DOI:10.19651/j.cnki.emt.2518802

多模态融合的输电线路部件多尺度检测方法*

周 景 赵 毅 刘 心
(华北电力大学控制与计算机工程学院 北京 102206)

摘 要:在输电线路无人机巡检航拍图像的关键部件检测任务中,针对单一模态检测方法精度低和小目标漏检率高的问题,提出了一种融合可见光图像和红外图像的多模态多尺度目标检测方法。首先,该网络构建了并行的双流特征提取主干,旨在同步处理可见光与红外图像,以充分利用前者丰富的色彩与纹理细节信息,以及后者卓越的成像稳定性与高对比度特性。其次,为实现跨模态信息的交互与互补,设计了多模态特征交互融合模块(MFIFM),该模块能动态地调整不同模态特征的融合权重,自适应地整合最具判别力的信息,有效缓解模态差异带来的信息冲突。此外,为提升对小目标部件的感知能力,提出了混合残差多尺度 Transformer(HRMS Transformer)模块嵌入到双流主干中,通过多头窗口注意力机制,层级式特征重组以及与残差相结合的策略,增强全局上下文信息提取能力。实验结果表明,该模型精度 mAP@50 和 mAP@50:95 较现有单模态方法分别提升 5.35%和 4.48%。验证了多模态融合技术在输电线路检测领域的有效性和可用性。

关键词:输电线路检测;多模态特征融合;Swing Transformer;注意力机制;双流主干网络;深度学习
中图分类号:TP391.41;TM75;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.6040

Multi-scale detection method for transmission line components based on multimodal fusion

Zhou Jing Zhao Yi Liu Xin
(School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)

Abstract: In the key component detection task of UAV inspection of aerial images for transmission lines, a multimodal multi-scale target detection approach is proposed to address the challenges of accuracy degradation and high miss rates for small targets in single-modal detection methods. This approach integrates visible light and infrared images. First, the network constructs a parallel two-stream feature extraction backbone designed to simultaneously process visible light and infrared images. This design fully utilizes the rich color and texture detail information from the visible light images, along with the superior imaging stability and high contrast characteristics of the infrared images. Next, to facilitate cross-modal information interaction and complementarity, a Multimodal Feature Fusion Interactive Module (MFIFM) is developed. This module dynamically adjusts the fusion weights of features from different modalities, adaptively integrating the most discriminative information and effectively mitigating conflicts arising from modality differences. Additionally, to enhance the perception of small target components, a Hybrid Residual Multi-Scale Transformer (HRMS Transformer) module is incorporated into the dual-stream backbone. By utilizing a multi-head attention mechanism, hierarchical feature reorganization, and a residual-based strategy, the model′s ability to extract global context information is significantly strengthened. Experimental results demonstrate that the model′s mean Average Precision (mAP) at IoU thresholds of 0.50 and 0.50:0.95 improves by 5.35% and 4.48%, respectively, compared to existing single-modal methods. These findings confirm the effectiveness and applicability of multimodal fusion technology in transmission line inspection.

Keywords: transmission line detection; multimodal feature fusion; Swing Transformer; attention mechanism; dual-stream backbone network; deep learning

0 引 言

输电线路作为电力系统的重要组成部分,其关键部件(如防震锤、绝缘子)的运行状态直接关系电网稳定性和供电安全。随着电网规模扩大与智能化巡检需求增长,无人机巡检凭借其高效性和广泛覆盖,逐渐替代传统的人工巡

收稿日期:2025-05-13
* 基金项目:国家自然科学基金(62301220)项目资助

检,成为现代输电线路巡检的主流手段。如何高效、精准地检测输电线路关键部件,已成为电力系统安全运维的核心挑战。因此,构建适应复杂场景的输电线路关键部件检测模型,对人机采集的海量图像数据进行自动化、智能化的分析处理技术具有重要意义。

近年来,深度学习技术的突破为输电线路检测研究提供了新的推动力。特别是卷积神经网络(convolutional neural networks, CNN)和 Transformer 架构,在计算机视觉任务中取得了巨大成功,并被广泛应用于输电线路部件检测。然而,当前大多数研究仍集中于单模态检测算法。在输电线路可见光图像检测方面,文献[1]通过先进的特征提取和多尺度融合技术改进 YOLOv8 模型,提高了在复杂背景和多目标场景下绝缘子的检测性能。文献[2]提出一种基于强化学习和 Transformer 的输电线路缺陷智能识别方法,通过 DetNet 特征提取、DQN 筛选重要区域和双线性注意力机制提升检测精度和鲁棒性。文献[3]针对小绝缘子检测场景,采用 YOLOv7-Tiny 进行轻量化改进,提升了小目标检测 mAP 和推理速度。但是可见光检测在恶劣天气和高曝光环境下精度易下降[4]。为此,部分学者研究红外图像以增强输电线路检测性能。例如,文献[5]通过优化红外图像特征提取网络与锚框比例改进 Faster R-CNN[6]。文献[7]在 YOLOv4-tiny 中引入全局信息聚合与红外图像特征增强融合网络。文献[8]结合实例分割提出了一种基于 Mask R-CNN 的红外图像绝缘子诊断方法。以上方法均实现了对红外图像检测速度与精度的提升。但是红外图像的低分辨率与模糊性仍限制了其在复杂环境下的检测效果[9]。

尽管上述研究取得了显著进展,但单一模态信息的固有局限性则不可避免地导致图像质量下降与特征信息缺失[10]。为此,一些学者尝试通过融合可见光(visible iight, VIS)和红外光(infrared light, IR)两种模态图像信息以应对这些挑战,例如:文献[11]通过结合 SIFT 特征提取与 RANSAC 配准的可见光-红外图像融合方法,可有效提升设备故障的识别与定位精度;文献[12]采用多输入模型和特征级融合的方法结合可见光图像的高分辨率和红外图像的高稳定性,提升电力线路检测的准确性;文献[13]改进了 Fast-SCNN 模型并引入可见光与增强对比度红外图像的四通道堆叠输入,实现了 6.7% 的性能提升。虽然这些方法融合了可见光和红外特征,但是仍然存在以下局限:1)简单的融合策略未能充分挖掘模态间的深层互补性和关联性;2)缺乏自适应融合机制,无法根据不同场景和目标特性动态调整各模态信息的贡献度,可能导致信息冗余或模态冲突;3)对于输电线路部件普遍存在的大尺度变化问题,尤其是在远距离拍摄时小目标特征微弱,现有融合方法可能未能有效增强多尺度特征表达。

针对上述问题,本文提出一种多模态[14]融合的双流[15]输电线路多尺度检测模型。鉴于 YOLO 系列在检测精度与推理速度方面相较于其他单一模态方法表现出更优的综合性能,因此本文在当前较先进的 YOLO 模型的基础上进行改进。主要工作为:1)设计了一个双流主干网络,通过交互特征提取分别捕获可见光图像的颜色与纹理信息,以及红外图像的特征,提高特征表示的丰富性与多样性;2)引入跨模态机制,设计多模态特征交互融合模块 MFIFM (multimodal feature interactive fusion module, MFIFM),动态分配可见光与红外特征权重,增强信息交互能力,同时减少模态差异带来的干扰;3)提出 HRMS Transformer (hybrid residual multi-scale transformer, HRMS Transformer)的多头窗口注意力机制,在双流主干网络中通过层级特征重组和残差网络构建多尺度上下文依赖关系,优化小目标检测效果。

1　方　　法

1.1　整体结构

在本文提出的多模态多尺度目标检测网络整体架构如图 1 所示,主要由双流主干网络(Backbone)、颈部网络(Neck)和检测头(Head)三部分组成。考虑到检测精度、推理速度和部署可行性,本研究选择在成熟且高效的 YOLOv8[16]目标检测框架基础上进行改进。

原始 YOLOv8 采用单流主干网络处理单模态输入。为有效利用可见光和红外图像的互补信息,设计了双流特征提取主干(dual-stream backbone),其中包含两个并行的分支,分别接收对齐后的 VIS 和 IR 图像作为输入,独立提取各自的多尺度特征图。为增强双流骨干网络的全局建模能力与长距离依赖关系建模能力,本文在其结构中嵌入 HMCS Transformer 模块。该模块通过引入多头窗口注意力机制和残差多层感知机,提升局部区域内特征建模能力;同时采用层级特征重组策略,强化高层语义信息与低层细节特征的跨层次交互;并结合残差网络结构,建立多尺度上下文依赖关系,实现局部与全局信息的高效融合,从而提升多尺度小目标的检测性能。提取的可见光与红外特征进一步输入多模态特征交互融合模块,该模块是实现跨模态信息有效整合的核心。该模块基于多尺度上下文注意力机制动态生成模态间交互融合权重,自适应调整融合策略,有效缓解模态差异带来的干扰,提升多模态信息交互效率和检测鲁棒性。

融合后的特征进入颈部网络(Neck),本研究沿用 YOLOv8 中高效的路径聚合网络与特征金字塔网络相结合的结构,通过双流特征金字塔网络(path aggregation network-feature pyramid network, PAN-FPN)结构实现不同尺度特征的有效融合。具体来说,颈部网络结合路径聚合网络(PAN)与特征金字塔网络(FPN),提升了多尺度特征的传递与融合能力,从而增强了对多尺度目标的检测能力。优化后的特征最终送入解耦检测头(decoupled head),在此结构中,目标分类与边界框回归任务被解耦,分别采用

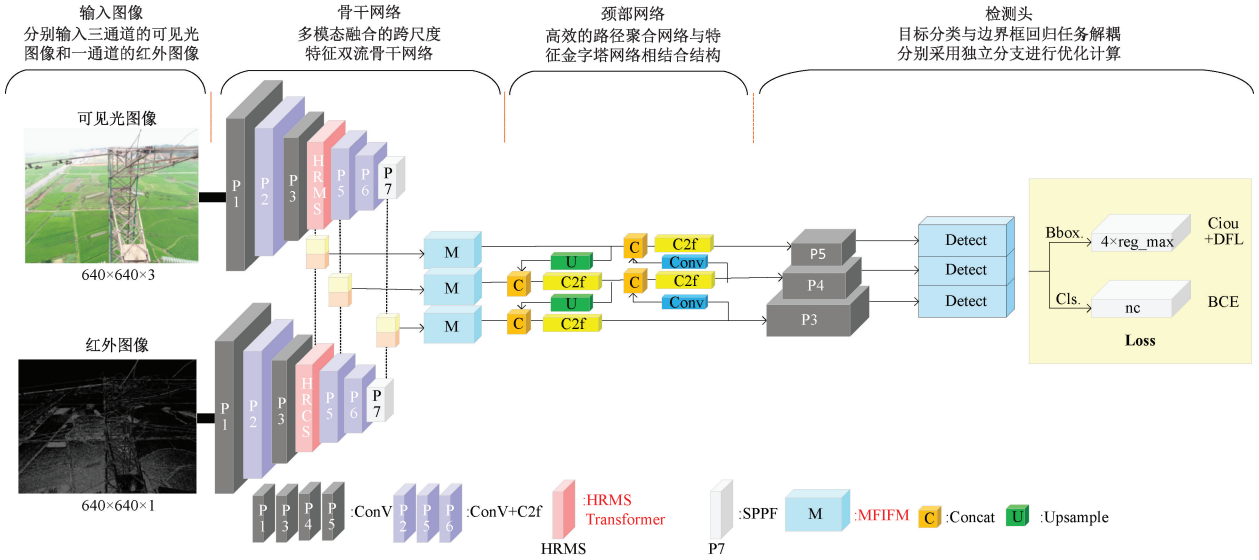


图 1　整体结构

Fig. 1　Overall structure

独立分支进行优化计算，进一步提升了模型的检测精度与鲁棒性。通过这一优化流程，网络能够更精确地处理不同尺度信息，提升在恶劣天气环境下目标检测的表现。

1.2　多模态特征交互融合模块设计

本章节提出了特征交互融合模块，如图 2 所示，用于有效融合可见光与红外特征，提高输电线路关键部位的检测精度。由于可见光和红外图像在光照、纹理等方面存在较大差异，直接进行特征拼接或加权平均可能导致信息损失或干扰。为了解决这一问题，MFIFM 通过动态调整两种模态特征的权重，使网络能够自适应地选取最有利的信息，同时削弱冗余或冲突特征的影响。

MFIFM 主要包括局部特征提取和全局信息整合两个部分。在局部特征提取方面，使用点卷积（pointwise convolution，PWConv），即 1×1 卷积，如图 3 所示，仅在通道维度上进行特征交互，不改变空间分辨率，确保计算高效的同时增强局部信息表达。随后，本研究通过瓶颈结构进一步精炼局部特征，以减少计算开销。在全局信息整合方面，MFIFM 计算整个特征图范围内的重要性权重，并自适应地分配给可见光和红外特征，使模型能够在不同场景下动态调整对不同模态的关注程度。

融合过程采用逐像素加权策略，计算得到的融合权重 $M(X,Y)$ 介于 0～1，确保融合后的特征能够在可见光和红外特征之间灵活切换。例如，在光照充足的环境下，网络可能更依赖可见光特征，而在低光或复杂天气条件下，红外特征的权重会相对提高。最终，MFIFM 生成的融合特征能够有效增强目标区域的特征表达，减少模态不一致带来的干扰，从而提升输电线路关键部位的检测性能。

通过瓶颈结构计算局部特征上下文 $L(X)\in R^{C\times H\times W}$，

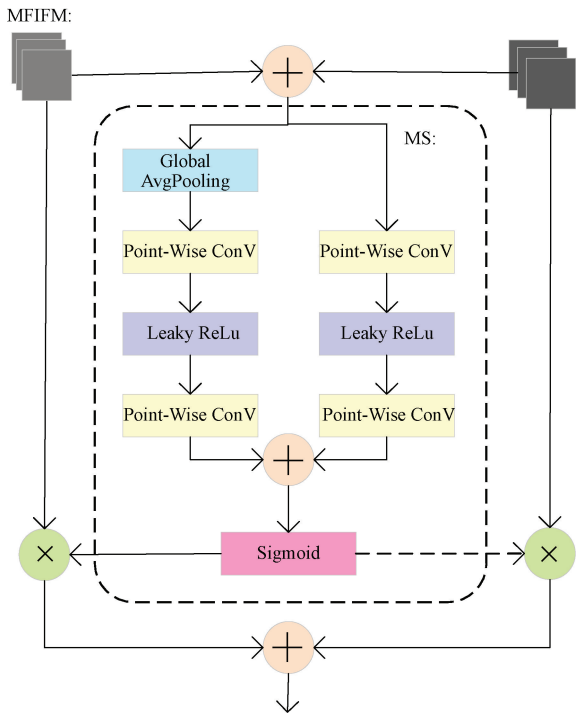


图 2　MFIFM 模块设计

Fig. 2　Design of the MFIFM module

其计算方式如下：

$$L(X)=B(PWConv2(\varphi(B(PWConv1(X))))) \quad (1)$$

其中，$PWConv1$ 和 $PWConv2$ 的卷积核尺寸分别为 $\frac{C}{r}\times C\times 1\times 1$ 和 $C\times\frac{C}{r}\times 1\times 1$。值得注意的是，$L(X)$ 与输入特征具有相同的形状，能够有效保留低层细节信息，增强局部特征表达。

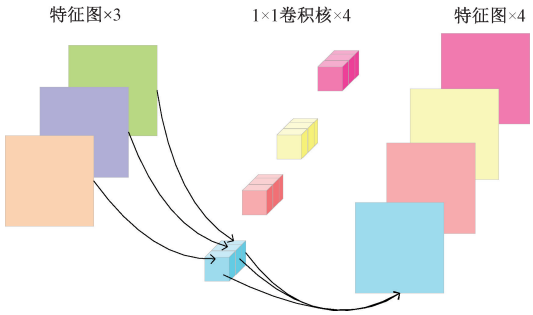


图 3　点卷积

Fig. 3　PWConv

在获得全局特征上下文 $g(\boldsymbol{X})$ 和局部特征上下文 $L(X)$ 后,MS(model structure)模块生成的精炼特征 $\boldsymbol{X}_0 \in \boldsymbol{R}^{C\times H\times W}$ 可以通过以下公式计算:

$$\boldsymbol{X}_0 = \boldsymbol{X} \otimes MS(\boldsymbol{X}) = \partial(L(\boldsymbol{X}) \oplus g(\boldsymbol{X})) \tag{2}$$

其中,$MS(\boldsymbol{X})$ 表示由 MS 组件生成的注意力权重,符号 $\oplus$ 表示广播加法,$\otimes$ 表示逐元素乘法。该机制能够自适应地调整可见光和红外信息的融合比例,以增强跨模态特征的互补性。

在多模态特征融合过程中,设可见光特征 $\boldsymbol{X} \in \boldsymbol{R}^{C\times H\times W}$ 和红外特征 $\boldsymbol{Y} \in \boldsymbol{R}^{C\times H\times W}$ 分别代表代表可见光和红外光的特征图,基于 MS 结构,MFIFM 特征融合可以表示为:

$$\boldsymbol{Z} = MS(\boldsymbol{X} \oplus \boldsymbol{Y}) \otimes \boldsymbol{Y} + (1 - MS(\boldsymbol{X} \oplus \boldsymbol{Y})) \otimes \boldsymbol{Y} \tag{3}$$

其中,$\boldsymbol{Z}$ 是融合后的特征,$MS(\boldsymbol{X} \oplus \boldsymbol{Y})$ 为 MS 组件生成的动态权重。为了增强信息互补性,本模块采用逐元素加法进行初始特征融合。虚线表示 $1 - MS(\boldsymbol{X} \oplus \boldsymbol{Y})$,用于动态调整模态信息的融合比例。值得注意的是,融合权重 $MS(\boldsymbol{X} \oplus \boldsymbol{Y})$ 和 $1 - MS(\boldsymbol{X} \oplus \boldsymbol{Y})$ 均介于 0 和 1 之间,这使得网络能够在 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间进行软选择或加权平均。

该 MFIFM 模块通过局部和全局特征融合,自适应调整不同模态特征的权重,有效缓解模态差异带来的信息不一致问题。点卷积用于提取局部特征信息,而 MS 结构则自适应生成融合权重,使模型能够在复杂环境下精准提取关键特征。

1.3　混合残差多尺度 Transformer 模块设计

本文在 Swin Transformer[17] 架构基础上提出了一种多尺度混合注意力模块(multi-scale hybrid attention, MSHA),作为 HRMS Transformer(如图 4 所示)的核心组成单元嵌入至双流骨干网络中。

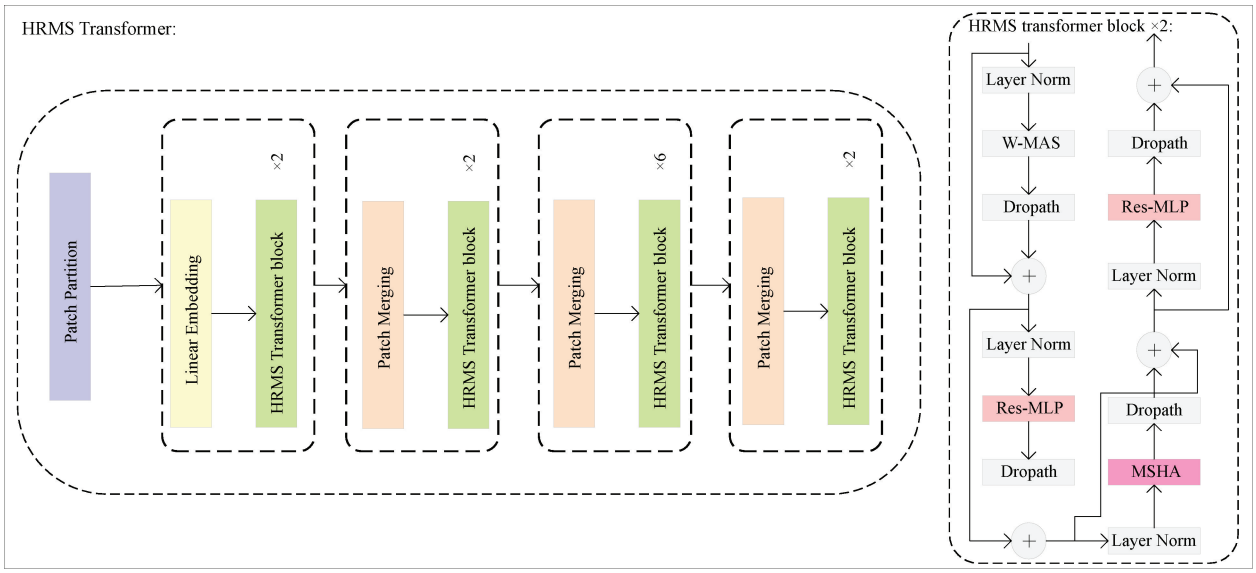


图 4　HRMS Transformer 设计

Fig. 4　Design of the HRMS Transformer

MSHA 通过融合窗口多头自注意力(window-based multi-head self attention,W-MSA)与移位窗口多头自注意力(shifted window multihead self-attention,SW-MSA),并引入横向与纵向长矩形窗口划分策略,显著增强了多尺度特征交互能力。在传统 SW-MSA 提升局部窗口信息流动性的基础上,MSHA 进一步通过 3 种窗口处理机制实现多尺度特征的有效建模,如图 5(a)所示。

1)局部窗口注意力(W-MSA)

W-MSA 采用固定大小的方形窗口划分特征图,通过标准的多头自注意力机制建模局部区域内的语义依赖,具有良好的小目标感知能力。该模块对于输电线路中如防震锤等细粒部件的识别具有重要作用。

2)条状窗口注意力(S-MSA)

条状窗口注意力机制旨在扩展模型的感受野以建模长距离依赖关系。具体而言,H-S-MSA,如图 5(b)所示,将输入特征图划分为多个横向长条窗口,而 V-S-MSA 则按纵向长条窗口划分。以 H-Strip-MSA 为例,其计算过程如下:

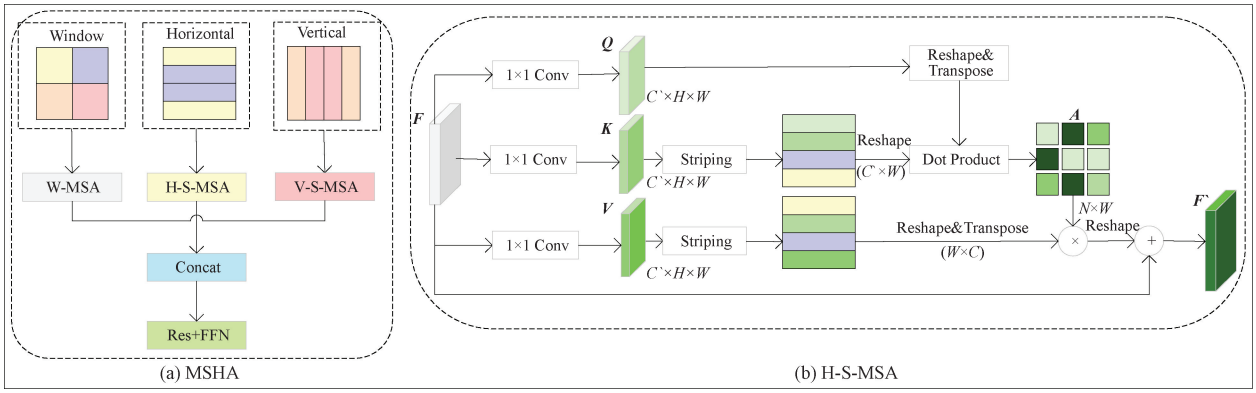


图 5　MSHA 模块设计

Fig. 5　Design of the MSHA module

设输入特征图为 $\boldsymbol{X}\in \mathrm{R}^{C\times H\times W}$，经过线性映射后获得查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)、值($\boldsymbol{V}$)向量：

$$\boldsymbol{Q}=\boldsymbol{X}W_{Q},\boldsymbol{K}=\boldsymbol{X}W_{K},\boldsymbol{V}=\boldsymbol{X}W_{V} \tag{4}$$

其中，$W_{Q},W_{K},W_{V}\in \mathrm{R}^{C\times C}$ 为可学习参数，C 为每个注意力头的维度。

将特征图划分为 N_S 个横向窗口，每个窗口大小为 $h_s\times W$，则注意力机制可表示为：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{5}$$

其中，d 为缩放因子，最终将所有窗口拼接恢复为原始尺寸。同理，V-S-MSA 沿列划分窗口，操作过程一致。

3)融合与残差连接

上述 3 类注意力机制的输出分别记为：($\hat{X}_W$，$\hat{X}_H$，$\hat{X}_V\in \mathrm{R}^{H\times W\times C'}$)，将三路注意力特征进行拼接，并通过 1×1 卷积映射回原始通道数：

$$\begin{aligned}&F=\mathrm{Concat}(\hat{X}_W,\hat{X}_H,\hat{X}_V)\in \mathrm{R}^{H\times W\times 3C'}\\&F'=Conv_{1\times 1}(F)\end{aligned} \tag{6}$$

为了提升非线性建模能力，MSHA 还包括前馈神经网络(feed-forward network，FFN)和残差连接，整体表达为：

$$Y=X+FFN(F') \tag{7}$$

通过引入三类注意力机制的协同建模，MSHA 在提升小目标检测精度的同时，有效捕捉了多尺度上下文信息和结构方向性特征，增强了模型对复杂场景中目标关系的理解与表达能力。

此外，为进一步增强 HRMS Transformer 架构在复杂场景中的深层次特征表达能力，本文设计引入了残差[18]多层感知机(residual multi-layer perceptron，Res-MLP)模块，如图 6(b)所示。传统 MLP 如图 6(a)所示，在深层网络中易出现梯度消失、非线性建模能力有限等问题，通过残差连接与深层 MLP 结构的融合，有效提升了特征变换能力与模型稳定性。一方面，残差连接机制允许深层特征在网络中跨层流动，缓解梯度消失，提升模型训练效率；另一方面，深层感知机结构增强了模型对复杂非线性特征关系的表达能力，提升其对不同尺度目标与复杂背景的适应性；同时，通过特征变换过程的优化，Res-MLP 显著提升了模型在面对多样化分布及噪声干扰数据时的泛化能力，强化了模型在真实输电线路场景下的检测稳定性。

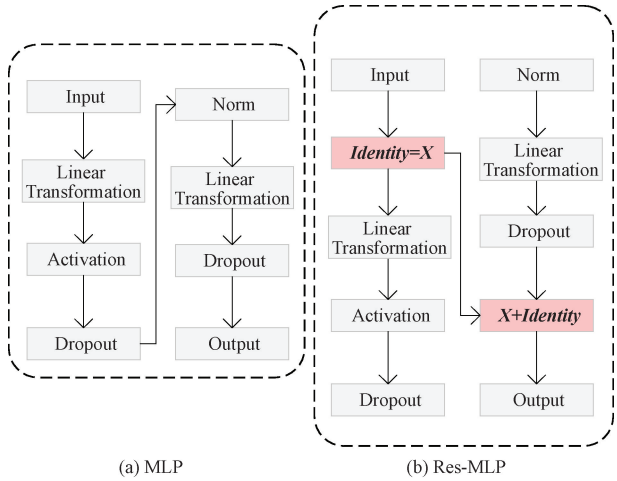


图 6　MLP 与 Res-MLP 对比

Fig. 6　Comparison between MLP and Res-MLP

改进后的 HRMS Transformer 结构结合了 MSHA 和 Res-MLP 两个关键模块，其计算过程可以表示为：

$$\begin{cases}\hat{z}_l=W-MSA(\mathrm{LN}(z_{l-1}))+z_{l-1}\\z_l=\mathrm{Res}-\mathrm{MLP}(\mathrm{LN}(\hat{z}_l))+\hat{z}_l\\{}_l\hat{z}_{l+1\,l+1}=\mathrm{MSHA}(\mathrm{LN}(z_l))+z_l\\z_{l+1}=\mathrm{Res}-\mathrm{MLP}(\mathrm{LN}(\hat{z}_{l+1}))+\hat{z}_{l+1}\end{cases} \tag{8}$$

通过融合 MSHA 模块与 Res-MLP 结构，本文提出的 HRMS Transformer 在保持局部特征提取能力的同时，能够有效增强全局信息交互能力。MSHA 模块提升了多尺度目标的适应性，而 Res-MLP 模块增强了特征转换的稳

定性与鲁棒性。

2　实验结果

2.1　数据集构建

本文所使用的数据集来源于某电网公司无人机拍摄的真实输电线路图像。该无人机拍摄的是同一地点的可见光图像与数据预处理后的红外光线图像,获取图像之后进行模态之间的对齐,使两种模态图像尺寸均为 640 pixel×640 pixel 大小。

该数据集由电网公司专业人员关键部位进行了认定与标注。其中包含了各种角度拍摄的输电线路关键部位的图片,场景覆盖如强光,暗光、薄雾极端天气或复杂背景,较好地模拟了实际检测环境。根据关键部位的类型,数据集被自定义划分为两类:绝缘子(insulator)、防震锤(hammer)。图像标注采用 LabelImg 标注工具,总计图像数量为 9 192 张。其中,训练集与测试集按照 7∶2∶1的比例划分,可见光图像与红外图像的训练集各包含 3 217 张图像,测试集各包含 919 张图像,验证集各包含 460 张图像,以保证多模态数据在模型训练与评估过程中的一致性与对齐性。各类型关键部位的实例数量占比如图 7 所示。

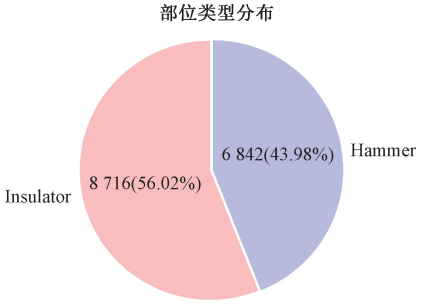


图 7　数据集划分

Fig. 7　Dataset partitioning

2.2　实验环境及参数设置

本研究在 Ubuntu 18.04 操作系统上进行,采用 Python 3.8 作为主要编程语言,深度学习框架选用 PyTorch 1.8.1,并结合 CUDA 11.1 与 cuDNN 8.0.5 实现 GPU 加速计算。硬件环境方面,配置包括双 RTX 2080 Ti 显卡(共 22 GB 显存)和 Intel Xeon Platinum 8255C 处理器,整体实验环境具备良好的并行计算与深度学习训练能力。

为保证模型有效训练,本文合理设置关键训练参数并通过预实验验证。基于 YOLOv8n 配置,结合中等规模数据集与轻量网络结构,初始及最终学习率均为 0.01,实现了收敛速度与精度的良好平衡。Batch size 设为 16,主要考虑到双流输入(可见光与红外图像)对显存的较大占用,同时兼顾训练的稳定性和 GPU 资源的有效利用。训练轮数(Epochs)设置为 200,确保模型充分学习数据特征而不过拟合。输入图像尺寸为 640 pixel×640 pixel,兼顾分辨率与计算效率,有利于提升检测精度。数据加载时采用 16 个 Workers,提升数据预处理和加载速度,缩短训练时间。Momentum 设置为 0.937,沿用 YOLOv8 推荐值,有助于加速收敛并避免震荡。Weight decay 设为 0.000 5,用于防止过拟合,提升模型泛化能力。Warmup epochs 设置为 3,缓解训练初期梯度波动,稳定训练过程,防止模型陷入次优解。同时,Warmup momentum 设为 0.8,在预热阶段控制动量,有效平滑参数更新。实验在经典配置基础上结合模型结构和数据分布特征进行适度调整,确保训练过程收敛稳定、性能可靠。具体参数配置如表 1 所示。

表 1　模型参数配置

Table 1　Model parameter configuration

参数	值
Epochs	200
Batch size	16
Image Size	640×640
Workers	16
Ir0	0.01
Irf	0.01
Momentum	0.937
Weight decay	0.000 5
Warmup epochs	3
Warmup momentum	0.8

2.3　评价指标

在本文中,模型的性能评估主要依赖于以下几个常用的评价指标:

1)精度(precision,P):精度衡量的是模型在预测为正类的样本中,实际为正类的比例。它反映了模型在正类预测上的准确性。公式如式(9)所示。

$$P = \frac{TP}{TP + FP} \tag{9}$$

其中,TP 为真阳性,FP 为假阳性。精度高意味着模型在预测为正样本时,错误率较低。

2)mAP@50(mean average precision at IoU=0.5):在 IoU 阈值为 0.5 时,计算各类别的平均精度(AP),并对所有类别的 AP 取平均。公式如式(10)所示。

$$mAP@50 = \frac{1}{C}\sum_{K=1}^{C} AP_K \tag{10}$$

其中,C 为类别数,AP_K 为类别 K 的平均精度。

3)mAP@50:95(mean average precision at IoU=[0.5∶0.95]):计算在多个 IoU 阈值下(从 0.5～0.95,每隔 0.05)得到的平均精度。公式如式(11)所示。

$$mAP@50{:}95 = \frac{1}{C}\sum_{K=1}^{C}\frac{1}{10}\sum_{i=1}^{10} AP_K(IoU_i) \tag{11}$$

4)召回率(recall, R):衡量模型检测到的正类目标占所有实际正类目标的比例。公式如式(12)所示。

$$TP = \frac{TP}{TP + FN} \tag{12}$$

其中，TP 为真阳性，FN 为假阴性。较高的召回率意味着模型能够检测到更多实际目标。

2.4　对比试验

为全面评估所提出模型的检测性能，本文在统一的数据集上与多种主流目标检测算法进行了对比实验。对比方法涵盖了单模态检测模型，包括 YOLOv5[19]、YOLOv8[16]、RT-DETR[20]、Faster R-CNN[6]、DETR、Mamba-YOLO[21]，以及多模态检测模型，如 cm-SSFT[22]、AMNet[23]和 RISNe[24]。上述所有模型均在相同的训练集上进行训练，训练轮数统一为 200 轮，且使用一致的评价指标体系以确保比较的公平性和可重复性。各方法的具体实验结果汇总如表 2 所示。

表 2　对比实验

Table 2　Comparative experiments　%

模型名称	精度 P	mAP@50	mAP@50:95	召回率 R
Faster RCNN	80.25	75.92	54.19	69.38
YOLOv5	88.93	85.74	60.08	75.81
DETR	89.16	86.35	61.86	77.54
RT-Detr	93.69	90.47	63.89	78.43
Mamba-YOLO	95.46	93.42	68.04	82.96
cm-SSFT	95.27	92.57	66.12	80.96
AMNet	95.32	92.90	67.34	81.91
RISNet	95.38	93.30	67.96	82.87
YOLOv8	93.49	90.51	64.97	80.13
本文模型	97.68	95.86	69.45	84.27

本文模型在 mAP@50、mAP@50:95 和召回率等关键指标上均超越了现有多种先进检测算法，取得了优异的整体性能。在 mAP@50 上，本文模型达到 95.86%，相较于单模态方法 YOLOv8 和 Mamba-YOLO 分别提升了 5.35%和 2.44%，充分验证了所提出方法在整体检测准确率方面的优势。这一性能提升得益于双流结构对可见光与红外图像的独立特征提取机制，为后续信息融合提供了更丰富、更互补的语义信息。

MFIFM 模块在本文模型中的作用至关重要，是性能提升的核心驱动力。该模块通过动态特征权重调整机制，实现了对可见光和红外模态的自适应融合，有效抑制了模态差异带来的干扰问题。MFIFM 不仅结合了局部与全局上下文信息，还采用多尺度融合策略，增强了特征的表达能力和判别性。在 mAP@50:95 指标上，本文模型达到 69.45%，较 AMNet 和 YOLOv8 分别提升了 2.11%和 4.48%，说明模型在复杂背景下、尤其是小目标检测任务中展现出更强的稳定性与鲁棒性。

HRMS Transformer 模块进一步强化了模型的多尺度建模能力，尤其在小目标检测方面起到关键作用。通过引入分层注意力机制与上下文关联增强策略，HRMS 模块能有效捕捉不同尺度下的显著特征区域，提升模型在边界模糊、目标重叠等复杂场景下的检测能力。在召回率 R 上，本文模型达到 84.27%，相比 RISNet 提升 1.4%，显著降低了漏检率，体现出模型在真实场景中对难检目标具备更强的感知能力。

2.5　消融实验

为了进一步验证本文模型性能的提升源于所提出的 HRMS Transformer 模块，MFIFM 模块以及双流骨干，并评估 HRMS Transformer 在小目标检测中的显著优势，本文设计了两组消融实验。

1）消融实验一：整体数据集

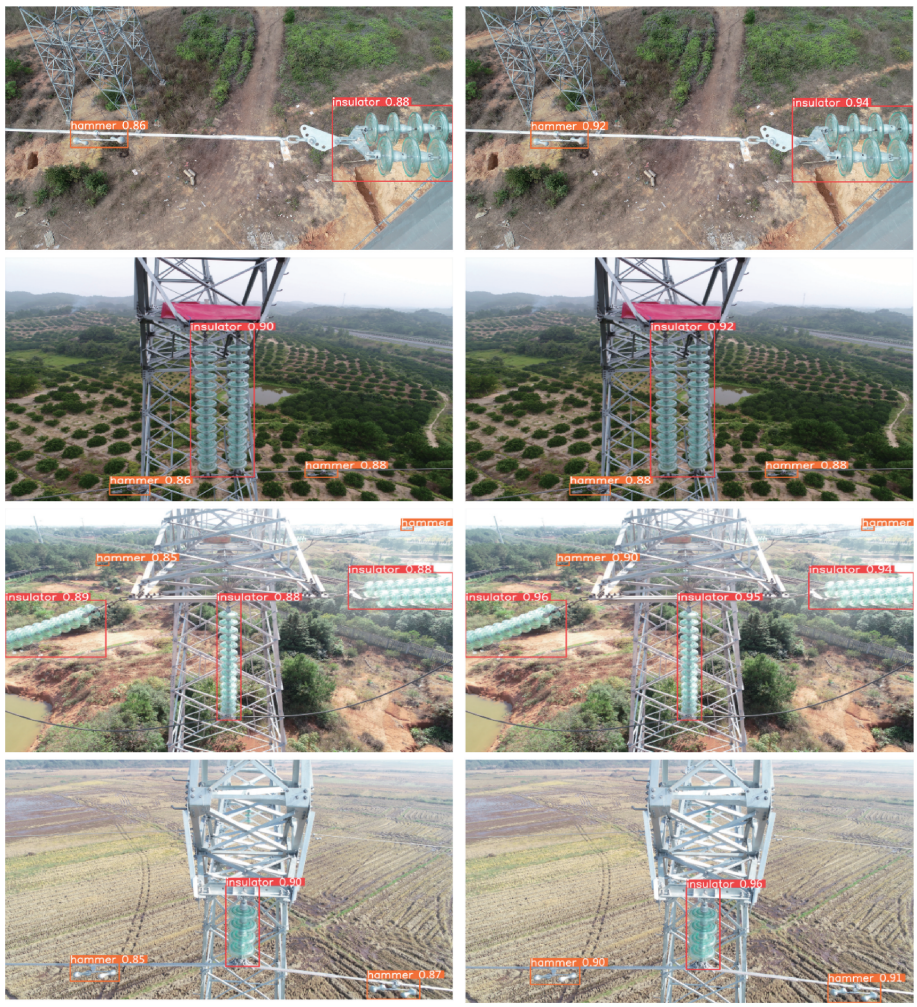
第一组消融实验在整体数据集上进行，分析各模块的组合顺序对模型性能的影响，并探讨其间可能存在的冗余或互补关系，以全面验证模块之间的协同效应和性能提升来源。实验结果如表 3 所示，实验效果图如图 8 所示。

表 3　消融实验一结果

Table 3　Results of ablation experiment Ⅰ　%

实验	HRMS Transformer	MFIFM	双流骨干网络	mAP		召回率
				@50	@5095	R
1	×	×	×	90.51	64.97	80.13
2	√	×	×	92.08	66.45	82.61
3	×	√	×	93.46	66.92	83.04
4	√	√	×	93.92	67.18	83.36
5	×	×	√	94.16	67.73	83.58
6	√	×	√	94.93	68.26	83.72
7	×	√	√	95.27	68.49	83.95
8	√	√	√	95.86	69.45	84.27

在第一组消融实验中，实验 1 作为基准模型，未引入任何改进模块，mAP@50 为 90.51%，mAP@50:95 为 64.97%，召回率为 80.13%。实验 2 和实验 3 分别在基准模型基础上引入 HRMS Transformer 模块和 MFIFM 模块，mAP@50 分别提升至 92.08%和 93.46%，表明前者在增强小目标特征表达能力方面具有明显优势，后者则在提升模态信息质量方面效果显著。实验 4 同时引入 HRMS 与 MFIFM 模块但未使用双流骨干网络，mAP@50 提升至 93.92%，mAP@50:95 为 67.18%，召回率为 83.36%，性能优于单一模块组合，验证了两模块在缺乏模态解耦结构条件下依然能够协同工作，具备一定互补性。实验 5 仅引入双流骨干网络，mAP@50 提升至 94.16%，说明对可见光与红外图像的解耦式特征提取为模型性能的提升奠定了坚实基础。在此基础上，分别加入 HRMS 模块（实验 6）和 MFIFM 模块（实验 7），mAP@50 进一步提升至 94.93%和 95.27%，mAP@50:95 分别提升至 68.26%和



(a) YOLOv8n　　(b) 本文模型

(b) Model in this paper

图 8　实验结果对比图

Fig. 8　Experimental comparison diagram

68.49%,召回率有小幅提升,表明两模块在双流结构下均具备良好的独立增益能力。进一步分析可知,HRMS 模块主要强化单模态内部的上下文建模与多尺度感知能力,适用于复杂背景下的小目标检测;而 MFIFM 模块通过引入局部与全局注意力机制及模态动态加权策略,实现多模态特征间的有效互补与信息整合。二者在处理层次、作用机制和功能定位上具有明确区分,表现出较强的互补性,未见显著冗余。实验 8 将三者联合引入,模型性能达到最优,mAP@50 为 95.86%,mAP@50:95 为 69.45%,召回率为 84.27%。相比任意两模块组合,整体性能仍有提升,且更为稳定,进一步验证了三模块在多模态建模任务中具有良好的协同作用,能够有效提升复杂场景下模型的检测性能与鲁棒性。

2) 消融实验二:小目标数据集

第二组消融实验则在整体数据集中筛选出小目标防震锤样本进行,旨在重点分析 HRMS Transformer 对小目标检测能力的影响,实验结果如表 4 所示。

表 4　消融实验二结果

Table 4　Results of ablation experiment II　%

模型名称	mAP@50	mAP@50:95	召回率 R
YOLOv8n	86.65	60.34	77.92
YOLOv8+HRMS Transformer	90.53	62.72	80.37

在第二组消融实验中,聚焦于小目标防震锤样本,重点分析 HRMS Transformer 对小目标检测能力的提升。实验结果显示,YOLOv8n(未使用 HRMS Transformer)的 mAP@50 为 86.65%,mAP@50:95 为 60.34%,召回率为 77.92%。而引入 HRMS Transformer 后,YOLOv8+HRMS Transformer 的 mAP@50 提升至 90.53%,mAP@50:95 为 62.72%,召回率为 80.37%。这一结果表明,

HRMS Transformer 在处理小目标时具有显著的性能提升，特别是在小目标的检测精度和召回率方面。HRMS Transformer 通过其优化的小目标特征提取和增强的跨尺度特征交互，增强了模型在低质量、小尺寸目标检测中的能力，从而有效减少了漏检和假阳性。

2.6　结果可视化分析

为进一步验证本文所提方法的有效性，并进一步提升模型的可解释性，本文选择在 MFIFM 特征融合层之后应用 Grad-CAM[25]（gradient-weighted class activation mapping，Grad-CAM）方法进行可视化分析。Grad-CAM 能够直观展示模型在处理多模态图像特征融合及小目标检测任务中的关注区域，图中的深色区域表示模型在该部分区域的关注程度较高，而浅色区域则表明模型对此部分的关注较少。具体的可视化结果如图 9 所示。

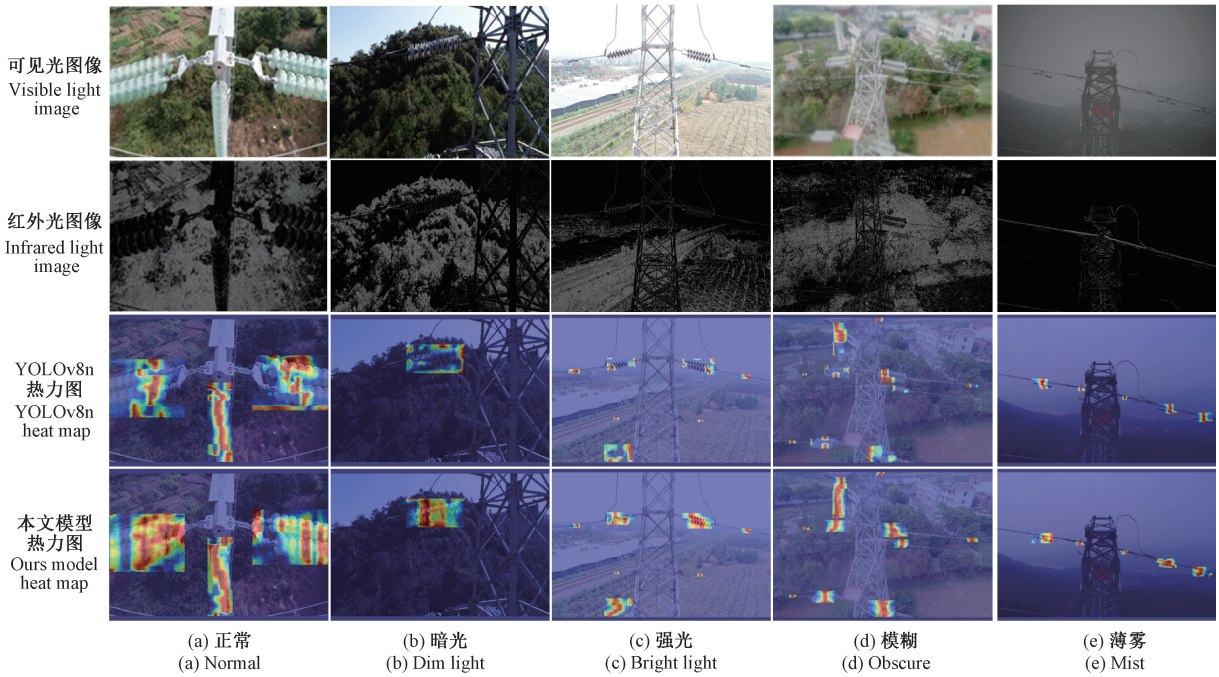


图 9　可视化结果对比

Fig. 9　Comparison of visualization results

由图 9 可视化结果可知，本文提出的模型在不同光照和环境条件下（正常、暗光、强光、模糊、薄雾），相较于单模态方法 YOLOv8n，能够更为精准地聚焦于关键目标区域。在正常光照条件下，本文模型对目标的关注区域分布更为集中且合理，充分体现了多模态信息融合在目标特征提取中的优势。在暗光环境下，YOLOv8n 出现了关注区域分散的情况，而本文模型仍能稳定锁定目标，表明其在低光照条件下具备更好的鲁棒性。在强光场景中，本文模型受光线干扰的影响较小，关注区域未出现明显偏移或误判，能够准确捕捉目标特征，而 YOLOv8n 的部分关注区域则出现错乱。在模糊图像条件下，本文模型通过自适应调整关注权重，有效突出目标的轮廓和关键特征，而 YOLOv8n 则存在关注区域泛化的问题，难以精准定位目标。在薄雾条件下，本文模型对目标的关注程度明显高于 YOLOv8n，能够更清晰地区分目标与背景，显著降低环境干扰对检测任务的影响。

综上所述，通过 Grad-CAM 可视化分析，进一步验证了本文所提方法在多模态图像特征融合及小目标检测任务中的有效性与优越性，显著提升了模型在复杂环境下的可解释性与检测性能。

3　结　　论

针对单一模态目标检测在输电线路关键部位检测中存在的精度下降与漏检率高等问题，本文提出了一种多模态融合的多尺度目标检测网络。构建双流主干网络能够充分利用可见光图像在颜色与细节信息上的优点，以及红外图像在高曝光环境下的成像稳定性，从而弥补了单一模态在多模态信息交互与特征表达上的局限性。设计多模态特征交互融合模块（MFIFM）通过引入局部与全局上下文建模机制，自适应调整不同模态特征的融合权重，显著增强了可见光与红外图像之间的特征互补性与信息协同表达能力。在此基础上，将所提出的混合残差多尺度 Transformer 模块（HRMS Transformer）嵌入双流主干架构，通过多尺度混合注意力机制提升小目标检测能力，同时结合残差结构增强特征的非线性建模与长距离依赖关系建模能力，有效强化复杂场景中的上下文理解与精细感知。实验结果表明，所提出的多模态目标检测网络在检测精度与漏检率方面均显著优于单一模态检测网络，有效提

升了检测性能。后续研究将进一步扩展输电线路的可见光与红外图像数据集规模,并引入更先进的多模态特征交互策略以及对于模型轻量化的考量,以持续优化模型的检测性能与泛化能力。

参考文献

[1] HE M, LI Q, DENG X, et al. MFI-YOLO: Multi-fault insulator detection based on an improved YOLOv8[J]. IEEE Transactions on Power Delivery, 2024, 39(1): 168-179.

[2] 李帷韬, 侯建平, 张倩, 等. 基于强化学习和 Transformer 的输电线路缺陷智能检测方法研究[J]. 高电压技术, 2023, 49(8): 3373-3384.
LI W T, HOU J P, ZHANG Q, et al. Research on intelligent detection method of transmission line defects based on reinforcement learning and Transformer[J]. High Voltage Engineering, 2023, 49(8): 3373-3384.

[3] WANG Q, ZHANG ZH, CHEN Q, et al. Lightweight transmission line fault detection method based on leaner YOLOv7-Tiny[J]. Sensors, 2024, 24(2): 565.

[4] 陈广秋, 温奇璋, 尹文卿, 等. 用于红外与可见光图像融合的注意力残差密集融合网络[J]. 电子测量与仪器学报, 2023, 37(8): 182-193.
CHEN G Q, WEN Q ZH, YIN W Q, et al. Attention residual dense fusion network for infrared and visible image fusion[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(8): 182-193.

[5] OU J, WANG J, XUE J, et al. Infrared image target detection of substation electrical equipment using an improved Faster R-CNN[J]. IEEE Transactions on Power Delivery, 2022, 38(1): 387-396.

[6] GIRSHICK R. Fast R-CNN[C]. IEEE International Conference on Computer Vision, 2015: 1440-1448.

[7] LI J, XU Y, NIE K, et al. PEDNet: A lightweight detection network of power equipment in infrared image based on YOLOv4-tiny[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-12.

[8] WANG B, DONG M, REN M, et al. Automatic fault diagnosis of infrared insulator images based on image instance segmentation and temperature analysis[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(8): 5345-5355.

[9] JIN L, TIAN ZH, AI J, et al. Condition evaluation of the contaminated insulators by visible light images assisted with infrared information [J]. IEEE Transactions on Instrumentation and Measurement, 2018, 67(6): 1349-1358.

[10] 许光宇, 陈浩宝, 张杰. 多路径生成对抗网络的红外与可见光图像融合[J]. 国外电子测量技术, 2024, 43(3): 18-27.
XU G Z, CHEN H B, ZHANG J. Multi-path generative adversarial network for infrared and visible image fusion[J]. Foreign Electronic Measurement Technology, 2024, 43(3): 18-27.

[11] CHEN SH, LUO Y, YIN J, et al. Application of visible light-infrared image fusion technology in power system fault detection[C]. 2023 Asia Conference on Computer Vision, Image Processing and Pattern Recognition. New York, NY, USA: Association for Computing Machinery, 2023: 8.

[12] ABOALIA H, HUSSEIN S, MAHMOUD A. Enhancing power lines detection using deep learning and feature-level fusion of infrared and visible light images [J]. Arabian Journal for Science and Engineering, 2025, 50(2): 987-999.

[13] CHOI H, KOO G, KIM B J, et al. Real-time power line detection network using visible light and infrared images[C]. 2019 International Conference on Image and Vision Computing New Zealand(IVCNZ). IEEE, 2019: 1-6.

[14] 邢致恺,何怡刚,姚其新. 基于多模态信息融合的变压器在线故障诊断方法[J]. 电子测量与仪器学报, 2024,38(9):95-103.
XING ZH K, HE Y G, YAO Q X. Transformer online fault diagnosis method based on multimodal information fusion [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(9): 95-103.

[15] 葛荣泽,武一. 基于阶段特征融合的图像融合行人检测[J]. 电子测量技术, 2024, 47(24): 103-109.
GE R Z, WU Y. Pedestrian detection in image fusion based on stage feature fusion [J]. Electronic Measurement Technology, 2024, 47(24): 103-109.

[16] SOHAN M, SAI RAM T, RAMI REDDY C V. A review on YOLOv8 and its advancements [C]. International Conference on Data Intelligence and Cognitive Informatics. Singapore: Springer, 2024: 529-545.

[17] LIU ZH, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.

[18] HE K, ZHANG X, REN SH, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[19] WU W, LIU H, LI L, et al. Application of local fully

convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image[J]. PLoS One, 2021, 16(10): e0259283.

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

[21] WANG ZH, LI CH, XU H, et al. Mamba YOLO: A simple baseline for object detection with state space model[J]. AAAI Conference on Artificial Intelligence, 2025, 39(8): 8205-8213.

[22] LU Y, WU Y, LIU B, et al. Cross-modality person re-identification with shared-specific feature transfer [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 13379-13389.

[23] ZHANG T, HE X, JIAO Q, et al. AMNet: Learning to align multi-modality for RGB-T tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 7386-7400.

[24] WANG Q, CHI Y, SHEN T, et al. Improving RGB-Infrared object detection by reducing cross-modality redundancy[J]. Remote Sensing, 2022, 14(9): 2020.

[25] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. IEEE International Conference on Computer Vision, 2017: 618-626.

作者简介

周景(通信作者),博士,副教授,硕士生导师,主要研究方向为人工智能、机器学习和电力大数据分析。

E-mail:zhoujing108@ncepu.edu.cn

赵毅,硕士研究生,主要研究方向为计算机视觉与电力部件检测。

E-mail:zhaoyi@ncepu.edu.cn

刘心,硕士研究生,主要研究方向为计算机视觉与电力故障检测。

E-mail:liuxin0710@ncepu.edu.cn