引用格式：LIU Yuhang, LI Wenbo, YANG Ke, et al. Parallel Network for Dual-band Image Fusion and Object Detection with Lightweight Deployment[J]. Acta Photonica Sinica, 2026, 55(8):0810002
刘羽航，李文博，杨轲，等. 双光图像融合与目标检测并行网络及轻量化部署[J]. 光子学报，2026，55(8)：0810002

# 双光图像融合与目标检测并行网络及轻量化部署

刘羽航[1,2,3]，李文博[1,2,3]，杨轲[1,2,3]，孙彬[1,2,3]，汪子君[1,2,3]，蒋大钢[1,2,3]
（1 电子科技大学 航空航天学院，成都 611731）
（2 自适应光学全国重点实验室，成都 610209）
（3 飞行器集群智能感知与协同控制四川省重点实验室，成都 611731）

**摘　要**：双光图像融合与目标检测因其在遥感、安防等领域的应用价值受到广泛关注，然而现有方法多采用串行级联结构，存在顺序依赖、难以平衡任务间关系、计算复杂度高等问题。提出一种并行网络实现双光图像融合与目标检测的多任务学习与轻量化部署。首先，设计伪孪生特征提取器与通道注意力特征融合模块，实现任务适配的高效特征提取；其次，引入跨尺度跳跃连接以增强融合细节表征，并采用自适应任务损失权重缓解任务间的优化冲突；最后，通过结构化剪枝与混合量化对网络进行压缩，得到适用于边缘设备的轻量化模型。在公开数据集$M^3FD$上的对比实验表明，所提方法在保持较高检测与融合性能的同时，显著降低了参数量与计算开销。嵌入式平台上的部署结果进一步验证了该模型在效率与精度间的良好平衡，体现了其在实际应用中的有效性与实用性。

**关键词**：双光图像；图像融合；目标检测；多任务学习；模型轻量化
**中图分类号**：TP391.4　　**文献标识码**：A
**doi**：10.3788/gzxb20265508.0810002

## 0 引言

近年来，双光探测通过结合可见光与热红外成像的优势，能够显著提升复杂光照如夜间、雾霾条件下的环境感知能力，在自动驾驶、安防监控、应急救援等领域展现出重要应用价值[1]。其中可见光波段的RGB图像能够提供丰富的颜色和纹理信息，而红外热（Thermal，T）成像技术则对温度敏感且不受光照影响。双光图像具有天然的互补性，大量研究致力于融合双光图像信息以提升任务性能，尤其在实际应用中，面向机器视觉的目标检测[2-4]和人类视觉的图像融合[5-7]需求往往同时存在：一方面，自动化系统需要利用双光目标检测技术，在复杂环境下实现对关注目标更精准的分类与定位；另一方面，决策人员需要通过图像融合技术将双光图像合成为肉眼可以观测的单幅图像以拓展视觉感知范围，将目标检测结果标注在融合图像上将同时满足机器视觉和人类视觉的探测需求。

早期的研究将图像融合与目标检测视为两个独立的过程分别进行优化，忽略了任务间的内在关联。最近，越来越多的研究通过串行级联方式加强任务间的联系。TANG Linfeng等[8]提出任务驱动的图像融合网络，与语义分割网络级联，有效弥合了图像融合与高层视觉任务之间的鸿沟；LIU Jinyuan等[9]提出目标感知双对抗学习网络（Target-aware Dual Adversarial Learning Network, TarDAL），并基于对齐的RGB和热红外图像建立了广泛认可的基准数据集$M^3FD$。然而，简单的级联操作不足以解决上下游任务间的冲突问题，

**基金项目**：国家自然科学基金（62020106011），四川省自然科学基金（2025ZNSFSC0002），四川省科技计划（2022YFG0050），成都市科技项目（2025－XT00－00024－GX）
**第一作者**：刘羽航，202521100516@std.uestc.edu.cn
**通讯作者**：孙彬，sunbinhust@uestc.edu.cn
**收稿日期**：2026-01-31；**录用日期**：2026-07-13

因此一些研究开始关注在级联结构中引入高层语义信息指导图像融合。SUN Yiming等[10]提出检测驱动的红外与可见光图像融合网络，利用从目标检测网络中学习到的目标特异性信息指导图像融合；ZHAO Wenda等[11]引入元特征嵌入模型，提出联合融合与检测的学习框架；WU Yuhui等[12]提出语义级融合网络，避免了人工设计融合规则的局限性；LIU Xiaowen等[13]提出语义驱动的耦合网络，通过共享跨模态特征解决特征异质性问题；TANG Linfeng等[14]聚焦像素级融合与高层视觉任务的关联，将分割模型中的语义特征嵌入融合网络，验证了像素级融合方法对高层视觉任务的有效性；CHEN Xiaoxuan等[15]通过交互式融合模块加强融合网络与检测网络之间的关联，确保生成的融合图像更适合检测任务；YANG Zengyi等[16]提出指令驱动融合，根据下游任务的相关指令自适应调整基础融合网络的输出特征。以上串行级联架构能够生成更符合下游任务需求的图像融合结果，然而这种架构的顺序依赖也存在一些局限性。图像融合作为底层视觉任务，以像素重建为核心，而目标检测作为高层视觉任务，更注重抽象的语义信息，级联框架难以在单一流程中同时平衡像素保真度与高层表征能力；其次，两阶段的推理过程会增加计算成本，较难满足资源受限的边缘设备的部署需求[17]。

本文提出一种面向双光感知的并行多任务网络及其轻量化部署方案。首先，设计并行的图像融合与目标检测多任务网络，并引入结构化剪枝，在保持性能的同时显著降低计算复杂度；其次从特征冲突与任务冲突两个层面设计优化策略，通过跨模态跳跃连接增强特征，减少融合过程中的信息损耗；采用自适应权重损失函数，动态调节不同任务在训练中的贡献，实现多任务平衡优化；最后面向资源受限的边缘计算场景，将模型部署于低成本低功耗的RK3588终端设备，结合混合量化技术，验证了方案在实际应用中的可行性。

## 1 理论方法

### 1.1 总体网络结构

现有的图像融合和目标检测研究大部分采用串行级联方式，即图像融合网络的输出作为目标检测网络的输入。级联网络能够生成更适用于检测任务的融合图像，但任务间存在顺序依赖关系且难以部署在资源受限的终端设备。本文提出一种并行网络实现双光图像融合和目标检测任务及其轻量化模型，该网络由两个浅层特征提取器、一个共享多模态特征融合模块[18]、一个检测子网和一个融合子网[19]组成。剪枝方法包括基于L2范数和欧式距离的结构化剪枝。

图1是并行网络PDFNet（Parallel Network for Object Detection and Image Fusion）[20]及结构化剪枝后的轻量化模型。模型采用YOLOv5s[21]前三层的伪孪生特征提取网络用于提取RGB与热红外图像的浅层特征，输出为RGB图像特征图$F_{\mathrm{RGB}}$和热红外图像特征图$F_{\mathrm{T}}$。$F_{\mathrm{RGB}}$和$F_{\mathrm{T}}$经设计的共享多模态特征融合模块处理后，得到融合特征图$F_{\mathrm{fus}}$。最后将经过上采样和跳跃连接的$F_{\mathrm{fus}}$以及原融合特征图$F_{\mathrm{fus}}$分别输入图像融合子网络与目标检测子网络，同步并行地完成图像融合与目标检测任务。本文选择YOLOv5作为目标检测器并与现有先进的融合网络进行对照实验。随后对PDFNet进行两阶段的剪枝以完成模型的压缩，为后续模型的部署应用奠定基础。

为了使两种不同模态的特征能够更好地利用互补的信息得到语义更加丰富的特征图，设计了通道注意力特征融合模块（Channel Attention Feature Fusion，CAFF）整合两种模态的特征图，生成可同时用于融合和检测任务的新特征图。如图1所示，CAFF模块由六个卷积层和一个通道注意力块[22]组成。通道注意力块包含两个全连接层和一个平均池化层，采用Softmax函数作为激活函数。在CAFF模块中，输入为$F_{\mathrm{RGB}}$和$F_{\mathrm{T}}$，为避免元素级操作中两种模态信息的冲突，先对$F_{\mathrm{RGB}}$以及$F_{\mathrm{T}}$进行简单像素级加法得到$F_{\mathrm{add}}$，接着利用通道注意力得到通道注意力权重$\boldsymbol{W}_{\mathrm{atten}}$对两种特征图进行加权。输出$F_{\mathrm{fus}}$的计算过程为

$$F_{\mathrm{add}}=\mathrm{Conv}\left(\mathrm{Conv}\left(F_{\mathrm{RGB}}\right)\right)\oplus\mathrm{Conv}\left(\mathrm{Conv}\left(F_{\mathrm{T}}\right)\right)\# \tag{1}$$

$$\boldsymbol{W}_{\mathrm{atten}}=\mathrm{Softmax}\left(\mathrm{Avgpool}\left(\mathrm{Fc}\left(\mathrm{Fc}\left(F_{\mathrm{add}}\right)\right)\right)\right)\# \tag{2}$$

$$F_{\mathrm{fus}}=\left(F_{\mathrm{RGB}}\otimes\boldsymbol{W}_{\mathrm{atten}}\right)\oplus\left(F_{\mathrm{T}}\otimes\boldsymbol{W}_{\mathrm{atten}}\right)\# \tag{3}$$

式中，Conv(·)表示卷积层，Avgpool(·)表示平均池化层，Fc(·)表示全连接层，$\oplus$表示元素级加法，$\otimes$表示元素级乘法。通过网络训练，通道注意力块能够学习到特征图中哪些通道与当前任务更相关，从而减轻元素级操作可能带来的负面影响。

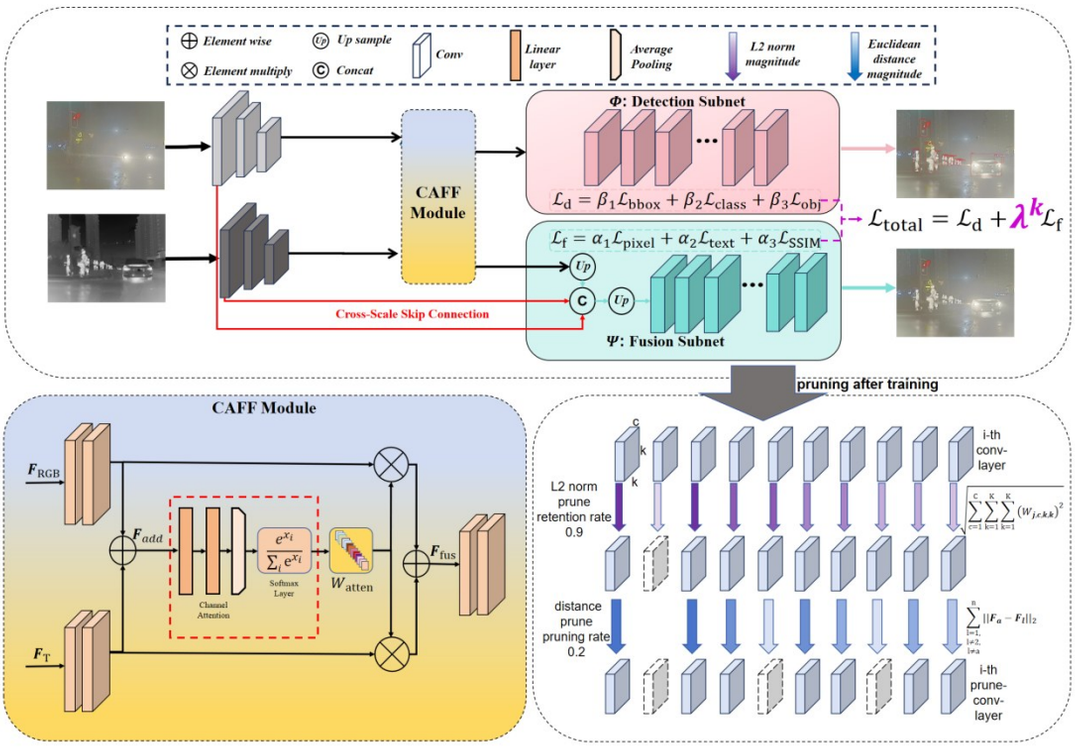


图1　PDFNet总体架构

Fig. 1　Overall architecture of PDFNet

在嵌入式设备端资源受限的背景条件下，为了将本文模型更有效地进行部署，需要对模型进行轻量化处理。在卷积神经网络中，滤波器是卷积层的卷积核。滤波器剪枝可定义为对卷积层中卷积核的结构化稀疏化操作。当对目标卷积层进行滤波器剪枝时，该层的输出特征图通道数将发生等量削减，进而在保持网络拓扑结构完整性的前提下，有效压缩模型参数规模并降低浮点运算量。HE Yang等[23]提出滤波器剪枝并非减去“不重要”的滤波器而是减去“最具可替代性的”冗余滤波器的观点。根据几何中值的思想，通过计算每一层中单个滤波器与其余滤波器的欧氏距离之和，所得值越小，代表该滤波器的信息可替代性越强，说明该滤波器越可以被剪枝，即基于几何中值的滤波器剪枝（Filter Pruning via Geometric Median，FPGM）。本文对训练后的PDFNet模型进行两次剪枝，首先对训练后的PDFNet采取基于L2范数的剪枝，筛选并掩码置零低范数的次要滤波器，保留高范数的核心滤波器；接着利用式（4）及公式（5）执行基于欧式距离的剪枝，将卷积层保留的$n$个核心滤波器记为$\{F_1, F_2, \cdots, F_a, \cdots, F_n\}$，其中$F_a$为第$a$个滤波器；$/\!/ F_a - F_l /\!/_2$为滤波器$F_a$与$F_l$之间的欧氏距离；滤波器$F_a$与其他核心滤波器的累计欧氏距离定义为$D(F_a)$。计算第$a$个滤波器与其余所有核心滤波器的累计欧氏距离，筛选出累计距离最小的高冗余滤波器并完成二次掩码置零，$a^*$就是累计距离最小、高冗余、二次待剪枝滤波器编号；最后对两次剪枝后的模型进行重新训练以达到恢复模型精度的目的。

$$D(F_a)=\sum_{l=1,l\neq a}^{n} /\!/ F_a - F_l /\!/_2 \tag{4}$$

$$a^*=\mathrm{argmin}_{a\in[1,n]} D(F_a) \tag{5}$$

## 1.2　多任务优化设计

并行的图像融合与目标检测多任务网络面临特征和任务两个层面的冲突问题。在特征层面，图像融合需要高分辨率的细节特征实现像素级重建，而目标检测任务需要复杂的语义信息实现分类及定位，两者在特征需求上存在一定冲突[24]；而在任务层面，图像融合作为底层视觉任务，其优化目标与目标检测这一高层语义任务在训练重点与梯度更新方向上也存在差异。为有效缓解上述冲突、实现多任务间的协同优化，本文引入跨尺度跳跃连接机制以增强不同层级特征的信息交互，同时采用自适应任务损失权重策略对各任务的训练过程进行动态平衡，从而提升模型整体的稳定性与性能表现。

如图1所示，CAFF模块输出的融合特征图$F_{fus}$是低分辨率特征，包含目标检测任务所需的高级语义特征，可直接送入目标检测子网络完成分类与定位任务。然而，图像融合任务需要高分辨率特征以实现像素级重建，$F_{fus}$无法满足这一需求。为解决该问题，对输入$F_{fus}$进行上采样，恢复特征图的空间分辨率，随后引入跨尺度跳跃连接，将上采样后的特征图与浅层特征提取器中第一个卷积层输出的高分辨率特征图进行通道维度拼接，利用拼接后的特征图输入卷积层重建融合图像。

并行多任务网络的损失函数由两部分组成：图像融合损失$\mathcal{L}_f$[25]和目标检测$\mathcal{L}_d$损失，分别定义为

$$\mathcal{L}_f = \alpha_1\mathcal{L}_{pixel} + \alpha_2\mathcal{L}_{text} + \alpha_3\mathcal{L}_{ssim} \tag{6}$$

$$\mathcal{L}_d = \beta_1\mathcal{L}_{bbox} + \beta_2\mathcal{L}_{class} + \beta_3\mathcal{L}_{obj} \tag{7}$$

式中，$\mathcal{L}_{pixel}$，$\mathcal{L}_{text}$和$\mathcal{L}_{ssim}$分别表示像素损失、纹理损失和结构相似性（Structural Similarity，SSIM）损失[26]，超参数$[\alpha_1, \alpha_2, \alpha_3]$设置为[4,4,2]。$\mathcal{L}_{pixel}$像素损失主要用于衡量融合图像与真实图像在像素级别上的差异，确保融合后的图像在像素强度上与目标图像接近；$\mathcal{L}_{text}$纹理损失通过计算图像梯度的差异来保持图像的纹理细节和高频信息；$\mathcal{L}_{ssim}$从亮度、对比度、结构三个维度衡量图像相似性，使融合图像更符合人眼视觉感知，保持图像整体结构。$\mathcal{L}_{bbox}$，$\mathcal{L}_{class}$和$\mathcal{L}_{obj}$分别表示定位损失、分类损失和置信度损失，超参数$[\beta_1, \beta_2, \beta_3]$与YOLOv5保持一致，设置为[0.05,0.5,1.0]。$\mathcal{L}_{bbox}$定位损失用于衡量预测边界框与真实边界框的位置偏差使预测框准确定位到目标位置；$\mathcal{L}_{class}$分类损失用于衡量预测类别与真实类别的差异使模型正确识别目标的类别；$\mathcal{L}_{obj}$置信度损失用于衡量预测框中是否包含目标的置信度使模型能够区分前景和背景，减少误检。在多任务学习中，当各项任务相关性较低时，训练难度较大[27]。图像融合与目标检测任务具有不同的训练难度和收敛速度，通过设计自适应损失权重$\lambda^k$以平衡两项任务：

$$\lambda^k = \begin{cases} \mathcal{L}_d^k / \left(T * \mathcal{L}_f^k\right) & , k = 1 \\ \mathcal{L}_d^{k-1} / \left(T * \mathcal{L}_f^{k-1}\right) & , k \geqslant 2 \end{cases} \tag{8}$$

$$T = 3 - 2(k-1)/K \tag{9}$$

式中，$k$表示训练轮次索引，$K$表示总训练轮次，$T$表示温度参数，用于控制融合任务权重的平滑度类似[28]。最终总损失定义为

$$\mathcal{L}_{total} = \mathcal{L}_d + \lambda^k\mathcal{L}_f \tag{10}$$

## 2 实验

### 2.1 实验设置

1）数据集：实验在$M^3FD$数据集[9]上进行。该数据集包含4 200对高分辨率对齐的红外和可见光图像，总共标注了6个类别的33 603个目标，具体类别为：人、汽车、公交车、摩托车、路灯和卡车。本文随机选择了3 360对图像作为训练集，剩余840对作为测试集。

2）实现细节：训练阶段，输入图像尺寸设置为640×640像素，批量大小为16，学习率采用余弦退火策略从$5\times10^{-3}$调整至$5\times10^{-4}$。模型采用随机梯度下降（SGD）优化器先训练300轮。剪枝后重新训练阶段，输入图像尺寸与批量大小不变，学习率从$1\times10^{-2}$调整至$1\times10^{-3}$同时采用随机梯度下降（SGD）优化器训练400轮，该部分实验基于四块NVIDIA GeForce RTX 2080 GPU实现。

3）部署方案：选用搭载RK3588芯片的Orangepi5Ultra嵌入式计算平台，对所设计的网络模型进行端侧部署与融合推理实验。通过在嵌入式端与桌面级GPU平台上采用一致的评价指标，对模型的推理性能与效果进行对比分析，从而系统性评估所提模型在实际端侧部署场景下的可行性。

4）评价指标：选择四项图像融合质量指标结构相似性SSIM，基于边缘保留的质量指标（Quality metric based on edge preservation，$Q^{abf}$），熵（Entropy，EN）和标准差（Standard Deviation，SD）来评估图像融合效果[29-30]。在整个测试集上计算mAP50作为目标检测性能的评估指标，并随机抽取105幅图像计算图像融合质量指标。

### 2.2 消融实验

设计了消融实验以验证网络中各个模块的有效性。如表1所示，$PDFNet_{d/f}$表示单独训练图像融合或目标检测任务，$PDFNet_c$表示使用拼接操作替代CAFF模块，$PDFNet_{ns}$表示不含跨尺度跳跃连接的网络，

$PDFNet_{ew}$代表使用等权重替代自适应任务损失权重。

**表 1　消融实验**

**Table 1　Ablation experiments**

| Method | Multitask learning | Channel attention | Skip connection | Adaptive loss weight |
|---|---|---|---|---|
| $PDFNet_{d/f}$ | × | √ | √ | × |
| $PDFNet_{c}$ | √ | × | √ | √ |
| $PDFNet_{ns}$ | √ | √ | × | √ |
| $PDFNet_{ew}$ | √ | √ | √ | × |
| PDFNet | √ | √ | √ | √ |

如表2所示，从消融实验结果来看，PDFNet、$PDFNet_{f}$及$PDFNet_{c}$在融合的各项指标上的表现较为接近，表明网络对具体的特征融合算子具有较强的鲁棒性，不同融合策略对整体性能影响有限，网络结构具有一定的鲁棒性。相比之下，$PDFNet_{ns}$的整体性能出现明显衰减，尤其在$Q^{abf}$指标的下降较为显著，表明跨尺度跳跃连接在保留源图像边缘与纹理信息方面发挥着关键作用。在目标检测结果方面，PDFNet与仅采用基础双模态结构的$PDFNet_{d}$相比，PDFNet性能提升明显，而$PDFNet_{ns}$与$PDFNet_{c}$模型性能均出现不同程度下降。上述结果说明，跨尺度跳跃连接结构能够有效增强多层级特征的信息流动，而通道注意力特征融合模块有助于突出关键语义信息、抑制冗余特征，二者的协同设计对模型整体性能具有积极贡献。最后，结合$PDFNet_{d}$、$PDFNet_{ew}$与PDFNet的性能对比结果可以得出，当对图像融合与目标检测任务采用默认的等权重联合训练策略时，网络难以平衡两项任务的优化目标，导致训练过程中参数发散。而通过对多任务损失权重进行合理调整，不仅能够有效缓解任务间的特征冲突问题，还可以在一定程度上实现任务间的相互强化，进一步提升模型的稳定性与综合性能。实验结果表明，融合任务所提取的多模态互补特征能够为检测任务提供更具判别性的输入表征，进而提升整体检测性能。

**表 2　消融实验结果**

**Table 2　Ablation experiments result**

| Model | SSIM | $Q^{abf}$ | EN | SD | mAP50 |
|---|---|---|---|---|---|
| $PDFNet_{d/f}$ | **0.709** | **0.659** | 6.830 | 34.952 | <u>0.883</u> |
| $PDFNet_{c}$ | <u>0.707</u> | 0.649 | 6.841 | 35.299 | 0.876 |
| $PDFNet_{ns}$ | 0.705 | 0.579 | <u>6.849</u> | <u>35.366</u> | 0.875 |
| $PDFNet_{ew}$ | — | — | — | — | — |
| PDFNet | 0.705 | <u>0.651</u> | **6.869** | **35.921** | **0.901** |

The best results are shown in bold, and the second-best results are underlined

最后通过可视化方式[31]对比$PDFNet_{ns}$与PDFNet模型的损失曲面特性。其中$PDFNet_{ns}$模型呈现出较明显的非凸特征，其等高线在不同方向移动时呈现不规则变化，而PDFNet模型则在中心区域展现出凸面轮廓的平坦特性，表明包含跨尺度跳跃连接的PDFNet模型更易优化，进一步验证了PDFNet模型设计的合理性。

### 2.3　对比实验

为验证所提算法的有效性与优越性，设计了系统性对比实验，选取当前领域内代表性先进图像融合算法TarDAL[9]、SwinFusion[24]、DetFusion[10]、MetaFusion[11]、PSFusion[14]、SDCFusion[13]及SeAFusion[8]作为对比基准，在统一实验环境下开展定性与定量的图像融合结果对比分析。在此基础上，引入目标检测模型YOLOv5s作为评估检测器，对各算法生成的融合图像进行目标检测性能对比，从下游视觉任务角度进一步验证融合结果的实用价值，从而全面论证所提算法在提升图像融合质量与增强目标感知能力方面的优势。

#### 2.3.1　图像融合定量与定性对比

由表3可知，所提PDFNet算法在多项关键评价指标上表现良好，与当前领域内先进的独立图像融合算法SwinFusion展现出高度的可比性。具体而言，在衡量图像结构保真度的结构相似度（SSIM）指标上PDF-

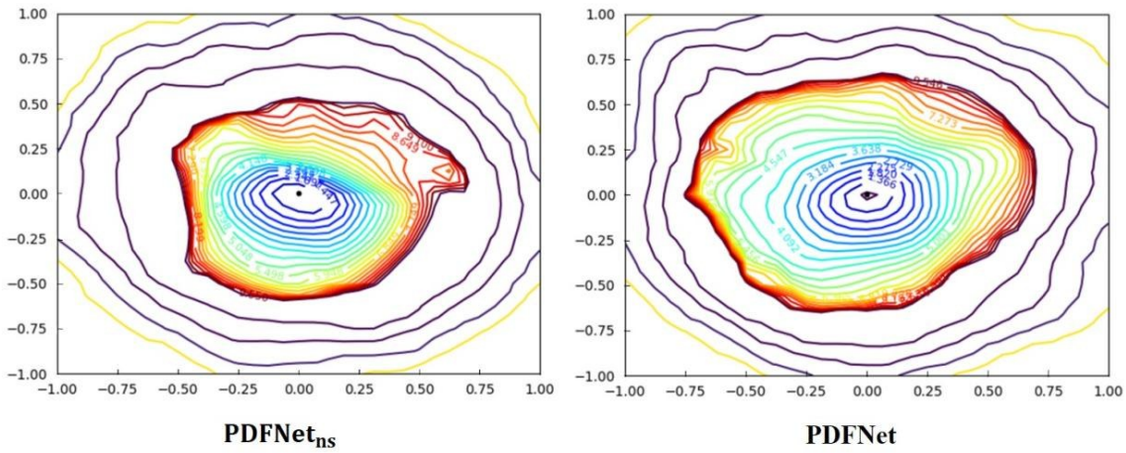


图 2 PDFNet$_{ns}$与PDFNet损失曲面的2D可视化

Fig. 2 2D visualization of the loss surface of PDFNet$_{ns}$ and PDFNet

Net达到0.705，相较于其他算法优势明显。而在反映图像信息丰富度的信息熵(EN)指标上，PDFNet取得6.869的成绩，略高于SwinFusion的6.839，表明本文设计的并行网络结构能够更充分地保留源图像中的有效信息。这一结果表明所设计的并行网络结构在兼顾检测任务的同时，具备良好的图像特征提取与保留能力，达到了与专用融合网络相当的性能水准。相较于TarDAL，PDFNet在SSIM和$Q^{abf}$上提升明显；值得注意的是，尽管TarDAL展现出较高的标准差(SD)，但结合其极低的SSIM值分析，该高SD值主要归因于图像失真或伪影带来的非自然对比度。而PDFNet在保证高结构保真度的同时兼具合理的对比度，得到了更优的图像质量。相较于其他融合算法PDFNet同样具有较均衡的性能。在$Q^{abf}$指标上，PDFNet虽略低于SDCFusion，但高于其他多数对比算法，表明其融合图像的色彩一致性和整体协调性更具优势；在EN指标上，PDFNet虽不及MetaFusion和PSFusion，但显著高于TarDAL，且与SwinFusion、SeAFusion等算法处于同一区间，保证了融合图像的信息完整性；在SD指标上，PDFNet处于合理范围，既避免了TarDAL、MetaFusion、PSFusion等算法过高SD值带来的图像失真，又兼顾了图像的对比度需求，实现了结构保真度与对比度的平衡，最终生成更符合人眼视觉特性和下游检测任务需求的融合图像。

**表 3 不同算法图像融合指标对比**

**Table 3 Comparison of image fusion metrics for different algorithms**

| Algorithm | SSIM | $Q^{abf}$ | EN | SD |
|---|---|---|---|---|
| TarDAL | 0.384 | 0.303 | 6.628 | 42.924 |
| SwinFusion | <u>0.704</u> | <u>0.660</u> | 6.839 | 35.486 |
| Detfusion | 0.646 | 0.502 | 6.987 | 37.935 |
| Metafusion | 0.602 | 0.389 | <u>7.340</u> | <u>45.882</u> |
| PSFusion | 0.626 | 0.593 | **7.366** | **49.115** |
| SDCFusion | 0.674 | **0.686** | 6.952 | 36.405 |
| SeAFusion | 0.686 | 0.614 | 6.875 | 35.661 |
| PDFNet | **0.705** | 0.651 | 6.869 | 35.921 |

The best results are shown in bold, and the second-best results are underlined.

图3呈现了多种代表性图像融合算法在森林、日间行人及夜间行人等多类典型场景下的融合结果可视化对比。观察可见，PDFNet生成的融合图像在整体视觉质量上表现出显著优越性，具体体现在更优的结构保真度、精细纹理细节的完整保留以及可见光与红外模态信息的自然融合等方面，而其他对比方法普遍存在过平滑、对比度不足、细节丢失或伪影干扰等问题，尤其是在Detfusion方法中出现了普遍的黑色伪影，这些伪影导致融合后的图像出现了大面积特征丢失，这对后续目标检测易产生极大干扰。而在第一行和第二行的森林场景中，PDFNet清晰还原了树木枝叶纹理、地面落叶及行人细节，红色框标注的感兴趣区域中目标轮廓锐利且背景层次丰富；相较之下，PSFusion和SeAFusion等方法则出现明显模糊或颜色失真。在第

三行和第四行的低照度夜间场景下，PDFNet有效抑制了光晕与噪声，同时平衡了可见光色彩信息与红外热显著性，使行人及车辆等关键目标边界清晰、背景过渡自然，而从第四行车辆上图案可以看出SwinFusion、DetFusion及PSFusion等方法则导致目标区域纹理细节消失。这些定性结果直观表明，PDFNet在复杂环境下的融合性能更具鲁棒性，为下游目标检测等视觉任务提供了高质量的输入基础，进一步验证了其在提升图像融合实用价值方面的有效性。

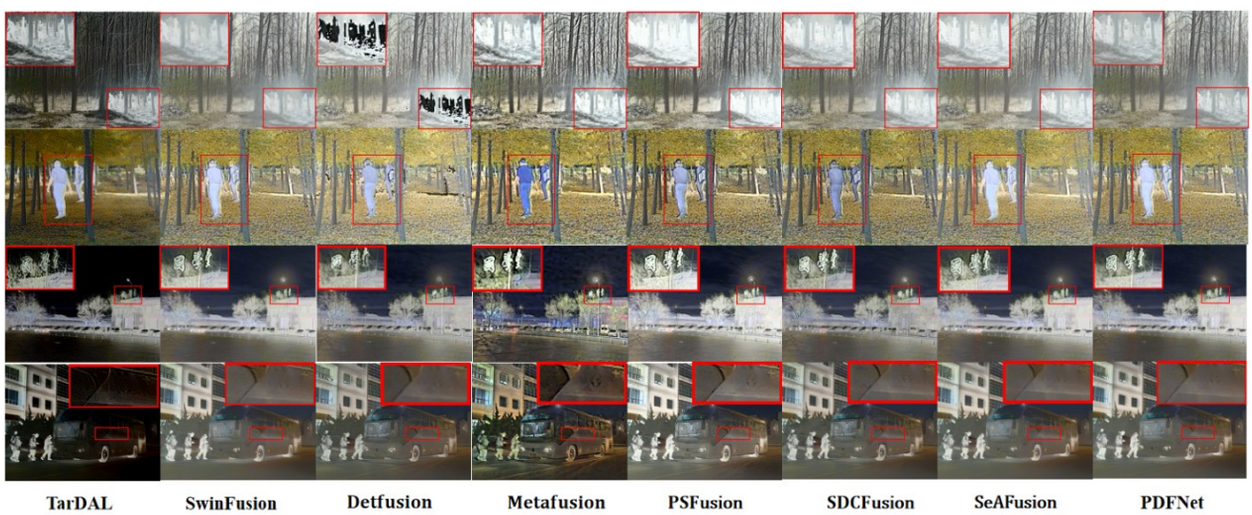


图 3　不同方法的融合结果

Fig. 3　The fusion results of different methods

2.3.2　目标检测定量与定性对比

表4给出了不同算法的融合图像在多类别目标检测任务上的性能对比结果。从整体性能来看，提出的PDFNet算法mAP50指标达到了0.901，相较于次优方法SwinFusion的0.869提升了3.2%，显著优于其他对比算法，充分验证了所提方法的有效性。在多模态方法之间的对比中，PDFNet算法相较于TarDAL算法在mAP50指标上进一提升了3.6%，并在所有类别如行人、车辆及路灯上取得了更高的检测精度。这一结果表明，相较于级联式融合策略，本文采用的并行特征提取与融合架构能够更充分地保留不同模态的判别信息，从而提升整体检测性能。

从各类别的检测精度分析，PDFNet在六个目标类别中均表现出色，其中在行人这一类别上取得了0.871的最高精度，相比表现次优的TarDAL(0.830)提升了4.1%，这表明PDFNet生成的融合图像能够更好地保留红外图像中的热辐射信息，从而增强了对行人目标的识别能力。在路灯类别上，PDFNet同样以0.912的精度大幅领先其他方法，较次优算法YOLOv5$_{rgb}$提升了1.2%，证明了本文方法在保持可见光图像细节信息方面的优势。此外，在汽车、公交车、摩托车和卡车等车辆类别上，PDFNet也均取得了最优或接近最优的检测精度，这些结果充分说明本文算法在融合多模态互补信息方面具有明显优势，能够有效兼顾不同类别目标的特征表达需求。

值得注意的是，基于RGB图像的YOLOv5$_{rgb}$和基于热红外图像的YOLOv5$_{t}$在不同类别上表现出明显的互补性：YOLOv5$_{rgb}$在车辆类别检测中表现较好，而YOLOv5$_{t}$在行人检测上更具优势，这一现象进一步说明了多模态图像融合的必要性。相比之下，PDFNet能够充分融合两种模态的互补信息，在所有类别上均保持稳定且优异的性能，体现出更强的鲁棒性和泛化能力，同时也体现了本文方法在融合图像质量与下游任务适配性方面的双重优势。

图4展示了不中同图像融合算法生成的融合图像在YOLOv5s目标检测器下的定性检测结果对比。首先，仅基于可见光模态的YOLOv5$_{rgb}$在面临光照退化及环境干扰时泛化能力较弱。特别是在图4第2行的雨夜低照度场景中，由于缺乏红外热辐射信息的辅助，RGB模型难以提取有效的目标特征，导致行人目标出现严重漏检，基本失效。相比之下，引入红外模态的融合算法显著提升了检测的鲁棒性。然而，现有的TarDAL算法在处理密集或极小目标时仍存在局限性，例如在第2行的密集人群与第3行远处的小目标，以及第

表 4 不同算法目标检测指标对比

Table 4 Comparison of object detection metrics for different algorithms

| Algorithm | Person | Car | Bus | Motorcycle | Lamp | Truck | mAP50 |
|---|---|---|---|---|---|---|---|
| $YOLOv5_{rgb}$ | 0.744 | 0.922 | 0.919 | 0.823 | 0.900 | 0.855 | 0.860 |
| $YOLOv5_{t}$ | 0.826 | 0.906 | 0.905 | 0.785 | 0.807 | 0.848 | 0.846 |
| TarDAL | 0.830 | 0.920 | 0.917 | 0.807 | 0.853 | 0.863 | 0.865 |
| SwinFusion | 0.731 | 0.912 | **0.942** | **0.968** | 0.762 | 0.801 | 0.869 |
| Detfusion | 0.779 | 0.916 | 0.929 | 0.832 | 0.773 | 0.841 | 0.845 |
| Metafusion | 0.759 | 0.912 | 0.935 | 0.844 | 0.766 | 0.856 | 0.845 |
| PSFusion | 0.785 | 0.922 | 0.92 | 0.862 | 0.765 | 0.843 | 0.85 |
| SDCFusion | 0.786 | 0.922 | 0.932 | 0.873 | 0.78 | 0.851 | 0.857 |
| SeAFusion | 0.779 | 0.921 | **0.942** | 0.83 | 0.788 | 0.844 | 0.851 |
| PDFNet | **0.871** | **0.935** | 0.937 | 0.881 | **0.912** | **0.871** | **0.901** |

The best results are shown in bold, and the second-best results are underlined

4行远处密集车辆场景中，均出现了小目标漏检的情况。

而几种先进的融合算法得到的融合图像在漏检方面有了明显改善。观察可知，在雾天、低照度及复杂路况等多类挑战性场景下，PDFNet融合图像呈现出更高的结构保真度和目标显著性，显著提升了检测器的感知能力，使其能够准确、完整地生成车辆、行人、交通标志及信号灯等关键目标的边界框；而其他对比方法则普遍存在细节模糊、过平滑等问题，导致部分场景出现误检或定位偏差，尤其是SwinFusion算法得到的融合图像在第1行以及第2行图像的场景下虽然检测更加全面，但是出现了许多误检以及行人定位偏差的情况。除此之外，在第3行的海滩远处小目标场景下，Metafusion以及PSFusion也出现了行人误检的现象，同时检测框出现了一定的偏差，而PDFNet虽然存在一定程度的漏检，但实现了较为精准的检测；类似地，在第2行的傍晚场景下，PDFNet显著抑制了光晕干扰并增强了目标轮廓，相较于TarDAL和SeAFusion等方法展现出更优的鲁棒性。这些可视化结果直观验证了PDFNet在图像融合质量提升与下游视觉任务感知增强方面的优越性，为其在实际智能感知系统中的应用提供了有力支撑。

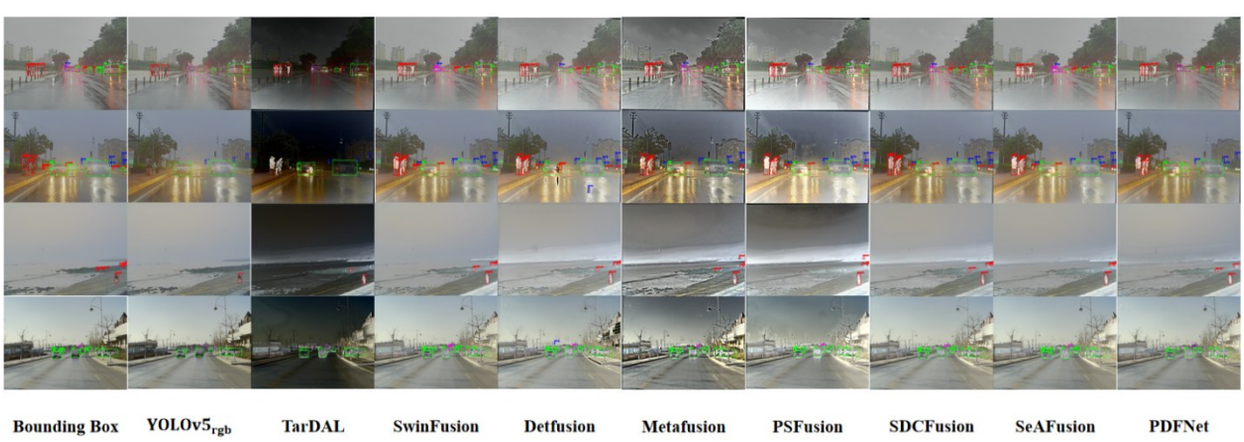


图 4 不同方法的检测结果

Fig. 4 The detection results of different methods

## 2.4 轻量化部署实验

为实现在资源受限边缘计算平台上的部署，从剪枝与量化两个层面对PDFNet模型进行网络结构与参数表示的优化，保持检测与融合性能的前提下，尽可能降低模型参数规模与计算复杂度，提升模型在边缘设备上的部署效率。

在剪枝层面，通过设置第一阶段剪枝的通道保留率(rate-norm)与第二阶段的通道剪枝率(rate-dist)控制剪枝模型的参数量与模型的尺寸大小。经过对剪枝模型的重新训练后，不同剪枝参数得到的剪枝模型指

标如表5所示。

**表5　不同剪枝率结果对比**

**Table 5　Comparison of results with different pruning rates**

| Rate−norm | Rate−dist | Number of parameters/MB | Model size/MB | GPU model inference time/ms | mAP50 |
|---|---|---|---|---|---|
| 1.0 | 0 | 8.80 | 17.592 | 22.9 | 0.901 |
| 0.93 | 0.1 | 6.34 | 12.802 | 19.8 | 0.898 |
| 0.9 | 0.1 | 5.28 | 11.604 | 18.1 | 0.893 |
| 0.9 | 0.15 | 4.59 | 9.361 | 16.3 | 0.884 |
| 0.9 | 0.2 | 2.94 | 6.282 | 15.7 | 0.872 |
| 0.9 | 0.55 | 0.96 | 2.314 | 11.1 | 0.837 |

由表5可知，经过多轮实验，随着剪枝强度的逐步增加，模型参数量与模型尺寸持续下降，GPU端推理时间相应缩短，表明所采用的剪枝策略能够有效降低模型复杂度并提升推理效率。在仅引入较轻度剪枝的情况下（rate-norm = 0.93，rate-dist = 0.1），模型参数量由原始模型的8.80 MB减少至6.34 MB，模型尺寸下降至12.802 MB，GPU 推理时间缩短至19.8 ms，而mAP50仅由0.901轻微下降至0.898，表明适度的滤波器剪枝对检测性能影响较小。进一步将rate-norm降至0.9同时保持rate-dist仍为0.1时，模型参数量与推理时间继续下降，mAP50维持在0.893，显示基于L2范数的滤波器剪枝在一定范围内具有较好的鲁棒性。当保持rate-norm = 0.9并逐步增大基于距离的剪枝比例时，模型压缩效果进一步增强。随着rate-dist从0.1提升至0.15，模型参数量与模型尺寸分别降至4.59 MB和9.361 MB，GPU推理时间缩短至16.3 ms，而mAP50仍保持在0.884，精度下降幅度相对可控。但当rate-dist继续增大至0.2及0.55时，虽然模型规模和推理时间进一步降低，但检测精度下降趋势明显，过强的结构剪枝会对模型特征表达能力造成不利影响。综合检测精度保持与模型压缩效率之间的权衡，最终剪枝模型PDFNet$_P$的配置为rate-norm = 0.9、rate-dist = 0.15，在此剪枝配置下，各卷积层的滤波器数量，即输出特征图通道数发生相应削减，模型剪枝前后卷积层滤波器数量变化情况如图5所示。

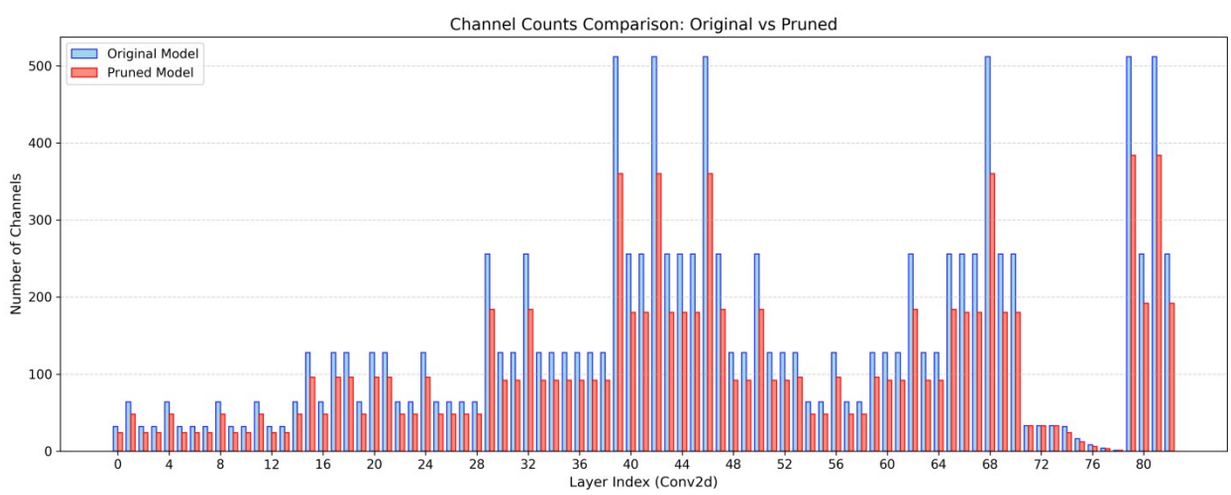


图5　剪枝前后滤波器数量对比

Fig. 5　Comparison of the number of filters before and after pruning

为验证剪枝的效果，对比剪枝前后模型在图像融合与目标检测任务中的性能，对应的定量指标如表6所示。

实验结果表明，本文的剪枝方式在显著降低模型复杂度的同时，对整体性能影响并不显著。综合表5可知，与原始模型PDFNet相比，剪枝后的PDFNet$_P$在参数量和模型规模方面均实现了大幅压缩，压缩比例接近48%，有效提升了模型的轻量化程度。在检测性能方面，PDFNet$_P$的mAP50从0.901略微下降至0.884，

表 6 剪枝前后模型对比

Table 6 Comparison of models before and after pruning

| Model | mAP50 | SSIM | $Q^{abf}$ | EN | SD |
|---|---|---|---|---|---|
| PDFNet | 0.901 | 0.705 | 0.651 | 6.869 | 35.921 |
| $PDFNet_P$ | 0.884 | 0.707 | 0.543 | 6.790 | 33.184 |

性能下降幅度较小，表明剪枝在保留主要判别能力的同时并未对目标检测精度造成明显破坏。在融合质量评价指标方面，$PDFNet_P$在SSIM指标上与原模型基本保持一致，说明剪枝后模型在结构相似性保持方面仍具备较强能力；而$Q^{abf}$、EN和SD指标出现一定程度的下降，反映出剪枝对边缘信息保持和信息量表达存在一定影响。总体而言，$PDFNet_P$在牺牲少量融合与检测性能的前提下，显著降低了模型参数规模与存储开销，在资源受限的嵌入式应用场景中具有更高的部署价值和实用性。

最后将剪枝前后的模型分别进行量化得到$PDFNet_Q$和$PDFNet_{PQ}$并部署在RK3588嵌入式设备。首先将PyTorch格式(.pt)转换为ONNX，再进一步转换为适配的RKNN格式。为避免全局INT8量化导致的精度损失，本文采用混合量化策略：筛选并标记量化后偏差较大的关键层，使其保持高精度计算而不参与量化，从而将精度控制在可接受范围。最终将RKNN模型部署至RK3588端侧，完成模型推理、检测头解码、后处理及指标评估，获得验证集上的推理结果。

表7对比了模型在未量化、采用混合量化以及经过剪枝后量化策略条件下各算子的性能差异。上述量化流程在保证模型结构与功能完整性的前提下，通过降低非关键算子的数值精度并简化计算过程，有效减少了模型推理阶段的计算开销与内存占用，为模型在嵌入式平台上的高效运行提供了有力支撑。

表 7 算子性能对比

Table 7 Operator performance comparison

| Operator performance | CPU/μs | NPU/μs | Total time/μs |
|---|---|---|---|
| Unquantified | 21 771 | 437 984 | 459 755 |
| Mixed quantization | 9 310 | 268 912 | 278 222 |
| Pruning and quantization | 6 416 | 159 796 | 166 212 |

表8对比了剪枝前后的量化模型在资源有限的RK3588嵌入式设备上的部署效果。剪枝后的$PDFNet_P$的检测速度为8.36 帧/s，相较于未剪枝模型$PDFNet_Q$的4.76 帧/s推理效率提升了75.6%。与此同时$PDFNet_{PQ}$的mAP50为0.879，仅较$PDFNet_Q$下降1.7%，精度损失处于可接受范围内。

表 8 剪枝前后部署效果对比

Table 8 Comparison of deployment performance before and after pruning

| Model | mAP50 | Frame/s |
|---|---|---|
| $PDFNet_Q$ | 0.896 | 4.76 |
| $PDFNet_{PQ}$ | 0.879 | 8.36 |

上述结果表明，通过剪枝与混合量化相结合的轻量化策略，模型在端侧部署场景中能够在较大幅度提升推理速度的同时，较好地保持目标检测性能。特别是在RK3588这类以NPU为主要计算单元的嵌入式平台上，剪枝后的模型在算子执行效率与整体推理延迟方面均表现出更优的适配性，验证了本文轻量化与部署方案在实际边缘计算应用中的可行性与有效性。

## 3 结论

本文提出一种用于双光图像融合与目标检测的并行多任务学习网络。与独立和级联架构的网络相比，该并行多任务网络平等对待底层图像融合任务和高层目标检测任务，在保证感知精度的同时，有效降低了计算冗余，为多模态光电信息的端侧协同处理提供了一种可行方案。在双光数据集和嵌入式平台上的实验结果表明，并行处理图像融合与目标检测任务能够同时提升两项任务的性能，验证了其在实际光电感知系

统中的部署可行性。

随着多模态技术的发展，图像融合与目标跟踪、语义分割等高级视觉任务的协同处理将成为重要趋势，并行多任务学习架构具有更好的任务扩展能力。下一步工作将围绕资源受限环境下的轻量化部署展开，深入研究在保持精度的同时，利用结构优化与推理策略改进进一步提升并行多任务网络的工程适用性。

## 参考文献


[1] SONG Kechen, ZHAO Ying, HUANG Liming, et al. RGB-T image analysis technology and application: A survey[J]. Engineering Applications of Artificial Intelligence, 2023, 120: 105919.

[2] GUO Cuixia, XU Yongtao, ZOU Zhanghuang, et al. Lightweight pedestrian vehicle detection algorithm based on visible and infrared bimodal fusion[J]. Acta Photonica Sinica, 2025, 54(6): 0610001.
郭翠霞，徐永涛，邹章煌，等. 基于可见和红外双模态融合的轻量化行人车辆检测算法[J]. 光子学报, 2025, 54(6): 0610001.

[3] YANG Chen, HOU Zhiqiang, LI Xinyue, et al. Object detection algorithm based on CNN-transformer dual modal feature fusion [J]. Acta Photonica Sinica, 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.

[4] YU Zhirui, YIN Zhanpeng, WANG Junyu, et al. Multimodal object detection method using adaptive dusion of infrared and visible features[J]. National Remote Sensing Bulletin, 2025 , 29(10): 3006-3019.
喻智睿, 尹展鹏, 王俊宇, 等. 可见光和红外特征自适应融合的多模态目标检测方法[J]. 遥感学报, 2025, 29 (10): 3006-3019.

[5] SUN Bin, YOU Hang, LI Wenbo, et al. Dual-band payload image fusion and its applications in low-altitude remote sensing[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(11): 531343.
孙彬, 游航, 李文博, 等. 双光载荷图像融合及其在低空遥感中的应用[J]. 航空学报, 2025, 46(11): 531343.

[6] HAO Yongping, CAO Zhaorui, BAI Fan, et al. Research on infrared visible image fusion and target recognition algorithm based on region of interest mask convolution neural network[J]. Acta Photonica Sinica, 2021, 50(2): 0210002.
郝永平, 曹昭睿, 白帆, 等. 基于兴趣区域掩码卷积神经网络的红外-可见光图像融合与目标识别算法研究[J]. 光子学报, 2021, 50(2): 0210002.

[7] ZHANG Xiaodong, WANG Shuo, GAO Shaoshu, et al. Infrared and visible image fusion method based on information enhancement and mask loss[J]. Acta Photonica Sinica, 2024, 53(9): 0910003.
张晓东, 王硕, 高绍姝, 等. 基于信息增强和掩码损失的红外与可见光图像融合方法[J]. 光子学报, 2024, 53(9): 0910003.

[8] TANG Linfeng, YUAN Jiteng, MA Jiayi. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network[J]. Information Fusion, 2022, 82: 28-42.

[9] LIU Jinyuan, FAN Xin, HUANG Zhanbo, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022: 5792-5801.

[10] SUN Yiming, CAO Bing, ZHU Pengfei, et al. Detfusion: a detection-driven infrared and visible image fusion network [C].Proceedings of the 30th ACM international conference on multimedia, 2022: 4003-4011.

[11] ZHAO Wenda, XIE Shigeng, ZHAO Fan, et al. Metafusion: infrared and visible image fusion via meta-feature embedding from object detection[C].Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 13955-13965.

[12] WU Yuhui, LIU Zhu, LIU Jinyuan, et al. Breaking free from fusion rule: A fully semantic-driven infrared and visible image fusion[J]. IEEE Signal Processing Letters, 2023, 30: 418-422.

[13] LIU Xiaowen, HUO Hongtao, LI Jing, et al. A semantic-driven coupled network for infrared and visible image fusion[J]. Information Fusion, 2024, 108: 102352.

[14] TANG Linfeng, ZHANG Hao, XU Han, et al. Rethinking the necessity of image fusion in high-level vision tasks: a practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity[J]. Information Fusion, 2023, 99: 101870.

[15] CHEN Xiaoxuan, XU Shuwen, HU Shaohai, et al. UIFD: a unified interactive network for image fusion and traffic object detection under low-light conditions[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(11): 19182-19196.

[16] YANG Zengyi, ZHANG Yafei, LI Huafeng, et al. Instruction-driven fusion of Infrared - visible images: Tailoring for diverse downstream tasks[J]. Information Fusion, 2025, 121: 103148.

[17] BAKIRCI M. Performance evaluation of low-power and lightweight object detectors for real-time monitoring in resource-constrained drone systems[J]. Engineering Applications of Artificial Intelligence, 2025, 159: 111775.

[18] DENG Jian, LI Mingyue, CHEN Yongxin, et al. Cross-guided feature fusion with intra-modality reweighting for multi-spectral pedestrian detection[C].2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022: 4864-4870.
[19] MA J, TANG L, XU M, et al. STDFusionNet: an infrared and visible image fusion network based on salient target detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.
[20] 杨轲.可见光与红外图像融合和目标检测多任务学习研究[D].成都:电子科技大学, 2024.
[21] JOCHER G. YOLOv5 by Ultralytics[CP/OL]. San Francisco: Ultralytics, 2020(2020-05)[2026-01-31]. https://github.com/ultralytics/yolov5.
[22] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[23] HE Yang, LIU Ping, WANG Ziwei, et al. Filter pruning via geometric median for deep convolutional neural networks acceleration[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 4340-4349.
[24] YANG C, LUO X, ZHANG Z, et al. KDFuse: a high-level vision task-driven infrared and visible image fusion method based on cross-domain knowledge distillation[J]. Information Fusion, 2025, 118: 102944.
[25] MA Jiayi, TANG Linfeng, FAN Fan, et al. SwinFusion: cross-domain long-range learning for general image fusion via swin transformer[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.
[26] WANG Zhou, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
[27] GONG Ting, LEE T, STEPHENSON C, et al. A comparison of loss weighting strategies for multi task learning in deep neural networks[J]. IEEE Access, 2019, 7: 141627-141632.
[28] LIU Shikun, JOHNS E, DAVISON A J. End-to-end multi-task learning with attention[C].Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 1871-1880.
[29] ZHANG Xingchen, DEMIRIS Y. Visible and infrared image fusion using deep learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10535-10554.
[30] SUN Bin, GAO Yunxiang, ZHUGE Wuwei, et al. Analysis of quality objective assessment metrics for visible and infrared image fusion[J]. Journal of Image and Graphics, 2023, 28(1): 144-155.
孙彬, 高云翔, 诸葛吴为, 等. 可见光与红外图像融合质量评价指标分析[J]. 中国图象图形学报, 2023, 28(1): 144-155.
[31] LI Hao, XU Zheng, TAYLOR G, et al. Visualizing the loss landscape of neural nets[J]. Advances in neural information processing systems, 2018, 31: 6389 - 6399.

# Parallel Network for Dual-band Image Fusion and Object Detection with Lightweight Deployment

LIU Yuhang[1,2,3], LI Wenbo[1,2,3], YANG Ke[1,2,3], SUN Bin[1,2,3], WANG Zijun[1,2,3], JIANG Dagang[1,2,3]

(1 *School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu* 611731, *China*)

(2 *National Laboratory on Adaptive Optics, Chengdu* 610209, *China*)

(3 *Aircraft Swarm Intelligent Sensing and Cooperative Control Key Laboratory of Sichuan Province, Chengdu* 611731, *China*)

**Abstract**: The purpose of this study is to address the limitations of cascade-based dual light object detection frameworks by developing a parallel multi-task learning network capable of jointly performing dual-band image fusion and object detection. Instead of enforcing a strict sequential dependency between fusion and detection tasks, the proposed approach aims to achieve collaborative optimization through shared feature learning while preserving task-specific representations. By mitigating the common error propagation and task interference in traditional cascade architectures, this study aims to leverage the advantages of dual-optical feature complementarity to improve detection accuracy and fusion quality in complex lighting scenarios, while maintaining deployability on resource-constrained edge computing platforms. To accomplish the stated objective, a novel parallel multi-task network architecture is designed, in which dual light image fusion and object detection are simultaneously learned within a unified framework. Unlike

conventional cascade-based approaches, the proposed architecture enables both tasks to share intermediate representations while retaining task-specific feature pathways, thereby reducing error accumulation caused by sequential processing. A pseudo-twin feature extraction module is first introduced to process RGB and thermal inputs separately, enabling the network to capture modality-specific characteristics while maintaining structural consistency between feature streams. This design ensures that each modality preserves its inherent representational advantages during early-stage feature encoding. To further enhance cross-modal representation learning, a channel-attention-based feature fusion module is incorporated, allowing the network to adaptively emphasize informative channels and suppress redundant or noisy features across modalities. Through learned channel-wise weighting, the fusion module facilitates effective inter-modal interaction and improves the discriminative capability of the shared feature space. Considering the heterogeneous feature requirements of the two tasks, the proposed network explicitly differentiates between shared and task-oriented feature representations. The shared backbone focuses on extracting robust cross-modal representations, while dedicated task heads are designed to accommodate the specific objectives of image fusion and object detection. For the image fusion branch, cross-scale skip connections are employed to preserve fine-grained spatial details and high-frequency structural information that are critical for generating perceptually informative fused images. These connections enable low-level texture features to be effectively propagated and integrated with higher-level semantic representations, thus enhancing visual quality and structural consistency in the fused outputs. In parallel, the object detection branch prioritizes higher-level semantic features and multi-scale contextual information to ensure robust object localization and classification under varying illumination and environmental conditions. To mitigate potential optimization conflicts arising from joint learning, an adaptive task loss weighting strategy is further designed. Instead of relying on fixed loss coefficients, this mechanism dynamically adjusts the relative contributions of fusion and detection losses during training based on task learning dynamics. By continuously balancing the optimization objectives of both tasks, the proposed strategy improves convergence stability, prevents dominance of a single task, and promotes coordinated multi-task learning throughout the training process.Comprehensive qualitative and quantitative experiments are conducted on the $M^3FD$ dual light dataset to evaluate the effectiveness of the proposed parallel multi-task learning framework. Experimental results demonstrate that the proposed method consistently outperforms baseline cascade-based models in both object detection accuracy and image fusion quality. In terms of detection performance, the proposed network achieves a relative improvement of 2.8% on visible-light images and 4.2% on thermal images compared with the baseline method. Qualitative evaluations further indicate that the fused images generated by the proposed approach exhibit clearer structural contours and enhanced target saliency, which are beneficial for downstream detection tasks. Furthermore, to reduce model complexity and improve deployment efficiency, structured pruning is applied to the trained network, followed by sufficient retraining to obtain a lightweight model. Experimental results show that the pruned model retains only 52.1% of the original parameters, while reducing the GPU inference latency from 22.9 ms to 16.3 ms, with a detection accuracy drop of only 1.7%. Compared with the baseline YOLOv5s model, the proposed lightweight network achieves a detection accuracy improvement of 2.4% while reducing the parameter count by approximately 4.5 million, demonstrating its superior compactness and efficiency. In addition to algorithmic performance, deployment experiments are carried out on an edge computing platform equipped with the RK3588 chip to assess real-world applicability. The results show that the proposed model can be successfully deployed and executed under limited computational and memory resources, while maintaining satisfactory detection accuracy. These research findings validate that the proposed parallel multi-task architecture achieves a good balance between performance and computational efficiency, making it suitable for practical complex lighting scenarios. The proposed parallel multi-task learning network effectively realizes the joint optimization of dual light image fusion and object detection, demonstrating superior performance and deployment feasibility on resource-constrained edge platforms.

**Key words**: Dual-band images; Image fusion; Object detection; Multi-task learning; Model lightweighting

**OCIS Codes**: 110.2970; 040.1880; 040.3060

**CSTR**: 32255.14.gzxb20265508.0810002

**Foundation item**: National Natural Science Foundation of China (62020106011), National Natural Science Foundation of Sichuan Province (2025ZNSFSC0002), Sichuan Science and Technology Program (2022YFG0050), Chengdu Science and Technology Project (2025-XT00-00024-GX)