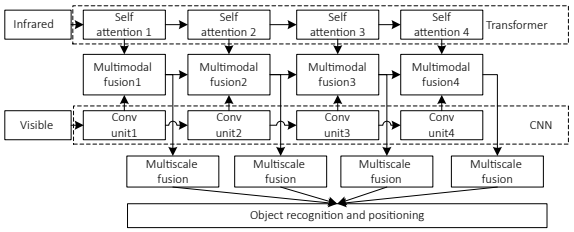
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 2 月, 第 53 卷, 第 2 期

融合 CNN 与 Transformer 网络的多模态目标识别定位研究



游春芝[1*]，吴永飞[2]，王　剑[1]

[1]山西医药学院基础医学部，山西 吕梁 032200；

[2]太原理工大学计算机科学与技术学院，山西 晋中 030600

摘要：为克服单模态目标识别定位方法在复杂环境下的固有局限，提出了一种基于 CNN-Transformer 框架的红外-可见光多模态目标识别定位网络。网络利用分组卷积单元构建 CNN 分支从局部角度挖掘可见光目标信息，同时以特征聚合的自注意力机制构建 Transformer 分支从全局角度提取红外目标特征。其次，设计多模态融合结构，以自适应加权策略实现特征互补。最后，引入多尺度特征动态调节机制，从空间与通道角度缓解特征冲突，为增强网络的尺度不变性。实验结果表明：所提网络能够充分挖掘并融合不同模态目标特征，与同类型网络相比，该网络也体现出更高的精确性和鲁棒性。

关键词：目标识别定位；红外可见光；CNN-Transformer；多模态互补；多尺度融合

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250210 | CSTR: 32245.14.oee.2026.250210 | 中图分类号：TP391.41 | 文献标识码：A

游春芝，吴永飞，王剑. 融合 CNN 与 Transformer 网络的多模态目标识别定位研究 [J]. 光电工程，2026，**53**(2): 250210

You C Z, Wu Y F, Wang J. Research on multimodal object recognition and localization integrating CNN and transformer networks[J]. *Opto-Electron Eng*, 2026, **53**(2): 250210

1　引　言

随着人工智能与计算机视觉技术的持续突破，目标识别定位作为智能感知系统的关键技术[1]，被广泛应用于智能安防、无人驾驶、工业检测和智慧医疗等领域[2-3]。传统目标识别定位方法大多依赖可见光图像数据进行检测[4]，该方式在光照充足、背景清晰的环境下表现良好，但在光照不均、雾霾或目标遮挡等复杂条件下，可见光成像易受干扰，进而导致目标特征模糊甚至丢失，使得检测性能显著下降。相比之下，红外成像通过捕捉物体自身辐射的热信息，具备光照不敏感、烟雾穿透性强等优势，适用于低照度或恶劣天气场景，但其分辨率较低、纹理和颜色信息缺乏，难以实现精细分类与精确定位[5]。可见，单一模态在实际应用中均存在固有局限，难以满足全天候、高鲁棒性的感知需求[6]。因此，如何将可见光图像的细节特征与红外图像的抗干扰能力充分融合，使目标在保持空间结构清晰的同时增强其与背景的可区分性，对于提升复杂环境下目标识别的准确性、推动全天候目标感知技术的发展都具有重要意义。

目前，针对红外-可见光多模态目标识别定位研究主要以深度学习技术为主，而生成式神经网络作为深度学习方向的研究热点，尽管在多模态特征融合以及挖掘深

收稿日期：2025-07-19；修回日期：2025-10-09；录用日期：2025-10-09；网络出版日期：2026-02-13

基金项目：山西省教育厅项目 (J20241687)；山西医科大学汾阳学院科研项目 (2022C21, 2022A02)

*通信作者：游春芝，919839697@qq.com

版权所有©2026 中国科学院光电技术研究所

层潜在特征等方面存在一定潜力[7-8]，但模型存在复杂度高、推理速度慢等局限，难以满足实时性要求高的场景，使得当下目标识别定位主流方法仍以 CNN 或 Transformer 网络为基础框架[9]。Peng 等人[10]将红外和可见光图像从输入角度进行通道拼接，再利用深层卷积神经网络配合 K-Means++聚类策略优化目标框，以提升目标识别定位的准确性与鲁棒性。该方法虽提升了单一模态目标检测性能，但多模态信息交互不够充分，网络性能提升相对有限。王凯等人[11]以 YOLOv7 卷积神经网络为基础搭建双分支特征提取主干网络分别提取红外和可见光目标特征后，再结合 FPN 特征金字塔结构将不同模态、不同尺度特征进行融合后实现多模态目标识别定位。该方法虽提升了不同模态特征间的互补性，但网络结构相对单一，对多模态特征挖掘的深度不够。Shen 等人[12]针对多光谱图像融合检测问题提出了一种基于双交叉注意力的自注意力网络架构，通过并列的 Transformer 结构分别提取红外和可见光特征后通过迭代交叉注意力引导不同模态特征进行融合，以提升多光谱目标识别定位的性能。该结构虽然实现了基于 Transformer 架构的多模态检测，但网络在背景干扰、目标遮挡等复杂场景下的检测效果较差，其泛化性和鲁棒性也相对较弱。Xiao 等人[13]针对多模态目标检测问题提出了一种 CNN-Transformer 级联跨模态目标识别定位网络，网络先利用 CNN 结构提取红外和可见光目标细节特征，再利用 Transformer 结构融合多模态特征以丰富特征多样性，进而提升目标识别准确性。该方式虽提高了模型对跨模态信息的利用率，但网络结构与不同模态图像特性耦合度较低，且网络过于复杂，效率较低。Nie 等人[14]提出了一种基于 Transformer 预测的端到端多模态目标检测网络，通过卷积结构分别提取红外和可见光特征后利用 Transformer 结构多模态特征融合后直接输出目标类别和位置，提高了目标预测模块的推理速度。该网络虽然通过移除后处理过程提升了网络效率，但在多模态特征差异处理方面存在不足，对小目标的识别效果较差。综上可见，现阶段的红外可见光多模态信息融合的目标识别定位技术仍存在诸多局限，模型在识别精度、鲁棒性以及复杂场景适应性等方面均有待进一步提升。

针对当前多模态目标识别定位方法存在的局限，本文以城市复杂环境下的目标检测任务以及电网设备识别任务作为应用场景，以红外模态图像和可见光模态图像作为研究对象，构建了一种基于 CNN-Transformer 架构的红外-可见光目标识别与定位一体化网络。论文主要贡献：1) 针对可见光模态图像，提出了基于特征关键度分组的 CNN 卷积特征提取结构，根据特征重要程度分配计算资源，引导网络聚焦目标关键特征；2) 针对红外模态图像，提出了基于特征聚合的自注意力机制，并构建了高效率 Transformer 结构从全局上下文角度提取图像中的显著目标特征，3) 针对多模态信息互补，设计了基于特征增强的自适应融合结构，利用注意力机制突出红外和可见光目标特征后，结合浅层多模态特征以丰富特征多样性。4) 针对多尺度融合，引入了尺度动态调整策略，根据特征尺度差异自适应调整融合比例，并从特征空间通道角度增强目标信息，抑制背景信息，以缓解不同尺度间的特征冲突问题。

2 CNN-Transformer 网络设计

2.1 网络基础框架

本文所构建的 CNN-Transformer 目标识别定位网络主要由四个模块组成：可见光 CNN 网络模块、红外 Transformer 网络模块、CNN-Transformer 融合结构以及多尺度特征融合过滤结构，框架结构如图 1 所示。其中，可见光 CNN 网络由多个不同感受野的卷积单元级联构成，通过滑动卷积操作从局部视角逐步提取目标在可见光图像中的浅层细节纹理以及深层语义特征；红外 Transformer 网络主要以多头自注意力机制为核心，配合特征拆分合并等操作从全局上下文角度提取不同维度下目标显著特征；CNN-Transformer 融合结构则将可见光模态下的卷积局部特征和红外模态下的自注意力全局特征进行融合，以实现目标在不同模态、不同视角下的信息融合互补，提升特征多样性；而多尺度特征融合过滤结构则将不同尺度的 CNN-Transformer 融合特征以特征金字塔池化策略进一步从空间和通道角度进行自适应加权融合，降低特征冲突的同时提升网络尺度不变性。最后，基于多尺度融合特征预测目标类别和位置，实现目标精准识别与定位。

2.2 可见光 CNN 网络

由于 CNN 网络的局部连接和权重共享特性与可见光图像丰富的空间细节及层次结构高度契合，其层次化架构不仅能高效捕捉局部强相关的纹理与边缘特征，还可逐层抽象出物体轮廓与场景语义[15]。因此，针对可见

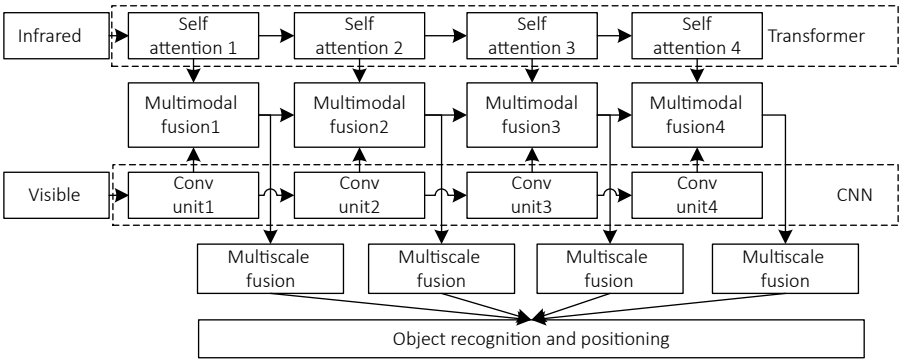


图 1 红外-可见光目标识别定位网络架构

Fig. 1 Infrared-visible light object recognition and localization network architecture

光图像，本文采用 CNN 网络作为特征提取的骨干结构，提取目标在可见光模态下的多层次特征，具体结构如图 2 所示。

CNN 骨干网络主要利用多个结构相同、维度不同的卷积单元来提取目标在不同尺度的特征，网络框架示意图如图 2(a) 所示。其中，卷积单元作为 CNN 网络的核心结构，由特征关键度分组模块和特征压缩提炼模块以并列拼接的方式构成，如图 2(b) 所示。特征关键度分组模块针对输入特征的每个通道特征，先通过最大值和平均值加权的方式来粗略计算其关键度 (即特征通道包含目标信息的多少)；然后，借助全连接操作来综合所有通道信息对关键度进行优化调整，并利用 Sigmoid 函数将结果归一化处理；最后，将关键度与对应通道特征加权以增强有效特征并抑制背景信息，同时，基于特征关键度大小，以式 (1)~式 (3) 对特征进行分组卷积，使关键度越大的组，经卷积输出后的特征通道越多，反之越少，以此实现网络对关键特征的自适应聚焦，详细结构如图 2(c) 所示。而特征压缩提炼模块主要利用 1×1 的点卷积操作结合 GELU 激活函数将特征从通道维度进行压缩提炼，并将提炼结果与分组卷积结果进行拼接，提升不同层次特征间的信息交互，降低微小目标在特征传递过程中的信息丢失概率。而对于不同维度的卷积单元之间采用了步长为 2 的池化与卷积操作以并列的方式进行下采样，缓解特征降维过程中的信息损失，结构如图 2(d) 所示。

$$N = S * C\,, \tag{1}$$

$$w_k = N - k + 1\,, \tag{2}$$

$$O_k = w_k M \Big/ \sum_{i=1}^{N} w_i\,, \tag{3}$$

式中：S 表示特征下采样次数；C 表示目标类别数；N 表示特征关键度分组数；w_k 表示第 k 组输出特征通道数比例；M 表示输入特征通道数；O_k 表示第 k 组输出特征通道数。

2.3 红外 Transformer 网络

考虑到红外图像低分辨率、低对比度、细节模糊等特点，结合 Transformer 网络全局上下文建模能力对局部细节的低敏感特性，可有效克服红外成像的固有局

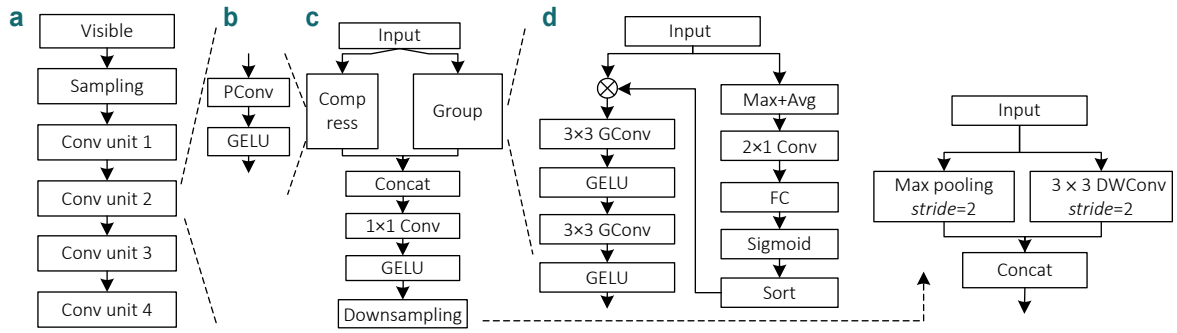


图 2 可见光 CNN 网络结构。(a) 基本结构；(b) 卷积单元；(c) 关键度分组卷积；(d) 混合采样

Fig. 2 Architecture of the visible-light CNN. (a) Basic structure; (b) Convolution unit; (c) Key degree grouped convolution; (d) Mixed sampling

限性，并能精准捕捉目标与背景间由热辐射差异形成的全局语义关联[16]。基于此，本文采用 Transformer 网络作为红外图像特征提取的骨干结构，利用基于特征聚合的多头自注意力机制为基础，提取目标在红外模态下的显著特征，详细结构如图 3 所示。

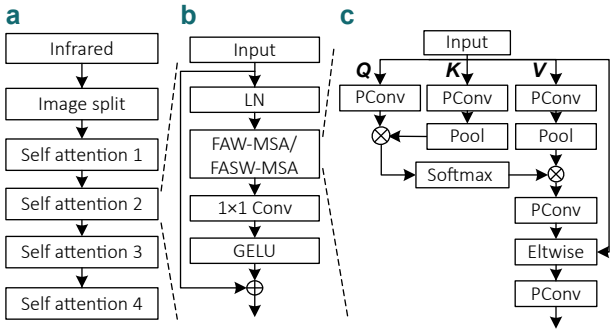


图 3 红外 Transformer 网络结构。(a) Transformer 结构；(b) 移位窗口结构；(c) 特征聚合自注意力

Fig. 3 Infrared Transformer network structure. (a) Transformer structure; (b) Shifted window structure; (c) Feature aggregation self-attention

Tranformer 网络主要以多头自注意力模块为基础，配合图像特征块拆分合并等操作实现对红外模态下的目标信息全局建模以及特征层次化提取，整体结构如图 3(a) 所示。考虑到标准的多头自注意力结构基于输入特征的所有数据进行计算，虽然能更好地全局建模，但也引入巨大的计算量，大幅降低了整体网络的计算效率。基于此，本文引入了移位窗口策略以及局部特征聚合机制设计了的基于特征聚合的移位窗口自注意力结构 (feature aggregation windows/shifted windows efficient multi-head self-attention, FAW-MSA/FASW-MSA)。其中，移位窗口策略主要借鉴 Swin-Transformer网络[17]，通过将特征以窗口形式进行划分，并配合移位方式跨窗口特征信息交互，降低模块的复杂度，框图结构如图 3(b) 所示；而局部特征聚合则针对自注意力结构中键矩阵和值矩阵，利用 2×2 的平均池化操作进行聚合，在降低计算量的同时也能更好地使网络聚集显著特征，聚合结构如图 3(c) 所示，计算方式如式 (4) 和式 (5)所示。

$$\boldsymbol{X}=\begin{cases}\boldsymbol{Q}: & PConv(\boldsymbol{f})\\ \boldsymbol{K}: & AvgPool(PConv(\boldsymbol{f}))\\ \boldsymbol{V}: & AvgPool(PConv(\boldsymbol{f}))\end{cases}, \tag{4}$$

$$\boldsymbol{Y}=PConv(Softmax(\boldsymbol{Q}*\boldsymbol{K})*\boldsymbol{V})+\boldsymbol{f}, \tag{5}$$

式中：$\boldsymbol{f}$ 表示红外特征；$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询、键、值特征矩阵；$\boldsymbol{X}$ 表示自注意力输入；PConv 表示卷积核为 1×1 的点卷积操作；AvgPool 表示核为 2×2 的平均池化操作；Softmax 表示归一化指数函数；$\boldsymbol{Y}$ 表示自注意力输出特征。

2.4 CNN-Transformer 特征融合

CNN-Transformer 特征融合结构主要是将同维度的可见光局部细节特征和红外全局语义特征进行自适应加权融合，以提升特征多样性，详细结构如图 4 所示。

CNN-Transformer 融合结构主要以相同尺度的卷积特征和自注意力特征作为输入，而为增强融合特征间的信息交互，进一步引入了浅层融合特征采样后的特征，如图 4(b) 所示。考虑到各模态输入数据尺度、量纲存在差异，融合结构先采用层归一化 (layer normalization, LN) 操作处理各模态数据。同时，为突出各模态关键特征，采用了如图 4(a) 所示专注于全局信息的通道注意力机制和如图 4(c) 所示专注于局部细节的空间位置注意力机制来分别突出红外和可见光特征中的关键信息，计算过程如式 (6) 和式 (7) 所示。而对于上一阶段的融合特征，

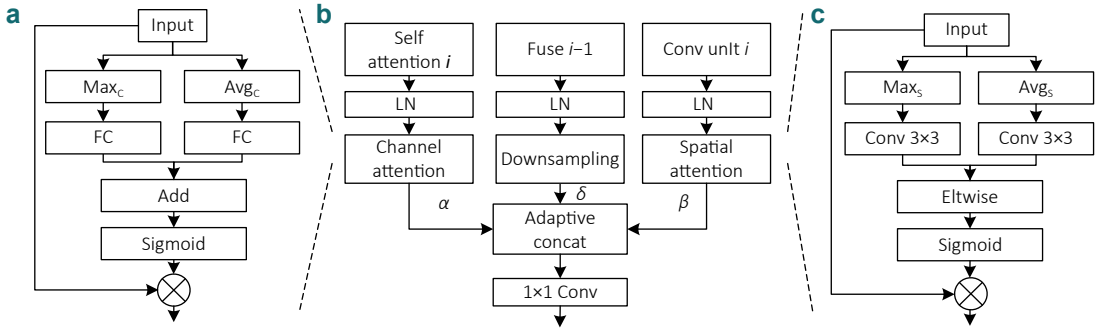


图 4 CNN-Transformer 多模态特征融合详细结构。(a) 通道注意力；(b) 特征融合结构；(c) 空间注意力

Fig. 4 Detailed architecture for CNN-Transformer multimodal feature fusion. (a) Channel attention; (b) Feature fusion structure; (c) Spatial attention

则利用最大池化和卷积降维方式来提炼关键特征，计算方式如式 (8) 所示。在特征融合阶段，引入了如式 (9) 所示的自适应拼接融合方案，通过可学习的权重参数动态调整各输入数据的贡献比例，从而实现跨模态特征的自适应对齐与互补增强。

$$A_{\mathrm{C}}(\boldsymbol{f}_{\mathrm{I}}^{i}) = \boldsymbol{f}_{\mathrm{I}}^{i} * \sigma(FC(Avg_{\mathrm{C}}(\boldsymbol{f}_{\mathrm{I}}^{i})) + FC(\mathrm{Max}_{\mathrm{C}}(\boldsymbol{f}_{\mathrm{I}}^{i}))), \quad (6)$$

$$A_{\mathrm{S}}(\boldsymbol{f}_{\mathrm{v}}^{i}) = \boldsymbol{f}_{\mathrm{v}}^{i} * \sigma\left(Eltwise\left(\begin{array}{c} Conv(Avg_{\mathrm{S}}(\boldsymbol{f}_{\mathrm{v}}^{i})) \\ Conv(\mathrm{Max}_{\mathrm{S}}(\boldsymbol{f}_{\mathrm{v}}^{i})) \end{array}\right)\right), \quad (7)$$

$$D(\boldsymbol{F}^{i-1}) = Concat(MaxPool(\boldsymbol{F}^{i-1}), DSConv(\boldsymbol{F}^{i-1})), \quad (8)$$

$$\boldsymbol{F}^{i} = PConv(Concat(\alpha A_{\mathrm{C}}(\boldsymbol{f}_{\mathrm{I}}^{i}), \beta A_{\mathrm{S}}(\boldsymbol{f}_{\mathrm{v}}^{i}), \delta D(\boldsymbol{F}^{i-1}))), \quad (9)$$

式中：$\boldsymbol{f}_{\mathrm{I}}^{i}$表示红外 Transformer 网络在 i 尺度的特征；$\boldsymbol{f}_{\mathrm{v}}^{i}$表示可见光 CNN 网络在 i 尺度的特征；A_{C} 表示通道注意力机制；A_{S} 表示空间注意力机制；$\boldsymbol{F}$ 表示 CNN-Transformer 融合特征；D 表示特征降维采样；α，β 和 δ 表示权重参数；Conv 表示卷积操作；FC 表示全连接；Eltwise 表示特征相加；σ 表示 Sigmoid 函数；Avg 表示均值计算；Max 表示最大值计算；MaxPool 表示最大池化；DSConv 表示深度可分离卷积；Concat 表示拼接操作。

2.5 多尺度特征融合过滤结构

CNN-Transformer 特征融合结构实现了跨模态信息互补，但未考虑不同尺度特征之间的信息交互。基于此，本文在多模态融合特征的基础上引入了特征金字塔结构，设计了如图 5 所示的多尺度特征融合结构以增强跨尺度特征的交互能力，提升模型对不同尺度目标的识别性能。

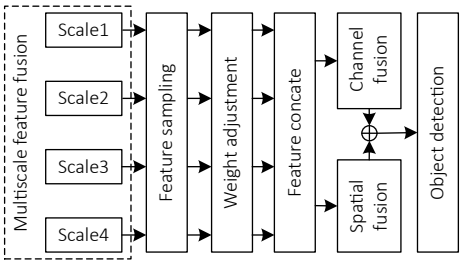


图 5 多尺度特征融合详细结构

Fig. 5 Multi-scale feature fusion detailed structure

多尺度融合结构以不同维度的 CNN-Transformer 融合特征作为输入，通过特征采样与通道压缩-扩张操作进行跨尺度信息融合后，采用共享权重的分类-定位联合预测头对目标类别与空间位置进行预测，进而实现目标分类与定位任务的一体化建模。考虑到多尺度特征融合时，特征维度差距越大时，特征冲突越大。基于此，为保障多尺度信息互补的同时降低特征间的冲突，本文根据特征维度差距大小来动态调整各尺度特征融合比例，并沿用 2.4 节中的通道和空间注意力结构来突出各尺度特征中的关键信息，从而达到强化目标特征的同时弱化背景信息的干扰。其中，特征融合比例调整方式如式 (10) 所示。

$$d_{j\to i} = d_i \mathrm{Min}(d_i, d_j) / \mathrm{Max}(d_i, d_j), \quad (10)$$

式中：d_i表示 i 尺度的特征通道数量；d_j表示 j 尺度的特征通道数量；Min 表示最小值；Max 表示最大值；$d_{j\to i}$表示 j 尺度采样到 i 尺度后的特征通道数量，即尺度差距越大，采样后的通道数量越少。

3 实验结果与数据分析

3.1 实验设置及评价指标

为验证所提方法的有效性，实验利用多个数据集对方法各模块进行了训练测试。实验过程中所依赖的组件如表 1 所示。实验数据集见表 2 所示，其中，KAIST、FLIR、RGB-T 为公开标准数据集，电网数据为实际机器人采集的电网设备图像数据 (主要包含绝缘子、断路器、冷控箱、变压器等目标)，在网络子模块验证时，主要以 KAIST 数据集为主，其他数据集主要用于和同类方法对比验证。网络训练过程中主要超参数设置如表 3 所示。

表 1 实验环境

Table 1 Experimental environment

软硬件	版本
操作系统	Linux 3.12
CPU	Intel Core i9-11900k
GPU	NVIDIA RTX 3070
编程语言	Python 3.11
深度学习框架	PyTorch 2.1.0

表 2 数据集详情

Table 2 Dataset details

数据集	总数	大小	类别数
KAIST	8342	512×512	3
FLIR	12241	640×512	4
RGB-T	9217	640×640	8
电网数据	4452	640×640	5

表 3 超参数设置

Table 3 Hyperparameter settings

参数名称	赋值
批大小	16
迭代轮数	200
优化器	Adam
学习率	0.001
动量	0.95
权重衰减	0.005
分类损失函数	Cross-Entropy
定位损失函数	DFL+CIoU

对于所提网络的性能衡量，本文采用了平均精度均值 (mean average precision, mAP)、网络参数量 (parameter, *P*) 以及浮点运算次数 (FLOPs, *F*) 等指标来对网络性能进行评估。

3.2 可见光 CNN 网络实验

针对可见光模态下 CNN 网络对目标特征提取的有效性验证，实验利用 KAIST 数据集中的可见光图像为训练测试数据，以 YOLO 网络分类定位一体化预测结构作为检测头，分别测试了不同分组卷积策略对网络性能的影响，并将 CNN 骨干网络与当前主流的主干网络进行了对比，实验结果如表 4 和表 5 所示。

表 4 不同卷积策略对比

Table 4 Comparison of different convolution strategies

卷积策略	mAP/%	*P*/M	*F*/G	FPS
标准卷积	83.6	22.8	120.2	78
固定分组卷积	81.7	18.1	103.5	93
深度可分离卷积	78.8	11.3	63.1	118
关键度分组卷积	83.1	16.9	94.3	101

表 5 CNN 网络对比

Table 5 Comparison of CNN network

CNN网络	mAP/%	*P*/M	*F*/G	FPS
YOLO[18]	83.9	25.4	125.3	74
ShuffleNet[19]	82.1	17.4	109.1	98
MobileNet[20]	79.5	10.7	60.2	123
本文网络	83.1	16.9	94.3	101

根据不同分组卷积策略对比结果可以看出，从标准卷积、固定分组卷积到深度可分离卷积的过程中，虽然网络参数量和浮点运算量显著降低、计算效率逐步提升，但网络对于复杂特征的学习能力也逐渐减弱，使得识别精度下降。而论文所提关键度分组策略在兼顾网络学习能力与计算复杂度的前提下，将主要计算资源集中在目标有效特征的挖掘上，并降低对背景信息的冗余计算，进而达到了效率与性能的协同优化。同时，与主流主干网络对比可见，YOLO、ShuffleNet 和 MobileNet 网络的核心模块分别对应标准卷积、分组卷积和深度可分离卷积，故 YOLO 网络在高精度目标识别定位任务中表现更优，MobileNet 网络则能更好地适用于边缘低性能设备。而本文所提 CNN 网络和 ShuffleNet 网络虽然都以平衡精度与效率为目的，但所提结构通过特征压缩提炼有效降低了跨层信息传递过程中的损耗，并通过特征动态分组机制来聚焦关键特征，故而使得网络整体性能相对传统分组策略更优。

3.3 红外 Transformer 网络实验

对于 Transformer 网络有效性验证，实验则基于 KAIST 数据集中的红外图像，以 Swin Transformer[17] 为基准网络，测试对比了引入特征聚合前后网络性能变化，结果如表 6 所示。

表 6 Transformer 网络结构对比

Table 6 Comparison of Transformer network

Transformer网络	mAP/%	*P*/M	*F*/G	FPS
Swin Transformer	79.8	33.5	173.2	69
本文网络	78.9	21.8	116.5	88

根据红外模态下 Transformer 网络的对比结果可以看出，自注意力结构主要以输入特征的全局上下文信息进行建模，其参数量和浮点运算量都相对更高，故网络效率相对于 CNN 网络较低。而引入特征聚合后，网络通过聚合了局部特征有效减少了冗余计算，同时增强了长程依赖的建模能力，使得网络效率有较大幅度提升，但特征聚合也破坏了部分信息，导致模型精度出现轻微下降。

3.4 模态与网络适配实验

针对 CNN 与 Transformer 网络对红外和可见光模态图像的适配性差异，实验对比测试了不同网络结构在两类模态图像上的性能表现，结果如表 7 所示。

表 7 模态与网络适配性测试结果

Table 7 Modal and network adaptability testing results

模态	网络架构	mAP/%	*P*/M	*F*/G	FPS
可见光	CNN	83.1	16.9	94.3	101
	Transformer	81.2	22.3	116.9	87
红外	CNN	78.2	16.6	94.0	102
	Transformer	78.9	21.8	116.5	88

根据网络架构与图像模态的交叉实验结果可见，CNN 网络通过局部卷积以及层次化降维机制可以较好地提炼出可见光目标特征，相比之下，Transformer 网络虽具备强大的长距离依赖建模能力，但在背景复杂的可见光场景中易受非目标区域干扰，导致特征提取的效果相对较差。而对于红外模态，CNN 受限于局部感受野难以充分建模跨区域的语义关联，而 Transformer 通过全局自注意力机制可有效捕捉低对比度场景下的热源分布特征，进而更精准地提取出红外目标特征。由此可见，CNN 网络更适用于可见光模态图像，而 Transformer 网络则更适配红外模态图像。

3.5 多模态融合实验

针对可见光局部卷积特征和红外全局自注意力特征的融合方式有效性测试，实验分别对比了特征相加、拼接、加权等策略，结果如表 8 所示，特征可视化效果如图 6 所示。

根据不同融合方式的对比结果可以看出，直接将两种模态的特征数据相加的方式虽然也能丰富目标特征信息，但也引入了较多噪声，如图 6(e) 中存在较多噪声，容易对小目标造成较大干扰，融合效果相对较差；多模态特征拼接的方式虽在一定程度上降低了噪声的引入，但增加了后续特征处理的计算量，影响网络整体效率；加权融合虽然降低了噪声以及过多计算量的引入，但对整个模态的特征进行加权叠加的方式相对粗糙，无法较好的实现特征互补，如图 6(g) 中增强特征的同时也增强了部分背景信息；而本文所提融合策略针对不同模态的特征，分别从通道和空间位置角度突出目标特征后，再结合浅层特征进行自适应加权融合，有效降低了背景噪声的引入，使得网络能够更好的在复杂环境中识别出目标特征。

3.6 多尺度特征融合实验

对于所提空间-通道融合过滤结构的有效性验证，实验将所提结构分别与 FPN[21]、ASFF[22] 多尺度融合结构进行了对比，结果如表 9 所示，可视化效果如图 7 所示。其中，mAPl 表示大目标的平均精度均值，大目标指像素面积>96^2；同理，mAPm 表示中目标，32^2<像素面积≤96^2；mAPs 为小目标，像素面积≤32^2。

根据不同多尺度融合结构对比结果可以看出，FPN

表 8 CNN-Transformer 特征融合方式对比

Table 8 Comparison of CNN-Transformer fusion methods

融合策略	mAP/%	*P*/M	*F*/G	FPS
叠加	86.8	32.1	167.5	72
拼接	87.5	33.9	176.3	68
加权	87.8	32.2	167.9	72
注意力	88.4	32.4	168.5	71

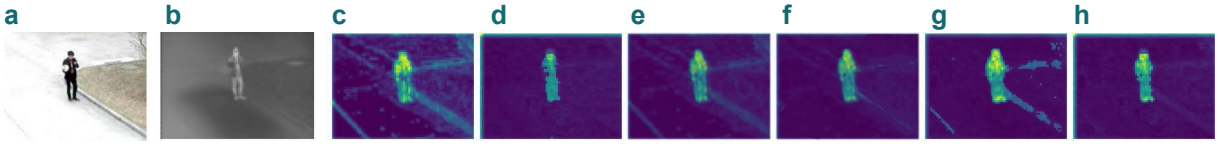


图 6 不同多模态融合方式可视化效果图。(a) 可见光；(b) 红外原图；(c) 可见光特征；(d) 红外特征；(e) 相加融合；(f) 拼接融合；(g) 加权融合；(h) 注意力融合

Fig. 6 Visualization effect diagram of different multi-modal fusion methods. (a) Visible light; (b) Infrared original image; (c) Visible light feature; (d) Infrared feature; (e) Additive fusion; (f) Mosaic fusion; (g) Weighted fusion; (h) Attention fusion

表 9 多尺度特征融合方式对比

Table 9 Comparison of multi-scale feature fusion methods

融合结构	mAP/%	mAPs/%	mAPm/%	mAPl/%	*P*/M	*F*/G	FPS
FPN	88.4	78.6	89.5	93.9	32.4	168.5	71
ASFF	88.8	78.9	90.0	94.3	32.7	168.9	70
本文	89.3	79.4	90.5	95.0	33.2	169.1	69

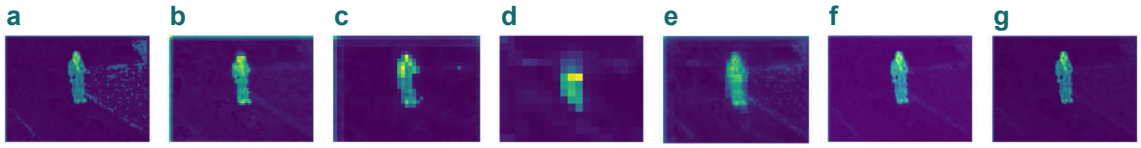


图 7 不同多尺度融合方式可视化效果。(a) 尺度 1；(b) 尺度 2；(c) 尺度 3；(d) 尺度 4；(e) FPN；(f) ASFF； (g) 本文

Fig. 7 Visualization effects of different multi-scale fusion methods. (a) Scale 1; (b) Scale 2; (c) Scale 3; (d) Scale 4; (e) FPN; (f) ASFF; (g) Ours

结构采用深层逐层融合浅层的方式相对简单高效，在一定程度上也提升网络尺度不变性，但其仅考虑相邻尺度间信息交互，且未考虑噪声干扰，使得网络检测精度相对较低，如图 7(e) 中存在较多背景信息；ASFF 结构通过将各尺度特征归一化维度后利用自适应参数进行加权叠加的方式虽然相对充分的综合了各尺度信息，但其仅对每个尺度的空间位置特征进行了调整，忽略了通道角度上的特征优化 (如 7(f) 中仍有少量噪声信息)，使得网络对不同尺度目标的识别性能提升相对有限；而本文所提多尺度融合结构在 ASFF 基础上兼顾了通道与空间位置维度的特征优化，并创新性地融入注意力机制，使网络能够更好地抑制噪声信息并捕捉各尺度特征中的关键信息，进而能更准确地识别定位出各尺度的目标。

3.7 同类型方法对比实验

针对所提方法与当前同类型的红外-可见光多模态目标识别定位方法性能对比，实验分别利用 KAIST、FLIR、RGB-T 以及实际电网设备数据集对各个方法进行训练测试，测试对比结果如表 10 所示。其中，文献 [11] 为双 CNN 架构；文献 [12] 为双 Transformer 架构；文献 [23] 为 CNN-Transformer 并列架构；文献 [13] 为 CNN 特征提取，Transformer 特征融合架构；文献 [14] 为 CNN 特征提取，Transformer 预测架构。

根据不同方法在各数据集上的测试结果可以看出，文献 [11] 和文献 [12] 仅依赖单一架构的网络模型在多模态特征提取挖掘上存在一定局限，特征多样性不足，进而制约了目标识别定位精度，如单一 CNN 架构过于关注局部特征，在复杂背景下容易产生误检情况；而单一 Transformer 架构过于追求全局特征，使得小目标的识别效果较差。文献 [23] 以两种网络分别特征提取的方式虽在一定程度上丰富了特征，但简单的网络结合无法充分发挥多模态互补性，且效率也较低，在复杂环境下的识别性能仍有较大提升空间。文献 [13] 将特征提取和特征融合分别用不同网络进行处理虽较好地综合了精度与效率，但从全局角度融合多模态特征的过程容易引入背景噪声，进而增加了误报概率。文献 [14] 虽移除了预测框非极大值抑制筛选过程，但对于重叠度高的目标无法较好地区分。而本文方法将红外和可见光模态图像特点与 CNN 和 Transformer网络特性深度结合，借助注意力机制精准聚焦关键特征信息，并运用自适应加权策略动态调整特征权重，使得多模态、多尺度特征能够充分融合与互补，进而达到更优的目标识别定位效果。

4 结 论

针对现阶段红外和可见光多模态目标识别定位方法

表 10 不同数据集下各方法测试结果

Table 10 Test results of various methods under different datasets

方法	P/M	KAIST			FLIR			RGB-T			电网数据		
		mAP/%	*F*/G	FPS	mAP/%	*F*/G	FPS	mAP/%	*F*/G	FPS	mAP/%	*F*/G	FPS
文献[11]	30.6	87.7	158.4	80	86.9	169.5	75	85.1	177.6	71	87.4	177.0	72
文献[12]	39.7	86.8	201.5	56	86.0	212.3	51	84.3	220.4	47	86.5	219.9	48
文献[23]	41.6	89.0	213.2	51	88.3	221.5	46	86.4	229.8	42	88.8	229.3	43
文献[13]	32.1	88.5	162.1	74	88.0	172.7	69	86.1	180.9	65	88.3	180.5	66
文献[14]	34.0	87.9	174.6	66	87.3	183.3	61	85.8	191.1	57	87.6	190.5	58
本文	33.2	89.3	169.1	69	88.9	179.8	64	87.1	188.1	60	89.1	187.6	61

在特征挖掘、多模态特征互补以及多尺度特征融合等方面存在的局限，本文将当前主流的神经网络架构与多模态图像特性深度结合，提出了一种融合 CNN 和 Transformer 网络的多模态目标检测方法。方法通过构建基于关键度分组的卷积单元和基于特征聚合的自注意力结构来分别提取目标在可见光模态和红外模态下的特征，保证特征提取效率的同时提升了所提特征的有效性。同时，设计注意力机制以及特征自适应动态调整策略，从特征通道和空间位置的角度增强不同模态、不同尺度特征间的信息交互，显著增强了特征多样性和尺度不变性。通过在多个数据集上的测试结果有效验证了论文所提方法的可行性和实用性，能够更精准地识别定位出目标。未来的研究工作将关注模型剪枝压缩以及极端复杂场景下的目标识别等场景，进一步提升网络的检测性能以及在恶劣环境下的适用性。

作者贡献

游春芝：研究资料的收集、方法设计，分析，论文写作；吴永飞、王剑：指导论文选题、论文修改。

利益冲突

所有作者声明无利益冲突

参考文献

1. Zhu Z W, Song X O, Cui W, et al. Review of visible and infrared image fusion for intelligent object detection[J]. *Comput Eng Appl*, 2025, **61**(17): 17−32.
   朱自文, 宋晓鸥, 崔巍, 等. 可见光-红外图像融合的目标检测综述[J]. 计算机工程与应用, 2025, **61**(17): 17−32.
2. Wu H X, Zhu Y, Li S Q. CDYL for infrared and visible light image dense small object detection[J]. *Sci Rep*, 2024, **14**(1): 3510.
3. Xiao Z J, Lai S Y, Qu H C. Multi-level refined UAV image target detection[J]. *Opto-Electron Eng*, 2025, **52**(5): 240287.
   肖振久, 赖思宇, 曲海成. 多层次精细化无人机图像目标检测[J]. 光电工程, 2025, **52**(5): 240287.
4. Wang D, Liu J, Liu R, et al. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection[J]. *Inform Fusion*, 2023, **98**: 101828
5. Luo X, Luo S J, Chen M Y, et al. MBFormer-YOLO: multibranch adaptive spatial feature detection network for small infrared object detection[J]. *IEEE Sens J*, 2024, **24**(12): 19517−19530.
6. Li Y, Yang H T, Kong Z, et al. A survey of pixel-level infrared and visible image fusion methods[J]. *Comp Eng Appl*, 2022, **58**(14): 40−50
   李扬, 杨海涛, 孔卓, 等. 像素级红外与可见光图像融合方法综述[J]. 计算机工程与应用, 2022, **58**(14): 40−50
7. Sengar S S, Hasan A B, Kumar S, et al. Generative artificial intelligence: a systematic review and applications[J]. *Multimed Tools Appl*, 2025, **84**(21): 23661−23700.
8. Yadav A B. An analysis on the use of image design with generative AI technologies[J]. *Int J Trend Sci Res Dev*, 2024, **8**(1): 596−599.
9. Li M L, Wang X X, Hou M X, et al. An object detection algorithm based on infrared-visible feature enhancement and fusion[J]. *Infrared Technol*, 2025, **47**(3): 385−394.
   李明禄, 王肖霞, 侯茂新, 等. 基于红外可见光特征增强与融合的目标检测算法[J]. 红外技术, 2025, **47**(3): 385−394.
10. Peng R H, Lai J, Yang X T, et al. Camouflaged target detection based on multimodal image input pixel-level fusion[J]. *Front Inform Technol Electron Eng*, 2024, **25**(9): 1226−1239.
11. Wang K, Lou S L, Ding X Z. Infrared and visible light fusion object detection algorithm based on YOLOv7[J]. *Laser Infrared*, 2025, **55**(5): 781−788.
    王凯, 娄树理, 丁晓震. 基于 YOLOv7 的红外与可见光融合目标检测算法[J]. 激光与红外, 2025, **55**(5): 781−788.
12. Shen J F, Chen Y F, Liu Y, et al. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection[J]. *Pattern Recognit*, 2024, **145**: 109913.
13. Xiao Y M, Meng F M, Wu Q B, et al. GM-DETR: generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Seattle: IEEE, 2024: 5541–5549. https://doi.org/10.1109/CVPRW63382.2024.00563.
14. Nie J Y, Sun H, Sun X, et al. Cross-modal feature fusion and interaction strategy for CNN-transformer-based object detection in visual and infrared remote sensing imagery[J]. *IEEE Geosci Remote Sens Lett*, 2024, **21**: 5000405.
15. Nani P, Das S, Dey S. Enhancing object recognition: a comprehensive analysis of CNN based deep learning models considering lighting conditions and perspectives[J]. *Evol Intell*, 2025, **18**(4): 72.
16. Yuan S, Qin H L, Yan X, et al. SCTransNet: spatial-channel cross transformer network for infrared small target detection[J]. *IEEE Trans Geosci Remote Sens*, 2024, **62**: 5002615.
17. Shi M L, Zheng D L, Wu T H, et al. Small object detection algorithm incorporating swin transformer for tea buds[J]. *PLoS One*, 2024, **19**(3): e0299902.
18. Vijayakumar A, Vairavasundaram S. Yolo-based object detection models: a review and its applications[J]. *Multimed Tools Appl*, 2024, **83**(35): 83535−83574.
19. Su J F, Yang M R, Tang X L. Integration of ShuffleNet V2 and YOLOv5s networks for a lightweight object detection model of electric bikes within elevators[J]. *Electronics*, 2024, **13**(2): 394.
20. Kamath V, A R. Investigation of MobileNet-Ssd on human follower robot for stand-alone object detection and tracking using Raspberry Pi[J]. *Cogent Eng*, 2024, **11**(1): 2333208.
21. Shi Z C, Hu J, Ren J, et al. HS-FPN: high frequency and spatial perception FPN for tiny object detection[C]//*Proceedings of the 39th AAAI Conference on Artificial Intelligence*. Philadelphia: AAAI, 2025: 6896–6904. https://doi.org/10.1609/aaai.v39i7.32740.
22. Pan C Y, Chen J Y, Huang R S. Medical image detection and

classification of renal incidentalomas based on YOLOv4+ ASFF swin transformer[J]. *J Radiat Res Appl Sci*, 2024, **17**(2): 100845.

23. Yang C, Hou Z Q, Li X Y, et al. Object detection algorithm based on CNN-transformer dual modal feature fusion[J]. *Acta Photonica Sin*, 2024, **53**(3): 0310001.
杨晨, 侯志强, 李新月, 等. 基于 CNN-Transformer 双模态特征融合的目标检测算法[J]. 光子学报, 2024, **53**(3): 0310001.

作者简介



【通信作者】游春芝 (1989-)，女，硕士，讲师，研究方向：图像处理，模式识别。

E-mail：919839697@qq.com

Research on multimodal object recognition and localization integrating CNN and transformer networks

You Chunzhi[1*], Wu Yongfei[2], Wang Jian[1]

[1] Department of Basic Medicine, Shanxi Medical University, Lvliang, Shanxi 032200, China;
[2] School of Computer Science and Technology, Taiyuan University of Technology, Jinzhong, Shanxi 030600, China

Abstract

Objective In complex real scenes (such as low illumination, haze, occlusion or strong light interference), single mode target recognition and localization methods face significant performance bottlenecks. Although the visible image is rich in texture and color details, it is highly sensitive to light conditions; Infrared image has all-weather working ability and strong penetrability by virtue of thermal radiation imaging, but it is limited by low resolution and lack of structural information. The inherent limitations of the two make it difficult for the single-mode system to meet the requirements of all-weather perception with high robustness and high accuracy. Therefore, how to deeply integrate the complementary advantages of infrared and visible modes, that is, combining the local detail expression ability of visible light with the global thermal semantic saliency of infrared, has become a key challenge to improve the performance of target detection in complex environments. This paper aims to build a new neural network architecture that can deeply mine, adaptively fuse and efficiently use multimodal features, in order to achieve higher accuracy, stronger robustness and better scale invariance of target recognition and localization.

Methods This paper proposes an infrared visible multimodal target recognition and positioning network based on CNN transformer heterogeneous fusion framework. The network is composed of four core modules: visible light oriented CNN feature extraction branch, infrared oriented transformer feature extraction branch, cross modal adaptive fusion module and multi-scale dynamic adjustment structure. In the visible light branching, the "feature criticality grouping convolution" mechanism is designed: the feature importance is evaluated by weighting the maximum and average values at the channel level, and the normalized criticality weight is generated after full connection layer optimization. Based on this, the number of convolution output channels is dynamically allocated - the key region gets richer characterization, and the background region is compressed, so as to guide the network to focus on the essential information of the target; At the same time, the compression extraction module composed of 1 × 1 point convolution and Gelu activation is introduced, and the hybrid down sampling strategy is adopted to reduce the computational overhead and the loss of micro target information. In the infrared branch, the feature aggregation shift window self-attention structure is constructed: the computational complexity is controlled by using the window shifting mechanism of swing transformer for reference, and 2 × 2 average pooling is applied to the key (k) and value (V) matrices in the self-attention for local aggregation, which not only significantly reduces the redundant calculation, but also strengthens the ability of global context modeling for low contrast heat source targets. Then, a cross modal fusion module is designed: firstly, the local features of

Foundation item: Project of Shanxi Provincial Department of Education (J20241687), and Scientific Research Project of Fenyang College, Shanxi Medical University (2022C21, 2022A02)

*Correspondence: You C Z, E-mail: 919839697@qq.com

CNN and the global features of transformer at the same scale are normalized, and then channel attention (highlighting semantic importance) and spatial attention (accurately locating the target area) are applied respectively, and shallow features are fused to enhance detail retention; Finally, the adaptive stitching of the three features is realized through the learnable weight parameters, and the complementary enhancement between modes is achieved. Finally, in order to alleviate the feature conflict in multi-scale fusion, a multi-scale dynamic adjustment mechanism is proposed: Based on the feature pyramid structure, the fusion proportion is dynamically adjusted according to the difference in the number of adjacent scale channels (the larger the scale difference, the smaller the fusion weight), and the channel spatial attention mechanism is combined to strengthen the target response, suppress the background interference, and significantly improve the unified detection ability of small, medium and large targets.

Results and Discussions Experiments is systematically validated on three public datasets—KAIST, FLIR, RGB-T—and a self-collected power grid equipment dataset (covering 5 types of electrical equipment). The results show that: 1) the proposed key-grouped convolution achieves 83.1% mAP on the visible light modality, with parameters and FLOPs reduced by 26% and 21% compared to standard convolutions, and inference speed increased by 29%, outperforming mainstream lightweight networks and demonstrating an excellent efficiency-accuracy balance. 2) After introducing feature aggregation in the infrared branch, FLOPs decreased from 173.2 G to 116.5 G, FPS increased from 69 to 88, and mAP only slightly dropped by 0.9%, verifying its efficiency. 3) The modality-network adaptability experiment clearly shows that CNN performs better on visible light (83.1% mAP) than Transformer (81.2%), while Transformer outperforms CNN on infrared (78.9% vs. 78.2%), strongly supporting the rationality of the heterogeneous dual-branch design. 4) In the comparison of multimodal fusion strategies, the proposed attention-guided adaptive fusion achieved 88.4% mAP on KAIST, significantly surpassing addition (86.8%) and concatenation (87.5%) methods; feature visualization confirms its effectiveness in suppressing background noise, particularly benefiting small object edge preservation. 5) The multi-scale fusion module, built on FPN and ASFF, introduces dynamic ratio adjustment and dual attention, achieving an overall mAP of 89.3%, with small object mAPs improved to 79.4%, validating its effectiveness in alleviating scale conflicts. 6) Comprehensive comparisons with five advanced methods (including dual CNNs, dual Transformers, CNN-Transformer cascades, etc.) across the four datasets show that our method achieved the best or second-best results on KAIST (89.3%), FLIR (88.9%), RGB-T (87.1%), and the power grid dataset (89.1%), with controlled computational cost (FLOPs ≈169 G, FPS ≈69), balancing accuracy, robustness, and practicality. In summary, this method systematically addresses the issues of insufficient multimodal feature complementarity and scale sensitivity through the deep synergy of modality characteristics, network architecture, and fusion strategies.

Conclusions Aiming at the problems of insufficient feature mining, weak modal complementarity and poor multi-scale adaptability in infrared visible multimodal target recognition and localization, this paper innovatively proposes a heterogeneous network architecture that deeply integrates the advantages of CNN and transformer. Through visible light oriented feature criticality grouping CNN, infrared oriented feature aggregation lightweight transformer, attention driven adaptive cross modal fusion and dynamic multi-scale adjustment mechanism, the extraction efficiency, fusion quality and scale invariance of multimodal features are systematically improved. A large number of experiments have verified the superior performance of this method on multiple real scene data sets, which not only leads significantly in detection accuracy, but also maintains good computational efficiency and environmental robustness. This study provides an effective technical path for highly reliable target perception in complex environments. Future work will focus on model compression (such as knowledge distillation, neural architecture search) to adapt edge devices, and explore strategies to enhance the generalization ability in extreme weather (such as rainstorm and smoke), so as to further promote the application of this technology in key fields such as power patrol, automatic driving, emergency rescue and so on.

Keywords: object recognition and localization; infrared and visible light; CNN-Transformer; multimodal complementarity; multiscale fusion

You C Z, Wu Y F, Wang J. Research on multimodal object recognition and localization integrating CNN and transformer networks[J]. *Opto-Electron Eng*, 2026, **53**(2): 250210; DOI: 10.12086/oee.2026.250210

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文