrating wavelet convolution and frequency-domain attention, YOLO-WF)。**方法**　首先,在骨干网络中构建了基于傅里叶频域增强的自注意力机制与门控机制模块(Fourier-based self-attention convolution module,CFSA)增强图像的特征,提升模型对关键信息的提取能力;其次,在特征提取模块设计了基于二级分解低频增强小波变换卷积(low-frequency enhanced wavelet transform convolution,LOWTC)模块,利用小波变换的多尺度特性扩展感受野,有效缓解传统卷积长距离依赖性不足的问题;最后,在提取浅层特征后增加针对小目标的检测头,提升模型对小目标的检测能力。**结果**　在VisDrone2019-DET(vision-based drone detection and tracking 2019—detection)、UAVDT(unmanned aerial vehicle benchmark object detection and tracking)和CARPK(car parking lot dataset)数据集上实验,结果表明提出的YOLO-WF模型比基线模型的APs(average precision of small objects)指标分别提高5.5%、3.08%和6.8%,达到19.9%、38.54%和33.3%。AP50(AP at IoU threshold 0.50)和APm(AP of medium objects)指标也均有提升,以VisDrone2019-DET为例,AP50和APm分别达到47.1%和40.3%,相比基线模型分别提高3.5%和3.0%,且参数量下降0.4%。**结论**　YOLO-WF通过频域—小波融合策略,显著提升了中小目标的检测精度,且未引入额外存储负担,可直接迁移至其他航拍检测任务。

关键词:深度学习;小目标检测;基于傅里叶频域增强的自注意力机制与门控机制模块(CFSA);低频增强小波变换卷积(LOWTC);YOLOv8

Small-object detection model integrating wavelet convolution and frequency-domain attention

Liu Xu[1], Song Peibo[1], Bao Fangxun[1*], Du Hongwei[2]

1. School of Mathematics, Shandong University, Jinan　250100, China; 2. Inspur General Software Co., Ltd., Jinan　250101, China

Abstract: **Objective**　Unmanned aerial vehicle (UAV) imagery is characterized by an extremely high density of small objects (often < 16 × 16 pixels) and frequent corruption by weather-related noise, such as haze, rain, and motion blur.

收稿日期:2025-07-14;**修回日期**:2025-11-10;**预印本日期**:2025-11-17

*通信作者:包芳勋　fxbao@sdu.edu.cn

基金项目:国家自然科学基金项目(62372269);山东省自然科学基金项目(ZR2022MF245,ZR2024QF158);山东省重点研发计划资助(2024TSGC1132)

Supported by: National Natural Science Foundation of China (62372269); Natural Science Foundation of Shandong Province, China (ZR2022MF245, ZR2024QF158); Key R&D Program of Shandong Province, China (2024TSGC1132)

These factors cause severe signal attenuation and geometric ambiguity, making small-object detection a persistent bottleneck for convolutional neural networks and Transformer detectors. Although two- and one-stage frameworks have achieved remarkable progress on natural-scene datasets, their accuracy decreases considerably on UAV benchmark datasets, including VisDrone2019-DET, UAV Benchmark Object Detection and Tracking (UAVDT), and Car Parking Lot (CARPK). This study aims to enhance small-object precision without increasing the model size or sacrificing real-time capability. **Method** To address the challenges above, this study proposes YOLO-WF, an enhanced YOLOv8 architecture that synergistically fuses wavelet convolution with frequency-domain attention. First, YOLO-WF embeds a composite Fourier self-attention module (CFSA) in the backbone. By applying 2D fast Fourier transform (FFT) to feature maps, CFSA estimates cross-channel covariance on the amplitude spectrum, generates a learnable frequency mask that amplifies harmonics beneficial for small-object detection, and suppresses high-frequency noise caused by haze, rainfall, and motion blur. The modulated spectrum is then transformed back to the spatial domain via inverse FFT, enabling the model to obtain a global receptive field and long-range dependencies without increasing convolution kernel sizes. Second, a low-frequency wavelet transform convolution module (LOWTC) is inserted at the feature extraction stage. Utilizing second-order Daubechies-4 wavelets, LOWTC decomposes the input features into approximation (A) and detail (D) subbands. The A subband, containing object shape and contextual information, is processed by dilated depth-wise convolutions to capture long-range dependencies. The D subband, which preserves edges and textures, is unchanged. This divide-and-conquer strategy not only enlarges the receptive field but also alleviates the long-distance modeling deficiency of standard convolutions while keeping parameter growth and computational overhead low, resulting in rich, robust feature representations for small and medium objects. Third, after shallow feature extraction, a dedicated P2 detection head for objects smaller than 32 pixels is added, together with a small-object label assignment strategy, to prevent large-instance interference and substantially improve the localization accuracy of tiny targets. Extensive experiments on VisDrone2019, UAVDT, and CARPK benchmarks demonstrated that YOLO-WF achieves substantial accuracy gains over the baseline model. **Result** All experiments were conducted on single NVIDIA GeForce RTX 4090 D (24 210 MiB) running Ubuntu 16.04. The software stack comprised Python 3.9 and PyTorch 2.6.0+cu124. Input images were resized to 640 × 640 pixels. The mini-batch size was set to 4 because of limited GPU memory. In accordance with the standard practice for fair comparison, no ImageNet pretrained weights were loaded; all models were trained from scratch for 100 epochs. On the VisDrone2019 dataset, the average precision (AP) metrics of AP50, APs (small objects), and APm (medium objects) reach 47.1%, 19.9%, and 40.3%, respectively. On the UAVDT dataset, the AP50, APs, and APm metrics are 81.56%, 38.54%, and 65.12%, respectively. Furthermore, on the CARPK dataset, the AP50, APs, and APm metrics are 94.3%, 33.3%, and 67.9%, respectively. These results indicate that the YOLO-WF model not only achieves high detection accuracy for small objects but also maintains good performance for objects of different sizes. In addition, comparisons with several typical detection methods showed that the YOLO-WF model has a substantial increase in average detection accuracy. For example, compared with the traditional YOLOv8 model, the YOLO-WF model achieves an average improvement of 6.5% in detection accuracy on the VisDrone2019 dataset and 2.44% on the UAVDT dataset. These improvements highlight the effectiveness of the proposed YOLO-WF model in enhancing the detection performance for small objects in UAV-captured images. Sequential module removal on VisDrone2019 shows that the detection head of small objects contributes the most (+3.6% APs), followed by CFSA (+2.7%) and LOWTC (+2.1%). Combining CFSA with LOWTC yields an extra +1.2%, indicating that frequency-domain enhancement and wavelet context are complementary. Replacing standard convolutions with GSConv recovers 0.3% lost by the additional detection head, validating the slimming strategy. **Conclusion** The proposed YOLO-WF model demonstrates notable improvements in the detection of small objects in UAV-captured images. By incorporating the CFSA module, LOWTC module, and a specialized detection head, the model effectively enhances the extraction of key features and improves the detection capability for small objects. Experimental results on multiple benchmark datasets validate the superior performance of the YOLO-WF model in terms of detection accuracy and recall. This study provides a promising solution for small-object detection in UAV-captured images, and the proposed model has the potential to be applied in various practical scenarios, such as surveillance, traffic monitoring, and environmental monitoring. Future work may focus on further optimizing the model architecture and exploring the integration of additional techniques to further

enhance the detection performance and adaptability of the model to different types of UAV-captured images and environmental conditions.

Key words: deep learning; small object detection; Fourier-based self-attention convolution module (CFSA); low-frequency enhanced wavelet transform convolution (LOWTC); YOLOv8

0 引　言

目标检测是计算机视觉的核心任务之一，在人们的日常生活，例如安防监控、农业发展、灾害规划，乃至军事领域，如防空预警、战场侦察等方面，都发挥着越来越重要的作用。无人机拍摄的图像是目标检测领域的一类重要研究对象（Wang等，2023b；Ma等，2022）。由于无人机拍摄视角高，其拍摄的图像中小目标占比很高；另外，拍摄图像背景复杂、雨雪尘埃等噪声因素繁杂的问题，导致对小目标的检测任务比较困难。因此，在考虑硬件平台资源有限的情况下提升小目标检测性能是无人机航拍场景中目标检测的核心问题之一。

传统的目标检测算法主要依赖人工设计提取图像的低层特征，对图像特征的提取能力有限，存在速度慢、精度低以及模型泛化能力差等问题，如Viola-Jones Detectors（Wang，2014）使用最直接的滑动窗口方式遍历图像中所有可能的位置和比例，以查看是否有任何窗口包含人脸。近年来，基于深度学习的目标检测算法已成为主流检测方法，其大致可以分为双阶段检测方法和单阶段检测方法两个范畴。R-CNN（region-based convolutional neural network）（Bharati和Pramanik，2020）是目标检测领域最早的双阶段检测方法之一，其核心思想是首先采用选择性搜索算法在图像上生成一组候选区域，这些候选区域被视为可能包含目标的区域；然后，每个候选区域对应的图像块会重新缩放为固定大小。使用在ImageNet数据集上预训练的卷积神经网络（convolutional neural network，CNN）模型（如AlexNet）提取特征；最后，针对每个候选区域，使用线性支持向量机（support vector machine，SVM）（Chandra和Bedi，2021）分类器进行目标类别的预测，判断该区域是否包含目标。由于R-CNN的检测速度过于缓慢，一些改进的检测模型如SPPNet（spatial pyramid pooling in deep convolutional networks for visual recognition）（He等，2015）、Faster R-CNN（Ren等，2017）等相继提出。双阶段方法虽然具有较高的准确率，但检测速度往往较慢。以YOLO（you only look once）系列模型为典型代表的单阶段（Redmon等，2016；Redmon和Farhadi，2017，2018）目标检测算法是将检测任务视为一种回归问题，将整幅图像划分为多个区域，同时预测每个区域的边界框和目标类别。后续的YOLO系列算法如YOLOv6（Li等，2022）、YOLOv7（Wang等，2023a）、YOLOv10（Wang等，2024a）虽然在检测速度上较双阶段检测器有了较大的提高，但其定位精度却有所下降，尤其是对一些小目标的定位精度。另外，鉴于Transformer（Vaswani等，2017）模型在自然语言处理方面表现出优异的效果，DETR（end-to-end object detection with Transformers）（Carion等，2020）模型将Transformer从自然语言处理领域引入到计算机视觉中，又为目标检测提供了一种新的研究思路。总之，无人机拍摄的自然场景下的图像中小尺寸目标占据目标的大部分，而能够体现小目标的位置和细节信息本来就非常稀少，现有基于卷积神经网络（CNN）的主流检测方法（潘晓英 等，2023；石争浩 等，2023），在池化过程中往往会将这些信息过滤掉，致使对小目标的检测效果并不尽如人意。

为解决上述问题，在YOLO模型现有的基础上，本文提出一种基于频域与浅层特征增强的小目标检测算法，称之为YOLO-WF（an enhanced YOLO model integrating wavelet convolution and frequency-domain attention）模型。首先，为了提升模型对关键信息的提取能力，构建了基于傅里叶频域增强的自注意力机制与门控机制模块（Fourier-based self-attention convolution module，CFSA），并嵌入主干网络，通过在不同的通道上计算注意力来高效聚合图像的全局和局部特征。然后，为了更精细地提取小目标的特征，设计了低频增强的小波变换卷积（low-frequency enhanced wavelet transform convolution，LOWTC）模块，它直接对输入图像进行小波分解以利用小波变换的多尺度特性扩展感受野。在频域中进行小核卷积能够在不显著增加参数的情况下获得接近全局的感受野，以较少的参数对小尺度特征进

行捕捉。最后，为了使模型更准确地识别小目标，在浅层特征提取后增加针对小目标的检测头，其特征融合部分结合LOWTC和CFSA的输出特征，可以有效提升对小目标的检测能力。

1 相关工作

1.1 YOLO系列检测模型

YOLO系列算法是通过将目标检测问题转化为回归问题，直接从全图预测边界框和类别概率，具有实时性和高效性。它们由骨干网络（backbone）、颈部（neck）和检测头（head）3部分组成，各部分可以专注于完成自己的特定任务。其中，骨干网络负责提取输入图像的特征，颈部网络负责将特征进一步融合和增强，而头部网络负责将颈部网络融合增强的特征进行解码。虽然前沿的YOLO系列算法在检测性能上优于其他单阶段目标检测算法，但对小目标的检测仍然存在检测精度不高的问题。此外，基于YOLO系列的一些改进模型也相继诞生，如YOLOv3（Xu和Wu，2020）通过引入Dense Net（Huang等，2017b）改进现有的特征提取网络，该方法在多尺度遥感目标检测中表现出良好的性能。YOLO-HR（improved YOLOv5 for object detection in high-resolution optical remote sensing images）（Wan等，2023）是一种用于高分辨率光学遥感目标识别算法，该算法采用多个检测头进行目标检测，并重复使用特征金字塔的输出特征，从而进一步提高检测性能。SODYOLOv8（enhancing YOLOv8 for small object detection in aerial imagery and traffic scenes）（Khalili和Smyth，2024）通过增强多路径融合、引入第四检测层、加入高效多尺度注意力模块（efficient multi-scale attention，EMA）以及采用新的PIoU（powerful intersection over union）损失函数提高小目标的检测精度。虽然基于深度学习的单阶段目标检测算法在无人机图像的目标检测方面取得了显著的成功，但有效检测多尺度和小目标仍然是一项重大挑战。

1.2 视觉Transformer

Transformer模型最初源自自然语言处理领域中的序列建模技术，它利用自注意力机制直接解决了传统CNN在全局信息捕捉方面的不足。该模型首先将输入的单词进行向量化，通过位置编码嵌入位置信息；然后，编码器将其作为输入，通过自注意力机制和前馈神经网络发送到下一个编码器，输出后，再经过全连接和softmax得到最终的输出。近年来，Transformer模型已成功扩展到视觉任务中，基础的视觉Transformer（vision Transformers，ViT）（Dosovitskiy等，2020）及使用移位窗口的分层视觉Transformer（hierarchical vision Transformer using shifted windows，swin Transformer）（Liu等，2021）将图像分解为一系列图像块（局部窗口），并学习它们之间的相互关系。其核心优势在于能够捕捉图像块之间的全局依赖关系，同时具备对输入内容的高度适应性。近几年的一些研究还结合了Transformer与CNN的优点（高丹丹 等，2023），通过动态整合特征信息增强长距离依赖建模能力。此外，门控机制的引入也能够选择性地保留重要信息并遗忘无关信息，从而更好地建模长距离依赖关系，例如长短期记忆网络（long short-term memory，LSTM）（Zhao等，2020）和门控循环单元（gated recurrent unit，GRU）（Wang和Li，2022）通过门控单元解决了传统循环神经网络（recurrent neural network，RNN）（Elman，1990）在长序列处理中的梯度消失问题。

Transformer的自注意力机制在处理高分辨率图像时面临挑战，其计算复杂度随图像块数量的增加呈二次方增长。随着神经网络的逐步加深，深层特征所提取到的语义信息已经不能涵盖小目标的完整信息，而与之相对的浅层特征却能很好地表征与小目标相关的信息。Jocher等人（2023）在YOLOv8原有的Bottleneck中加入改进的Transformer模块，将Transformer与傅里叶变换相结合，然后利用门控机制进一步增强特征，不仅能高效处理Transformer中计算矩阵乘法的问题，还能有效提升小目标检测能力。

1.3 小波卷积

自20世纪80年代以来，小波变换得到广泛的应用。作为一种强大的信号处理和分析工具，已广泛用于处理各种任务的神经网络结构中。Waveletsrnet（Huang等，2017a）和DWSR（deep wavelet prediction for image super-resolution）（Guo等，2017）模型对输入图像的小波高频系数进行预测，重构出更高分辨率的输出。CWNN + MRF（convolutional-wavelet neural networks and Markov random field）方法（Duan等，2017）和小波池化（Williams和Li，2018）使用小波变换作为CNN中的池化算子。Swagan（Gal等，2021）方法、基于Wavelet分数的生成模型（Guth等，2022）和小波扩散模型（Phung等，2023）在生成模型

中使用小波变换来增强生成图像的视觉质量，并提高计算性能。大感受野的小波卷积层 WTConv（wavelet convolutions）（Finder 等，2024）通过结合小波变换扩展感受野，使卷积神经网络能够在不显著增加参数的情况下获得接近全局的感受野。WTConv 层利用小波分解将输入分成不同频带，允许卷积层在低频和高频分量上分别进行处理，增强了模型对低频成分（即形状特征）的响应。与传统方法中卷积核尺寸增大导致参数和计算量指数级增长不同，WTConv 实现了参数的对数增长，使得在大感受野的情况下保持参数效率。

上述工作表明，对输入特征的低频分量与高频分量分别进行卷积操作，能够提取更加丰富、具有判别性的特征表示。受此启发，本文首先对输入的图像进行二级小波分解，并对各级频域特征分别进行卷积，然后利用小波逆变换重建图像，再与基础卷积层的输出进行残差连接。这样通过小波变换可以在不显著增加参数的情况下扩大感受野，使网络能够捕捉到更多上下文信息，并且能够提取图像的多尺度特征，有助于网络更好地理解图像内容。通过将小波域卷积的输出与基础卷积层的输出进行残差连接，可以在保留原始特征的同时引入小波变换带来的优势。总之，这可以在保持参数数量相对较少的情况下，获得较大的感受野。

2　研究方法

2.1　YOLO-WF 模型

本节将从特征提取、注意力机制以及多尺度特征融合等角度对模型进行优化。图 1 给出了 YOLO-

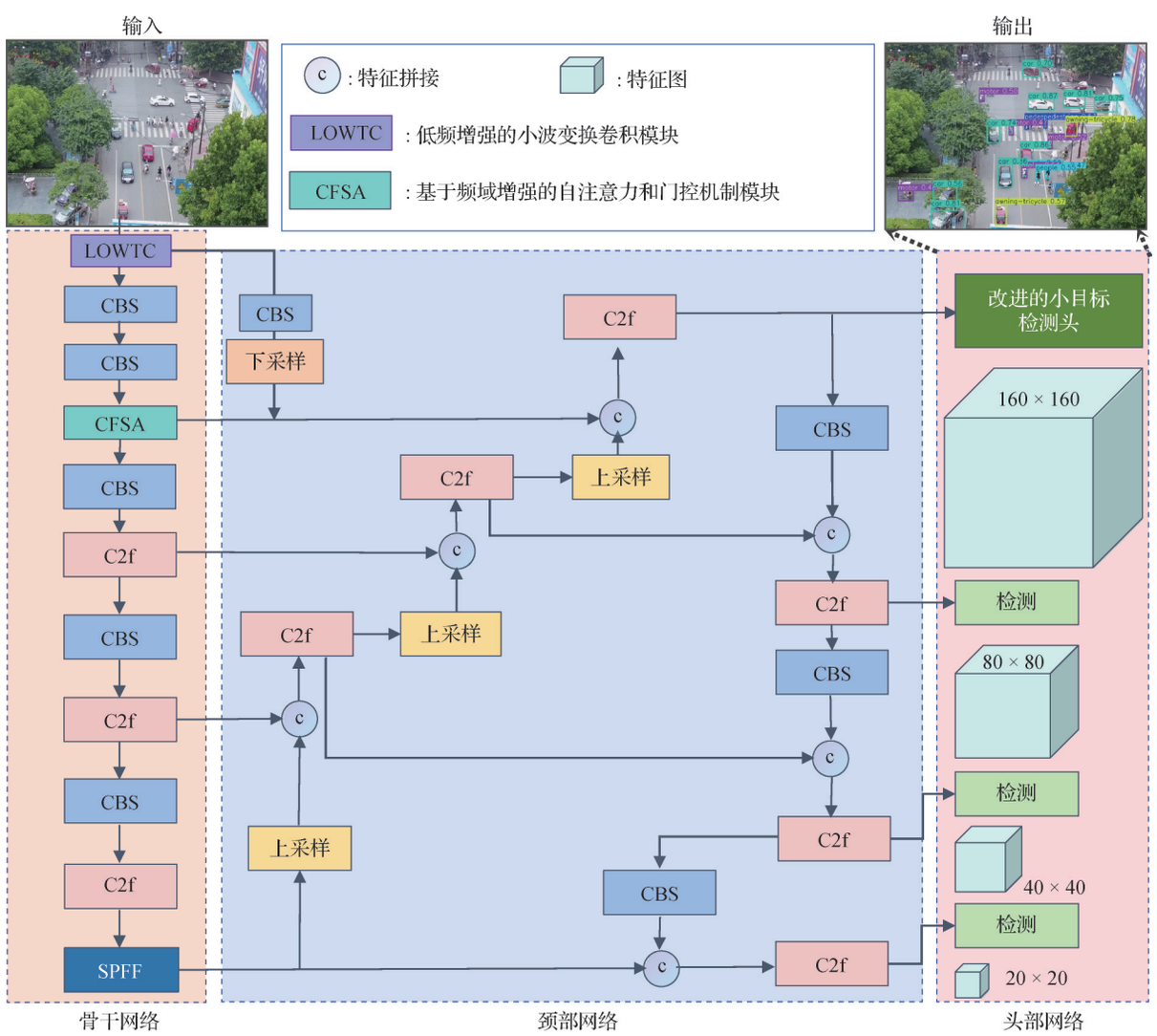


图 1　YOLO-WF 模型结构图

Fig. 1　The structural diagram of the YOLO-WF module

WF模型的整体框架，首先LOWTC模块将输入图像进行二级小波分解，对频域分别进行卷积以提取更加丰富细致的特征信息；其次将C2f（cross stage partial bottleneck with 2 convolutions, faster implementation）替换成基于频域增强的自注意力和门控机制模块CFSA，进一步提取特征；然后将LOWTC和CFSA提取的特征以及深层语义信息进行特征拼接。用一个新加入的小目标检测头对融合后的特征进行针对小目标的预测。

图1中CBS（conv + batch normalization + SiLU）模块由3个主要部分组成：二维卷积层Conv2d、二维批量归一化层BatchNorm2d和一个激活函数层SiLU（sigmoid linear unit）。CBS先通过Conv2d层提取特征，再使用BatchNorm2d层进行归一化以加速训练并提高稳定性，最后经激活函数层SiLU引入非线性，进一步增强模型的表达能力和稳定性。C2f通过一个CBS模块进行特征提取后，通过一个Split操作将特征图按通道数分成浅层特征和深层特征两部分，其中深层特征先经由Bottleneck模块进行特征增强，再进一步提取更具判别性的深层语义特征。将所有处理后的特征图与浅层特征进行特征拼接后，通过CBS模块将输出特征变为与输入特征相同的维度。C2f中的Bottleneck模块通过CBS模块进行特征提取和处理，然后将处理后的特征与输入特征进行残差连接，有助于缓解深度网络中的梯度消失问题，使得网络能够更有效地学习和训练。而SPFF（spatial pyramid pooling fusion）模块则通过CBS模块进行特征提取，以及3个二维最大池化层MaxPool2d对特征图进行不同尺度的池化操作，最后将所有池化后的特征图进行特征拼接，并通过另一个CBS模块使特征恢复到与输入特征的相同维度。基于频域增强的自注意力机制与门控机制模块CFSA、低频增强的小波变换卷积模块LOWTC和新增的小目标检测头的特征融合模块的网络结构如图2—图5所示。

2.2 基于频域增强的自注意力和门控模块CFSA

受到图像恢复任务中Restormer模块（Zamir等，2022）和FSAF（feature selective anchor-free）模块（Zhu等，2019）的启发，本文提出CFSA模块，旨在充分挖掘Transformer的强大潜力，构建更为高效的特征提取模型，以减少无人机拍摄图像中噪声污染等因素对特征提取的不良影响。CFSA结构如图2所示。它由基于傅里叶频域的自注意力模块（Fourier frequency-based self-attention module，FFAS）和门控机制模块（gatede DConv，GDC）两部分组成。

FFAS模块将生成的$\boldsymbol{Q}$和$\boldsymbol{K}$分别进行局部分块后再进行局部傅里叶变换，在频域中进行点乘，将点乘后的结果进行傅里叶逆变换并将特征转到空域，恢复到原始分辨率后再与$\boldsymbol{V}$进行矩阵相乘，最后再将其输出结果与输入特征进行残差连接，增强了特征的表达能力；门控深度可分离卷积网络GDC对FFAS的输出特征进一步处理，通过门控机制和双线性变换控制特征的流动与融合，选择性增强或抑制不同特征。此处与Restormer中的GDFN（gated-dconv feed-forward network）模块不同的是，将3×3卷积换成了自定义的深度可分离卷积DSConv。首先经过一个深度卷积对每个通道单独进行卷积操作，不进行跨通道的融合，然后使用1×1卷积对深度卷积的输出进行通道混合。这种操作大大降低了计算复杂度，改进后的GFLOPs（giga floating-point operations）由原来的357.00降低为196.20。整体模块的工作原理如下：

FFAS模块首先将输入特征图$\boldsymbol{X}$归一化得到张量$\boldsymbol{Y}$，通过1×1的卷积获取空间上下文信息，通过3×3的深度可分离卷积获取通道上下文信息，生成$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，具体为

$$\begin{cases}\boldsymbol{Q} = W_d^Q W_p^Q Y \\ \boldsymbol{K} = W_d^K W_p^K Y \\ \boldsymbol{V} = W_d^V W_p^V Y\end{cases} \tag{1}$$

式中，$W_p^{(\cdot)}$是1×1的卷积，$W_d^{(\cdot)}$是3×3深度卷积。

然后，将$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$进行填充得到$\hat{\boldsymbol{Q}}$、$\hat{\boldsymbol{K}}$、$\hat{\boldsymbol{V}}$，目的是将其宽高都变为离其最近的4的整数倍，为后续傅里叶分块打下基础。接着，对$\hat{\boldsymbol{Q}}$和$\hat{\boldsymbol{K}}$执行局部分块，并在分块后的傅里叶频域中进行元素的点乘。对$\hat{\boldsymbol{Q}}$和$\hat{\boldsymbol{K}}$进行傅里叶变换后再相乘等同于$\hat{\boldsymbol{Q}}$和$\hat{\boldsymbol{K}}$进行矩阵乘法（卷积）后再傅里叶变换。上述计算过程定义为

$$\begin{aligned}\hat{\boldsymbol{X}} &= \boldsymbol{W}_q f_{\text{Attention}}\left(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{K}}, \hat{\boldsymbol{V}}\right) + \boldsymbol{X} \\ f_{\text{Attention}}\left(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{K}}, \hat{\boldsymbol{V}}\right) &= \hat{\boldsymbol{V}} \cdot f_{\text{softmax}}\left(\hat{\boldsymbol{K}} \cdot \hat{\boldsymbol{Q}}/\alpha\right)\end{aligned} \tag{2}$$

式中，$\boldsymbol{X}$是输入特征图，$\hat{\boldsymbol{X}}$是输出特征图，$\boldsymbol{W}_q$是线性投影矩阵。$\hat{\boldsymbol{Q}} \in \mathbf{R}^{HW \times C}$；$\hat{\boldsymbol{K}} \in \mathbf{R}^{C \times HW}$；$\hat{\boldsymbol{V}} \in \mathbf{R}^{HW \times C}$是由归一化后的输入特征图张量$\boldsymbol{Y}$（维度是$\mathbf{R}^{H \times W \times C}$）变换得到的。$\alpha$是一个可学习的缩放参数，在模型训练过程中通过反向传播自动更新，用于动态调整注

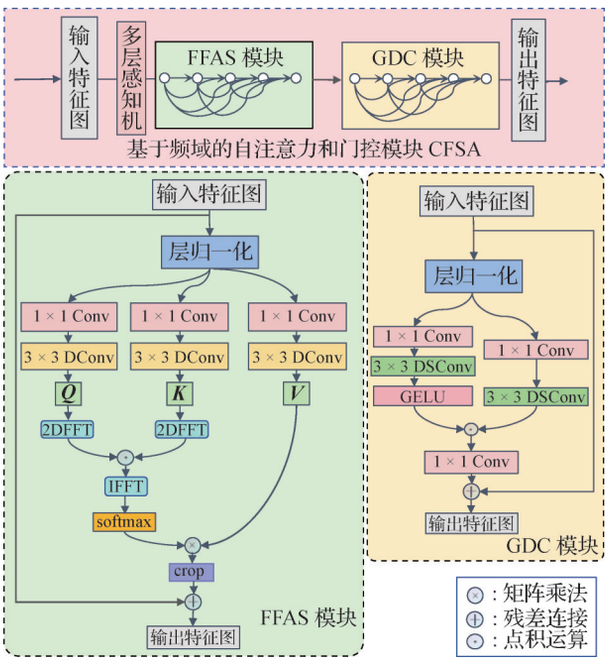


图 2　CFSA 模块结构图

Fig. 2　The structural diagram of the CFSA module

意力计算中的缩放比例。

而后，GDC 对 FFAS 的输出特征 $\hat{\boldsymbol{X}}$ 进一步处理，通过门控机制和双线性变换控制特征流动和融合，选择性增强或抑制不同特征。GDC 的计算过程可表达为

$$\hat{\boldsymbol{Y}} = W_p^0 f_{\text{Gating}}\left(\hat{\boldsymbol{X}}\right) + \hat{\boldsymbol{X}}$$

$$f_{\text{Gating}}\left(\hat{\boldsymbol{X}}\right) = \boldsymbol{\Phi}\left(W_d^1 W_p^1\left(LN\left(\hat{\boldsymbol{X}}\right)\right)\right) \odot W_d^2 W_p^2\left(LN\left(\hat{\boldsymbol{X}}\right)\right) \tag{3}$$

式中，$\hat{\boldsymbol{X}}$ 是 FFAS 模块的输出特征图，$\hat{\boldsymbol{Y}}$ 是本模块输出特征图。⊙是用来计算元素相关性的点积运算，即哈达玛积。$\boldsymbol{\Phi}$ 是非线性激活函数 GELU。LN 代表层归一化。$W_p^{(\cdot)}$ 是 1 × 1 的卷积运算，$W_d^{(\cdot)}$ 是 3 × 3 深度可分离卷积运算。

最后，将 CFSA 模块嵌入 C2f 中的 Bottleneck 层中，将之前提取到的浅层特征通过自注意力机制进一步进行处理。通过增强对关键信息的提取能力，可以提高 CFSA 模块对小目标的检测精度。

2.3　低频增强小波变换卷积模块 LOWTC

图像信号分为高频部分和低频部分。轮廓、颜色边缘等都是图像的高频信号，而大面积背景等都是低频信号。对于整幅图像而言，低频信息可以帮助区分小物体与背景，减少误检。为此，本小节设计了基于低频增强的小波变换卷积模块 LOWTC。输入的图像首先经过该模块进行二级小波分解，进一步提取图像的低频信息。由于二级分解意味着将图像分解为更粗略的近似部分（空域分辨率降低）和更精细的特征部分（频域分辨率提升），因此可以同时捕捉全局特征和局部细节，特别适合处理复杂图像。经此操作后，模型对输入图像的细节和语义信息的提取能力将进一步增强。

小波卷积层能够利用小波分解将输入信号分成不同的频带，使卷积操作能分别在不同的频带分量上进行处理。LOWTC 模块通过二级小波分解将输入的图像特征分解为不同频率成分，其中，低频分量保留图像的全局特征信息，高频分量保留图像的细节特征信息。其实现细节如图 3 所示。

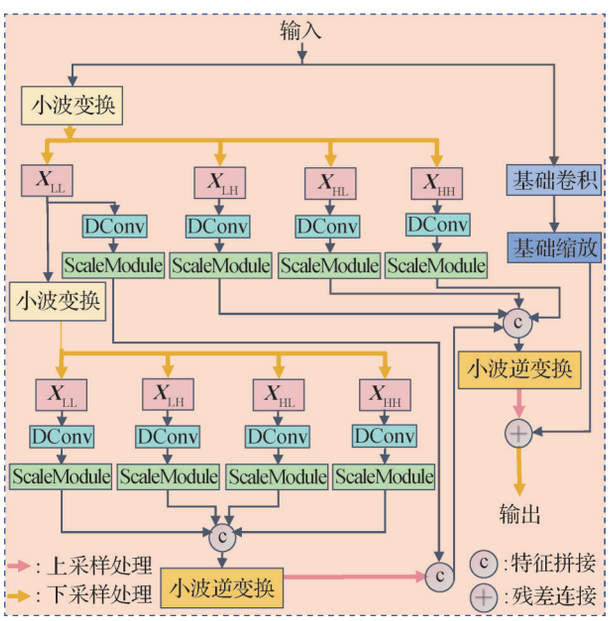


图 3　LOWTC 模块结构图

Fig. 3　The structural diagram of the LOWTC module

首先，定义小波卷积核，它们的数值矩阵计算式为

$$f_{\text{LL}} = \frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix},\quad f_{\text{LH}} = \frac{1}{2}\begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix}$$

$$f_{\text{HL}} = \frac{1}{2}\begin{bmatrix}1 & 1\\ -1 & -1\end{bmatrix},\quad f_{\text{HH}} = \frac{1}{2}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix} \tag{4}$$

式中，f_{LL} 用于提取低频分量，f_{LH}、f_{HL}、f_{HH} 分别用于提取垂直方向、水平方向、对角线方向的高频分量，对输入特征进行小波变换（wavelet transform，WT）。WT 的过程可表示为

$$\begin{bmatrix}\boldsymbol{X}_{\text{LL}} & \boldsymbol{X}_{\text{LH}}\\ \boldsymbol{X}_{\text{HL}} & \boldsymbol{X}_{\text{HH}}\end{bmatrix} = \begin{bmatrix}f_{\text{LL}}*\boldsymbol{X} & f_{\text{LH}}*\boldsymbol{X}\\ f_{\text{HL}}*\boldsymbol{X} & f_{\text{HH}}*\boldsymbol{X}\end{bmatrix} \tag{5}$$

式中，$\boldsymbol{X}$是输入特征图，*代表卷积运算，$\boldsymbol{X}_{\mathrm{LL}}$、$\boldsymbol{X}_{\mathrm{LH}}$、$\boldsymbol{X}_{\mathrm{HL}}$、$\boldsymbol{X}_{\mathrm{HH}}$代表经过小波变换后得到的4个频域的特征图，分别是输入特征图的低频分量、水平、垂直和对角线方向的高频分量，每一个特征图都有输入特征图$\boldsymbol{X}$一半的分辨率。

其次，通过递归分解将本层低频子带作为下层的输入，将其分成下一层的低频子带（LL）和高频子带（LH，HL，HH），对每个子带进行卷积操作，得到4个频域的特征。该过程可表示为

$$\boldsymbol{X}_{\mathrm{LL}}^{i},\boldsymbol{X}_{\mathrm{LH}}^{i},\boldsymbol{X}_{\mathrm{HL}}^{i},\boldsymbol{X}_{\mathrm{HH}}^{i}=WT\left(\boldsymbol{X}_{\mathrm{LL}}^{i-1}\right) \tag{6}$$

式中，$\boldsymbol{X}_{\mathrm{LL}}^{i}$、$\boldsymbol{X}_{\mathrm{LH}}^{i}$、$\boldsymbol{X}_{\mathrm{HL}}^{i}$、$\boldsymbol{X}_{\mathrm{HH}}^{i}$分别代表第$i$层（即下一层）的4个频域的特征图，$\boldsymbol{X}_{\mathrm{LL}}^{i-1}$代表第$i-1$层（即本层）的低频特征图。$WT$代表对本层的低频特征图进行小波变换。

然后，对各层的4个频域特征图分别进行深度可分离卷积（DConv），处理后低频部分的频率分辨率提高，而空间分辨率降低。因此，经过缩放模块（ScaleModule）调整输出，得到各层每一个子带的特征图，将它们拼接后得到各层小波卷积处理的特征图。再通过逆小波变换将这个特征图重新组合成一个完整的特征图作为各层的输出。需要说明的是：第i层的输出特征图与第$i-1$层小波变换处理后的低频特征图拼接后才得到第$i-1$层的低频特征图。其中，逆小波变换的卷积核即反卷积后的小波卷积核。上述过程可表述为

$$\begin{aligned}\boldsymbol{X}_{\mathrm{LL}}^{i-1}&=IWT(\boldsymbol{X}^{i})=\begin{bmatrix}\boldsymbol{f}_{\mathrm{LL}}^{-1} & \boldsymbol{f}_{\mathrm{LH}}^{-1}\\ \boldsymbol{f}_{\mathrm{HL}}^{-1} & \boldsymbol{f}_{\mathrm{HH}}^{-1}\end{bmatrix}*\boldsymbol{X}^{i}\\ \boldsymbol{X}^{i}&=\begin{bmatrix}SM(DC(\boldsymbol{X}_{\mathrm{LL}}^{i})) & SM(DC(\boldsymbol{X}_{\mathrm{LH}}^{i}))\\ SM(DC(\boldsymbol{X}_{\mathrm{HL}}^{i})) & SM(DC(\boldsymbol{X}_{\mathrm{HH}}^{i}))\end{bmatrix}\end{aligned} \tag{7}$$

式中，DC代表深度可分离卷积运算，SM代表特征缩放处理，IWT代表小波反卷积操作，由小波反卷积核（式（4）的转置矩阵）与$\boldsymbol{X}^{i}$经过卷积运算得到。

最后，将对原始特征图进行基础卷积后的输出特征图与逆小波变换的输出特征图进行残差连接，得到最终的特征图。如此可以结合小波变换和卷积操作，实现对图像特征的多尺度提取和融合，从而提高模型的特征表达能力。

该过程可以用图4描述：对低频子带递归进一步提取图像的浅层特征，通过逐层分解LL子带，可以看到在第2层上使用3×3卷积等于在第0层上进行12×12的卷积，这能有效地扩大感受野，在保留图像主要结构信息的同时，提取更丰富的语义信息。

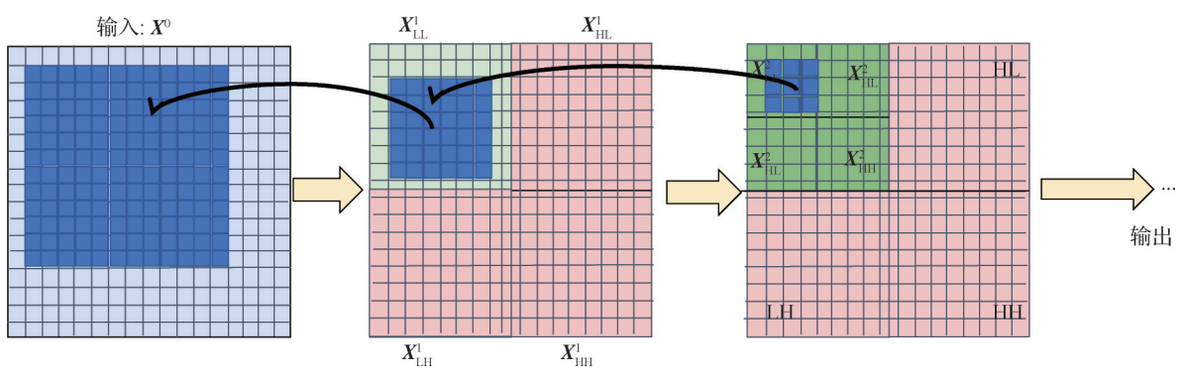


图4　二级分解低频带过程

Fig. 4　Two-level decomposition low-frequency band process

2.4　小目标检测头

在无人机拍摄的图像中，小目标通常占据整幅图像的很小的一部分像素。原有的YOLOv8模型结构，随着卷积的迭代，特征图的通道不断加深，所含有的深层语义信息逐渐丰富，小目标的位置和细节信息却逐渐被过滤掉。为了更好地保留原始图像中的小目标信息，本小节将在卷积操作中提取的浅层特征图后添加一个新的检测头，如图5中浅绿色部分所示。该特征图的分辨率为160×160像素，相对于分辨率为640×640像素的原始输入图像而言，可以更好地利用原始图像中的局部信息，从而提高小目标的检测能力。

3　实验与结果分析

3.1　实验配置

本文所有实验均在型号为NVIDIA GeForce RTX 4090 D，24 GB，操作系统为Ubuntu16.04的计

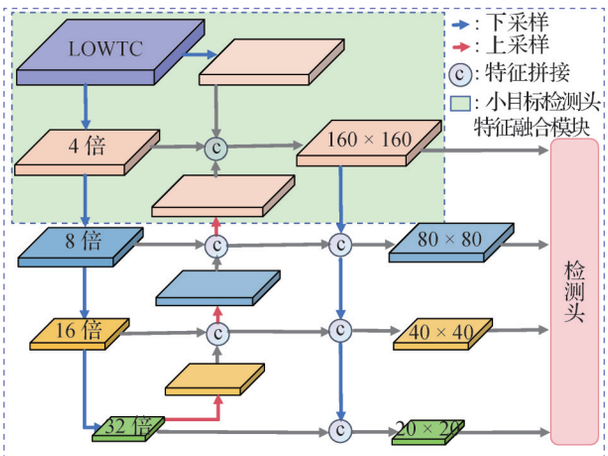


图 5　小目标检测头特征融合模块

Fig. 5　Small object detection head feature fusion module

算机上进行，使用 Python-3.9 软件环境和 18 torch-2.6.0+cu124 的深度学习框架。训练和验证时图像大小设为 640 × 640 像素，由于计算资源有限，将 batchsize 设置为 4，在使用 YOLO 系列模型训练时不使用预训练权重。其他对比实验的模型均按照其开发者文档进行配置。

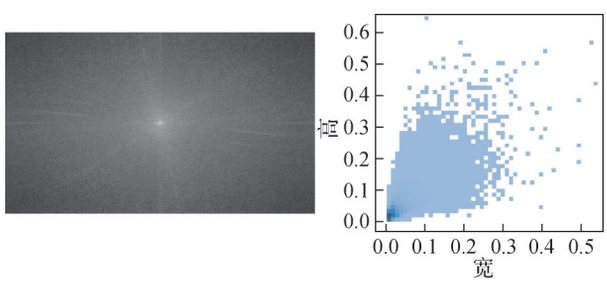
所有实验将在 VisDrone2019、UAVDT（unmanned aerial vehicle benchmark object detection and tracking）和 CARPK（car parking lot dataset）3 个公开数据集上进行。

3.1.1　数据集

天津大学等团队所拍摄和标注的 VisDrone2019 数据集（Zhu 等，2022）是一个开源的、大型的无人机视角的数据集。训练集 6 471 幅，验证集 548 幅，测试集 1 610 幅。包括 12 个类别，最终选取 10 个类别［0：pedestrian，1：people，2：bicycle，3：car，4：van，5：truck，6：tricycle，7：awning-tricycle，8：bus，9：motor］。其类别分布如图 6 所示。VisDrone2019 数据集图像中的目标大多为小目标（分辨率范围：（0，32］）和中等目标（分辨率范围：（32，96］），且其频域图显示低频信息含量丰富。

UAVDT 数据集（Du 等，2018）是无人机在多样化和复杂背景下捕捉的视频序列，视频以 30 帧/s 的速度录制，并以 1 080 × 540 像素的 JPEG（joint photographic experts group）格式存储，共约 80 000 帧，是一个用于无人机目标检测与跟踪的综合性数据集，涵盖了多种复杂场景，包括城市街道、乡村道路和自然环境。其高分辨率的图像和详细的标注信息为研究人员提供了丰富的研究资源。在对比实验中，UAVDT 数据集的高分辨率图像和多样化的场景显著提高了目标检测的准确性。

CARPK 数据集（Hsieh 等，2017）是一个专注于停车场场景的车辆检测与计数的数据集。覆盖了 4 个不同的停车场，涵盖了多种停车场环境，包括不同的光照条件、车辆密度和背景变化。包含 1 573 幅图像，这些图像由 Phantom 3 Professional 无人机在约 40 m 的高度拍摄，分辨率为 1 280 × 720 像素。数据集中标注了近 90 000 辆汽车，每辆车都通过边界框进行了精确标注，标注信息包括边界框的左上角和右下角坐标。



(a) 图像的频域图　　(b) 目标宽高分布（颜色越深，数目越多）

图 6　研究数据集 VisDrone2019 情况说明

Fig. 6　The description of the VisDrone2019 dataset ((a) frequency domain image; (b) the distribution of the target's width and height)

3.1.2　评价指标

1）精确度和召回率。精确度（precision）即被预测为正（标签为 1）的样本中有多少是正确的。召回率（recall）即真正为正（标签为 1）的样本中有多少被正确预测为正。

2）准确率。准确率（accuracy）是模型预测结果的标签与真实标签一致的比例。

3）平均精确率。平均精确率（average precision，AP）是准确率—召回率（precision-recall，PR）曲线下面积的度量。

4）参数量、计算量和每秒帧数。模型的参数量（Params）即模型中可训练参数的总数，计算量（GFLOPs）表示模型在一次前向传播中所需的浮点运算次数，每秒帧数（frames per second，FPS）是指模型在单位时间内能够处理的图像帧数。

3.2　对比实验

为了验证所提模型的优越性和有效性，选择

ATSS(adaptive training sample selection)(Zhang 等, 2020)、FCOS(fully convolutional one-stage object detection)(Tian 等, 2019)、Faster R-CNN、DINO(DETR with improved denoising anchor boxes)(Zhang 等, 2022)、YOLOv5、YOLOv8、YOLOv9(Wang 等, 2024b)、TPH-YOLOv5(improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios)(Zhu 等, 2021)、CZDet(cascaded zoom-in detector for high resolution aerial images)(Meethal 等, 2023)、RTMDet(an empirical study of designing real-time object detectors)(Lyu 等, 2022)、SODYOLOv8(enhancing YOLOv8 for small object detection in traffic scenes)(Khalili 和 Smyth, 2024)和 YOLOv12(Tian 等, 2025)多种经典检测算法或优秀检测模型,在3个数据集上从客观数据和视觉效果两个方面进行比较。表1—表3分别给出了不同检测算法在3个数据集上的实验结果。

表1 VisDrone2019数据集上的对比实验结果

Table 1 Comparison results on the VisDrone2019 dataset

模型	AP50/%	AP/%	APs/%	APm/%	Params/M	GFLOPs	FPS/(帧/s)
ATSS	24.34	13.97	5.42	23.05	32.1	256.2	50.6
FCOS	24.91	14.30	5.49	23.64	47.1	98.3	47.77
Faster R-CNN	26.57	16.35	6.20	29.67	41.4	370	50.4
DINO	35.59	19.52	10.61	30.03	47.6	500.1	19.94
YOLOv5l	38.33	21.69	13.27	31.95	46.2	107.8	73.9
YOLOv8l	40.60	24.50	14.40	37.80	43.7	165.7	67.9
YOLOv9m	44.17	26.37	16.75	39.27	25.3	102.1	33.63
TPH-YOLOv5	40.67	23.82	14.96	33.66	60.4	237.3	111.2
CZDet	43.62	24.49	17.33	33.00	**19.8**	**25.03**	55.25
RTMDet	37.61	22.23	12.04	34.63	52.3	129.8	46.32
SODYOLOv8	45.10	26.60	16.86	38.99	45.2	226	**116.8**
YOLOv12l	43.00	26.39	16.53	39.46	26.5	89.75	50.86
YOLO-WF(本文)	**47.10**	**28.60**	**19.90**	**40.30**	43.52	238.44	83.04

注:加粗字体表示各列最优结果。

从表1可以看出,本文模型YOLO-WF在AP50(IoU阈值为0.5)、AP、APs(小目标)和APm(中目标)这4个关键指标上都取得最佳结果。其中AP50、AP和APs指标分别达到47.10%、28.60%和19.90%,显著高于其他模型。尤其是在APs指标上,提出的模型比次优模型CZDet高出2.57%,这充分说明提出的模型在捕捉小目标的特征方面的优势。其次是SODYOLOv8、YOLOv9、CZDet和YOLOv12模型,4个关键指标也都取得不错的结果。其他模型则表现一般,在不同的指标上各有差异。从参数量(Params)来看,YOLO-WF参数量为43.52 M,与YOLOv8的43.7 M相近,在所有的对比模型中处于中间位置,但YOLO-WF的整体性能更优。尽管本文提出的模型GFLOPs为238.44,在所有对比模型中处于较高位,但每秒处理的帧数FPS指标除低于SODYOLOv8和TPH-YOLOv5外,明显高于其他检测模型。

如表2所示,由于UAVDT数据集是从原视频中抽取逐帧图像获得,其相似图像较多,因此各个模型的检测结果都较其在VisDrone2019数据集上的表现显著提升。其中,SDOYOLOv8和YOLOv8模型分别在AP和APm指标上达到最高,提出的模型在这两个指标上仅次于这两个模型,但AP50和APs指标是所有对比模型中最高的,分别达到81.56%和38.54%,明显高于上述这两个模型,比次优的YOLOv8和SODYOLOv8分别高出2.44%和2.22%。

表 2　UAVDT 数据集上的对比实验结果

Table 2　Comparison results on the UAVDT dataset

/%

模型	AP50	AP	APs	APm
ATSS	51.32	26.24	18.38	49.23
FCOS	61.67	32.42	23.35	59.10
Faster R-CNN	43.21	25.60	16.09	55.00
DINO	69.25	36.79	28.70	61.10
YOLOv5l	77.76	43.61	35.97	64.72
YOLOv8l	79.12	44.23	35.46	**67.11**
YOLOv9m	71.58	37.29	29.51	59.10
TPH-YOLOv5	77.84	37.46	29.93	58.13
CZDet	63.00	35.37	28.21	59.17
RTMDet	69.06	38.56	30.45	60.72
SODYOLOv8	73.44	**48.67**	36.32	63.90
YOLOv12l	62.80	35.90	28.50	56.70
YOLO-WF(本文)	**81.56**	45.63	**38.54**	65.12

注：加粗字体表示各列最优结果。

表 3　CARPK 数据集上的对比实验结果

Table 3　Comparison results on the CARPK dataset

/%

模型	AP50	AP	APs	APm
ATSS	80.9	49.8	3.0	51.4
FCOS	74.2	40.4	0.8	41.9
Faster R-CNN	79.9	49.6	3.3	51.1
DINO	94.4	52.5	27.9	62.1
YOLOv5l	**96.0**	58.2	24.6	67.8
YOLOv8l	95.1	58.2	26.5	67.9
YOLOv9m	95.0	56.0	23.5	65.0
TPH-YOLOv5	82.0	58.8	6.2	60.6
CZDet	68.5	34.9	6.93	36.1
RTMDet	82.0	**61.9**	3.0	63.9
SODYOLOv8	95.8	56.5	27.8	66.1
YOLOv12l	95.2	58.0	24.0	67.8
YOLO-WF(本文)	94.3	58.6	**33.3**	**67.9**

注：加粗字体表示各列最优结果。

从表 3 可以看出，YOLOv5l（Bochkovskiy 等，2020）模型在 AP50 指标上表现最佳，达到 96.0%，显示出其在检测任务中对目标的高召回率。RTMDet 在 AP 指标上表现突出，达到 61.9%。其他多数检测模型在不同的指标上各有优劣。但 APs 指标表明，ATSS、FCOS、Faster R-CNN、TPH-YOLOv5、CZDet 以及 RTMDet 模型对小目标的检测是失败的。YOLO-WF 模型虽然在 AP50 和 AP 两个指标上不是最优的，但 APs 和 APm 指标却是最高的，分别为 33.3% 和 67.9%。尤其是 APs 指标明显高于其他所有检测模型，比次优的 DINO 模型高出 5.4%。CARPK 数据集上的对比实验说明，所提 YOLO-WF 模型在复杂场景下同样能够表现出优异的检测性能，具有较好的鲁棒性。

综上所述，本文提出的 YOLO-WF 检测模型在众多先进模型中具有明显的竞争优势，特别在小目标检测任务中表现卓越。

为了更直观地说明改进模型的性能，选取 FCOS、DINO、YOLOv5、YOLOv8、YOLOv9 以及 YOLO-WF 模型分别在 VisDrone2019、UAVDT 和 CARPK 数据集上进行检测结果的可视化。

图 7 展示了 VisDrone2019 数据集上的可视化结果，其中每幅图像中的大黄框展示检测图像局部视野（小黄框）的放大图。DINO、FCOS、YOLOv5、YOLOv8、YOLOv9 模型分别检测到选中图像中的完整行人数量为：11、13、11、10、12。相比之下，YOLO-WF 模型成功检测出了图像中 17 个完整的行人，漏检数量显著降低。这表明其在小目标检测方面相较于其他检测模型具有显著的竞争优势。

图 8 给出了在 UAVDT 数据集上的可视化结果。可以看出，所有检测模型在处理视野近处的车辆时，均能取得较好的检测效果。但对于视野远处的极小目标，除本文提出的模型外，其他模型均未能提供令人满意的结果。图 8 每幅图像右上角的黄框展示了视野远端局部的放大图像。YOLOv5 模型仅仅检测到选中区域中的 1 辆汽车；DINO、FCOS、YOLOv8、YOLOv9 模型分别检测到选中区域中的汽车数量为：9、6、7、7。相比之下，YOLO-WF 模型成功检测出了选中区域中的 10 辆汽车。视野远端的物体尺寸仅几像素大小，即使肉眼也难以分辨，检测这些物体是一项极具挑战性的任务，本文提出的 YOLO-WF 模型较好地完成了这一任务。

图 9 展示了在 CARPK 数据集上的可视化结果。可以看出，提出的 YOLO-WF 模型在 CARPK 上，仍然

图7　VisDrone2019数据集上不同模型对比检测可视化结果

Fig. 7　Comparison of detection visualization results of different models on the VisDrone2019 dataset

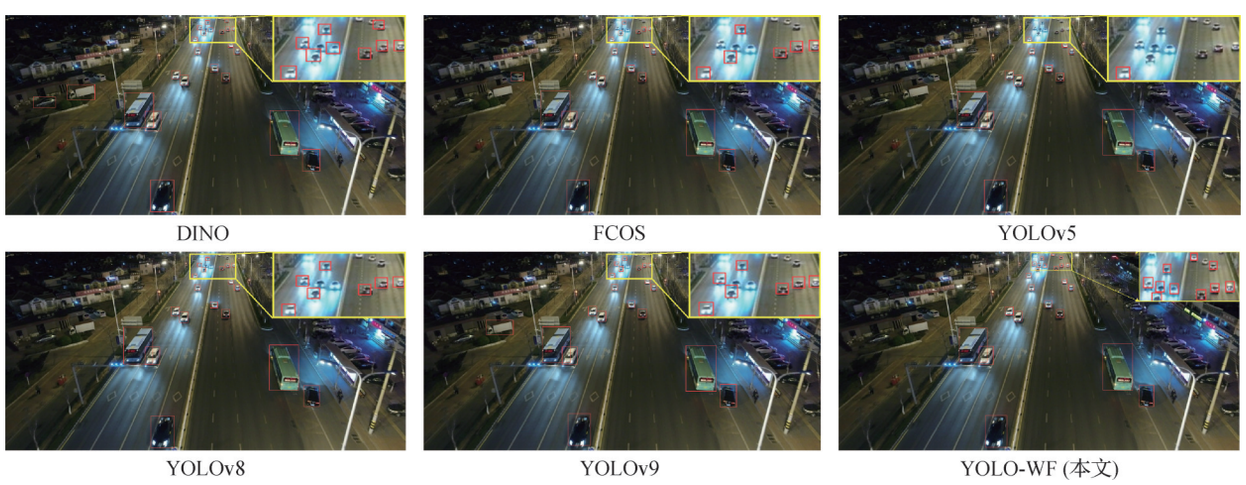


图8　UAVDT数据集上不同模型对比检测可视化结果

Fig. 8　Comparison of detection visualization results of different models on the UAVDT dataset

表现出优越的检测性能。对于非YOLO系列的模型DINO和FCOS而言，显然DINO能够检测出更多准确的目标，图9中只有一车漏检，而FCOS却有大量的车辆漏检；对于YOLO系列模型，YOLOv5和YOLOv9模型均有不同程度的漏检，在暗光环境下的检测效果欠佳。相比而言，作为基线模型的YOLOv8却表现出较好的检测效果，但存在很多错检目标。所提YOLO-WF模型能优化基线模型的检测结果，虽然错检目标现象依然存在，但却正确检测出了图像中的所有车辆。这是其他所有对比模型都未能具有的性能。

3个数据集上的对比实验可视化结果表明：在以小物体为主要检测对象的视觉任务场景中，相对于现有的经典和优秀模型，本文提出的YOLO-WF模型具有更好的检测性能和更强的适应性。

3.3　消融实验

为了验证改进模型各个模块的有效性，在VisDrone2019数据集上进行了消融实验。图10展示了消融实验训练过程中的准确率、召回率，以及IoU阈值分别为0.5(AP50)和0.5～0.95(AP50-95)时的平均准确率曲线图。可以看出，在加入所有模块后，模型的性能较基础模型有显著提升。

表4展示了添加不同模块的消融实验的结果。表中第1行数据表示在原始的YOLOv8模型上的结果；第2行数据表示在YOLOv8模型的backbone部分的C2f中加入基于傅里叶频域增强的自注意力模块

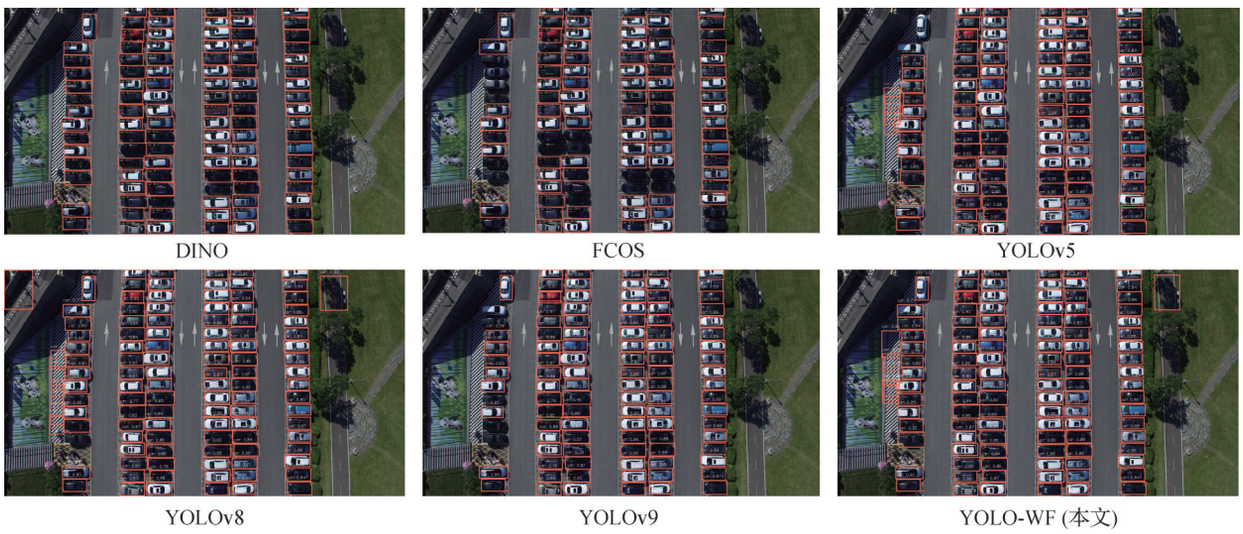


图 9　CARPK 数据集上不同模型对比检测可视化结果

Fig. 9　Comparison of detection visualization results of different models on the CARPK dataset

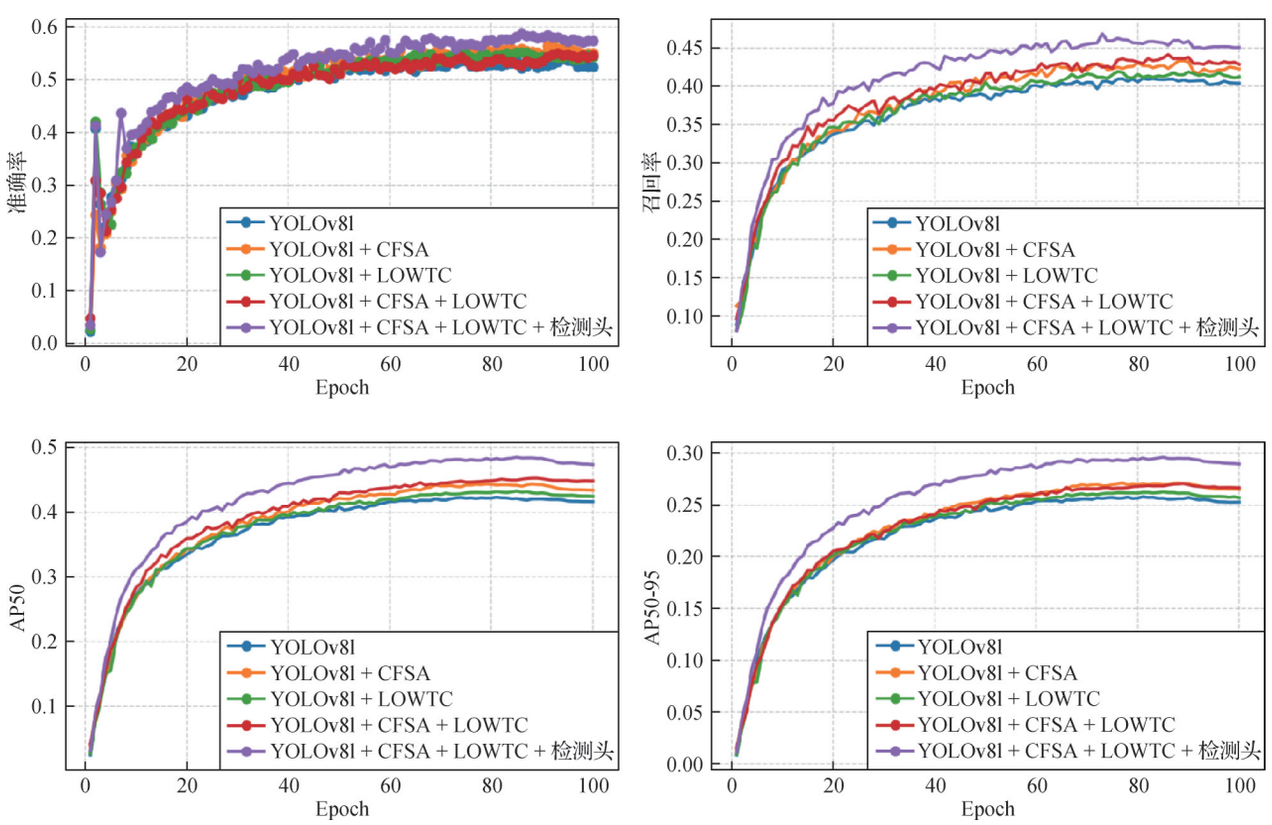


图 10　消融实验中准确率、召回率以及 AP50 和 AP50-95 的对比结果

Fig. 10　Comparison results of accuracy, recall, AP50, and AP50-95 in the ablation study

后的结果；第 3 行数据表示单独加上小目标检测头的结果；第 4 行数据表示加入递归低频的小波变换卷积 LOWTC 模块后的结果；第 5 行数据表示同时加入 CFSA 和 LOWTC 模块后的结果；第 6 行表示将 LOWTC 模块提取的特征融合进新加的小目标检测头后的结果；第 7 行表示将 CFSA 模块提取的特征融合进新加的小目标检测头后的结果；最后一行数据表示将 LOWTC 和 CFSA 模块提取的特征都融合到小目标检测头后的结果。

加入的 CFSA 模块、LOWTC 模块、频域增强和特征融合模块（即将 LOWTC 和 CFSA 提取的特征融合到检测头的子模块，图 5 所示）及小目标检测头都有

表4 消融实验结果

Table 4 Ablation study results

模型				指标					
YOLOv8	+CFSA	+检测头	+LOWTC	AP50/%	AP/%	APs/%	APm/%	Params/M	GFLOPs
√	–	–	–	40.6	24.5	14.2	37.8	43.69	**165.74**
√	√	–	–	40.9	25.0	14.8	37.1	44.31	196.92
√	–	√	–	43.3	25.3	17.2	35.4	**42.89**	206.29
√	–	–	√	40.8	24.4	14.7	37.1	43.69	165.88
√	√	–	√	42.0	25.4	15.4	38.6	47.44	357.41
√	–	√	√	46.5	28.2	19.7	39.3	42.89	207.27
√	√	√	–	46.6	28.4	19.8	39.9	43.50	237.60
√	√	√	√	**47.1**	**28.6**	**19.9**	**40.3**	43.52	238.44

注：加粗字体表示各列最优结果。“√”表示选择相应模块，“–”表示未选择。

助于提高模型的检测性能。通过对比单独加入不同模块的性能指标，发现加入小目标检测头对算法检测性能的提升最为显著，AP50提升2.7%，*AP*指标提升0.8%；单独加入CFSA和LOWTC模块后，虽然各项指标的变化不明显，但并不能说明这两个模块对提升整个检测模型的性能效果不佳。当这两个模块和添加的小目标检测头模块同时嵌入检测网络协作工作时，整个模型的检测性能比单独添加小目标检测头有了明显提升，AP50、AP、APs和APm分别提高了3.8%、3.3%、2.7%和4.9%，和基线模型YOLOv8相比，AP50、AP和APs指标提升更为显著，分别提高了6.5%、4.1%和5.7%，并且在保证对中等大小目标检测精度的同时，还有一定程度的提升，APm指标提高了2.5%，二者的贡献主要体现在将其输出特征融合到小目标检测头，融合了LOWTC和CFSA两个模块所学习的更精细的输出特征后，模型检测性能的提升更为显著。

改进模块在YOLO系列模型中展现出良好的泛化性能，能够显著提升检测精度。表5给出了在YOLOv5、YOLOv8及YOLOv11上引入该模块后的实验结果，三者指标均明显提高，验证了所提改进的通用性与有效性。

3.4 消融实验可视化分析

图11是消融实验的可视化结果。基础模型YOLOv8在加入CFSA、小目标检测头以及LOWTC后对小目标的检测能力均有不同程度的提升，尤其是在加入小目标检测头后对小物体的检测能力有显著提升，检测头在融合CFSA和LOWTC的输出特征后对小目标的检测能力更是进一步提升，这是因为其融合了CFSA和LOWTC提取到的目标的更精细的特征信息，YOLOv8模型检测出选中图像区域中的1辆汽车，相比之下，所提YOLO-WF模型检测出了7辆，显著提升了YOLOv8对小目标的检测性能。

表5 泛化能力实验结果

Table 5 Generalization ability experimental results

/%

模型	AP50	AP	APs	APm
YOLOv5l	38.33	21.69	13.27	31.95
YOLOv5l + 改进模块	45.90	27.70	19.00	35.30
YOLOv8l	40.60	24.50	14.40	37.80
YOLO-WF (本文)	47.10	28.60	19.90	40.30
YOLOv11l	40.40	24.30	14.30	36.90
YOLOv11l + 改进模块	47.00	28.20	19.70	39.20

4 结 论

由于无人机高空拍摄的图像数据集上的小目标众多，且图像背景复杂、噪声因素多，现有模型对小目标的检测效果不佳。本文以YOLOv8为基础，首先在骨干网络中构建基于傅里叶频域的注意力机制模块CFSA来高效聚合图像的全局和局部特征，在不同的频域上计算自注意力，有效提升模型对关键信息的提取能力；其次设计了基于二级分解低频的

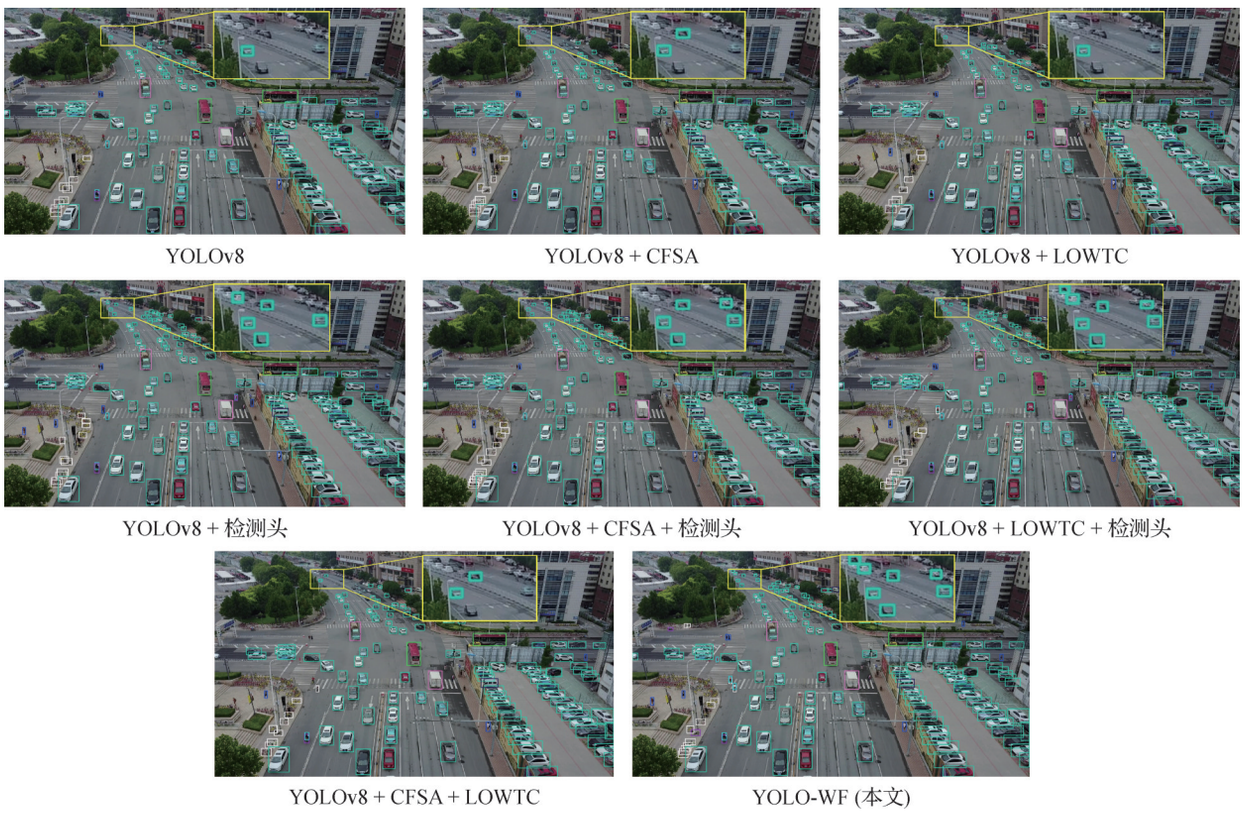


图 11　消融实验可视化结果

Fig. 11　Visualization of ablation study results

哈尔小波卷积模块LOWTC并嵌入网络中，它通过小波变换的多尺度特性扩展感受野，使传统CNN能够在不显著增加参数的情况下获得接近全局的感受野，有效缓解传统卷积的长距离依赖问题；最后针对小目标，增加一个融合CFSA和LOWTC模块的输出特征后的小目标检测头，提升了模型对小目标的检测能力。在VisDrone2019、UAVDT、CARPK3个数据集上分别与经典和优秀的目标检测模型进行了对比实验。结果表明，所提模型无论是在正常环境，还是在密集小目标、遮挡目标、黑暗环境、高曝光环境下都具有明显的竞争优势。

由于改进的YOLO-WF模型增加了CFSA注意力模块和LOWTC模块，导致模型的复杂度相较于基线模型YOLOv8有所提升，这也是现有所提模型的不理想之处。在实际应用中，模型的选择不仅取决于检测精度，还需考虑效率和资源消耗。因此，在保障检测精度的同时优化模型的资源消耗将是我们未来的重点工作之一。例如针对小目标检测，可以将大目标检测头进行剪枝以减少模型的复杂度和计算量；当需要部署至无人机端时，可考虑将参数改用半精度类型（FP32→FP16/INT8）以减少模型的内存占用和计算量。

参考文献（References）

Bharati P and Pramanik A. 2020. Deep learning techniques——R-CNN to mask R-CNN: a survey//Das A K, Nayak J, Naik B, Pati S K, Pelusi D, eds. Computational Intelligence in Pattern Recognition. Singapore, Singapore: Springer: 657-668 [DOI: 10.1007/978-981-13-9042-5_56]

Bochkovskiy A, Wang C Y and Liao H Y M. 2020. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. [2025-07-14]. https://arxiv.org/pdf/2004.10934.pdf

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]

Chandra M A and Bedi S S. 2021. Survey on SVM and their application in image classification. International Journal of Information Technology, 13(5): 1-11 [DOI: 10.1007/s41870-017-0080-1]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2020. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2025-07-14]. https://arxiv.org/pdf/2010.11929.pdf

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

Duan Y P, Liu F, Jiao L C, Zhao P and Zhang L. 2017. SAR image segmentation based on convolutional-wavelet neural network and Markov random field. Pattern Recognition, 64: 255-267

Elman J L. 1990. Finding structure in time. Cognitive Science, 14(2): 179-211 [DOI: 10.1207/s15516709cog1402_1]

Finder S E, Amoyal R, Treister E and Freifeld O. 2024. Wavelet convolutions for large receptive fields//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 363-380 [DOI: 10.1007/978-3-031-72949-2_21]

Gal R, Hochberg D C, Bermano A and Cohen-Or D. 2021. SWAGAN: a style-based wavelet-driven generative model. ACM Transactions on Graphics, 40(4): #134 [DOI: 10.1145/3450626.3459836]

Gao D D, Zhou D W, Wang W J, Ma Y and Li S S. 2023. Lightweight super-resolution via grouping fusion of feature frequencies. Journal of Computer-Aided Design and Computer Graphics, 35(7): 1020-1031 (高丹丹, 周登文, 王婉君, 马钰, 李珊珊. 2023. 特征频率分组融合的轻量级图像超分辨率重建. 计算机辅助设计与图形学学报, 35(7): 1020-1031) [DOI: 10.3724/SP.J.1089.2023.19524]

Guo T T, Mousavi H S, Vu T H and Monga V. 2017.Deep wavelet prediction for image super-resolution//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, USA: IEEE: 1100-1109, [DOI: 10.1109/CVPRW.2017.148]

Guth F, Coste S, De Bortoli V and Mallat S. 2022. Wavelet score-based generative modeling//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #35

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]

Hsieh M R, Lin Y L and Hsu W H. 2017. Drone-based object counting by spatially regularized regional proposal network//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4165-4173 [DOI: 10.1109/ICCV.2017.446]

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017b. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Huang H B, He R, Sun Z N and Tan T N. 2017a. Wavelet-SRNet: a wavelet-based CNN for multi-scale face super resolution//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1698-1706 [DOI: 10.1109/ICCV.2017.187]

Jocher G, Chaurasia A and Qiu J. 2023. Ultralytics YOLOv8 [EB/OL]. [2025-07-14]. https://docs.ultralytics.com/models/yolov8/#citations-and-acknowledgments

Khalili B and Smyth A W. 2024. SOD-YOLOV8——enhancing YOLOv8 for small object detection in aerial imagery and traffic scenes. Sensors, 24(19): #6209 [DOI: 10.3390/s24196209]

Li C Y, Li L L, Jiang H L, Weng K H, Geng Y F, Li L, et al. 2022. YOLOv6: a single-stage object detection framework for industrial applications [EB/OL]. [2025-07-14]. https://arxiv.org/pdf/2209.02976.pdf

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Lyu C Q, Zhang W W, Huang H A, Zhou Y, Wang Y D, Liu Y Y, et al. 2022. RTMDet: an empirical study of designing real-time object detectors [EB/OL]. [2025-07-14]. https://arxiv.org/pdf/2212.07784.pdf

Ma J R, Hu Z W, Shao Q Q, Wang Y C, Zhou Y Q, Liu J Y, et al. 2022. Detection of large herbivores in UAV images: a new method for small target recognition in large-scale images. Diversity, 14(8): #624 [DOI: 10.3390/d14080624]

Meethal A, Granger E and Pedersoli M. 2023. Cascaded zoom-in detector for high resolution aerial images//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2023: 2046-2055 [DOI: 10.1109/CVPRW59228.2023.00198]

Pan X Y, Jia N X, Mu Y Z and Gao X R. 2023. Survey of small object detection. Journal of Image and Graphics, 28(9): 2587-2615 (潘晓英, 贾凝心, 穆元震, 高炫蓉. 2023. 小目标检测研究综述. 中国图象图形学报, 28(9): 2587-2615) [DOI: 10.11834/jig.220455]

Phung H, Dao Q and Tran A. 2023. Wavelet diffusion models are fast and scalable image generators//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10199-10208 [DOI: 10.1109/CVPR52729.2023.00983]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2025-07-14]. https://arxiv.org/pdf/1804.02767.pdf

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards

real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/tpami.2016.2577031]

Shi Z H, Wu C W, Li C J, You Z Z, Wang Q and Ma C C. 2023. Object detection techniques based on deep learning for aerial remote sensing images: a survey. Journal of Image and Graphics, 28(9): 2616-2643 (石争浩, 仵晨伟, 李成建, 尤珍臻, 王泉, 马城城. 2023. 航空遥感图像深度学习目标检测技术研究进展. 中国图象图形学报, 28(9): 2616-2643) [DOI: 10.11834/jig.221085]

Tian Y J, Ye Q X and Doermann D. 2025. YOLOv12: attention-centric real-time object detectors//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. San Diego, USA: OpenReview.net

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wan D H, Lu R S, Wang S L, Shen S Y, Xu T and Lang X L. 2023. YOLO-HR: improved YOLOv5 for object detection in high-resolution optical remote sensing images. Remote Sensing, 15(3): #614 [DOI: 10.3390/rs15030614]

Wang A, Chen H, Liu L H, Chen K, Lin Z J, Han J G, et al. 2024a. YOLOv10: real-time end-to-end object detection//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #3429

Wang C Y, Bochkovskiy A and Liao H Y M. 2023a. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7464-7475 [DOI: 10.1109/CVPR52729.2023.00721]

Wang C Y, Yeh I H and Mark Liao H Y. 2024b. 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 G, Chen Y F, An P, Hong H Y, Hu J H and Huang T G. 2023b. UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. Sensors, 23(16): #7190 [DOI: 10.3390/s23167190]

Wang Q Q and Li C B. 2022. Incident detection and classification in renewable energy news using pre-trained language models on deep neural networks. Journal of Computational Methods in Sciences and Engineering, 22(1): 57-76 [DOI: 10.3233/JCM-215594]

Wang Y Q. 2014. An analysis of the Viola-Jones face detection algorithm. Image Processing on Line, 4: 128-148 [DOI: 10.5201/ipol.2014.104]

Williams T L and Li R. 2018. Wavelet pooling for convolutional neural networks//Proceedings of 2018 International Conference on Learning Representations. Vancouver, Canada: ICLR

Xu D Q and Wu Y Q. 2020. Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection. Sensors, 20(15): #4276 [DOI: 10.3390/s20154276]

Zamir S W, Arora A, Khan S, Hayat M, Khan F S and Yang M H. 2022. Restormer: efficient transformer for high-resolution image restoration//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5718-5729. [DOI: 10.1109/CVPR52688.2022.00564]

Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, et al. 2022. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection [EB/OL]. [2025-07-14]. https://arxiv.org/pdf/2203.03605.pdf

Zhang S F, Chi C, Yao Y Q, Lei Z and Li S Z. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9756-9765 [DOI: 10.1109/CVPR42600.2020.00978]

Zhao H H, Xue W C, Li X B, Gu Z X, Niu L and Zhang L Q. 2020. Multi-mode Neural network for human action recognition. IET Computer Vision, 14(8): 587-596 [DOI: 10.1049/iet-cvi.2019.0761]

Zhu C C, He Y H and Savvides M. 2019. Feature selective anchor-free module for single-shot object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 840-849 [DOI: 10.1109/CVPR.2019.00093]

Zhu P F, Wen L Y, Du D W, Bian X, Fan H, Hu Q H, et al. 2022. Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11): 7380-7399 [DOI: 10.1109/TPAMI.2021.3119563]

Zhu X K, Lyu S C, Wang X and Zhao Q. 2021. 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]

作者简介

刘旭，女，硕士研究生，主要研究方向为计算机视觉和深度学习。E-mail: 923374286@qq.com

包芳勋，通信作者，男，教授，博士生导师，主要研究方向为函数逼近、分形、CAGD和图像处理。E-mail: fxbao@sdu.edu.cn

宋佩博，男，博士研究生，主要研究方向为计算机视觉和深度学习。E-mail: 202311875@mail.sdu.edu.cn

杜宏伟，男，工程师，主要研究方向为生成式AI和计算机视觉。E-mail: duhongwei@inspur.com