中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-17

论文引用格式:Ma Xiangyang, Chen Junying, Yan Mengming, Liang Dong. Task-Driven Dual-Modal Adaptive Fusion Network for Low-Illumination Object Detection[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260187.(马向阳, 陈俊英, 闫孟明, 梁栋. 面向低光照场景的任务驱动双模态自适应融合检测网络[J/OL]. 中国图象图形学报, XXXX:1-17. DOI: 10.11834/jig.260187.)[DOI:10.11834/jig.260187]

面向低光照场景的任务驱动双模态自适应融合检测网络

马向阳,陈俊英,闫孟明,梁栋

西安建筑科技大学信息与控制工程学院,陕西　西安　710055

摘　要:针对低光照场景下可见光与红外双模态目标检测中存在的模态干扰、特征选择性不足及边界框回归适应性差等问题,本文构建了任务驱动双模态自适应融合检测网络(Task-driven Dual-modal Adaptive Fusion Detection Network, TDAFNet)。在特征提取阶段,设计动态局部注意力模块(Dynamic Local Attention, DLA),采用恒等与增强双分支并行结构,通过动态非线性压缩机制和自注意力机制抑制低光照噪声,强化目标边缘与纹理表达。在跨模态交互阶段,构建任务感知双模态融合模块(Task-Aware Dual-modal Fusion, TADF),利用CLIP文本编码器提取类别语义作为查询原型,与可见光及红外特征的共享键值对进行交互注意力计算,并辅以多尺度空洞卷积增强语义相关性,实现语义引导的自适应融合。同时,提出自适应归一化EIoU损失(AN-EIoU),引入尺寸归一化、中心偏移及长宽比三个调节因子,对中心距离项和宽高误差项进行结构化调节。在M3FD数据集上,模型的mAP@0.5和mAP@0.5:0.95分别达到89.3%和62.6%;在LLVIP数据集上,对应指标为97.8%和68.3%。融合质量评估中,所提方法在源图像相关差异(SCD)、视觉信息保真度(VIF)、基于梯度的融合质量(Q_{abf})等多项指标上均优于对比方法。该方法提升了低对比度、弱纹理及遮挡场景下的检测精度与定位能力,为低光照环境下的目标检测提供了一种有效解决方案。

关键词:双模态目标检测;跨模态特征融合;动态局部注意力;自适应归一化EIoU损失;低光照场景

Task-Driven Dual-Modal Adaptive Fusion Network for Low-Illumination Object Detection

Ma Xiangyang, Chen Junying, Yan Mengming, Liang Dong

College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China

Abstract: Visible light images captured under low-light conditions suffer from low contrast, while infrared images lack detailed textures and are susceptible to thermal noise interference. Existing dual-modal fusion methods fail to effectively suppress cross-modal interference, adopt task-agnostic feature selection strategies, and exhibit poor adaptability of bounding box regression to multi-scale and non-square targets. To address the above issues, this paper proposes a Task-driven Dual-modal Adaptive Fusion Detection Network (TDAFNet) based on YOLOv8, with coordinated improvements implemented from three perspectives: feature enhancement, semantic fusion, and loss function optimization. First, a Dynamic Local Attention (DLA) module is proposed for the feature extraction stage, adopting a parallel structure consisting of an

收稿日期:2026-04-09;**修回日期**:2026-07-31

基金项目:陕西省自然科学基础研究计划(2025JC-YBMS-662)

Supported by: Fund Project: Shaanxi Province Natural Science Foundation Research Program (2025JC-YBMS-662)

identity branch and an enhancement branch. The identity branch preserves original spatial information, whereas the enhancement branch adaptively modulates channel responses via learnable nonlinear activation functions and Layer Normalization to suppress abnormal activations. Adaptive average pooling and depthwise separable convolutions are utilized to generate spatial attention, which is integrated with channel attention to construct a self-attention triplet for highlighting target edges and textures. The outputs of the two branches are concatenated to enrich feature representations. Second, a Task-Aware Dual-modal Fusion (TADF) module is designed between the Backbone and Neck. Visible and infrared features are compressed along the channel dimension for concatenation, followed by layer normalization and projection into a shared Key/Value space. Meanwhile, category labels are fed into a frozen CLIP text encoder to extract semantic embeddings, which generate modality-specific Queries through dual-branch linear projection. Cross-attention is computed between modality-specific Queries and shared Key-Value pairs. Afterwards, a receptive field attention enhancement module leverages multi-scale dilated convolutions and direction-aware attention to produce context-enhanced features and boost feature discriminability. Finally, an Adaptive Normalized EIoU (AN-EIoU) loss function is put forward. A size normalization factor is introduced to balance gradients of targets with extreme aspect ratios; a center offset factor dynamically adjusts centroid penalties according to target area; an aspect ratio factor strengthens sensitivity to non-square objects. The three factors jointly optimize centroid distance and width-height errors to accelerate localization convergence. In terms of fusion quality, on the M3FD dataset, the proposed method achieves a Spatial Frequency (SF) of 15.2, a Source Image Correlation Difference (SCD) of 1.88, and a Gradient-Based Fusion Quality Metric (Q_{abf}) of 0.67. On the LLVIP dataset, it attains an Entropy (EN) of 7.4, a Visual Information Fidelity (VIF) of 0.47, and a Q_{abf} of 0.71. Demonstrating the superiority of the fused results in clarity, information integration and edge preservation. For detection performance, ablation experiments on the M3FD dataset show that individually embedding DLA, TADF and AN-EIoU raise the baseline mAP@0.5 from 80.8% to 84.4%, 85.2% and 83.5%, respectively. The combination of DLA and TADF reaches 88.6%, and the integration of all three modules yields the optimal performance. The complete TDAFNet obtains an mAP@0.5 of 89.3% and an mAP@0.5-0.95 of 62.6% on M3FD, and 97.8% and 68.3% on LLVIP, surpassing other state-of-the-art dual-modal detection methods. The model has a parameter count of 12.6 M, computational cost of 56.6 GFLOPs, and an inference speed of 41.3 frames per second (FPS). Although its inference speed is lower than YOLOv8n, it achieves substantial accuracy improvements with much lower computational overhead than other dual-modal detection models. Visualization results verify that the proposed network reduces missed detections and delivers more precise localization in scenarios with low contrast, occlusion and weak textures. In conclusion, the synergistic cooperation of DLA, TADF and AN-EIoU in TDAFNet mitigates the challenges of cross-modal interference, task-insensitive feature selection and insufficient regression adaptability in low-light dual-modal object detection, and improves detection accuracy and localization robustness. Extensive experiments validate its effectiveness and generalization capability across diverse scenarios, offering a novel solution for low-light object detection.

Key words: Dual-modal Object Detection; Cross-modal Feature Fusion; Dynamic Local Attention (DLA); Adaptive Normalized EIoU Loss (AN-EIoU); Low-light Scenes

论文引用格式:DOI:10.11834/jig.260187

0 引　言

近年来,随着多模态技术的发展,在低光照场景中有效融合多源信息以提升检测鲁棒性成为研究热点(Yi等,2025)。传统基于可见光图像的检测方法在光照充足时表现优异,但在低光照、夜间或恶劣天气条件下性能下降。红外图像可在这些条件下稳定成像,但受限于分辨率和纹理细节(Tang等,2023),难以单独支撑高精度检测。因此,利用可见光与红外图像互补特性的双模态融合(Zhang等,2023),成为提升复杂环境中目标识别性能的重要方向。

目标检测方法经历了从传统手工特征到深度学习模型的发展历程。早期方法多基于HOG、LBP等人工设计特征(Li等,2024),对复杂背景、尺度变化及光照波动的适应能力较为有限。随着深度学习技术的发展,Fast R-CNN、Mask R-CNN等双阶段检测器在检测精度上取得了显著进展(Girshick等,2015;He等,2017),而SSD、YOLO系列等单阶段检测器则

在检测速度与精度之间实现了较好的平衡(Liu等，2016；Zhang等，2017；Chen等，2022)。这些研究成果为可见光与红外双模态检测网络的构建奠定了重要基础。

围绕可见光与红外融合检测，现有研究主要从跨模态特征交互、网络结构优化及融合策略改进等方向展开。代表性方法包括双分支特征分解网络、通道注意力机制与多光谱通道特征融合模块等(Xu等，2024；Chen等，2022；Cao等，2021)。此外，Ghost卷积、张量分解、BoT结构及全局与局部注意力融合等技术也被用于降低模型复杂度或增强特征表达能力(Han等，2020；Yin等，2021；Liu等，2022；Xue等，2023；Sun等，2024)。上述研究在双模态特征建模领域取得了一定进展，但在低光照场景中，仍面临目标尺度变化、样本分布不均、遮挡干扰及复杂背景等挑战(Sun等，2023；Aleerxia，2025；代小波等，2025)。尤其在低光照条件下，可见光特征退化与模态信息不均衡会进一步加剧检测难度，相关方法仍存在改进空间(Tang等，2022；Liu等，2021)。具体而言，现有方法仍存在以下关键问题：由于可见光与红外图像在成像机制和语义表达上存在固有差异，直接融合易引入冗余信息并加剧模态间干扰；同时，不同空间区域与语义层级的模态有效性往往不同，而现有方法对任务相关信息的选择性仍有待增强；此外，复杂场景中目标尺度与形状差异较大，常规边界框回归损失对小目标及极端长宽比目标的适应性也亟需提升。

针对上述挑战，本文构建了一种任务感知的可见光与红外融合目标检测网络，主要工作包括：

1)针对可见光与红外图像的模态差异与特征冗余干扰问题，构建显式双分支特征建模结构，结合非线性门控机制、多尺度注意力和跨模态对齐模块，实现可见光与红外特征的解耦与动态调控，以缓解模态冗余和特征错配问题。

2)针对不同空间区域和语义层级上的模态有效性不一致的问题，设计任务感知融合机制，引导双模态特征在空间和语义层面进行选择性融合，使网络能够根据检测任务强化相关模态信息。

3)针对常规边界框回归损失对小目标和极端长宽比目标适应性不足的问题，提出尺度与形状自适应的归一化EIoU损失函数，即AN-EIoU，通过尺度归一化、中心偏移调节和长宽比自适应约束，增强小目标及复杂形状目标边界框回归的适应性。

1 TDAFNet算法设计

1.1 TDAFNet网络

针对低光照场景下多尺度目标检测困难以及可见光与红外信息利用不足的问题，本文提出一种基于双模态融合的TDAFNet目标检测模型。该模型以YOLOv8为基础框架，从特征增强与双模态融合两个方面进行改进，整体结构如图1所示。

在特征提取阶段，本文在Backbone中融入DLA(Dynamic Local Attention)机制，通过动态局部注意力对特征响应进行自适应调节，并结合多路径特征扩展策略，增强特征表达能力，提高模型对多尺度目标及复杂背景的感知能力。

在跨模态特征交互阶段，在Backbone与Neck之间设计TADF(Task-Aware Dual-modal Fusion)，通过自适应权重分配实现可见光与红外特征的动态融合，提升融合特征的判别能力。

因此，TDAFNet结构方面的主要改进体现在两个方面：一是利用DLA提升融合前单模态特征质量，二是利用TADF实现任务语义引导下的双模态自适应融合。该设计能够在保持高效检测框架的基础上，更充分挖掘可见光与红外图像的互补信息，从而提高低光照场景下目标检测的准确性和鲁棒性。

1.2 DLA模块

为提升双模态融合前输入特征的稳定性与结构表达能力，本文在Backbone中引入动态局部注意力模块DLA。该模块旨在抑制低光照噪声和模态差异带来的不稳定响应，同时增强目标边缘、纹理及显著区域的特征表达。DLA首先对输入特征进行非线性变换，在保留特征相对强度的同时抑制异常值，提供更为丰富的初始特征表示；随后进行双分支异构处理，将输入张量解耦为两个子空间，一个子空间进行恒等映射以保持特征拓扑不变性，另一个子空间进行特征增强，最终将两个子空间的结果拼接输出。DLA结构如图2所示。

DLA的核心在于引入动态非线性压缩机制与自注意力机制，以提升复杂场景下特征表达的稳定性和判别性。动态非线性压缩机制通过自适应调节特征响应，抑制低光照、噪声干扰及跨模态差异引起的异常激活，从而缓解特征分布波动。自注意力机制

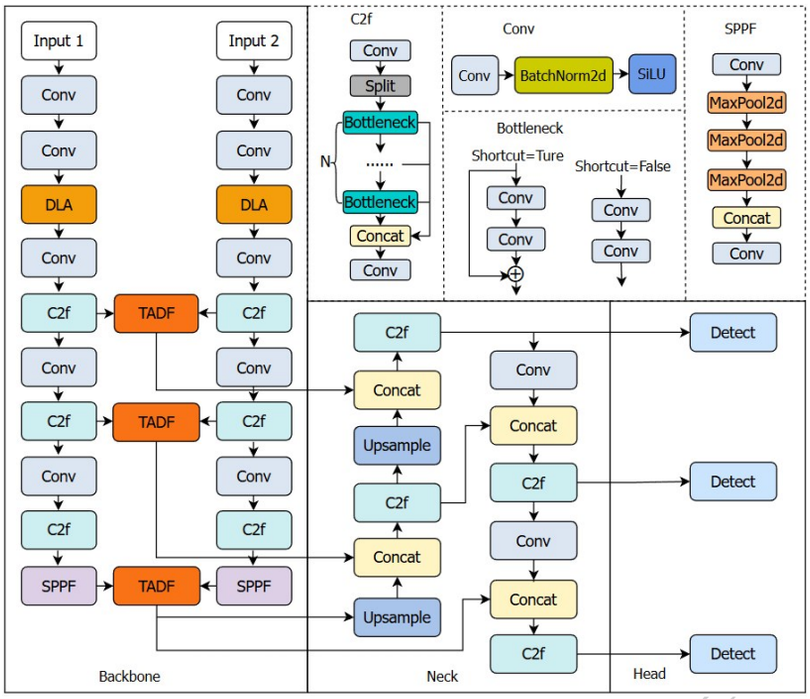

图1　TDAFNet网络结构图

Fig. 1　TDAFNet Network Architecture Diagram

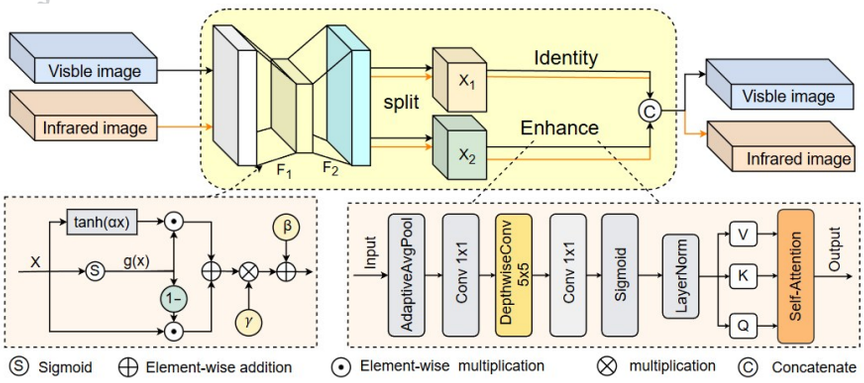

图2　DLA结构图

Fig. 2　DLA structure chart

则进一步建模空间位置间的依赖关系，对目标相关区域和有效上下文信息进行重标定，强化边缘、纹理及显著区域等判别性表达。恒等分支用于保留原始特征的结构信息，增强分支则通过非线性压缩与自注意力建模突出任务相关响应。两条分支经拼接融合后，在保持信息完整性的同时提升融合前单模态特征的表达质量。

在非线性激活设计中，DLA借鉴PReLU的参数化理念，引入可学习参数，使得激活函数在训练过程中能依据数据分布和任务需求自适应地调整负值区间响应，增强非线性表达能力。

与此同时，结合Layer Normalization的可学习仿射变换机制，在特征维度设定可训练的平移及缩放参数，以此动态调节特征分布，提高特征的稳定性与可控性。在双分支结构进行特征处理之前引入上述机制，借助可调非线性激活函数实施非线性压缩，如公式(1)所示：

$$\mathrm{DyT}(\boldsymbol{x})=\gamma\cdot[g(\boldsymbol{x})\cdot\tanh(\alpha\boldsymbol{x})+(1-g(\boldsymbol{x}))\cdot\boldsymbol{x}]+\beta. \tag{1}$$

式中，$\boldsymbol{x}$表示输入特征张量；α、β、γ为可学习参数，分别控制非线性压缩程度、特征的输出幅度与中心偏移；可调型非线性压缩函数如公式(2)所示：

$$\tanh(\alpha x)=\frac{e^{\alpha x}-e^{-\alpha x}}{e^{\alpha x}+e^{-\alpha x}}. \tag{2}$$

式中，$\tanh(\alpha x)$是$\tanh(x)$的改进，通过α可学习因子控制激活强度；$g(x)$为门控函数，如公式(3)所示：

$$g(x)=\sigma(W_g * x+b_g). \tag{3}$$

式中，W_g为可学习的卷积权重；b_g为偏置项；σ激活函数对输出范围进行限制，将权重归一化至[0,1]区间。

特征经非线性增强后进行分割，一部分进行恒等映射，另一部分进行注意力显著性增强。借助Adaptive Average Pooling提取整体空间上下文信息，并利用深度可分离卷积进一步构建空间相关性，最终通过Sigmoid激活函数生成如下权重A_1，如公式(4)所示：

$$A_1=\sigma(\mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{5\times5}(\mathrm{Conv}_{1\times1}(\mathrm{AdaptiveAvgPool}(X_2))))). \tag{4}$$

式中，X_2为分割后输入的特征信息，$A_1\in[0,1]^{H\times W}$表示每个空间位置的重要性权重。该注意力权重将用于调制输入特征，从而引导后续单头自注意力模块聚焦关键区域。

为进一步增强注意力模块对显著区域的聚焦能力，把输出的空间注意力权重当作引导信息，融入通道注意力的输入里。首先对输入特征进行归一化处理，接着依据空间注意力权重，在显著区域强化归一化特征表达，在非显著区域保留原始特征，进而得到引导后的融合结果X''，如公式(5)所示：

$$X''=A_1\odot\mathrm{LayerNorm}(X_2)+(1-A_1)\odot X_2. \tag{5}$$

式中，$\odot$表示逐元素乘法，$A_1\in[0,1]^{B\times1\times H\times W}$在通道维度进行广播，借助残差式门控机制增强模型在语义区域与背景区域之间的表达差异。基于融合后的特征X''构建自注意力模块的Query、Key、Value三元组，如公式(6)所示：

$$Q=X''W^Q,K=X''W^K,V=X''W^V. \tag{6}$$

式中，W^Q，W^K，W^V为可学习的线性映射矩阵。通过自注意力计算公式，获得注意力增强后的特征，如公式(7)所示：

$$\mathrm{Attention}(Q,K,V)=\mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{7}$$

最终，将经过注意力增强处理后的特征与分割后未经处理的原始特征进行拼接融合，以保留多语义层次的特征信息。

DLA模块在轻量化框架中引入更为灵活的非线性激活、通道注意力和空间建模机制，以此增强特征表达能力，保持信息完整性，降低模型复杂度，进而有效提升后续双模态的融合效果。

1.3 TADF模块

在低光照场景中，可见光图像具有较丰富的纹理和颜色信息，但易受光照变化影响。红外图像能够稳定突出目标热辐射区域，但细节和结构信息相对不足。为充分利用两种模态的互补性，本文设计任务感知双模态融合模块(Task-Aware Dual-modal Fusion，TADF)，用于在检测类别语义引导下实现可见光与红外特征的自适应交互与融合。该模块将类别文本作为任务先验，引导模型关注与目标检测任务相关的模态信息和空间区域，从而提升融合特征的任务相关性和判别能力。

TADF模块通过将目标类别名称构建为文本输入，并利用CLIP文本编码器获取语义嵌入，将其作为查询向量(Query)，与由双模态特征构建的键(Key)和值(Value)进行交互。借助注意力机制，实现语义引导的特征匹配，从而对不同模态特征进行自适应加权融合。通过上述设计，TADF模块能够根据任务语义动态调整可见光与红外特征的贡献比例，从而提升模型在低光照场景下的检测性能。整体结构如图3所示。

设输入的可见光特征图为F_{vis}，红外特征图为F_{ir}，为充分整合两种模态的底层视觉信息，首先在通道维度对两种特征图进行通道压缩与融合，如公式(8)所示：

$$F_{fused}=Conv_{1\times1}(Concat_{channel}(F_{vis},F_{ir})). \tag{8}$$

随后，对融合特征图进行层归一化处理，并将其展平为序列，如公式(9)所示：

$$F=Flatten(LayerNorm(F_{fused})). \tag{9}$$

接着，构建一个能够被不同任务语义引导的统一图像语义空间。采用共享式注意力键值映射策略，将经过归一化与展平处理后的融合特征序列，分别投影到注意力机制所需的键(Key)和值(Value)空间中，如公式(10)所示：

$$K=FW_K,V=FW_V. \tag{10}$$

式中，W_K，W_V为可学习的线性投影矩阵。通过统一构造的键值对，模型可在后续阶段接收来自不同任务语义引导下生成的查询向量。

类别标签被视为任务先验的语义指引，通过引

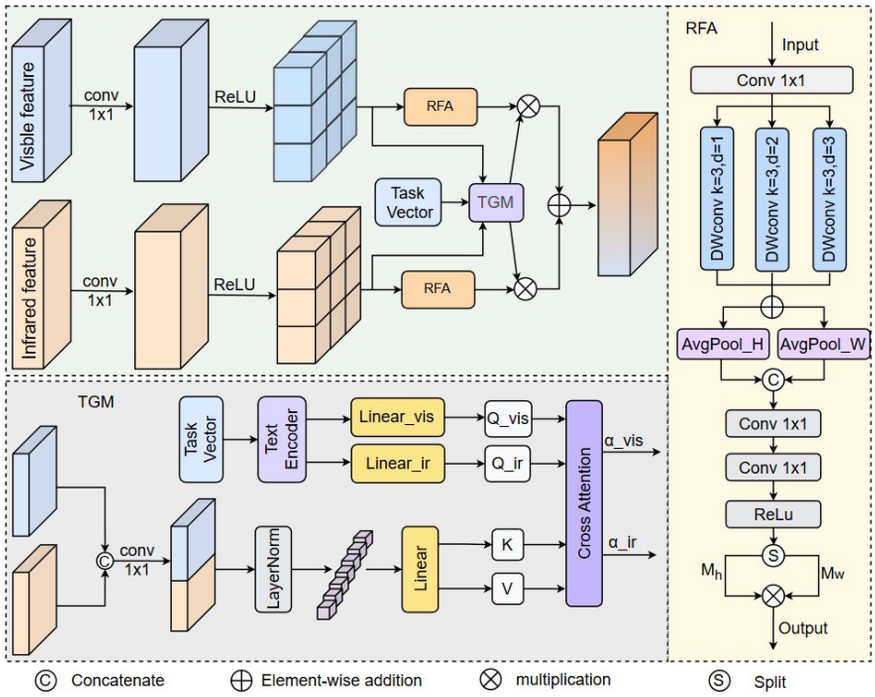

图3　TADF结构图

Fig. 3　TADF structure chart

入文本编码器对类别信息进行嵌入表达，以此构造任务相关的查询向量(Query)。将类别名称构成输入文本序列$\boldsymbol{T}$，将其输入预训练语言编码器以获得类别语义嵌入表示，如公式(11)所示：

$$\boldsymbol{T}_{enc}=\text{TextEncoder}(\boldsymbol{T}). \tag{11}$$

式中，$\boldsymbol{T}_{enc}$表示类别的高阶语义特征，可作为任务查询的原型表示。CLIP模型仅用于对类别标签进行语义建模，其文本编码器作为固定的语义先验提取模块参与特征构建过程，并不参与网络的端到端训练。为避免引入额外的训练不稳定性与计算开销，CLIP文本编码器在训练过程中保持参数冻结，仅用于提供稳定的类别语义表示。同时，该语义表示并不直接参与目标分类或边界框回归过程，而是通过构建任务查询向量参与注意力计算，从而引导跨模态特征的交互方式，使融合过程能够根据检测任务语义动态调整不同模态特征的响应程度。

考虑到红外和可见光两种模态需要具有不同的关注模式，进一步采用双通路线性投影策略，将任务嵌入分别映射为两种模态下的独立查询向量，如公式(12)所示：

$$\boldsymbol{Q}_{vis}=\boldsymbol{T}_{enc}\boldsymbol{W}_{vis},\boldsymbol{Q}_{ir}=\boldsymbol{T}_{enc}\boldsymbol{W}_{ir}. \tag{12}$$

式中，$\boldsymbol{W}_{vis}$，$\boldsymbol{W}_{ir}$为两种模态对应的独立线性映射矩阵。获得任务引导的查询向量后，借助共享键值对进行交互，分别计算两个模态分支的注意力输出，如公式(13)所示：

$$\begin{aligned}a_{vis}&=\text{Softmax}\left(\frac{\boldsymbol{Q}_{vis}\boldsymbol{K}^{\top}}{\sqrt{d}}\right)\cdot\boldsymbol{V},\\ a_{ir}&=\text{Softmax}\left(\frac{\boldsymbol{Q}_{ir}\boldsymbol{K}^{\top}}{\sqrt{d}}\right)\cdot\boldsymbol{V}\end{aligned} \tag{13}$$

为提升模型对空间结构与上下文的建模能力，在RFA语义增强模块中，通过多尺度空洞卷积融合方向感知注意机制，生成上下文增强的特征表达，如公式(14)所示：

$$\boldsymbol{F}_1=\sum_{i=1}^{3}DWConv_{k=3,d=i}(Conv_{1\times1}(\boldsymbol{X})). \tag{14}$$

式中，$\boldsymbol{X}$为输入特征图，d为空洞率。采用三组具有不同空洞率的深度可分离卷积进行融合，从而获取不同感受野下的上下文信息。为引入空间方向的结构信息，RFA模块采用方向感知的池化机制，分别沿水平方向与垂直方向对$\boldsymbol{F}_1$进行全局平均池化操作，生成方向特征映射，如公式(15)所示：

$$\boldsymbol{F}_{dir}=Concat\left(AvgPool_W(\boldsymbol{F}_1),AvgPool_H(\boldsymbol{F}_1)\right). \tag{15}$$

方向注意力权重的生成过程如下：首先，对方向感知特征$\boldsymbol{F}_{dir}$运用两个串联的1×1卷积和ReLU激活函数，以此实现通道压缩与非线性映射，从而得到中

间特征 $\boldsymbol{F}_{attn}$。随后，将该中间特征在通道维度上进行分割，得到两个方向注意力图 $\boldsymbol{M}_h$、$\boldsymbol{M}_w$，它们分别表示高度和宽度维度的响应分布。最终，通过融合方向权重，增强特征图中关键空间位置与结构方向的响应能力，如公式(16)所示：

$$\boldsymbol{F}_{out} = \boldsymbol{F}_{in} \odot (\boldsymbol{M}_h \otimes \boldsymbol{M}_w). \tag{16}$$

TADF模块通过任务向量驱动的TGM模块得到融合权重，以此增强模态融合对任务的敏感性与判别性；再通过RFA模块完成多尺度上下文聚合，获取更丰富的背景上下文信息，同时利用方向感知能力强化网络对关键语义区域的响应。该模块充分挖掘了不同模态间的互补性与结构信息，最终提升了融合特征的表达能力与鲁棒性。

1.4 损失函数优化

CIoU损失函数通过度量预测框与目标框之间的差异，解决IoU在边界框无重叠时梯度消失的问题。CIoU公式如(17)所示：

$$\mathcal{L}_{CIoU} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + av. \tag{17}$$

式中，IoU表示预测框与真实框的交并比；ρ表示两中心点之间的欧氏距离；c表示包围预测框和真实框的最小闭包框的对角线长度；α表示调节因子，控制宽高比一致性项v在总损失的影响；v为度量预测框与真实框宽高比之间差异的项。

尽管CIoU损失函数在提升定位精度和收敛速度方面相较于IoU具有明显优势，但其仍存在不足。首先，中心点距离项在训练过程中常常主导梯度更新，致使在高IoU阶段优化效果减弱。其次，宽高比惩罚项所依赖的动态因子数值较小，约束作用有限，难以充分调整预测框的形状。此外，CIoU对于低IoU或困难样本缺乏专门的加权机制，从而影响其学习能力和鲁棒性。

为缓解CIoU损失函数中角度惩罚项对宽高回归约束弱、梯度难以精确指引的问题，EIoU将宽高误差显式拆解为欧氏差值形式，从而实现梯度方向更明确、优化更高效的边界框拟合。然而，EIoU在处理小目标、极端比例目标或预测框不稳定时，仍存在宽高归一方式粗糙、各项权重固定的问题，导致其在复杂目标场景下性能受限。

为进一步提升EIoU损失函数对多尺度目标和复杂长宽比目标的适应能力，本文在其宽高解耦回归思想的基础上，引入尺度自适应和长宽比自适应调节机制，构建AN-EIoU损失函数。对于第i个预测框及其匹配的真实框，设w_i、h_i分别表示预测框的宽和高，w_i^{gt}、h_i^{gt}分别表示对应真实框归一化后的宽和高。下列公式中的w_i^{gt}、h_i^{gt}均指同一匹配真实框的宽度和高度。

首先，引入尺寸归一化因子$\omega(i)$，用于平衡不同长宽比目标在宽高误差项中的影响：

$$\omega(i) = \sqrt{\frac{w_i^{gt}}{h_i^{gt}}} + \epsilon. \tag{18}$$

式中，ϵ为平滑项，用于避免极小目标或分母过小导致的数值不稳定问题。该因子基于真实框长宽比构建，能够在保留目标形状差异信息的同时，缓解极端长宽比对宽高回归过程造成的梯度波动。

其次，引入中心偏移调节因子$\lambda(i)$，用于自适应调整中心点距离误差的权重：

$$\lambda(i) = \frac{1}{1 + w^{gt} \cdot h^{gt}}. \tag{19}$$

其中，$w^{gt} \cdot h^{gt}$表示真实框的归一化面积。对于小目标，$\lambda(i)$取值相对较大，可增强中心点偏移误差的约束作用；对于大目标，$\lambda(i)$取值相对较小，可避免中心距离项对整体损失产生过强影响，从而提升多尺度目标训练的均衡性。

进一步地，引入长宽比调节因子$\mu(i)$，用于增强模型对非正方形目标形状误差的感知能力：

$$\mu(i) = 1 + \left| \frac{w^{gt} - h^{gt}}{w^{gt} + h^{gt}} \right|. \tag{20}$$

当目标框接近正方形时，$\mu(i)$接近1，宽高误差项保持常规约束强度；当目标呈现细长或非规则比例形态时，$\mu(i)$随长宽差异增大而增大，从而提高宽高误差项权重，使模型更加关注复杂长宽比目标的形状回归精度

综合上述调节因子，本文提出的AN-EIoU损失函数定义如下：

$$\mathcal{L}_{AN-EIoU} = (1 - IoU) + \lambda(i)\frac{\rho^2}{c^2} + \mu(i)\left[\left(\frac{w_i - w_i^{gt}}{\omega(i)}\right)^2 + \left(\frac{h_i - h_i^{gt}}{\omega(i)}\right)^2\right]. \tag{21}$$

式中，第一项$1 - IoU$用于衡量预测框与真实框的重叠程度；第二项为经$\lambda(i)$加权的中心点距离误差项，用于增强不同尺度目标的中心定位能力；第三项为经$\omega(i)$和$\mu(i)$调节的宽高误差项，用于提高模型对不同长宽比目标的形状回归能力。

与现有动态损失函数主要依据IoU值、样本难度或训练阶段对整体损失进行统一加权的方式不同，AN-EIoU从目标框自身几何属性出发，对中心距离项和宽高误差项进行结构化调节。其中，$\lambda(i)$基于目标面积实现尺度感知加权，使小目标获得更强的位置约束；$\mu(i)$基于长宽比偏离程度增强细长目标和非规则比例目标的形状约束；$\omega(i)$则对宽高误差进行平滑归一化，降低极端长宽比对优化稳定性的影响。因此，AN-EIoU在保留EIoU宽高解耦优势的同时，进一步提升了损失函数对目标尺度差异和形状差异的自适应建模能力，有助于提高模型在多尺度及复杂长宽比目标检测任务中的定位精度与收敛稳定性。

2 实验与结果分析

2.1 实验数据集

M3FD数据集（Liu等，2022）聚焦于复杂交通环境下多类别目标的检测，尤其针对雾天、夜晚等视野受限条件下的应用需求。该数据集包含4，200对配准的可见光与红外图像，覆盖白天、阴天和夜间三类典型光照环境。目标类别包括行人、汽车、公交车、摩托车、卡车和信号灯，共计六类。在图像场景中，存在大量因车灯、信号灯导致的过曝区域，还有浓雾、光照变化等干扰因素。这使得该数据集在目标定位、类别识别及环境适应性方面具备较高的实用价值。

LLVIP数据集（Jia等，2021）是一套面向低光照条件下行人检测任务的双模态图像数据集，共包含30976张图像，即15488对已配准的可见光与红外图像，采集自24个夜间场景和2个白天场景。场景中普遍存在光照不足、遮挡、背景干扰等复杂情况，使得该数据集在弱光环境下的目标检测研究中具有较强的代表性和挑战性。

M3FD强调多类别、多尺度目标在恶劣环境中的可检测性与完整性；LLVIP则突出夜间弱光下的模态互补问题，适用于单类目标在特殊光照条件下的分析。两者互为补充，能够为低光照与复杂气象条件下的视觉感知任务提供数据支撑。

2.2 实验环境

实验基于PyTorch深度学习框架构建算法，其中PyTorch版本为2.4.1，CUDA版本是12.1，实验环境如表1所示。

模型训练的超参数设置如下：训练轮数为800，批次大小为16，输入图像尺寸为640×640像素，初始学习率为0.01，最终学习率为0.0001。

表1 实验环境设置

Tab. 1 Experimental environment setup

名称	参数
操作系统	Ubuntu 20.04
GPU	NVIDIA GeForce RTX 2080Ti
处理器	Intel(R) Xeon(R) Gold6240 CPU@2.60GHz
内存	12GB
Python	3.11

2.3 评价指标

为从图像层面对双模态融合结果的有效性进行辅助验证，本文采用多种客观评价指标对融合质量进行定量分析。具体包括信息熵（EN，Entropy）、空间频率（SF，Spatial Frequency）、源图像相关差异（SCD，Source Image Correlation Difference）、视觉信息保真度（VIF，Visual Information Fidelity）、基于梯度的融合质量指标（Q_{abf}，Edge-Based Fusion Quality Metric）以及结构相似性指数（SSIM，Structural Similarity Index）。其中，EN用于衡量融合图像的信息丰富度，SF反映图像清晰度与细节变化强度，SCD表征融合结果对双源互补信息的整合能力，VIF从人类视觉感知角度评价融合结果对有效视觉信息的保真程度，Q_{abf}用于衡量源图像边缘与梯度信息向融合图像传递的质量，而SSIM则用于评估融合结果对源图像结构信息的保持能力。上述指标从不同侧面对融合结果进行刻画，可为后续检测性能提升的原因分析提供支撑。

采用平均检测精度均值（mAP，Mean Average Precision）作为模型的评价指标。mAP主要用于综合量化检测结果在所有类别上检索精度（precision）与召回率（recall）之间的平衡。本质上是对不同类别的平均精度（AP，Average Precision）取算术平均，用以反映模型在多类别检测任务上的整体表现。AP和mAP的计算如公式（22）和（23）所示：

$$AP = \int_0^1 p(r)\mathrm{d}r . \tag{22}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}. \tag{23}$$

式中，$p(r)$为P-R曲线，N为类别总数。

2.4 融合质量对比分析

为从图像层面进一步验证本文所提出任务驱动双模态自适应融合机制的有效性，本文在M3FD与LLVIP数据集上对融合结果进行了客观指标与视觉效果评价，并与Tardal（Liu等，2022）、PIAFusion（Tang等，2022）、CDDFuse（Zhao等，2023）、IRFS（Wang等，2023）、TDFusion（Bai等，2025）及DEYOLO-n（Chen等，2025）等方法进行对比。所列方法基于其公开代码或论文描述，在统一实验环境下进行复现。所有方法均采用一致的数据预处理流程、评价指标及测试集划分，以保证对比结果的公平性与客观性。

本文提出的融合方法主要在特征层实现可见光与红外信息的融合，不直接输出像素级融合图像。由于红外与可见光图像融合任务通常缺乏统一的融合真值标准，因此未采用融合真值图像进行监督。模型训练主要依赖目标检测的标注信息（包括类别标签与边界框）；在融合质量评价阶段，则以配准后的可见光图像与红外图像为源图像，采用EN、SF、SCD、VIF、Q_{abf}、SSIM等指标开展客观评价。

表2 M3FD数据集上的融合质量结果

Table 2 Fusion quality results on the M3FD dataset

Method	EN	SF	SCD	VIF	Q_{abf}	SSIM
Tardal（Liu等，2022）	6.87	7.63	1.29	0.27	0.3	0.71
PIAFusion（Tang等，2022）	6.28	10.5	1.47	0.39	0.48	0.69
CDDFuse（Zhao等，2023）	7.28	9.9	1.47	0.39	0.48	0.74
IRFS（Wang等，2023）	6.95	12.3	1.6	0.37	0.57	0.71
TDFusion（Bai等，2025）	6.99	14.49	1.83	0.4	0.65	0.72
DEYOLO-n（Chen等，2025）	7.02	13.9	1.78	0.42	0.63	0.72
Ours	7.05	15.2	1.88	0.43	0.67	0.73

同时，考虑到无参考指标或依赖源图像的评价指标难以全面反映融合图像的视觉质量，本文进一步对融合结果进行了视觉对比分析，从目标显著性、纹理细节、边缘结构及场景信息完整性等方面进行综合评价，以辅助验证所提出特征级融合机制的有效性。模型训练完成后，从TADF模块提取双模态融合特征，经通道压缩、上采样及归一化映射生成可视化融合图像，用于融合质量评价。该过程仅用于融合效果分析，不参与目标检测训练，也未引入额外的像素级融合监督。

从表2的结果可以看出，在M3FD数据集上，本文方法在SCD、SF及Q_{abf}等指标上取得较优结果，表明其在跨模态信息整合与结构细节保持方面具有良好性能。这主要得益于DLA与TADF模块的协同作用，DLA在融合前增强单模态特征的结构表达并抑制噪声，而TADF通过任务语义引导实现双模态特征的自适应融合，从而提升边缘与结构信息的保持能力。

为进一步展示不同方法在复杂低光照场景下的融合效果，本文选取M3FD数据集中的典型道路场景进行可视化对比，结果如图4所示。为更清晰地比较不同方法在目标边缘保持方面的差异，图4进一步对右侧行人区域进行了局部放大和箭头标注。从图中可以看出，可见光图像能够提供道路、车辆、树木及背景结构等纹理信息，但受雨雾和低照度影响，远处目标及边缘细节不够清晰；红外图像虽能突出车辆和行人等热辐射显著目标，但场景纹理与颜色信息相对缺失。Tardal、PIAFusion、CDDFuse、IRFS、TDFusion、DEYOLO-n等对比方法虽然能够融合部分红外显著信息，但在放大区域中仍存在行人边界模糊、局部亮度扩散或目标轮廓不完整等问题。相比之下，本文方法在保持行人目标显著性的同时，能够获得更加完整的人体轮廓和更清晰的目标边缘，并较好地保留场景结构信息，使融合结果具有更高的目标辨识度。

从表3的结果可以看出，在LLVIP低光照数据

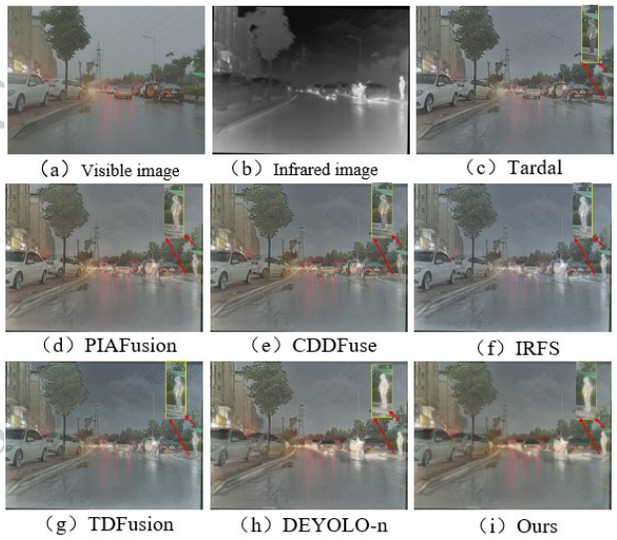

图4　M3FD数据集上不同方法融合结果的视觉对比示例

Fig. 4　Visual comparison examples of fusion results of different methods on the M3FD dataset

集上，本文方法在EN、VIF及Q_{abf}等指标上同样表现较优，说明在弱光条件下具有较好的信息表达能力与视觉一致性。其原因在于，低光照条件下可见光图像容易出现纹理退化和目标不清晰的问题，而红外图像能够提供稳定的目标响应，TADF能够动态提升红外特征的融合权重，以弥补可见光信息退化的影响。同时DLA模块进一步增强特征表达的稳定性，使融合结果在复杂弱光环境下保持更可靠的目标信息和结构信息。

综合两个数据集的结果可以看出，本文方法能够根据不同场景下的模态信息质量，自适应调整融合策略，在提升信息表达能力的同时强化与检测任务相关的判别性特征，从而为后续目标检测提供更加有效的输入。

为进一步验证本文方法在弱光行人场景中的融合表现，本文选取LLVIP数据集中的典型夜间样本进行融合结果可视化对比，结果如图5所示。

为突出不同方法在行人目标保持方面的差异，图5对左上方行人区域进行了局部放大和箭头标注。从图中可以看出，在夜间低照度条件下，可见光图像中的行人目标亮度较低，目标与背景的对比度较弱；红外图像能够突出行人的热辐射区域，但对交通信号灯、道路边缘及建筑背景等可见光细节的呈现不足。Tardal、PIAFusion、CDDFuse、IRFS、TDFusion、DEYOLO-n等方法虽能增强行人目标响应，但在放大区域中仍存在行人轮廓模糊、边界扩散或局部结构细节不清晰等问题。相比之下，本文方法能够在突出行人目标的同时，更好地保持人体边缘和目标结构，使融合结果兼具红外目标显著性与可见光细节表达能力。

上述不同场景下的融合结果表明，本文方法能够根据低光照场景下各模态信息质量进行自适应融合，从而为后续检测任务提供更具判别性的融合表示。

2.5　消融实验

为验证本文所提出的改进模型的合理性，并评估各模块对双模态检测性能的贡献，本文在M3FD和LLVIP两个数据集上开展了一系列消融实验。实验采用不同模块组合的方式，逐步引入DLA、TADF和AN-EIoU模块，并以“√”标注所采用的改进项。需要指出的是，DLA与TADF主要作用于双模态特征建模与融合过程，而AN-EIoU用于优化检测阶段的边界框回归。实验结果如表4所示。

从表4可以看出，在单模块配置下，DLA、TADF

表3　LLVIP数据集上的融合质量结果

Table 3　Fusion quality results on the LLVIP dataset

Method	EN	SF	SCD	VIF	Q_{abf}	SSIM
Tardal(Liu等,2022)	6.32	7.42	1.04	0.27	0.22	0.58
PIAFusion(Tang等,2022)	7.1	14.8	1.55	0.39	0.68	0.69
CDDFuse(Zhao等,2023)	7.2	13	1.45	0.45	0.6	0.7
IRFS(Wang等,2023)	6.92	13.3	1.5	0.41	0.61	0.64
TDFusion(Bai等,2025)	7.36	16.38	1.75	0.46	0.62	0.7
DEYOLO-n(Chen等,2025)	7.3	13.6	1.7	0.45	0.64	0.69
Ours	7.4	16	1.86	0.47	0.71	0.69

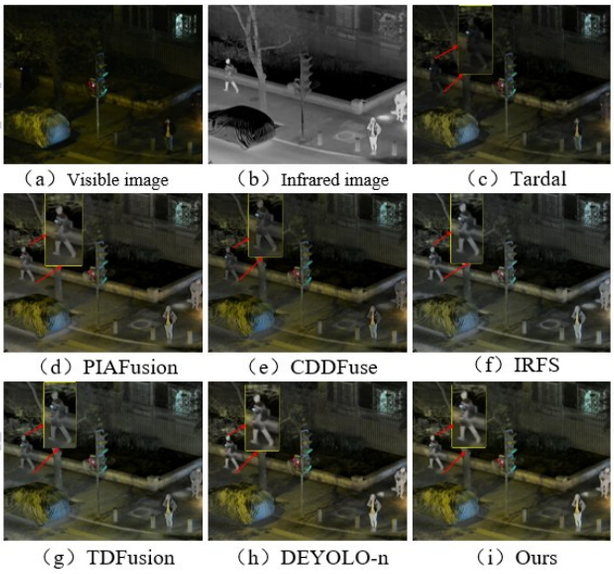

图5　LLVIP数据集上不同方法融合结果的视觉对比示例

Fig. 5　Visual comparison examples of fusion results of different methods on the LLVIP dataset

和AN-EIoU均能在不同程度上提升模型检测性能。其中，仅引入DLA后，模型在M3FD和LLVIP数据集上的性能均有所提升，说明该模块能够有效增强融合前单模态特征表达。结合1.2节的结构分析可知，DLA通过动态非线性压缩稳定特征响应，并借助自注意力机制建模空间依赖关系，共同提升融合前特征质量。TADF单独引入后，模型在两个数据集上的检测精度也得到提升，说明任务语义引导的双模态融合能够增强模型对低光照目标的表征能力。引入AN-EIoU主要作用于边界框回归过程，与基线模型相比，单独引入AN-EIoU后，检测精度均有所提升，说明该损失函数能够在一定程度上改善目标定位精度。

双模块组合进一步体现了模块间的互补特性。引入DLA与TADF后，模型在两个数据集上的检测性能均显著提升，表明空间特征增强与双模态融合之间具有良好的协同作用。同时，TADF与AN-EIoU在低光照场景下表现更优，说明融合优化与定位优化的结合有助于提升复杂环境中的检测能力。

表4　本文算法在不同数据集上的消融实验结果

Tab. 4　Ablation experiment results of the proposed algorithm on different datasets

Models			mAP@0.5(%)	mAP@0.5(%)	mAP@0.5-0.95(%)	mAP@0.5-0.95(%)
DLA	TADF	AN-EIoU	M3FD	LLVIP	M3FD	LLVIP
			80.8	95.4	54.3	58.4
√			84.4	95.9	55.8	60.2
	√		85.2	96.2	56.2	61.6
		√	83.5	96.1	55.6	60.5
√	√		88.6	96.5	62.2	62.7
√		√	86.9	96,3	61.0	61.6
	√	√	87.3	97.0	61.2	64.1
√	√	√	89.3	97.8	62.6	68.3

当三个模块联合使用时，模型在M3FD和LLVIP数据集上均取得最佳性能。结果表明，DLA、TADF和AN-EIoU分别作用于特征表达、跨模态融合和目标定位阶段，通过分层协同设计有效提升了双模态目标检测的整体性能。

为进一步验证DLA模块内部组件设计的合理性，本文在完整模型基础上对DLA模块进行了内部消融实验。实验中固定使用TADF与AN-EIoU，仅调整DLA模块的内部结构，其余训练参数、数据划分、输入尺寸及评价方式均保持一致。去除动态非线性压缩机制的模型记为DLA-1，去除自注意力机制的模型记为DLA-2，完整的DLA模块同时包含动态非线性压缩机制与自注意力机制。实验结果如表5所示。

从表5可见，完整DLA模块在M3FD和LLVIP数据集上的检测结果相对更优。与完整DLA相比，去除动态非线性压缩机制的DLA-1在两组数据集上的mAP指标均有一定下降，这表明该机制有助于缓

解低光照与噪声干扰导致的异常特征响应，提升特征表达稳定性。而去除自注意力机制的DLA-2同样造成检测性能下降，说明自注意力机制能够有效建模空间依赖关系，增强目标边缘与显著区域的判别性表达。

表5　M3FD和LLVIP数据集上的DLA模块消融实验结果

Tab. 5　Ablation results of the DLA module on the M3FD and LLVIP datasets

Models	mAP@0.5(%)	mAP@0.5(%)	mAP@0.5-0.95(%)	mAP@0.5-0.95(%)
	M3FD	LLVIP	M3FD	LLVIP
DLA	89.3	97.8	62.6	68.3
DLA-1	88.7	97.5	62.2	67.6
DLA-2	88.4	97.3	62.0	67.4

综合来看，动态非线性压缩机制与自注意力机制分别从特征响应调节和空间关系建模两个维度对DLA模块形成补充，二者结合可进一步提升融合前单模态特征的表达质量。

2.6　TADF注意力可视化分析

为进一步验证TADF模块中任务语义引导机制的有效性，本文对任务语义引导下的注意力响应进行可视化分析。结果如图6所示。

具体而言，将检测类别文本作为任务提示输入文本编码器，得到对应类别语义嵌入，并将其作为任务查询向量与双模态视觉特征进行注意力交互，获得任务相关的空间响应。随后，将注意力权重映射回图像空间并进行归一化处理，得到不同文本提示下的注意力热力图。

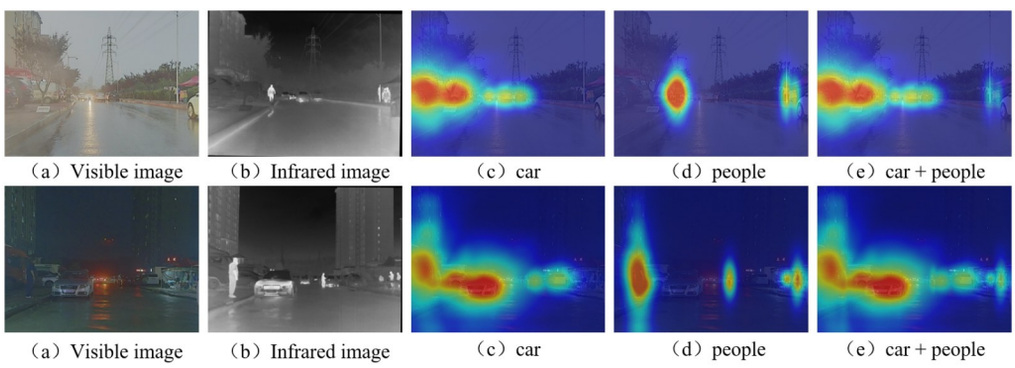

图6　不同任务文本提示下TADF模块的注意力热力图

Fig. 6　Attention heatmaps of the TADF module under different task prompts

从图6可见，当文本提示为“car”时，注意力响应主要集中于车辆轮廓、车灯及道路中的车辆分布区域，说明模型可针对车辆检测任务增强车辆相关的纹理与热辐射显著特征；当文本提示为“people”时，注意力响应更多聚焦于行人所在位置及其热辐射显著区域，表明TADF模块能够借助红外模态在低光照条件下对人体目标的突出表达能力；当文本提示为“car+people”时，注意力区域同时覆盖车辆与行人目标区域，说明任务语义可引导模型根据不同目标类别动态调整关注区域。

上述可视化结果表明，TADF模块并非对可见光与红外特征进行固定权重融合，而是能在任务语义引导下，自适应地聚焦与当前检测目标相关的模态信息及空间区域。该结果不仅进一步验证了任务感知双模态融合机制的有效性，也从可解释性角度揭示了TADF模块提升低光照场景目标检测性能的原因。

2.7　对比实验

为进一步验证所提出的TDAFNet在目标检测性能方面的优势，选取了近年来的双模态检测模型CFT(Fang等，2021)，CSAA(Cao等，2023)，MS-DETR(Xing等，2024)，CFMW(Li等，2024)及DEYOLO-n

进行比较。为确保比较的客观性和公平性，所有方法的检测头均采用YOLOv8的检测头，并使用相同的超参数配置重新训练以获取mAP值。实验结果如表6和表7所示。

表6　M3FD数据集上的对比试验结果

Tab. 6　Comparative experimental results on the M3FD dataset

Method	Input	mAP@0.5(%)	mAP@0.5-0.95(%)
YOLOv8n	RGB+IR	79.2	52.8
CFT(Fang等,2021)	RGB+IR	87.9	60.7
CSAA(Cao等,2023)	RGB+IR	85.1	58.2
MS-DETR(Xing等,2024)	RGB+IR	88.9	61.4
CFMW(Li等,2024)	RGB+IR	89.0	62.1
DEYOLO-n(Chen等,2025)	RGB+IR	86.6	58.9
Ours	RGB+IR	89.3	62.6

从表6和表7的实验结果分析，所提出的TDAFNet算法在M3FD和LLVIP两个数据集上均实现了较优的检测性能。在M3FD数据集上，TDAFNet的mAP@0.5和mAP@0.5:0.95分别为89.3%和62.6%。与性能较好的CFMW方法相比，分别提升了0.3%和0.5%；相较于DEYOLO-n方法，提升幅度为2.7%和3.7%。整体来看，该方法在该数据集上取得了最优结果，但相较于部分先进方法的提升幅度相对有限。在LLVIP数据集上，TDAFNet的mAP@0.5和mAP@0.5:0.95分别达到97.8%和68.3%，在所有对比方法中表现最佳。与次优的DEYOLO-n方法相比，分别提升了1.0%和2.9%，表明该方法在不同场景下具有较好的适应能力。

表7　LLVIP数据集上的对比试验结果

Tab. 7　Comparative experimental results on the LLVIP dataset

Method	Input	mAP@0.5(%)	mAP@0.5-0.95(%)
YOLOv8n	RGB+IR	93.6	64.1
CFT(Fang等,2021)	RGB+IR	93.1	58.8
CSAA(Cao等,2023)	RGB+IR	92.3	55.8
MS-DETR(Xing等,2024)	RGB+IR	94.2	62.1
CFMW(Li等,2024)	RGB+IR	94.8	64.6
DEYOLO-n(Chen等,2025)	RGB+IR	96.8	65.4
Ours	RGB+IR	97.8	68.3

从结果来看，虽然近年来提出的双模态检测方法均在一定程度上提升了检测精度，但这些方法在高IoU阈值(mAP@0.5:0.95)下的性能仍存在不足。而TDAFNet在该指标上表现出明显优势，表明其在高精度定位和细粒度目标识别方面具有更强的能力。

这一性能优势主要源于任务驱动的双模态特征深度融合策略与细粒度特征增强模块的协同效应。通过充分挖掘可见光和红外图像的互补信息，丰富了特征表达的多样性与完整性；同时，改进后的特征增强模块有效抑制了模态间冗余信息的干扰，提高了在低光照场景下对双模态特征的综合利用能力，进一步优化了检测结果的准确性和稳定性。

2.8　计算复杂度比较

为进一步分析所提方法的计算开销，本文统计了模型参数量(Params)、计算复杂度(GFLOPs)及推理速度，结果见表8。为确保推理速度评价的可比性，FPS测试在与第2.2节一致的实验环境中进行，

测试设备为NVIDIA GeForce RTX 2080Ti GPU，输入图像尺寸统一设为640×640，batch size为1。对于TADF模块中引入的CLIP文本编码器，本文仅将其用作类别语义先验提取模块，推理阶段的类别文本语义特征已预先计算并缓存，因此表中FPS主要反映双模态图像输入后的特征提取、跨模态融合及检测预测耗时，不包含离线文本编码过程。

从表8可见，与基线模型YOLOv8n相比，本文方法引入DLA与TADF模块后，参数量由4.2M增至12.6M，GFLOPs从18.1升至56.6，推理速度则从179.5帧/秒降至41.3帧/秒，计算开销有所增加。但相较于CFT、CFMW等其他双模态方法，本文方法的参数规模与计算复杂度仍处于较低水平。

表8　计算开销比较

Table 8　Comparison of computational cost

Method	Params (M)	GFLOPs	mAP@0.5(%)	FPS（帧/S）
YOLOv8n	4.2	18.1	79.2	179.5
CFT	206.1	240.5	87.9	—
CSAA	86.6	88.2	85.1	—
MS-DETR	67.2	133.4	88.9	—
CFMW	181.3	259.1	89.0	—
DEYOLO-n	8.3	22	86.6	136.1
Ours	12.6	56.6	89.3	41.3

进一步分析表明，参数量与计算量的增长主要源于双模态特征融合及特征增强模块的引入，不过这些模块在结构设计上相对轻量，因此在提升特征表达能力的同时，避免了计算成本的显著上升。综合精度与复杂度结果可知，本文方法在计算开销可控的前提下，实现了检测性能的有效提升，说明所提模型在实际应用中具备较好的可行性。

2.9　检测结果可视化对比

为了直观评估改进模型在双模态低光照场景中的鲁棒性与泛化能力，本文在相同参数配置下对检测结果进行了可视化对比分析。具体而言，分别使用基线模型YOLOv8和本文提出的方法，对M3FD与LLVIP两个数据集中的不同样本进行目标检测，并通过结果对比分析两者的性能差异。相关可视化结果如图7所示。

从图7可以看出，原始模型在低对比度、弱纹理或存在遮挡的低光照场景中存在目标漏检的问题。而改进后的模型，在目标边界不清晰或背景干扰较强的场景下，能够保持较为稳定的检测性能。与原始模型相比，改进模型的漏检情况明显减少。

3　结　论

本文针对低光照场景下多尺度目标检测的挑战以及可见光与红外模态信息融合不充分的问题，提出了TDAFNet双模态目标检测模型。通过在特征增强、双模态融合和损失函数设计等核心环节的系统性改进，有效提升了模型在复杂光照条件下的检测性能。具体而言，本文的主要贡献与结论如下：

（1）在特征提取方面，引入DLA机制，通过动态局部注意力与多路径特征扩展策略的协同作用，显著增强了Backbone对多尺度目标和复杂背景的特征表达能力，为后续融合与检测提供了更丰富的判别性特征。

（2）在跨模态融合方面，设计的TADF方法实现了可见光与红外特征的自适应动态融合，有效解决了传统融合方法中信息利用不充分、融合权重固定等问题。融合质量对比结果验证了所提融合策略在保持纹理细节和结构信息方面的优势。

（3）在损失函数优化方面，提出的AN-EIoU损失通过尺度归一化与自适应权重机制，提升了模型对多尺度目标及非规则几何目标的定位精度，与所提特征增强和融合模块形成了有效互补。

实验结果表明，TDAFNet在M3FD和LLVIP两个具有挑战性的低光照双模态数据集上，检测精度优于现有对比方法，且融合质量指标与主流融合方法相比具备竞争力。这充分验证了本文模型在低光照环境下联合利用可见光与红外信息进行高精度目标检测的有效性和鲁棒性。未来的工作将探索模型轻量化策略以提升部署效率，并进一步研究时序信息融合以应对视频流中的连续检测任务。

（a） YOLOV8模型在M3FD数据集上的检测结果示例

（a） Example of detection results of the YOLOV8 model on the M3FD dataset

（b） TDAFNet模型在M3FD数据集上的检测结果示例

（b） Example of detection results of the TDAFNet model on the M3FD dataset

（c） YOLOv8模型在LLVIP数据集上的检测结果示例

（c） Example of detection results of the YOLOv8 model on the LLVIP datase

（d） TDAFNet模型在LLVIP数据集上的检测结果示例

（d） Example of detection results of the TDAFNet model on the LLVIP dataset

图 7 不同方法在M3FD数据集和LLVIP数据集上的检测效果对比

Fig. 7 Comparison of detection performance of different methods on the M3FD and LLVIP datasets

参考文献

Yi X P, Ma Y, Li Y S, Xu H, Ma J Y. 2025. Artificial intelligence facilitates information fusion for perception in complex environments. The Innovation, 6(4): 100814 [DOI: 10.1016/ j.xinn.2025.100814]

Tang L F, Zhang H, Xu H and Ma J Y. 2023. Deep learning-based image fusion: a survey. Journal of Image and Graphics, 28(1): 3-36 [DOI: 10.11834/jig.220422] (唐霖峰, 张浩, 徐涵, 马佳义. 2023. 基于深度学习的图像融合方法综述. 中国图象图形学报, 28(1): 3-36) [DOI: 10.11834/jig.220422]

Zhang J L, Gong W H, Jia X. 2023. Object detection algorithm with dual-modal rectification fusion based on self-guided attention. Pattern Recognition and Artificial Intelligence, 36(9): 793-805 [DOI: 10.16451/j.cnki.issn1003-6059.202309003] (张惊雷, 宫文浩, 贾鑫. 2023. 基于自引导注意力的双模态校准融合目标检测算法. 模式识别与人工智能, 36(9): 793-805). DOI: 10.16451/j.cnki.issn1003-6059.202309003

Li Z H, Dong Y S, Shen L C, Liu Y F, Pei Y H, Yang H T, Zheng L T, Ma J W. 2024. Development and challenges of object detection: a survey. Neurocomputing, 598: 128102 [DOI: 10.1016/j.neucom.2024.128102]

Girshick R. 2015. Fast R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-

1448[DOI:10.1109/ICCV.2015.169]

He K M, Gkioxari G, Dollár P, Girshick R. 2017. Mask R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988[DOI:10.1109/ ICCV.2017.322]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. 2016. SSD: Single Shot MultiBox Detector//Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9905. Cham: Springer International Publishing: 21-37[DOI:10.1007/978-3-319-46448-0_2]

Zhang X Y, He Y X, Yang Z Y. 2021. Pedestrian detection based on simplified YOLOv3. International Core Journal of Engineering, 7(12): 21-28[DOI:10.6919/ICJE.202112_7(12).0003]

Chen Z X, Tian S W, Yu L, Zhang L Q, Zhang X Y. 2022. An object detection network based on YOLOv4 and improved spatial attention mechanism. Journal of Intelligent & Fuzzy Systems, 42(3): 2359-2368[DOI:10.3233/JIFS-211648].

Shen Y, Huang C H, Huang F, Li J, Zhu M J, Wang S. 2021. Research progress of infrared and visible image fusion technology. Infrared and Laser Engineering, 50(9): 202004[DOI: 10.3788/IRLA20200467](沈英,黄春红,黄峰,李杰,朱梦娇,王舒. 2021. 红外与可见光图像融合技术的研究进展. 红外与激光工程,50(9):202004[DOI:10.3788/IRLA20200467]

Xu J and He X. 2024. DAF-Net: a dual-branch feature decomposition fusion network with domain adaptive for infrared and visible image fusion[EB/OL].[2025-11-03]. https://arxiv.org/pdf/ 2409. 11642.

Chen S Q, Wang C Q, Zhou Y J. 2022. A pedestrian detection method based on YOLOv5s and image fusion. Electro-Optic Control, 29(7): 96-101, 131[DOI:10.3969/j.issn.1671-637X.2022.07.018](陈世权,王从庆,周勇军. 2022. 一种基于YOLOv5s和图像融合的行人检测方法. 电光与控制,29(7):96-101),131.[DOI:10.3969/j.issn.1671-637X.2022.07.018]

Cao Z W, Yang H H, Zhao J, Guo S H and Li L Q. 2021. Attention fusion for one-stage multispectral pedestrian detection. Sensors, 21(12): 4184[DOI:10.3390/s21124184]

Han K, Wang Y, Tian Q, Guo J, Xu C and Xu C. 2020. GhostNet: more features from cheap operations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 1577-1586[DOI: 10.1109/CVPR42600.2020.00165]

Sun W C, Hu X G, Yang Z F, Meng F Y and Sun W. 2024. Optimization of infrared-visible road target detection by fusing GPNet and image multiscale features. Journal of Jilin University (Engineering and Technology Edition), 54(10): 2799-2806[DOI: 10.13229/j.cnki.jdxbgxb.20230474](孙文财,胡旭歌,杨志发,孟繁雨,孙微. 2024. 融合GPNet与图像多尺度特性的红外-可见光道路目标检测优化方法. 吉林大学学报(工学版),54(10):2799-2806). DOI:10.13229/j.cnki.jdxbgxb.20230474.

Yin M, Sui Y, Liao S and Yuan B. 2021. Towards efficient tensor decomposition-based DNN model compression with optimization framework//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 10669-10678[DOI:10.1109/CVPR46437.2021.01053]

Xue Z, Zhang L L and Liu J. 2023. The multiObject detection method of fused images based on improved YOLOv7. Journal of Ordnance Equipment Engineering, 44(6): 166-172[DOI: 10. 11809/bqzbgcxb2023.06.023](薛震,张亮亮,刘吉. 2023. 基于改进YOLOv7的融合图像多目标检测方法. 兵器装备工程学报,44(6):166-172). DOI:10.11809/bqzbgcxb2023.06.023.

Liu J, Shang J, Liu R and Fan X. 2022. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. IEEE Transactions on Circuits and Systems for Video Technology, 32(8): 5026-5040[DOI: 10.1109/TCSVT. 2022. 3144455]

Sun Y, Hou Z Q, Yang C, Ma T G and Fan J L. 2023. Object detection algorithm based on dual-modal fusion network. Acta Photonica Sinica, 52(1): (0110002[DOI:10.3788/gzxb20235201.0110002]孙颖,侯志强,杨晨,马泰刚,范九伦. 2023. 基于双模态融合网络的目标检测算法. 光子学报,52(1):0110002. DOI: 10.3788/gzxb20235201.0110002.

Aleerxia. 2025. Review on infrared small target detection based on deep learning. Computer Science and Application, 15(6): 220-230[DOI:10.12677/csa.2025.156172](阿勒尔呷. 2025. 基于深度学习的红外小目标检测方法综述. 计算机科学与应用,15(6):220-230). DOI:10.12677/csa.2025.156172.

Dai X B, Ren S Y, He X H, Qing L B and Chen H G. 2025. Visible-infrared fusion object detection based on feature enhancement and alignment fusion. Chinese Journal of Engineering, 47(12): 2554-2565[DOI:10.13374/j.issn2095-9389.2025.03.26.004](代小波,任思羽,何小海,卿琳波,陈洪刚. 2025. 基于特征增强和对齐融合的可见光-红外图像融合目标检测. 工程科学学报,47(12):2554-2565). DOI:10.13374/j.issn2095-9389.2025. 03.26.004.

Tang S Y, Gong R H, Wang Y, Liu A S, Wang J K, Chen X Y, Yu F W, Liu X L, Song D, Yuille A, Torr P H S and Tao D C. 2021. RobustART: benchmarking robustness on architecture design and training techniques[EB/OL].[2025-11-03]. https://arxiv.org/ pdf/2109.05211.

Liu A S, Liu X L, Yu H, Zhang C, Liu Q and Tao D C. 2021. Training robust deep neural networks via adversarial noise propagation. IEEE Transactions on Image Processing, 30: 5769-5781[DOI:10.1109/TIP.2021.3082317]

Liu J Y, Fan X, Huang Z B, Wu G Y, Liu R S, Luo Z X and Zhong W J. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 5802-5811[DOI:10.1109/CVPR52688.2022.00571]

Jia X Y, Zhu C, Li M Z, Tang W Q and Zhou W L. 2021. LLVIP: a

visible-infrared paired dataset for low-light vision//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Montreal, QC, Canada: IEEE: 567-576 [DOI: 10. 1109/ICCVW54120.2021.00389]

Tang L F, Yuan J T, Zhang H, Jiang X Y and Ma J Y. 2022. PIAFusion: a progressive infrared and visible image fusion network based on illumination aware. Information Fusion, 83-84: 79-92 [DOI: 10. 1016/j.inffus.2022.03.007]

Zhao Z X, Bai H W, Zhang J S, Zhang Y L, Xu S, Lin Z D, Timofte R and Van Gool L. 2023. CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE: 5906-5916 [DOI: 10.1109/CVPR52729.2023.00572]

Wang D, Liu J Y, Liu R S and Fan X. 2023. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. Information Fusion, 98: 101828 [DOI: 10. 1016/j.inffus.2023.101828]

Bai H W, Zhang J S, Zhao Z X, Wu Y F, Deng L, Cui Y, Feng T and Xu S. 2025. Task-driven image fusion with learnable fusion loss// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 7457-7468 [DOI: 10.1109/CVPR52734.2025.00699]

Chen Y S, Wang B R, Guo X Y, Zhu W B, He J S, Liu X B , et al. 2024. DEYOLO: dual-feature-enhancement YOLO for cross-modality object detection//Pattern Recognition. Lecture Notes in Computer Science. Cham: Springer: 236-252 [DOI: 10.1007/978-3-031-78447-7_16]

Fang Q Y, Han D P and Wang Z K. 2022. Cross-modality fusion transformer for multispectral object detection [EB/OL]. [2025-11-03]. https://doi.org/10.2139/ssrn.4227745.

Cao Y, Bin J, Hamari J, Blasch E and Liu Z. 2023. Multimodal object detection by channel switching and spatial attention//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, BC, Canada: IEEE: 403-411 [DOI: 10.1109/CVPRW59228.2023.00046]

Xing Y H, Yang S, Wang S, Zhang S Z, Liang G Q and Zhang X W. 2024. MS-DETR: multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization. IEEE Transactions on Intelligent Transportation Systems, 25(12): 20628-20642 [DOI: 10.1109/TITS.2024.3450584]

Li H Y, Hu Q, Zhou B J, Yao Y, Lin J C, Yang K L and Chen P. 2024. CFMW: cross-modality fusion mamba for robust object detection under adverse weather [EB/OL]. [2025-11-03]. https://arxiv. org/ pdf/2404.16302.

作者简介

马向阳，男，陕西西安人，硕士研究生，主要从事机器视觉及多源图像融合方面的研究。 E-mail: caotan2023@xauat.edu.cn

陈俊英，通信作者，女，内蒙古丰镇人，博士，副教授，硕士生导师。2004年于西安交通大学获得硕士学位，2010年于西安交通大学获得博士学位，现为西安建筑科技大学信息与控制工程学院教师，主要从事计算机视觉及机器学习方面的研究。 E-mail: chenjy@xauat.edu.cn

闫孟明，男，陕西西安人，硕士研究生，主要从事机器视觉及姿态识别方面的研究。 E-mail: 1287646723@qq.com

梁栋，男，河北省行唐县人，博士。2021年于西安电子科技大学获得博士学位，现为西安建筑科技大学信息与控制工程学院教师，主要从事计算机视觉及自然语言处理和机器学习方面的研究。 E-mail: dliang@xauat.edu.cn