异构模型多层次蒸馏的红外-可见光图像融合

张 棋[1],宋 红[1*],李金夫[2],马士瀚[1],林毓聪[2],杨 健[2]
(1. 北京理工大学计算机学院,北京 100081;2. 北京理工大学光电学院, 北京 100081)

摘 要: 知识蒸馏可将复杂教师网络的表征能力迁移至轻量学生网络,有效提升模型性能与部署效率. 然而,现有基于知识蒸馏的多模态图像融合方法常忽视师生网络的特征表示、模态偏好异构性及多模态图像的固有差异,导致知识传递低效、语义对齐不足及融合性能退化. 针对上述问题,本文提出基于异构模型多层次知识蒸馏的红外与可见光图像融合方法,创新性设计跨层级知识传递机制,在特征层通过注意力引导红外显著性目标与可见光纹理的精准迁移;在关系层通过相似性关系匹配与拓扑结构对齐优化跨模态语义适配;在输出层通过响应约束确保融合结果的视觉一致性与语义完整性,缓解了师生网络模态偏好不匹配导致的信息失衡. 此外,构建适配任务特性的轻量化CNN-Transformer双分支学生网络,兼顾全局信息建模与局部细节感知,增强对异构知识的接收整合能力. 在MSRS、RoadScene、TNO和M3FD数据集上的实验结果表明,所提方法在三种结构差异显著的教师模型的指导下,互相关系数(Correlation Coefficient,CC)、峰值信噪比(Peak Signal-to-Noise Ratio,PSNR)、空间频率保持度(Sum of the Correlations of Differences,SCD)和结构相似性指数(Structural Similarity Index Measure,SSIM)四项指标均优于教师模型及现有方法,且模型参数量仅为0.077 2 M,服务器上推理时间仅为31.22 ms,在提升融合性能与蒸馏鲁棒性的同时,实现了融合网络的轻量化与实时性;同时模型在Jetson AGX Xavier边缘平台上推理时间仅为250.31 ms,具备良好的边缘部署能力与实际应用价值.

关键词: 红外与可见光图像融合;知识蒸馏;异构模型;轻量化设计;特征对齐
基金项目: 北京市自然科学基金(No.L242024);国家自然科学基金(No.U22A2052)
中图分类号: TP391.4;TH701 **文献标识码:** A **文章编号:** 0372-2112(2025)12-4250-17
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250764

Infrared-Visible Image Fusion via Heterogeneous Multi-Level Distillation

ZHANG Qi[1], SONG Hong[1*], LI Jin-fu[2], MA Shi-han[1], LIN Yu-cong[2], YANG Jian[2]
(1. *School of Computer Science and Technology*, *Beijing Institute of Technology*, *Beijing* 100081, *China*;
2. *School of Optics and Photonics*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: Knowledge distillation transfers the representation capability of a complex teacher network to a lightweight student network, thereby enhancing model performance and deployment efficiency. However, existing knowledge distillation-based multimodal image fusion methods often neglect the heterogeneity of feature representations and modality preferences between teacher and student networks, as well as the inherent differences across modalities. This limitation results in inefficient knowledge transfer, insufficient semantic alignment, and degraded fusion performance. To address these issues, we propose an infrared and visible image fusion method based on heterogeneous model multi-level knowledge distillation. Specifically, a cross-layer knowledge transfer mechanism is designed: at the feature layer, attention is utilized to guide the precise transfer of infrared salient targets and visible-light textures; at the relationship layer, similarity-based relational matching and topological structure alignment are employed to enhance cross-modal semantic adaptation; and at the output layer, response constraints are applied to ensure both visual consistency and semantic integrity of the fused results, alleviating the information imbalance caused by mismatched modality preferences between teacher and student networks. In addition, we construct a task-adaptive lightweight CNN-Transformer dual-branch student network that simultaneously models global information and captures local details, thereby enhancing its ability to receive and integrate heterogeneous knowledge. Experimental results on the MSRS, RoadScene, TNO, and M3FD datasets demonstrate that under the guidance of

收稿日期:2025-09-04;录用日期:2025-12-08;责任编辑:覃怀银
*通讯作者:宋红

three teacher models with significantly different architectures, the proposed method outperforms both the teacher models and state-of-the-art approaches in terms of correlation coefficient (CC), peak signal-to-noise ratio (PSNR), sum of the correlations of differences (SCD) and structural similarity index measure (SSIM) metrics, while requiring only 0.077 2 M parameters and achieving 31.22 ms inference time on a server platform. Moreover, the model maintains an inference time of 250.31 ms on the Jetson AGX Xavier edge platform, indicating strong suitability for edge deployment and practical applications.

Key words: infrared-visible image fusion; knowledge distillation; heterogeneous models; lightweight design; feature alignment

Foundation Item(s): Beijing Natural Science Foundation (No.L242024); National Natural Science Foundation of China (No.U22A2052)

1　引言

在现代计算机视觉与智能感知领域,红外与可见光图像融合(Infrared and Visible Image Fusion,IVIF)作为一种关键的信息增强技术,已广泛应用于安防监控、军事侦察、遥感监测等重要场景[1]. 红外图像能够捕捉物体热辐射特征,在复杂光照或恶劣天气条件下仍能清晰呈现目标轮廓;可见光图像则凭借丰富的纹理细节与色彩信息,精准展现场景的外观特征. 二者融合可生成兼具目标显著性与场景细节完整性的高质量融合结果,为后续目标检测[2]、跟踪[3]、分割[4]等高层视觉任务提供更为可靠的数据支撑.

随着深度学习方法的不断发展,图像融合技术逐步从依赖人工设计规则的传统方法[5,6]转向以端到端神经网络为代表的数据驱动范式[7,8]. 早期研究广泛采用卷积神经网络(Convolutional Neural Network,CNN)[9]进行多模态图像特征提取与整合,通过渐进式特征学习机制实现了融合性能的显著提升. 为进一步增强对图像全局结构和语义关系的建模能力,研究者引入基于自注意力机制的Transformer架构[10]显式建模长程依赖,有效提升了融合结果的语义一致性与结构完整性. 与此同时,为兼顾不同模态特征的互补优势与局部细节保持能力,一些融合模型采用CNN-Transformer混合或双分支架构,分别处理局部纹理与全局上下文,进一步推动了融合质量的改善. 然而,这些深度融合模型通常存在结构复杂、参数量大、计算开销高的问题,尤其是引入Transformer或复杂混合架构的模型,往往庞大的参数量和高昂的计算成本,导致其难以在计算资源受限的终端设备[11]中实现高效推理与实时部署.

知识蒸馏(Knowledge Distillation,KD)作为提升深度神经网络实用价值的核心技术,自Hinton等人[12]提出以来已成为模型压缩与性能优化的重要手段. 通过建立“教师-学生”网络架构,将复杂教师模型的知识迁移至轻量级学生模型,在大幅降低模型参数量与计算复杂度的同时,保留了教师模型的优异性能,为平衡融合质量与部署效率提供了新思路. 即借助预训练教师网络的学习经验,指导学生网络掌握跨模态特征表示与融合策略. 然而,当前基于知识蒸馏的融合方法仍面临诸多挑战,制约了其性能潜力的充分发挥.

如图1所示,目前多数知识蒸馏融合方法[13]依赖输出层或特征层的简单对齐实现知识传递,在处理结构差异显著的师生模型时暴露出明显局限. 为规避特征语义错位与梯度不稳定等风险,部分方法[14,15]强制要求师生模型保持高度同构性,虽然简化了蒸馏过程,却极大限制了教师模型的选择范围,同时压缩了学生网络的轻量化设计空间,难以平衡性能与效率. 针对这一问题,已有研究[16,17]开始探索异构师生网络蒸馏方案,试图突破同构框架的约束,但在实际应用中仍存在三方面问题亟待解决:

(1)师生特征表示异构性被忽视. 教师网络通常基于大规模数据集训练,其特征空间与面向特定融合任务设计的学生网络存在显著差异,直接蒸馏易导致知识传递失真,无法充分发挥教师的指导价值.

(2)多模态固有差异未有效处理. 红外与可见光图像在成像机理、模态特性上存在本质区别,教师网

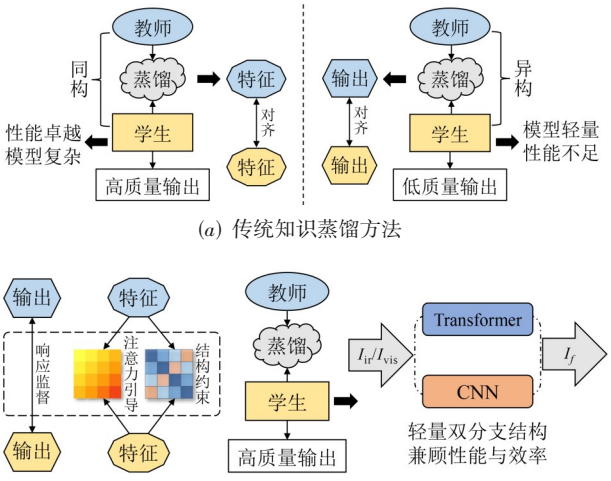


图1　基于知识蒸馏的红外与可见光图像融合方法对比

络形成的模态关注偏好往往与学生网络的学习需求不匹配,导致蒸馏过程中模态信息失衡,融合结果易出现细节丢失或伪影.

(3)蒸馏策略单一化限制了知识传递效率. 现有方法多聚焦于输出响应或浅层特征的模仿学习,缺乏对教师网络深层语义知识与跨模态关联信息的有效挖掘,难以支撑复杂场景下的高质量融合需求,造成知识传递低效与性能退化.

针对上述问题,本文提出了一种基于异构模型多层次知识蒸馏的红外与可见光图像融合网络,从综合利用教师输出响应、中间特征与语义关系的蒸馏策略和具备跨结构适配能力的双分支学生结构两个方面,缓解了师生结构不一致带来的特征异构性影响,实现了在学生结构固定情况下对多种异构教师稳健且高鲁棒性知识迁移.

本文主要贡献如下:

(1)提出了一种面向异构师生模型的多层次知识蒸馏框架,有效缓解了师生网络模态偏好不匹配导致的信息失衡问题,并克服了单一蒸馏策略对深层语义信息表征受限与利用不足的局限.

(2)设计了一种异构双分支轻量化学生网络,在保持轻量化与高效推理特性的同时,兼顾全局语义建模与局部细节保真,从而增强其对异构知识的接收与整合能力.

(3)在 MSRS、RoadScene、TNO 和 M^3FD 四个公开数据集上对所提方法与当前最先进的融合方法进行了全面比较,并进一步在多种教师模型指导下验证了框架的通用性与稳定性. 实验结果表明,所提方法在主观视觉效果和客观评价指标上均取得更优表现,同时在推理速度与计算效率方面展现出显著优势.

2 相关工作

2.1 红外与可见光图像融合

红外与可见光图像融合旨在结合多源图像的互补信息,生成兼具结构显著性与细节清晰度的融合图像. 早期方法多基于小波变换[18]、拉普拉斯金字塔[19]和主成分分析[20]等数学变换方法,通过多尺度分解与手工融合规则对底层信息进行组合. 然而,此类方法大多局限于处理底层视觉特征,难以应对复杂场景中的语义整合. 随着深度学习的发展,基于卷积神经网络的方法在细节建模方面展现出较强能力. Li 等人[21]提出 DenseFuse 网络,结合卷积层与密集块提取深层特征;Zhang 等人[22]提出 IFCNN,通过共享特征提取器实现端到端模态融合;Ram 等人[23]提出无监督 CNN 融合框架 DeepFuse,通过端到端训练提升结构与细节保留效果. 然而,这类方法仍受限于卷积的局部感受野,难以建模远程依赖与全局结构. 当处理目标跨区域分布或者边缘不连续时,常出现显著性信息削弱等问题,严重影响融合质量.

为缓解卷积方法在远程建模上的局限,Transformer 被引入图像融合任务中,其在建模长距离依赖与复杂结构关系方面展现出显著优势. ViT[24]首次将 Transformer 架构引入视觉任务,通过将图像划分为固定大小的 patch 输入,摆脱了对卷积的依赖;Ma 等人[25]提出了 SwinFusion,基于层次化的 Transformer 设计滑动窗口与自注意力机制,融合全局语义与局部细节;Li 等人[26]提出了 MixFuse,利用混合注意力机制将特征融合与提取合并成单阶段,实现了更精细的融合. 基于 Transformer 的方法在结构感知与远程建模方面优于传统 CNN,但在细节保持方面却存在不足,高层建模机制常忽视低层边缘与纹理特征,导致融合图像细节表现不稳定.

为同时兼顾结构建模与细节保持,部分研究引入双分支融合架构,结合 Transformer 的全局建模能力与 CNN 的局部感知优势,协同提取红外显著目标与可见光纹理信息. 该类方法通常采用并行分支提取异构特征,在融合模块中实现信息交互与重建. Li 等人[27]提出的 DCTNet 设计了一种异构双分支多级级联网络,结合卷积与变换域特征实现双模态的高效互补融合;Xu 等人[28]提出 DAF-Net,结合双分支架构与多核最大均值差异实现特征空间对齐,从而有效提升融合质量;Zhao 等人[29]提出的 CDDFuse 则采用 Transformer 与 INN 双分支结构,分别提取全局低频和局部高频特征,并通过相关性驱动的损失函数实现跨模态特征分解与互补融合,在结构与细节保持上均取得了优异表现. 此类方法在性能上通常表现出色,但其模型结构普遍较为复杂,计算路径长、参数量大,训练与推理阶段的资源消耗显著增加,难以在实时或资源受限的场景中部署,限制了其实用性.

2.2 知识蒸馏

知识蒸馏最初由 Hinton 等人[12]提出,旨在通过性能更强的教师模型监督轻量学生模型学习,在保持低参数量的同时提升表达能力. 作为一种高效模型压缩手段,蒸馏技术已广泛应用于图像分类、目标检测等视觉任务中. 其核心在于构建合理的指导信号,引导学生模型有效吸收教师模型的表征能力. 最初的研究多聚焦于响应蒸馏[12],即通过教师模型最终输出的软标签[30]指导学生学习. 该类方法实现简单、通用性强,在教师与学生模型结构相似的情况下表现稳定. 但其仅关注最终输出,忽视了教师模型的中间语义与结构建模,难以适应复杂任务.

为弥补响应蒸馏的不足,后续研究进一步拓展至

更深层次的知识迁移,应用最为广泛的为特征蒸馏[31]与关系蒸馏[32]. 特征蒸馏通过对齐中间特征实现更细粒度的监督,Wang等人[33]在目标检测任务蒸馏了教师的锚点特征;Wang等人[34]将学生检测器的中间特征与教师检测器的不同检测头进行跨头特征蒸馏,有效提升了目标检测性能;Liu等人[35]提出特征差异蒸馏,以差异化策略对齐不同层级特征. 这类方法更适合处理表征复杂的视觉任务,但其高度依赖特征空间一致性,在结构差异较大时易失效,针对特征的映射操作难以弥合语义差异,影响迁移稳定性.

相比之下,关系蒸馏通过建模样本间或通道间的相对关系实现结构无关的知识迁移,具备一定的结构无关性. Ni等人[36]引入空间与层间双重关系蒸馏;Yang等人[37]利用跨图像语义关系提升结构建模能力. 这类方法具备更强的泛化能力与鲁棒性,但对关系建模策略的设计较为依赖,不同任务和结构对关系表示的敏感性差异显著,且大规模关系构建易引入噪声与冗余,影响指导信号稳定性,其可控性和通用性仍有待提升.

2.3　图像融合中的知识蒸馏

近年来,部分研究开始尝试将知识蒸馏机制引入图像融合任务,以提升轻量网络在保持融合质量前提下的表达能力与建模效率. Mi等人[38]提出KDE-GAN利用蒸馏传递多模态特征,结合可解释模块提升性能与透明度;Deng等人[39]在MMDRFuse中设计多阶段蒸馏机制,动态引导学生逐步学习多模态信息,在超轻量模型下仍保持一定的性能. 这些方法验证了知识蒸馏在融合任务中的有效性,但其蒸馏机制仍存在较大局限. 当前多数方法依赖教师的输出响应或中间特征监督,侧重结果一致性或显著性区域重构,难以应对师生结构差异带来的表达不匹配,导致融合性能下降. 此外,缺乏适用于异构结构的通用蒸馏框架,语义偏移问题频发,影响蒸馏稳定性与泛化能力. 为此,本文提出一种基于异构模型多层次知识蒸馏的红外与可见光图像融合网络,有效提升了异构条件下学生模型的结构建模能力与细节保持能力.

3　提出的方法

3.1　方法概述

本文方法主要由两部分构成:一是面向多模态特征对齐的多层次知识蒸馏框架(Multi-Level Knowledge Distillation,MLKD). 该方法从特征层注意力引导、关系层相似性结构匹配与输出层响应约束三个层级构建互补蒸馏路径,协同约束学生网络的学习过程,实现对教师网络知识的高效迁移;二是CNN-Transformer异构双分支学生网络,两个分支分别侧重于全局结构建模与局部细节感知,从而在蒸馏学习中提供更强的异构知识承载能力与表达能力. 在蒸馏过程中,教师只需提供解码前的中间特征和最终输出结果,MLKD将从多个层次对这些信息进行处理并迁移给学生.

3.2　面向异构网络的多层次蒸馏方法

MLKD的整体结构如图2所示,不同于传统蒸馏策略需要固定的教师模型,MLKD的教师是可变的,教师和学生只需提供中间特征F_{Tea}、F_{Stu}以及输出结果I_{Tea}、I_{Stu}. 其中,F_{Tea}、F_{Stu}常取模型解码前的中间特征. 随后,MLKD将从3个层级实现知识迁移.

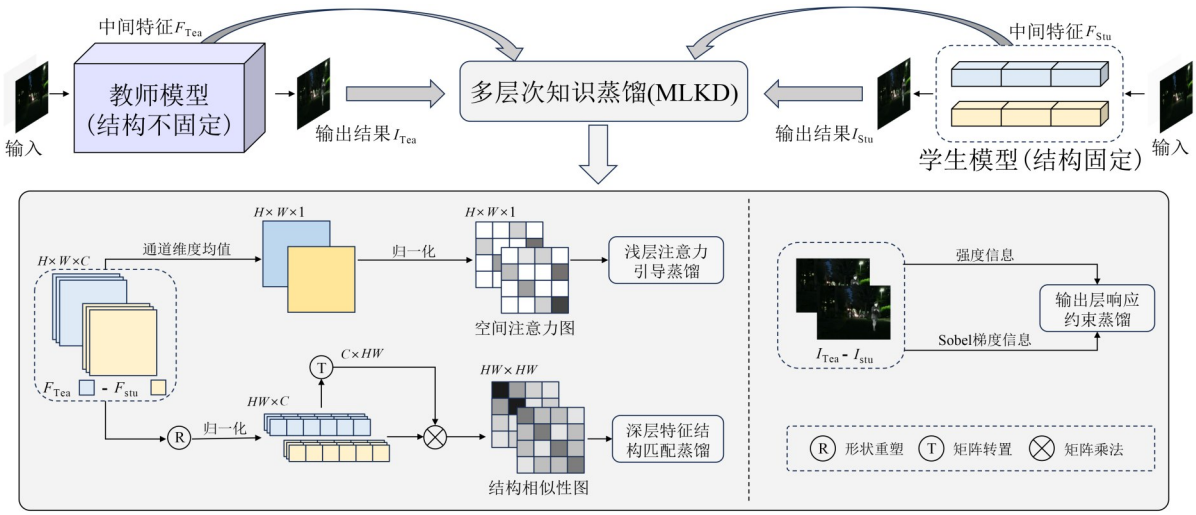


图2　本文提出的基于异构模型多层次知识蒸馏的红外与可见光图像融合架构

在特征层注意力引导中,MLKD通过蒸馏师生中间特征的空间注意力图实现知识传递. 具体而言,首先对教师与学生网络中间层提取的特征图沿通道维度计算均值,得到二维空间注意力图. 对该注意力图进行归一

化处理后展平成一维向量，公式表达如下：

$$\boldsymbol{A}_{\text{norm}}=\frac{\text{Flatten}\left(\frac{1}{C}\sum_{C=1}^{C}F_C\right)}{\left\|\text{Flatten}\left(\frac{1}{C}\sum_{C=1}^{C}F_C\right)\right\|_2} \tag{1}$$

其中，$F_C\in\mathbb{R}^{H\times W\times C}$为教师或学生提取的中间特征图；$\boldsymbol{A}_{\text{norm}}$表示对应的归一化注意力矩阵. 随后，计算教师与学生注意力向量之间的余弦相似度损失，形式为

$$L_{\text{att}}=1-\frac{\boldsymbol{A}_{\text{norm}}^{\text{Tea}}\cdot\boldsymbol{A}_{\text{norm}}^{\text{Stu}}}{\|\boldsymbol{A}_{\text{norm}}^{\text{Tea}}\|\cdot\|\boldsymbol{A}_{\text{norm}}^{\text{Stu}}\|} \tag{2}$$

该损失引导学生网络聚焦与教师相似的空间区域，提升空间感知能力.

在关系层相似性结构匹配中，MLKD对表示师生特征各区域结构关系分布的相似性矩阵[40]进行蒸馏. 首先对师生的中间特征图进行双线性插值下采样处理，然后展平成二维矩阵. 对矩阵中每一行执行L2归一化：

$$\boldsymbol{X}_{\text{norm}}=\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2} \tag{3}$$

其中，$\boldsymbol{X}\in\mathbb{R}^{N\times C}$表示展平后的特征矩阵；$N$为像素点的数量；$C$为通道数. 为了降低特征匹配的计算复杂度，下采样的目标空间尺寸为32 × 32，此时N = 32 × 32 = 1 024. 在该分辨率下，所生成的相似性矩阵规模适中，不会引入额外的计算瓶颈. 归一化矩阵与其转置相乘得到相似性矩阵：

$$\boldsymbol{S}=\boldsymbol{X}_{\text{norm}}\cdot\boldsymbol{X}_{\text{norm}}^{\text{T}} \tag{4}$$

教师和学生分别计算出相似性矩阵$\boldsymbol{S}^{\text{Tea}}$和$\boldsymbol{S}^{\text{Stu}}$，蒸馏损失定义为两者逐元素平方差：

$$L_{\text{sim}}=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\boldsymbol{S}_{i,j}^{\text{Tea}}-\boldsymbol{S}_{i,j}^{\text{Stu}}\right)^2 \tag{5}$$

该损失促使学生模型准确捕捉图像空间结构关系，增强全局一致性.

在输出层响应约束中，MLKD则是以教师与学生的融合结果为蒸馏对象. 考虑到图像的感知质量在很大程度上依赖于强度分布的合理性与边缘细节的清晰度，本文从像素强度和梯度信息两个维度监督学生输出. 强度损失L_{int}采用教师和学生的输出图像，取二者像素绝对值；同时，应用Sobel算子提取输出图像梯度，计算梯度特征间的绝对值，即为梯度损失L_{grad}. 输出响应损失L_{out}可表示为

$$L_{\text{int}}=\frac{1}{HW}\sum_{i,j}\left|I_{i,j}^{\text{Tea}}-I_{i,j}^{\text{Stu}}\right| \tag{6}$$

$$L_{\text{grad}}=\frac{1}{HW}\sum_{i,j}\left|\mathcal{G}(I^{\text{Tea}})_{i,j}-\mathcal{G}(I^{\text{Stu}})_{i,j}\right| \tag{7}$$

$$L_{\text{out}}=L_{\text{int}}+\lambda_{\text{grad}}L_{\text{grad}} \tag{8}$$

其中，I^{Tea}、$I^{\text{Stu}}\in\mathbb{R}^{H\times W\times 1}$分别表示教师和学生的输出图像；$\mathcal{G}(\cdot)$表示Sobel梯度算子；$H$和$W$表示图像的高度和宽度；$\lambda_{\text{grad}}$为梯度损失权重系数，以强化细节恢复和边缘保持.

3.3 异构双分支学生网络架构

异构双分支学生网络结构如图3所示，学生网络包含两条功能互补的子分支：由Restormer模块[41]（ReSTormer Block，RSTB）串联而成的全局建模分支负责捕捉跨模态的长距离结构关系与高层语义特征，增强模型的结构理解能力；由细节感知卷积模块（Detail-Aware Convolution Block，DACB）串联而成的局部感知分支则专注于纹理、边缘等高频细节的感知与增强，提升融合图像的清晰度与细节表现. 两路分支提取出的全局信息和局部信息再由RSTB统一建模，最终生成融合图像. 双分支网络所提取的结构化特征在网络内部具备清晰的语义分层，也为知识蒸馏提供了良好的表示基础.

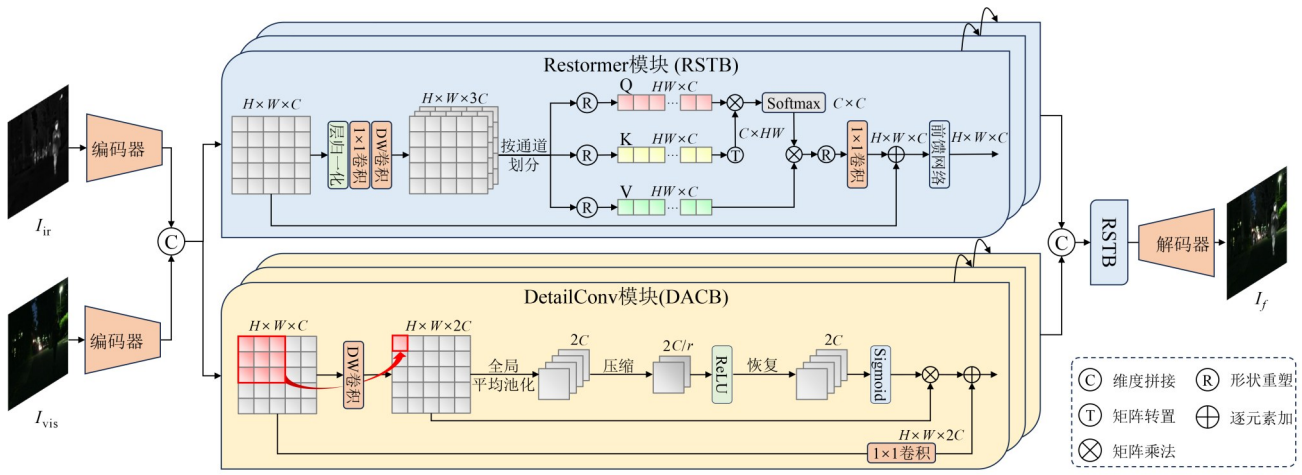


图3 本文设计的异构双分支学生网络结构

3.3.1 全局建模分支

为了充分挖掘不同模态间的全局信息与结构关系,增强它们的深层语义交互,本文在学生模型中引入全局建模分支,以增强模型的全局建模能力和对结构信息的感知能力. 首先,红外图像I_{ir}与可见光图像I_{vis}通过编码器提取浅层特征;随后,沿通道进行拼接,得到融合特征$F_0 \in \mathbb{R}^{H\times W\times C}$,其过程可表示为

$$F_0=\mathrm{Concat}\left(E_{\mathrm{ir}}(I_{\mathrm{ir}}),\ E_{\mathrm{vis}}(I_{\mathrm{vis}})\right) \tag{9}$$

其中,I_{ir}和I_{vis}分别表示红外图像与可见光图像;E_{ir}和E_{vis}分别表示对应的编码器. 拼接后的特征送入由3个RSTB串联而成的全局建模分支中,渐进式提取全局特征,其过程为

$$F_g=\mathrm{RSTB}_3\left(\mathrm{RSTB}_2\left(\mathrm{RSTB}_1(F_0)\right)\right) \tag{10}$$

其中,$F_g \in \mathbb{R}^{H\times W\times C}$是由全局建模分支提取出的全局信息特征. 每个RSTB块由归一化层、多头自注意力机制、门控前馈网络和残差连接组成. 与标准ViT结构不同,Restormer采用了轴向注意力机制,具体而言,首先通过深度可分离卷积[42]将特征的通道数扩张3倍,而后按通道维度进行拆分和变形,得到了与输入特征通道数相同的Q、K、V,并继续完成自注意力的计算,建模空间位置间的长距离依赖关系. 这种设计的计算开销要显著低于标准的二维多头自注意力,并提升了全局建模的效率和精度. 门控前馈网络则对通道维度的特征表达进行动态调节,使不同模态的信息在融合过程中得到更加有效的表征与融合. 每个RSTB中将注意力头数设为4,通道数设为32,参数规模显著低于传统Transformer架构,同时保留了较强的特征建模能力.

3层RSTB模块以残差结构逐层堆叠,逐渐增强融合特征的语义完整性与结构一致性,全局建模分支最终输出的特征F_g携带丰富的高层全局语义信息,作为主干表示的一部分与局部感知特征进行进一步融合. 为后续的融合过程提供了完整且均衡的多模态特征表达基础.

3.3.2 局部感知分支

除了全局结构信息外,局部纹理细节和边缘信息的精准表达同样不可或缺. 为此,本文在学生网络中引入局部感知分支,专门面向纹理细节和边缘特征的增强需求. 该分支由3层DACB串联而成,融合了深度可分离卷积与通道注意力机制[43]两类轻量而高效的局部建模策略,在保持模型轻量的同时实现高质量的局部感知与细节强化.

与全局建模分支相同,局部感知分支同样以浅层拼接特征$F_0 \in \mathbb{R}^{H\times W\times C}$为输入,依次经过3个模块进行特征提取与细节强化. 整体过程可表示为

$$F_1=\mathrm{DACB}_3\left(\mathrm{DACB}_2\left(\mathrm{DACB}_1(F_0)\right)\right) \tag{11}$$

其中,$F_1 \in \mathbb{R}^{H\times W\times C}$是由局部感知分支提取出的局部信息特征. 每个DACB模块首先采用深度可分离卷积对输入进行处理. 随后,模块引入通道注意力机制对特征响应进行动态调节. 具体而言,首先对输出特征进行全局平均池化,获得通道统计向量$\boldsymbol{s} \in \mathbb{R}^C$;随后,该向量依次通过两层全连接网络,第一层采用ReLU激活函数,第二层则采用Sigmoid激活函数,生成归一化权重$\boldsymbol{s}_{\mathrm{norm}} \in \mathbb{R}^C$;最终,原始通道特征按权重进行重标定,实现对显著通道的增强和冗余通道的抑制. 该过程可表示为

$$F_{\mathrm{out}}=\sigma\left(\boldsymbol{W}_2\cdot\delta\left(\boldsymbol{W}_1\cdot\mathrm{GAP}(F_{\mathrm{in}})\right)\right)\cdot F_{\mathrm{in}} \tag{12}$$

其中,GAP(·)表示全局平均池化;$\boldsymbol{W}_1$和$\boldsymbol{W}_2$为全连接层权重;$\delta(\cdot)$为ReLU激活函数;$\sigma(\cdot)$为Sigmoid函数. 此外,为增强特征流动稳定性,缓解深层网络中梯度衰减问题,DACB模块内部引入残差连接,将输入与注意力加权输出进行逐元素相加,提升特征传递效率并保留原始结构信息.

三个DACB模块按上述结构层层堆叠,使输入特征在多个语义层次上反复经历卷积提取与通道重标定的复合处理,从而实现局部纹理细节与边缘结构的逐级提炼与强化,增强模型对高频复杂信息的建模能力.

最终,局部感知分支输出的特征F_1与全局建模分支提取的高层语义特征F_g在通道维度上进行拼接,形成融合特征$F_f \in \mathbb{R}^{H\times W\times 2C}$:

$$F_f=\mathrm{Concat}(F_g,F_1) \tag{13}$$

3.3.3 统一建模与解码

为充分整合全局建模分支与局部感知分支提取的互补特征信息,本文在融合阶段引入统一建模模块,以进一步提升融合表示的耦合能力与语义一致性. 该模块采用单层RSTB构建,能够在保持轻量计算的同时,通过轴向注意力机制对融合特征F_f中的跨通道与跨空间依赖关系进行建模与重组,其处理过程可表示为

$$F_f=\mathrm{RSTB}\left(\mathrm{Conv}(F_f)\right) \tag{14}$$

其中,Conv(·)表示1×1卷积,用于恢复F_f的通道数至C. RSTB模块利用多头轴向注意力机制与门控前馈网络协同作用,能有效提升联合特征的结构对齐能力和表达完整性,为高质量图像重建奠定基础.

融合特征经过建模后送入解码模块进行最终图像重建. 该模块由两层卷积操作与中间的LeakyReLU激活函数组成,输出端使用Sigmoid函数将图像值归一化至[0,1]范围,整体过程可描述为

$$I_f=\sigma\left(\mathrm{Conv}_2\left(\mathrm{LReLU}\left(\mathrm{Conv}_1(F_f)\right)\right)\right) \tag{15}$$

其中,I_f表示融合图像;LReLU(·)表示LeakyReLU激活函数. 在训练过程中,统一建模后的融合特征F_f还被传

递到知识蒸馏中，以促进全局-局部协同表征在轻量学生模型中的高效迁移，进一步增强整体模型在复杂融合场景下的表现力与鲁棒性.

3.4 损失函数设计

本文总损失由融合任务损失与蒸馏损失共同构成，同时优化学生网络的图像融合能力与其对教师网络知识的继承能力. 总损失函数定义如下：

$$L_{\text{total}} = L_{\text{task}} + \alpha_{\text{kd}} L_{\text{kd}} \tag{16}$$

任务损失 L_{task} 由强度损失与梯度损失两部分组成. 强度损失用于衡量融合图像与输入图像在像素层面的相似程度，定义如下：

$$L_{\text{int}} = \frac{1}{N}\sum_{i=1}^{N}\left|\max(I_{\text{vis},i}, I_{\text{ir},i}) - I_{f,i}\right| \tag{17}$$

其中，I_{vis}、I_{ir} 和 I_f 分别表示输入的可见光图像、红外图像和融合结果图像；N 为图像中像素点的数量. 梯度损失通过Sobel算子提取图像边缘信息，对融合图像在结构保留方面进行约束，具体表达为

$$\mathcal{G}(I_{\max})_{i,j} = \max\left(\mathcal{G}(I_{\text{vis}})_{i,j}, \mathcal{G}(I_{\text{ir}})_{i,j}\right) \tag{18}$$

$$L_{\text{grad}} = \frac{1}{HW}\sum_{i,j}\left|\mathcal{G}(I_{\max})_{i,j} - \mathcal{G}(I_f)_{i,j}\right| \tag{19}$$

其中，$\mathcal{G}(\cdot)$ 表示使用Sobel卷积核计算以得到图像的梯度图；$\mathcal{G}(I_{\max})$ 表示通过对可见光图像与红外图像的梯度图进行逐像素最大值操作所得的源图像梯度图；H 和 W 分别表示图像的高度和宽度. 综合上述两项，任务损失表示为

$$L_{\text{task}} = L_{\text{int}} + \beta_{\text{grad}} L_{\text{grad}} \tag{20}$$

蒸馏损失 L_{kd} 用于约束学生网络教师网络知识的多层次学习，由注意力图特征损失 L_{att}、相似性图特征损失 L_{sim} 以及输出响应损失 L_{out} 组成：

$$L_{\text{kd}} = \lambda_{\text{att}} L_{\text{att}} + \lambda_{\text{sim}} L_{\text{sim}} + \lambda_{\text{out}} L_{\text{out}} \tag{21}$$

其中，λ_{att}、λ_{sim} 和 λ_{out} 为各损失项对应的权重系数，用于平衡各项损失占比，从而约束网络生成符合人眼视觉的高质量融合图像.

4 实验结果与分析

4.1 实验细节

4.1.1 数据集

本文在四个公开的红外与可见光数据集MSRS[44]、TNO[45]、RoadScene[46]及M^{3}FD[47]上对所提方法的有效性进行了验证. 其中，MSRS数据集涵盖街道、人行道、自然环境等多种典型户外场景；RoadScene数据集聚焦于昼夜交替下的城市交通场景；TNO数据集主要包含军事场景；M^{3}FD数据集涉及多种城市街景，且包含烟雾遮挡等复杂环境样本. 本文基于MSRS训练集的1 083对图像完成模型训练，并分别在MSRS的361对测试图像、RoadScene的221对测试图像、TNO的25对测试图像以及M^{3}FD的300对测试图像上开展丰富的对比实验与泛化实验.

4.1.2 实验设置

本文选取CDDFuse[29]、SHIP[48]、EMMA[49]这三种典型方法作为教师模型，它们均为近年来性能优异的代表性方法，且在融合范式上各具特色——CDDFuse采用特征分解的多分支协同建模，SHIP从频域角度引入高阶交互融合策略，EMMA利用自然成像等变先验实现语义层约束. 三种教师覆盖了特征层、频域层与语义层三种经典融合思路，能够充分验证所提方法对不同复杂结构教师网络的适应性与有效性.

(1)训练数据处理方面，将640 × 480训练集图像对裁剪成128 × 128的图像对，并随机选择了1 600对用于训练，批量大小设置为4.

(2)训练参数配置方面，采用Adam优化器[50]对学生模型进行参数更新，引入OneCycleLR学习率动态调度策略[51]，网络训练迭代轮数设为200，初始学习率为1×10^{-5}，调度器将学习率从初始值逐步升至最大值1×10^{-4}，随后采用余弦退火策略(cosine annealing)平滑下降至1×10^{-6}. 损失函数设计采用任务损失与蒸馏损失的组合形式，各损失项的权重系数设置如下：任务损失权重β_{grad}与输出响应蒸馏损失权重λ_{grad}均为10，总损失中蒸馏损失系数α_{kd}和蒸馏损失中各项子损失系数λ_{att}、λ_{sim}、λ_{out}均设为1.

(3)实验环境方面，所有核心实验均在搭载4块GeForce RTX 3080 Ti GPU的服务器上完成，计算复杂度分析在Jetson AGX Xavier边缘计算平台上完成，设备功率设置为MAXN. 模型训练与评估均在Ubuntu 18.04操作系统下基于PyTorch实现.

4.1.3 对比方法与评价指标

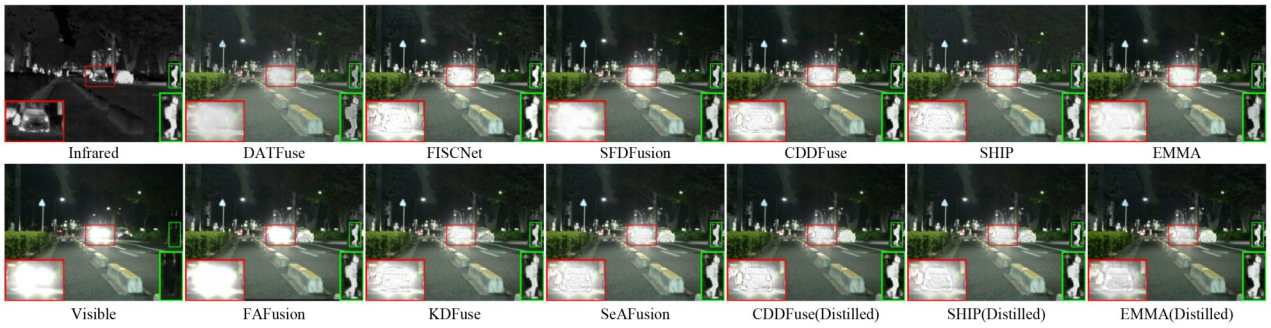
除3种教师模型方法外，本文额外选择了6种先进方法进行对比，包括DATFuse[52]、FISCNet[53]、FAFusion[54]、KDFuse[17]、SFDFusion[55]和SeAFusion[56]，以充分证明所提方法的有效性.

为全面量化评估融合图像的质量，本文采用了四种主流图像融合评价指标进行分析：互相关系数(Correlation Coefficient，CC)[57]、峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)[58]、空间频率保持度(Sum of the Correlations of Differences，SCD)[57]和结构相似性指数(Structural Similarity Index Measure，SSIM)[59]. 其中，CC衡量融合图像与源图像之间的线性相关程度，直观反映融合结果对源图像结构信息的保留能力；PSNR衡量融合图像与源图像的信噪比，量化融合结果的重构质量；SCD衡量融合图像与源图像之间差分图的相关性分析，表征融合图像对源图像边缘和结构的保持性能；

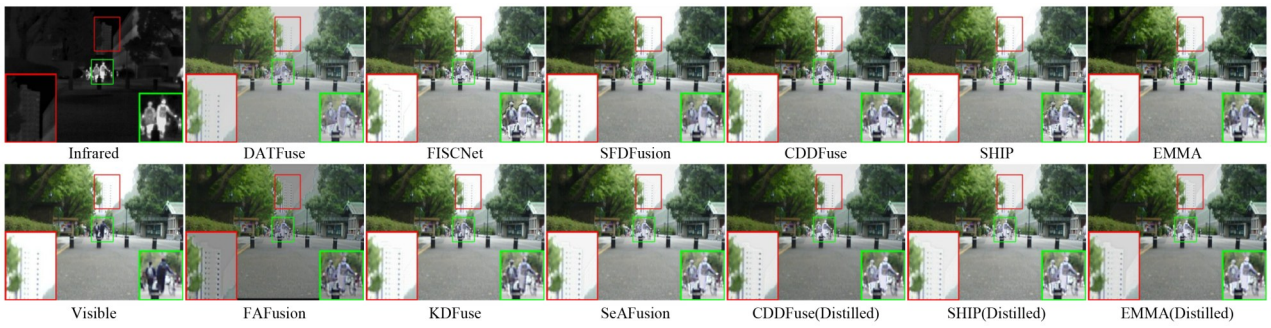
SSIM 则从亮度、对比度和结构相似性三个方面评价图像间的相似性. 上述指标均与图像质量呈正相关,即指标数值越高,表明融合效果越优.

4.2　对比实验结果与分析

本节分析了所提方法与最先进方法在 MSRS 数据集上的实验结果. 如图 4 所示,绿框为显著性目标,红框为纹理细节,通过放大进行详细比较. 定量结果如表 1 所示,加粗字体与下划线字体表示指标的最高值与次高值,特别地,为了更清晰地凸显蒸馏前后教师与学生的指标差异,在每个学生指标中增加了该组师生各指标差异值,绿色表示学生较教师有所上升,红色表示有所下降. 其中,CDDFuse(Distilled)、SHIP(Distilled)、EMMA(Distilled)分别表示每个教师对应的学生模型.



(*a*) MSRS 数据集中夜晚场景的定性对比图



(*b*) MSRS 数据集中白天场景的定性对比图

图 4　所提方法与对比方法在 MSRS 数据集上的定性对比图

4.2.1　定性分析

图 4(*a*)展示了 MSRS 数据集中夜间道路场景的对比结果. 可见光图像在车灯干扰下几乎丢失车辆纹理,而红外图像能清晰呈现整体结构. 在 10 种融合方法中,DATFuse、FAFusion 和 SFDFusion 未能保留纹理,KDFuse 和 EMMA 仅有少量,FISCNet、SeAFusion、CDDFuse 和 SHIP 稍优,但仍存在纹理缺失或边缘不连续等问题. 相比之下,三种学生模型效果更佳:CDDFuse(Distilled)在继承教师边缘信息的基础上展现了更丰富的细节纹理;SHIP(Distilled)和 EMMA(Distilled)分别突出更优的边缘结构和对比度,与教师优势保持一致. 图 4(*b*)展示了 MSRS 数据集中白天街道场景的对比结果. 可见光图像在强光的影响下,远处楼房细节严重丢失,而红外图像仅能提供边缘及部分纹理信息. 在融合结果中,DATFuse、SFDFusion、FAFusion 和 KDFuse 对远处楼房的结构恢复较差,边缘模糊,细节不足;FISCNet、SeAFusion 与 CDDFuse 虽能体现出一定的楼房轮廓,但细节纹理丢失,整体感不佳;SHIP 和 EMMA 在近处目标的表现稍优,但对远处细节的恢复仍显不足. 相比之下,CDDFuse(Distilled)在楼房边缘与纹理的恢复方面要优于原模型,能够较为完整地重建出远处结构;SHIP(Distilled)在保持较高整体对比度的同时,进一步增强了边缘的连贯性,使得楼房轮廓更加自然清晰;EMMA(Distilled)则在场景整体亮度和对比度的还原上更优,使红外与可见光信息的融合更为均衡. 整体来看,学生模型不仅继承了各自教师模型的优势,还通过异构双分支结构实现了局部细节与全局语义的协同建模,使融合效果在细节保真度和视觉感知上表现良好.

4.2.2　定量分析

表 1 中展示了 MSRS 数据集上的定量指标对比结果,学生模型整体表现出全面且稳定的提升,在 CC、PSNR、SCD 和 SSIM 四个指标中均取得了最佳或次佳值. 具体来看,CDDFuse(Distilled)在 CC 和 SSIM 指标上

表1 所提方法与对比方法在MSRS与RoadScene上的定量对比表

方法	MSRS数据集				RoadScene数据集			
	CC	PSNR	SCD	SSIM	CC	PSNR	SCD	SSIM
DATFuse[52]	0.598 7	14.587 6	1.410 2	1.228 8	0.605 9	14.134 8	1.145 6	1.292 4
FISCNet[53]	0.595 3	15.739 1	1.550 7	1.315 3	0.568 0	13.283 5	1.257 5	1.252 6
FAFusion[54]	0.511 1	14.561 0	1.109 4	1.323 0	0.492 2	12.917 7	0.830 9	1.169 4
KDFuse[17]	0.601 3	16.190 4	1.616 6	1.388 6	0.580 2	14.176 5	1.310 9	1.274 2
SFDFusion[55]	0.592 7	16.217 7	1.605 9	1.404 2	0.560 7	13.573 9	1.311 1	1.238 6
SeAFusion[56]	0.608 7	16.140 6	**1.685 2**	1.394 6	0.613 7	13.583 4	1.558 7	1.280 7
CDDFuse[29]	0.600 7	16.281 6	1.620 9	1.397 4	0.628 7	14.008 7	<u>1.711 8</u>	1.311 6
CDDFuse(Distilled)	<u>0.618 5</u> (+0.017 8)	<u>16.784 9</u> (+0.503 3)	1.645 6 (+0.024 7)	<u>1.415 6</u> (+0.018 2)	**0.640 0** (+0.011 3)	14.352 0 (+0.343 3)	**1.721 7** (+0.009 9)	**1.356 4** (+0.044 8)
SHIP[48]	0.594 6	15.928 3	1.511 9	1.309 7	0.573 3	13.745 2	1.291 8	1.267 5
SHIP(Distilled)	0.611 5 (+0.016 9)	15.950 2 (+0.021 9)	1.631 9 (+0.120 0)	1.358 6 (+0.048 9)	0.615 3 (+0.042 0)	<u>14.561 0</u> (+0.815 8)	1.516 8 (+0.225 0)	1.313 9 (+0.046 4)
EMMA[49]	0.597 0	16.319 1	1.629 4	1.397 6	0.616 9	13.912 9	1.639 5	1.291 1
EMMA(Distilled)	**0.623 4** (+0.026 4)	**16.848 7** (+0.529 6)	<u>1.669 3</u> (+0.039 9)	**1.424 8** (+0.027 2)	<u>0.633 1</u> (+0.016 2)	**14.632 3** (+0.719 4)	1.673 2 (+0.037)	<u>1.353 8</u> (+0.062 7)

注:表中加粗部分与下划线部分分别表示指标的最高值与次高值,绿色部分表示学生模型的指标相较于教师模型的上升值.

实现了持续提升,表明其对红外与可见光源图像的相关性与整体结构保留能力更强;EMMA(Distilled)在PSNR和融合均衡性上优势显著,能够生成更接近高质量参考图像的融合结果;SHIP(Distilled)在SCD上实现了大幅度提升,凸显了其在细节和边缘纹理保持方面的能力.这些结果说明,学生模型不仅在全局图像质量上有所增强,而且在纹理、边缘、对比度等多维度特征表现上也更为均衡.值得注意的是,三种不同教师模型训练出的学生均展现了优于教师的性能,这主要得益于本文所提出的多层次蒸馏机制在多个维度上的有效特征迁移.MLKD在特征层、关系层与输出层的联合约束使得学生模型能够充分吸收教师模型的高层语义与显著性分布信息,同时,学生模型以更紧凑的双分支结构实现了全局与局部特征的协同建模,使其在特征表达上更加高效与聚焦,从而获得了更高的融合质量与更强的泛化能力.这种结构–蒸馏协同优化的机制实现了跨结构高效知识迁移,使得学生具备了超越教师模型的综合表现.

4.3 泛化实验结果与分析

本节分析了所提方法与近年先进方法在RoadScene、TNO和M^3FD数据集上的泛化实验结果,定性对比结果如图5、图6和图7所示,定量对比结果如表1和表2所示.

图5 所提方法与对比方法在RoadScene数据集上的定性对比图

4.3.1 定性分析

图5展示了RoadScene数据集夜晚城市道路场景的对比结果.此时可见光图像因强光干扰而纹理模糊,对融合方法的鲁棒性提出更高要求.对比结果显示,DATFuse、SFDFusion与KDFuse几乎无法体现车辆信息;SHIP与FISCNet虽有所改善,但前者边缘不自然,后者过度强调细节;SeAFusion与EMMA边缘较弱,FAFusion纹理清晰但对比度过低,CDDFuse在整体与细节间取得一定平衡但仍不够清晰.相比之下,三种学生模型在车辆与行人等关键区域均展现出更清晰的纹理和更自然的整体视觉效果,显著优于其他方法,充分证明了所提方法在复杂光照场景下的融合优势.图6展示了TNO数据集中的军事场景.可见光图像因烟雾遮挡几乎丧失了士兵信息,而红外图像则完整呈现其轮廓与

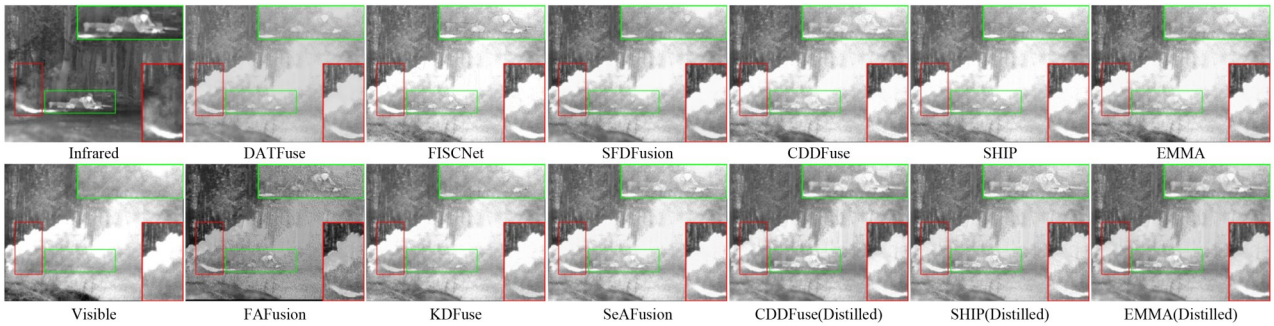


图6　所提方法与对比方法在TNO数据集上的定性对比图

图7　所提方法与对比方法在M^3FD数据集上的定性对比图

表2　所提方法与对比方法在TNO与M^3FD上的定量对比表

方法	TNO数据集				M^3FD数据集			
	CC	PSNR	SCD	SSIM	CC	PSNR	SCD	SSIM
DATFuse[52]	0.497 7	13.219 2	1.448 4	1.306 4	0.494 6	12.396 5	1.286 7	1.365 4
FISCNet[53]	0.477 8	13.094 9	1.557 0	1.278 2	0.458 0	12.607 6	1.276 2	1.322 4
FAFusion[54]	0.410 5	13.326 0	1.289 6	1.142 6	0.408 3	13.527 2	1.161 3	1.228 3
KDFuse[17]	0.479 9	13.812 6	1.554 5	1.297 1	0.474 9	13.356 3	1.362 2	1.383 1
SFDFusion[55]	0.474 9	13.704 4	1.526 1	1.292 3	0.457 4	13.229 3	1.229 9	1.388 3
SeAFusion[56]	0.515 8	13.737 9	1.717 4	1.315 6	0.524 6	12.945 3	1.585 6	1.388 2
CDDFuse[29]	0.512 5	13.800 2	1.760 5	1.340 7	0.535 5	13.542 8	1.647 5	1.412 2
CDDFuse(Distilled)	**0.539 1** (+0.026 6)	14.049 8 (+0.249 6)	**1.772 7** (+0.012 2)	**1.387 6** (+0.046 9)	**0.563 7** (+0.028 2)	14.337 2 (+0.794 4)	**1.730 5** (+0.056 0)	**1.443 1** (+0.030 9)
SHIP[48]	0.480 4	13.445 3	1.548 9	1.298 4	0.466 0	13.149 6	1.310 9	1.363 5
SHIP(Distilled)	0.522 4 (+0.042 0)	13.614 5 (+0.169 2)	1.671 7 (+0.122 8)	1.344 3 (+0.045 9)	0.537 7 (+0.071 7)	13.472 3 (+0.322 7)	1.598 6 (+0.287 7)	1.402 6 (+0.039 1)
EMMA[49]	0.493 5	14.043 2	1.664 1	1.309 4	0.502 6	13.604 9	1.494 2	1.383 2
EMMA(Distilled)	0.533 7 (+0.040 2)	**14.249 8** (+0.206 6)	1.726 5 (+0.062 4)	1.376 4 (+0.067 0)	0.560 0 (+0.057 4)	**14.487 3** (+0.882 4)	1.694 6 (+0.200 7)	1.434 2 (+0.051 0)

注:表中加粗部分与下划线部分分别表示指标的最高值与次高值,绿色部分表示学生模型的指标相较于教师模型的上升值.

纹理. 对比方法中,SFDFusion仅体现了极少的士兵信息,DATFuse与KDFuse仅在士兵头部与手部等高热源位置显现出极少量细节,难以辨识整体形态;FAFusion虽然在士兵局部纹理上稍有改善,但整体对比度显著下降,使图像过暗模糊;SHIP与EMMA保留的红外细节有限,士兵纹理仍显不清晰;FISCNet、SeAFusion与CDDFuse在位置感知上稍优,能勾勒出士兵大致轮廓,但边缘与细节层次不足. 相比之下,三种学生模型在烟雾环境下展现了最优的融合效果,完整保留了士兵的红外结构信息,不仅清晰呈现身体亮度层次,还能分辨手中枪械的细节,整体对比度与清晰度均优于教师及其他方法. 图7展示了M^3FD数据集中含烟雾遮挡的城市场景. 可见光图像中间区域因烟雾而模糊,红外图像则清晰呈现人像与楼房细节. 对比结果显示,DATFuse

与SFDFusion仅保留了红外中的人像信息，楼房几乎缺失；FAFusion虽能显现部分楼房，但整体对比度过低；SHIP、FISCNet、SeAFusion和KDFuse仅保留少量楼房边缘，难以重现完整结构；EMMA与CDDFuse表现稍好，但细节仍不足，建筑边缘模糊. 相比之下，三种学生模型能够有效补充烟雾遮挡区域的红外细节，清晰地再现了红外图像中的人像、楼房目标和可见光纹理，使结果更加符合人眼的视觉感受，展现出显著优势.

4.3.2 定量分析

表1与表2中展示了所提方法在RoadScene、TNO和M^3FD三个数据集上的指标对比结果. 在RoadScene数据集中，复杂的街道场景呈现剧烈变化的光照与背景，学生模型仍然能够在CC与SSIM上优于教师模型，说明其在不同环境下保持结构信息和全局感知能力的稳定性；在TNO数据集中，学生模型在CC与PSNR等指标上均有显著提升，表现出在红外主导的场景中依然能兼顾纹理与亮度信息；在M^3FD数据集中，学生模型更是全面超越教师与现有方法，在4项指标上均实现了明显提升，EMMA(Distilled)与CDDFuse(Distilled)在细节清晰度和全局一致性上展现出卓越优势.

综上所述，本文方法在不同场景下均能生成卓越的融合结果，其定性与定量性能均优于对比方法. 这充分证明了本文方法具有良好的实际应用前景.

4.4 消融实验

为了验证所提方法中各组成模块的有效性，本节在MSRS数据集上进行了系统性的消融实验. 实验以CDDFuse作为教师模型，分别验证不同蒸馏机制与学生网络结构对整体性能的影响. 具体而言，蒸馏部分包括：MLKD、输出响应约束的响应蒸馏(Response Distillation，RD)、基于特征空间的注意力蒸馏(Attention-based Feature Distillation，AFD)、基于结构关系的相似性蒸馏(Structure-aware Similarity Distillation，SFD)；在学生网络部分，分别移除了双分支(Dual-Branch，DB)、全局建模分支(Global Modeling Branch，GMB)与局部感知分支(Local Perception Branch，LPB)，以验证其对融合性能的贡献. 实验结果如表3所示.

从表3中可以观察到，移除蒸馏与结构组件均会导致性能下降. 其中，若同时去除蒸馏与双分支，性能大幅降低，各项指标远低于完整模型，说明蒸馏与双分支结构是性能提升的关键. 对比各蒸馏机制可知，单独移除AFD或SFD时性能均有所下降，表明注意力引导与结构关系对齐在特征迁移中起到互补作用；而移除RD时，SCD与SSIM均下降，说明输出层响应约束对于保证全局一致性与细节保持尤为重要. 在结构方面，去除双分支同样导致多项指标退化，验证了局部细节与全局语义的协同建模对融合任务的重要性；进一步对比GMB与LPB的结果发现，二者缺失均会造成性能下降，但LPB的缺失影响更为明显，说明局部纹理的建模对于图像融合任务中的细节保持具有更大贡献.

表3 在MSRS数据集上的消融实验结果

配置	RD	AFD	SFD	GMB	LPB	CC	PSNR	SCD	SSIM
w/o MLKD+DB						0.387 9	15.666 6	0.226 8	1.025 7
w/o MLKD				√	√	0.605 9	15.991 7	**1.647 5**	1.383 3
w/o RD		√	√	√	√	0.611 8	16.357 6	1.637 3	1.406 6
w/o AFD	√		√	√	√	0.614 5	16.566 9	1.644 4	1.402 4
w/o SFD	√	√		√	√	0.613 7	16.564 4	1.627 9	1.408 2
w/o DB	√	√	√			0.610 1	16.341 8	1.634 9	1.399 6
w/o GMB	√	√	√		√	0.611 2	16.566 5	1.643 0	1.399 4
w/o LPB	√	√	√	√		<u>0.616 6</u>	<u>16.623 9</u>	1.641 7	<u>1.410 1</u>
本文方法	√	√	√	√	√	**0.618 5**	**16.784 9**	<u>1.645 6</u>	**1.415 6**

注：表中加粗部分与下划线部分分别表示指标的最高值与次高值.

中间特征的消融可视化结果如图8所示. 所提方法的中间特征展示了清晰的局部细节与全局结构，且亮红区域表明其能够精准地捕捉到显著性目标. 相比之下，去掉MLKD及双分支结构后特征细节非常模糊，模型关注区域有限，仅在局部产生弱响应，融合特征表现出显著退化；单独去掉MLKD后特征呈“泛响应”，大量背景被误激活，说明缺乏教师的关键引导；去掉RD后模型能够识别出显著性目标，但因缺乏输出层软标签约束，整体响应能量下降，纹理细节欠缺；去掉SFD则出现局部过聚焦，目标虽高亮但不精准，表明特征结构关系建模受损；去掉AFD后特征显著目标丢失且边界模糊，说明缺失教师关注信息的引导. 结构方面，去掉DB结构后模型对教师复杂特征的适配能力下降，细节表达受损，注意力聚焦能力显著削弱；去掉GMB时，模型仅依赖LPB，其输出仍能保留清晰的纹理、边缘与小目标细节，但全局光照、空间层次与轮廓结构明显弱化，表明全局建模能力减弱；在去掉LPB时，模型依托GMB仍能稳定捕获场景的整体布局与全局结构，但局

部纹理锐度下降,边缘细节模糊,小目标难以清晰呈现,进一步验证了 LPB 在局部细节强化中的不可替代性.

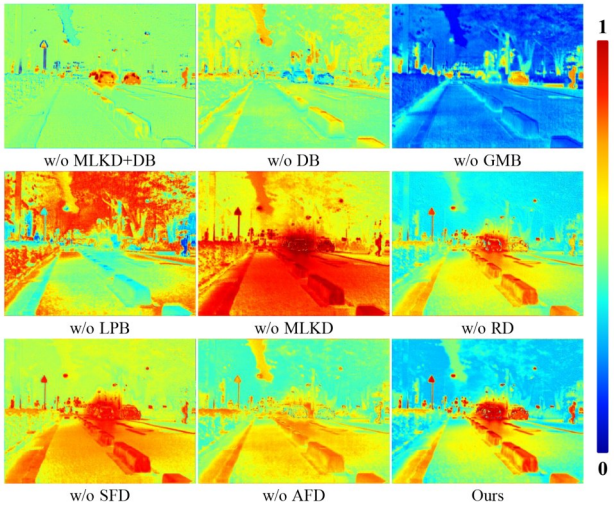


图 8　消融实验中间特征的可视化结果

综合来看,异构双分支结构通过全局与局部互补建模,实现了场景结构与细节纹理的协同表达;多层次蒸馏机制则在输出响应、结构关系与注意力分布三个维度上建立了教师与学生间的深层关联,使学生模型在轻量化的同时保持了高判别力与特征一致性. 值得注意的是,学生模型在部分场景下的性能甚至优于教师模型,这得益于蒸馏框架引入的多维特征约束与学生自身的分支互补机制:通过抑制教师中过度集中的冗余表示,并强化跨层次的显著性对齐,学生模型获得了更具泛化性与融合适应性的特征表达,从而在融合质量与计算效率之间实现了高效平衡,解决了因师生网络的特征表示异构性而导致的知识传递低效、语义对齐不足等问题,保证了最优的融合性能.

4.5　计算复杂度分析

为了进一步验证本文所提出方法在实际应用中的部署效率与轻量化优势,本节对各对比方法进行了计算复杂度方面的评估与分析. 具体而言,选取 MSRS 测试集中的第一组图片(分辨率为 640 × 480)用于测试,包括推理时间(Time)、浮点计算量(FLOPs)和模型参数量(Params),从三个角度综合评价融合方法,值越低说明模型越轻量. 同时,为了更全面地评估模型的实际部署效果,本节分别在服务器平台 GeForce RTX 3080 Ti GPU 和边缘计算平台 Jetson AGX Xavier 上开展实验.

4.5.1　服务器平台结果分析

在服务器平台上的计算复杂度实验结果如表 4 所示,本文方法在拥有较高融合质量的同时,推理时间位列第 5,FLOPs 和 Params 位列第 4,且相比最优方法均差距不大,仍保持较好的竞争力. 相比之下,DATFuse 的 FLOPs 和 Params 最低,SFDFusion 推理时间最短,但二者融合效果欠佳;SeAFusion 的复杂度较低,但融合结果缺少纹理细节;EMMA 推理时间较短,但 Params 较高,不利于边缘部署. 本文方法推理时间约 30 ms,大部分计算开销集中于全局分支自注意力模块,该模块虽耗时,但在建模长距离依赖关系方面具有不可替代优势,融合结果对比也验证了该设计有效性. 整体来看,本文方法在计算复杂度与融合性能之间实现了良好的平衡,在保证高质量融合的同时仍保持较低的资源消耗.

表 4　所提方法与对比方法在服务器上的计算复杂度对比

方法	Time/ms	FLOPs/G	Params/M
DATFuse[52]	25.797 6 ± 0.054 9	**0.555 4**	**0.001 1**
FISCNet[53]	72.454 2 ± 0.684 8	100.541 7	0.285 4
FAFusion[54]	73.512 9 ± 0.143 7	382.187 5	1.246 9
KDFuse[17]	174.791 3 ± 0.272 2	76.543 4	0.249 2
SFDFusion[55]	**8.607 9 ± 0.039 7**	4.280 7	0.014 0
SeAFusion[56]	13.836 8 ± 0.067 9	5.099 5	0.016 7
CDDFuse[29]	293.275 4 ± 44.131 1	568.687 1	1.787 4
SHIP[48]	66.158 5 ± 0.687 9	156.839 1	0.526 0
EMMA[49]	25.152 5 ± 0.504 7	41.538 4	1.516 1
本文方法	31.220 4 ± 0.089 4	22.614 8	0.077 2

注:表中加粗部分与下划线部分分别表示指标的最低值与次低值.

4.5.2　边缘计算平台结果分析

在 Jetson AGX Xavier 边缘计算平台上的计算杂度实验结果如表 5 所示. 相较于表 3,由于计算性能受限,所有方法的推理时间均出现了大幅增长,其中 KDFuse 和 CDDFuse 最为严重,推理时间均超过 2 000 ms,这种延时已无法满足对实时性要求较高的边缘应用. SeAFusion 推理时间虽仅有 77 ms,但其融合结果在细节纹理和目标对比度方面表现仍存在不足. 本文方法在保证融合质量的同时,推理时间仅为 250 ms 左右. 该结果表明,所提方法展现出了良好的边缘部署潜力,能够满足实时性与计算负载的双重需求.

4.6　异构双分支学生对教师局限的补偿分析

为了验证当教师模型在全局或局部信息建模上存在缺陷时,学生的异构双分支结构是否能够有效进行补偿,本节额外选取了 RFN-Nest[60] 和 SwinFusion[25] 两种具有代表性的方法作为教师模型进行了实验. RFN-Nest 是基于 CNN 的早期经典融合方法,在建模长程依赖和全局结构上存在局限;SwinFusion 是基于 Transformer 的早期经典融合方法,在局部细节和纹理捕捉上表现不足. 本节使用这两种融合方法作为教师模型,训练了对应的学生模型,并可视化其中间特征,与 CDDFuse、SHIP、EMMA 三种全局-局部建模兼顾的教师及其对应学生进行对比,以直观分析异构双分支学生

表5 所提方法与对比方法在边缘平台上的计算复杂度对比

方法	Time/ms	FLOPs/G	Params/M
DATFuse[52]	312.630 6 ± 0.898 4	**0.555 4**	**0.001 1**
FISCNet[53]	687.099 4 ± 3.054 5	100.541 7	0.285 4
FAFusion[54]	901.833 2 ± 22.117 3	382.187 5	1.246 9
KDFuse[17]	2 114.436 0 ± 4.947 4	76.543 4	0.249 2
SFDFusion[55]	133.582 8±27.850 3	4.280 7	0.014 0
SeAFusion[56]	**77.046 0 ± 19.202 0**	5.099 5	0.016 7
CDDFuse[29]	2 066.007 9 ± 6.717 4	568.687 1	1.787 4
SHIP[48]	637.867 3 ± 2.881 8	156.839 1	0.526 0
EMMA[49]	81.868 7 ± 1.839 6	41.538 4	1.516 1
本文方法	250.311 4 ± 7.171 3	22.614 8	0.077 2

注：表中加粗部分与下划线部分分别表示指标的最低值与次低值.

对教师局限的补偿效果.

可视化结果如图9所示，RFN-Nest虽然展现了较清晰的树木、路沿等纹理细节，但整体的特征响应较弱，说明在全局建模上存在较大局限；而SwinFusion在局部细节捕捉上表现欠佳，车辆和路沿等小目标的纹理信息不够清晰；其余三种教师在全局与局部建模上要更加均衡. 经过蒸馏训练的双分支学生模型能够有效缓解这些缺陷：RFN-Nest蒸馏后的学生在全局结构的亮度和形状上得到了显著增强，且学生很好地继承了教师在纹理细节上的优势；SwinFusion蒸馏后的学生在局部细节和小目标的响应上明显提升，整体特征更加均衡；其余三种教师蒸馏后的学生在全局-局部信息融合上表现更加稳定，特征响应更为均衡，整体视觉结构更加完整. 这一结果充分证明了异构双分支的设计能够让学生模型实现对教师模型局限的有效补偿，全局分支负责捕捉整体结构与长程依赖，局部分支专注于纹理与细节信息，两者协同作用使学生在多层次特征上获得更均衡、更完整的表征能力，从而缓解教师在特定信息建模上的不足.

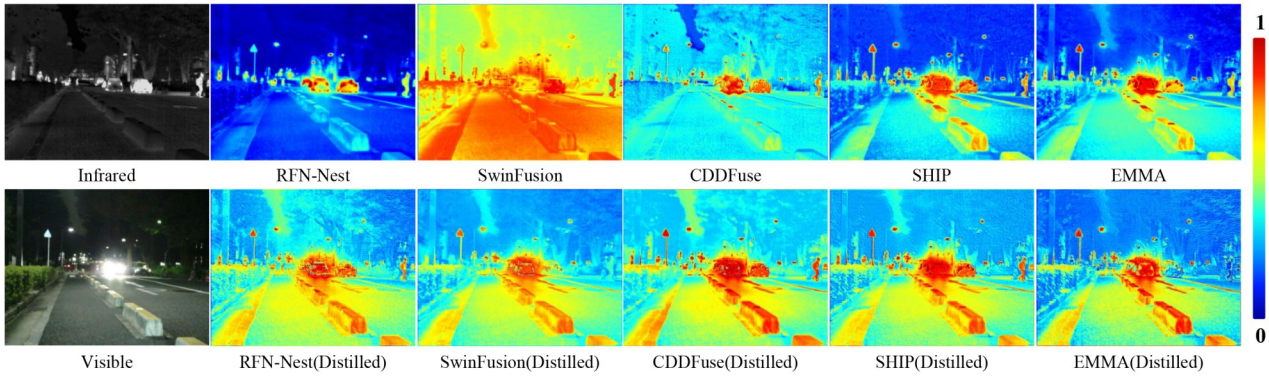


图9 5种教师模型及对应学生模型的中间特征可视化

4.7 目标检测实验结果与分析

为进一步验证所提方法在实际应用场景中的有效性，本节基于M³FD检测数据集进行了目标检测实验. 该数据集涵盖了城市道路、夜间监控、郊区道路等多种复杂真实场景，包含行人、车辆、公交车等6种典型目标，能够充分证明所提方法在实际下游任务中的适用性. 目标检测结果图与mAP@0.5对比结果分别如图10和表6所示.

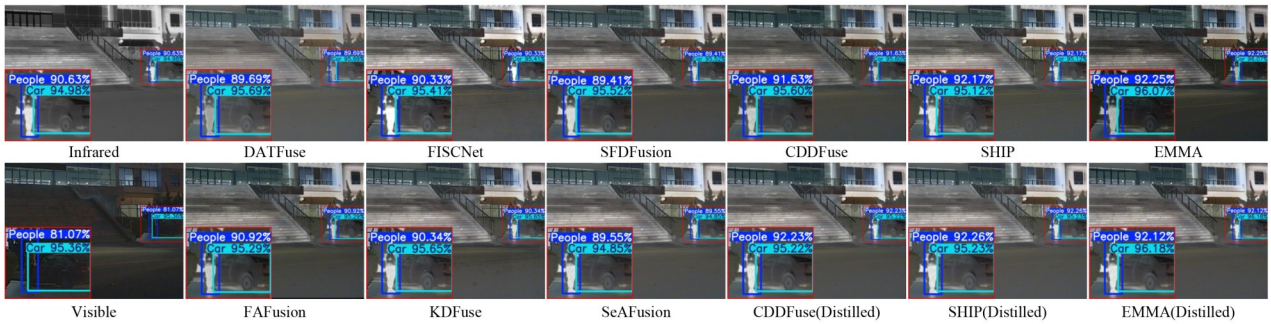
4.7.1 实验设置

在本节实验中，M³FD数据集被按照8∶1∶1的比例拆分为训练集、验证集和测试集. 检测器采用基于COCO数据集预训练的YOLOv5s[61]，训练迭代轮数设为100，训练过程中初始学习率设为0.01，批大小为32，采用SGD(Stochastic Gradient Descent)优化器，动量参数设为0.937，权重衰减系数为5×10^{-4}.

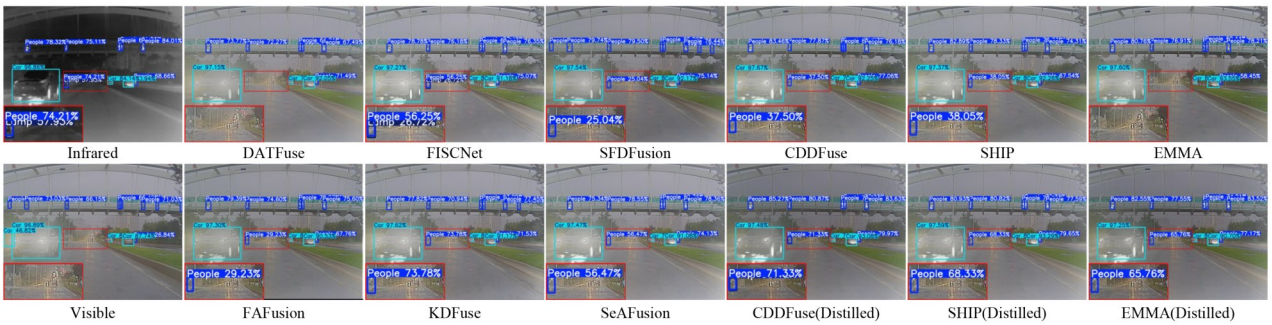
4.7.2 实验结果分析

图10(a)展示了M³FD数据集中夜间监控场景的检测结果对比. 可以看到，在低光照环境下，单独使用可见光图像时，行人检测精度仅为81.07%，而红外图像由于受光照影响较小，检测精度提升至90.63%. DATFuse、KDFuse等方法在融合后的检测精度仍维持在90%左右，未能显著提升对弱光目标的感知能力. 相比之下，SHIP(Distilled)达到最高检测精度92.26%，其余两个学生也均超过92%. 这一结果表明，所提方法能够有效提升模型在夜间场景下的目标识别能力. 图10(b)展示了M³FD数据集中城市道路自动驾驶场景的检测结果. 该场景中同时存在多辆汽车与小型行人目标，环境复杂且遮挡较严重. 原始红外与可见光图像对车辆的检测表现较好，但对小目标行人的检测精度明显不足. FAFusion、FISCNet等方法在融合后虽有一定提升，但多数行人检测精度仍低于80%，部分甚至不足60%. 相比之下，所提三种蒸馏学生模型的行人检测精度多数超过80%，大目标车辆的检测精度超过97%，小目标车辆超过93%. 这充分说明所提方法在复杂交通场景中的优越检测性能与鲁棒性.

从表6可以看出，所提方法在M³FD数据集上的目



(*a*) M^{3}FD数据集中夜间监控场景的检测对比图



(*b*) M^{3}FD数据集中自动驾驶场景的检测对比图

图10　所提方法与对比方法在M^{3}FD数据集上的检测结果对比

表6　所提方法与对比方法在M^{3}FD数据集上的mAP@0.5(%)对比

单位:%

方法	People	Car	Bus	Lamp	Motorcycle	Truck	mAP
DATFuse[52]	81.45	91.19	84.60	78.54	75.47	73.64	80.81
FISCNet[53]	82.84	92.36	88.88	86.51	78.03	**80.79**	84.90
FAFusion[54]	82.85	92.12	90.34	82.32	74.49	78.03	83.36
KDFuse[17]	82.12	<u>92.60</u>	90.68	<u>86.15</u>	78.80	76.93	84.55
SFDFusion[55]	81.29	90.73	84.87	80.40	76.54	72.61	81.07
SeAFusion[56]	82.54	92.08	89.21	85.16	78.88	<u>79.79</u>	84.61
CDDFuse[29]	83.14	92.59	89.15	85.81	79.41	79.33	84.90
SHIP[48]	82.78	92.63	90.51	**86.92**	78.91	78.16	84.99
EMMA[49]	81.48	92.35	90.34	84.68	73.92	79.32	83.68
CDDFuse (Distilled)	<u>84.43</u>	92.00	88.86	84.89	80.14	78.68	84.83
SHIP (Distilled)	84.00	**92.99**	<u>90.98</u>	84.41	<u>81.85</u>	77.60	<u>85.30</u>
EMMA (Distilled)	**84.61**	92.59	**92.03**	85.23	**82.65**	75.16	**85.38**

注:表中加粗部分与下划线部分分别表示指标的最高值与次高值.

标检测性能整体优于多数现有融合方法. 相比原始教师模型,蒸馏后的学生在多个类别检测精度上均有所提升,在行人与摩托车等细小或低对比度目标上表现尤为突出. 其中,EMMA(Distilled)取得了最高的mAP85.38%,相较原始EMMA提升约1.7个百分点,说明蒸馏过程有效增强了融合特征的可分辨性. SHIP(Distilled)也在多数类别上实现性能增益,CDDFuse(Distilled)的mAP虽略有下降,但也高于多数对比方法,验证了所提蒸馏机制在不同结构下的通用性与有效性. 总体而言,所提方法使融合特征更具判别力,有助于提升下游检测任务的准确性与稳健性,在实际场景中有较好的应用价值.

5　结束语

本文针对红外与可见光图像融合任务中异构师生模型知识蒸馏存在的特征空间不匹配与迁移效率低下问题,提出了一种基于异构模型多层次知识蒸馏的融合方法. 通过引入特征层注意力引导、关系层结构相似性对齐与输出层响应约束的多层次蒸馏机制,有效提升了跨模型知识传递的效率和鲁棒性;同时,所设计的异构双分支学生网络结合了CNN的局部细节感知优势与Transformer的全局语义建模优势,增强了特征表示与知识学习的稳定性. 在MSRS、RoadScene、TNO及M^{3}FD四个公开数据集上的实验结果表明,该方法在多项评价指标上均显著优于教师模型及最先进的融合方法,并在保持轻量化特性的前提下实现了性能与蒸馏稳定性的协同提升. 尽管所提方法在融合性能与效率方面表现优异,有效提升了下游目标检测任务的准确率,但当前融合蒸馏策略与下游任务的适配性仍有待联合优化. 未来将重点研究动态权重蒸馏机制与多任务协同学习框

架,旨在进一步增强模型的泛化能力与实际应用价值.

参考文献

[1] MA J Y, MA Y, LI C. Infrared and visible image fusion methods and applications: A survey[J]. Information Fusion, 2019, 45: 153-178.

[2] 周非, 舒浩峰, 白梦林, 等. 生成对抗网络协同角度异构中心三元组损失的跨模态行人重识别[J]. 电子学报, 2023, 51(7): 1803-1811.
ZHOU F, SHU H F, BAI M L, et al. Cross-modal person re-identification based on generative adversarial network coordinated with angle based heterogeneous center triplet loss[J]. Acta Electronica Sinica, 2023, 51(7): 1803-1811. (in Chinese)

[3] 汪进中, 戴顺, 张秀伟, 等. 无人机视角多源目标检测数据集UAV-RGBT及算法基准[J]. 电子学报, 2025, 53(3): 686-704.
WANG J Z, DAI S, ZHANG X W, et al. UAV-RGBT multispectral object detection dataset and algorithm benchmark[J]. Acta Electronica Sinica, 2025, 53(3): 686-704. (in Chinese)

[4] LIU J Y, LIU Z, WU G Y, et al. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 8081-8090.

[5] JIN X, JIANG Q, YAO S W, et al. A survey of infrared and visual image fusion methods[J]. Infrared Physics & Technology, 2017, 85: 478-501.

[6] BHATARIA K C, SHAH B K. A review of image fusion techniques[C]//2018 Second International Conference on Computing Methodologies and Communication. Piscataway: IEEE, 2018: 114-123.

[7] ZHANG X C, DEMIRIS Y. Visible and infrared image fusion using deep learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10535-10554.

[8] WANG R C, ZHOU Z F, LI S H, et al. Advances and challenges in infrared-visible image fusion: A comprehensive review of techniques and applications[J]. Artificial Intelligence Review, 2026, 59(1): 18.

[9] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2023-08-02) [2025-10-10]. https://arxiv.org/abs/1706.03762.

[11] KHAN W Z, AHMED E, HAKAK S, et al. Edge computing: A survey[J]. Future Generation Computer Systems, 2019, 97: 219-235.

[12] HINTON G. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09) [2025-10-10]. https://arxiv.org/abs/1503.02531.

[13] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[14] HSU C C, NI C C, LEE C M, et al. CSAKD: Knowledge distillation with cross self-attention for hyperspectral and multispectral image fusion[EB/OL]. (2024-06-28) [2025-10-10]. https://arXiv.org/abs/2406.19666.

[15] YUE C K, ZHANG Y, YAN J H, et al. Diffusion mechanism and knowledge distillation object detection in multimodal remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 4408314.

[16] XIAO W X, ZHANG Y F, WANG H B, et al. Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 5004015.

[17] YANG C J, LUO X Q, ZHANG Z C, 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.

[18] LI S T, KANG X D, HU J W. Image fusion with guided filtering[J]. IEEE Transactions on Image Processing, 2013, 22(7): 2864-2875.

[19] BURT P J, KOLCZYNSKI R J. Enhanced image capture through fusion[C]//1993 (4th) International Conference on Computer Vision. Piscataway: IEEE, 2002: 173-182.

[20] KUMAR S S, MUTTAN S. PCA-based image fusion[J]. Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XII, 2006, 6233: 62331T.

[21] LI H, WU X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.

[22] ZHANG Y, LIU Y, SUN P, et al. IFCNN: A general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99-118.

[23] PRABHAKAR K R, SAI SRIKAR V, BABU R V. DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 4724-4732.

[24] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2025-10-10]. https://arXiv.org/abs/2010.11929.

[25] MA J Y, TANG L F, FAN F, 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] LI J F, SONG H, LIU L, et al. MixFuse: An iterative mix-attention transformer for multi-modal image fusion[J]. Expert Systems with Applications, 2025, 261: 125427.

[27] LI J F, LIU L, SONG H, et al. DCTNet: A heterogeneous dual-branch multi-cascade network for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 5030914.

[28] XU J, HE X. DAF-net: A dual-branch feature decomposition fusion network with domain adaptive for infrared and visible image fusion[EB/OL]. (2024-09-18)[2025-10-10]. https://arXiv.org/abs/2409.11642.

[29] ZHAO Z X, BAI H W, ZHANG J S, et al. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 5906-5916.

[30] ZHOU H L, SONG L C, CHEN J J, et al. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective[EB/OL]. (2021-02-01) [2025-10-10]. https://arXiv.org/abs/2102.00650.

[31] HEO B, KIM J, YUN S, et al. A comprehensive overhaul of feature distillation[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 1921-1930.

[32] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 3962-3971.

[33] WANG T, YUAN L, ZHANG X P, et al. Distilling object detectors with fine-grained feature imitation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 4928-4937.

[34] WANG J B, CHEN Y M, ZHENG Z H, et al. CrossKD: Cross-head knowledge distillation for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16520-16530.

[35] LIU K, ZHANG Y Y, ZHANG J Y, et al. DFD: distilling the feature disparity differently for detectors[C]//Proceedings of the 41st International Conference on Machine Learning. New York: ACM, 2024: 32421-32430.

[36] NI Z L, YANG F K, WEN S Z, et al. Dual relation knowledge distillation for object detection[EB/OL]. (2023-06-01)[2025-10-10]. https://arxiv.org/abs/2302.05637.

[37] YANG C G, ZHOU H L, AN Z L, et al. Cross-image relational knowledge distillation for semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 12309-12318.

[38] MI J, WANG L F, LIU Y, et al. KDE-GAN: A multimodal medical image-fusion model based on knowledge distillation and explainable AI modules[J]. Computers in Biology and Medicine, 2022, 151: 106273.

[39] DENG Y L, XU T Y, CHENG C Y, et al. MMDRFuse: Distilled mini-model with dynamic refresh for multi-modality image fusion[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 7326-7335.

[40] TUNG F, MORI G. Similarity-preserving knowledge distillation[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 1365-1374.

[41] ZAMIR S W, ARORA A, KHAN S, et al. Restormer: Efficient transformer for high-resolution image restoration[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5718-5729.

[42] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1800-1807.

[43] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[44] TANG L F, YUAN J T, ZHANG H, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware[J]. Information Fusion, 2022, 83: 79-92.

[45] TOET A. The TNO multiband image data collection[J]. Data in Brief, 2017, 15: 249-251.

[46] XU H, MA J Y, LE Z L, et al. FusionDN: A unified densely connected network for image fusion[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12484-12491.

[47] LIU J Y, FAN X, HUANG Z B, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5792-5801.

[48] ZHENG N S, ZHOU M, HUANG J, et al. Probing synergistic high-order interaction in infrared and visible image fusion[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 26374-26385.

[49] ZHAO Z X, BAI H W, ZHANG J S, et al. Equivariant multi-modality image fusion[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 25912-25921.

[50] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2017-01-30)[2025-10-10]. https://arxiv.org/abs/1412.6980.
[51] SMITH L N, TOPIN N. Super-convergence: Very fast training of neural networks using large learning rates[C]// Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. SPIE, 2019: 2520589.
[52] TANG W, HE F Z, LIU Y, et al. DATFuse: Infrared and visible image fusion via dual attention transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3159-3172.
[53] ZHENG N S, ZHOU M, HUANG J, et al. Frequency Integration and Spatial Compensation Network for infrared and visible image fusion[J]. Information Fusion, 2024, 109: 102359.
[54] XIAO G B, TANG Z M, GUO H L, et al. FAFusion: Learning for infrared and visible image fusion via frequency awareness[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 5015011.
[55] TANG L F, YUAN J T, MA J Y. 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.
[56] HU K, ZHANG Q L, YUAN M X, et al. SFDFusion: An efficient spatial-frequency domain fusion network for infrared and visible image fusion[EB/OL]. (2024-10-30)[2025-10-10]. https://arxiv.org/abs/2410.22837.
[57] LIU J Y, WU G Y, LIU Z, et al. Infrared and visible image fusion: From data compatibility to task adaption[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2349-2369.
[58] HUYNH-THU Q, GHANBARI M. Scope of validity of PSNR in image/video quality assessment[J]. Electronics Letters, 2008, 44(13): 800-801.
[59] WANG Z, 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.
[60] LI H, WU X J, KITTLER J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images[J]. Information Fusion, 2021, 73: 72-86.
[61] JOCHER G, CHAURASIA A, STOKEN A, et al. Ultralytics/yolov5:v7.0 - yolov5 sota realtime instance segmentation[EB/OL]. (2022-11-22)[2025-10-10]. https://github.com/ultralytics/yolov5/discussions/10258.

作者简介

张 棋 男,2004年3月出生于辽宁省铁岭市. 现为北京理工大学计算机学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: zq@bit.edu.cn

马士瀚 男,1998年12月出生于山东省枣庄市. 现为北京理工大学计算机学院博士研究生. 主要研究方向为计算机视觉.
E-mail: mashihan@bit.edu.cn

宋 红 女,1977年10月出生于陕西省西安市. 现为北京理工大学计算机学院教授、博士生导师. 获中国电子学会科技进步奖一等奖、吴文俊人工智能科技进步奖一等奖等奖项6项. 在国内外发表学术论文100余篇. 主要研究方向为计算机视觉.
E-mail: songhong@bit.edu.cn

林毓聪 男,1993年12月出生于广西壮族自治区南宁市. 现为北京理工大学光电学院特聘副研究员. 国内外发表学术论文10余篇,牵头承担国家自然科学基金青年科学基金项目,作为项目骨干参与多项国家级项目. 主要研究方向为多模态医学数据智能分析.
E-mail: linyucongbit@bit.edu.cn

李金夫 男,1990年8月出生于湖北省咸宁市. 现为北京理工大学博士后. 国内外发表学术论文10余篇,主持国家重点研发计划子课题、北京市自然科学基金、四川省科技支撑计划等国家/省部级项目. 主要研究方向为多模态图像融合与目标检测.
E-mail: jinfuli@bit.edu.cn

杨 健 男,1977年10月出生于云南省楚雄州. 现为北京理工大学光电学院教授、博士生导师. 获国家技术发明奖二等奖、教育部技术发明奖一等奖等省部级以上科研奖励20余项. 国内外发表学术论文300余篇. 主要研究方向为计算机视觉. 中国电子学会会员编号:E190013149S.
E-mail: jyang@bit.edu.cn