

文章编号:1007-2780(2026)05-0734-20

基于预训练 SAM 的语义保持红外-可见光 多模态融合语义分割方法

陈悦, 许锋*, 李昊文

(中国刑事警察学院 公安信息技术与情报学院, 辽宁 沈阳 110854)

摘要: 针对红外-可见光图像融合中长期存在的视觉融合质量与语义任务精度之间难以兼顾的问题, 本文提出一种基于语义先验的多模态语义分割增强框架。现有红外-可见光融合方法虽然能够在一定程度上提升融合图像的视觉质量, 但往往会削弱图像中的语义信息表达, 从而影响下游语义分割任务的性能。为解决这一问题, 本文引入预训练的 SAM (Segment Anything Model) 作为冻结的语义先验模型, 在融合过程中对多模态特征学习进行引导, 以保持语义信息的一致性。在所提框架中, RGB 图像与红外图像首先部分输入 SAM 编码器以获取多层次视觉特征表示, 同时通过专门设计的融合模块提取不同模态之间的互补特征。在各个特征阶段, SAM 提取的特征通过语义保持分支进行进一步优化, 并引入语义感知跨模态增强模块进行特征强化。该模块通过融合文本标签嵌入信息, 提高模型对不同类别目标的响应能力, 并有效提升目标边界区域的分割精度。实验结果表明, 所提方法在语义信息保持能力、融合精度以及视觉重建质量等方面均取得了显著提升, 验证了该框架在红外-可见光语义分割任务中的有效性与稳定性。

关键词: 多模态语义分割; 红外-可见光图像融合; 预训练 SAM; 特征调制; 门控融合机制; 语义先验; 跨模态学习

中图分类号: TP391.41 **文献标识码:** A **doi:** 10.37188/CJLCD.2026-0051 **CSTR:** 32172.14.CJLCD.2026-0051

Semantic-preserving multimodal fusion framework for enhanced infrared-visible image segmentation using pretrained SAM

CHEN Yue, XU Feng*, LI Haowen

(School of Public Security Information Technology and Intelligence,
Criminal Investigation Police University of China, Shenyang 110854, China)

Abstract: To address the long-standing challenge of balancing visual fusion quality with semantic task accuracy in infrared-visible image fusion, this paper proposes a multimodal semantic segmentation enhancement framework based on a semantic prior. While existing infrared-visible fusion methods can improve the visual quality of the fused image to some extent, they often weaken the semantic information representation

收稿日期: 2026-03-21; 修订日期: 2026-04-15.

基金项目: “十三五”国家重点研发技术项目 (No. 2017YFC0821004); 2022 辽宁省教育厅基本科研重大攻关项目 (No. LJKZZ20220007); 中央高校基本科研业务重大项目培育计划

Supported by National Key R&D Program of China (“13th Five-Year Plan”) (No. 2017YFC0821004); 2022 Major Fundamental Research Project of Liaoning Provincial Department of Education (No. LJKZZ20220007); Major Project Cultivation Program of the Fundamental Research Funds for the Central Universities

*通信联系人, E-mail: 7677759@qq.com

within the image, thus affecting the performance of downstream semantic segmentation tasks. To solve this problem, this paper introduces a pre-trained Segment Anything Model (SAM) as a frozen semantic prior model to guide multimodal feature learning during the fusion process, maintaining the consistency of semantic information. In the proposed framework, RGB and infrared images are first partially input into the SAM encoder to obtain multi-level visual feature representations. Simultaneously, a specially designed fusion module extracts complementary features between different modalities. At each feature stage, the features extracted by SAM are further optimized through a semantic preservation branch, and a semantically aware cross-modal enhancement module is further introduced for feature strengthening. This module improves the model's responsiveness to different target categories by fusing text label embedding information and effectively enhances the segmentation accuracy of target boundary regions. Experimental results show that the proposed method has achieved significant improvements in semantic information preservation, fusion accuracy, and visual reconstruction quality, verifying the effectiveness and stability of the framework in infrared-visible semantic segmentation tasks.

Key words: multimodal semantic segmentation; infrared-visible fusion; pretrained SAM; feature modulation; gated fusion; semantic prior; cross-modal learning

1 引言

随着计算机视觉技术的飞速发展^[1-4],多模态图像融合与语义分割已成为解决复杂视觉任务的关键技术手段^[5-7]。多模态图像融合通过整合不同传感器或图像源的信息,显著提升了感知系统在复杂环境下的鲁棒性,尤其在夜间监控、自动驾驶及军事目标识别等挑战性场景中展现出独特优势^[5,8-9]。在众多融合任务中,红外与可见光图像融合是最具代表性的研究方向之一^[1-2,10]。红外图像在弱光及雾霾条件下仍能提供可靠的目标检测信息,而可见光图像则富含纹理细节与色彩信息^[5,11]。通过有机整合这两种互补的模式信息,能够获得更为完整的场景表征,从而有效促进目标识别和语义分割等下游任务的性能提升^[12-14]。

尽管红外-可见光图像融合技术已取得显著进展,现有方法仍面临一个核心挑战:如何在视觉质量与任务性能之间实现有效平衡^[15-17]。当前众多融合方法主要聚焦于视觉增强,如对比度提升与清晰度优化,但在很大程度上忽视了语义分割所需的深层语义信息^[5,18]。与此相对,传统语义分割模型虽然注重像素级分类精度,但由于模态差异及语义不一致性的影响,往往难以在融合图像上取得理想的分割效果^[8-9]。因此,如何有效平衡图像融合的视觉质量与分割精度,仍然是

红外-可见光多模态感知领域亟待解决的核心科学问题^[13-14]。

既往研究主要集中于基于深度学习的像素级或特征级融合方法以提升融合质量^[4,15,19-20]。随着深度学习技术的发展,卷积神经网络与基于Transformer的模型显著改善了融合性能^[4]。然而,这些方法大多依赖于固定的网络架构和预设的融合策略,限制了其对高层语义信息进行显式建模的能力^[5]。因此,近期研究开始尝试将语义引导引入融合过程,以更好地支持语义分割等下游任务^[21]。

在此背景下,大规模预训练模型,尤其是SAM (Segment Anything Model),为多模态融合与语义分割提供了有前景的研究方向^[8]。作为视觉基础模型,SAM通过大规模预训练学习了丰富且结构化的视觉表示,在不同场景下展现出强大的语义泛化能力^[22]。与传统多尺度或浅层特征融合方法^[23-24]相比,SAM提供了更强的语义表示能力,使融合过程不仅能考虑底层纹理,还能整合更高层次的上下文信息。然而,尽管SAM在各类视觉任务中取得了成功^[8],如何有效地将其语义先验融入红外-可见光图像融合与分割任务中仍有待深入探索。现有的多模态融合方法在多阶段特征交互过程中,往往缺乏明确的机制来保持和传递预训练的语义知识,这可能导致融合结果中的语义退化。

针对上述挑战,本文提出了一种多模态语义分割增强框架,利用SAM的多阶段视觉表示作为语义先验,引导红外与可见光模态间的特征融合。在该方法中,RGB图像和红外图像通过两条互补的路径进行处理:一条路径使用预训练的SAM编码器提取多级视觉特征,另一条路径采用专用融合模块捕获互补的模态信息。在每个阶段,SAM导出的特征通过双分支结构进行精化,该结构包含一个保持全局语义结构的语义保持分支,以及一个通过轻量级适配器集成多模态融合特征的自适应调制分支。两个分支的输出通过可学习的门控融合机制进行组合,并通过语义感知的跨模态模块进一步增强,该模块引入文本标签嵌入以强化类别表示并改善分割边界。

本文的主要贡献总结如下:

(1)提出了一种多模态语义分割增强框架,将预训练的SAM知识融入红外-可见光图像融合中,利用SAM的多阶段视觉表示作为语义先验。该方法在融合过程中提升了语义一致性,同时增强了分割精度和视觉质量。

(2)设计了一种多阶段特征交互机制,包含一个语义保持分支。语义保持分支保留了SAM固有的结构信息,减少了融合过程中的语义漂移。

(3)引入了一个语义感知的跨模态模块,该模块融合了文本编码器的嵌入表示,以增强类别级表征并改善分割边界预测。

在多个基准数据集上的大量实验表明,所提框架在语义性能和融合质量指标上均持续优于当前最先进的基于融合的分割方法。

总体而言,本文工作为将预训练的视觉基础模型融入多模态融合任务提供了一个实用框架,并为提升红外-可见光语义分割的语义可靠性和视觉一致性提供了一个可行的研究方向。

2 相关工作

2.1 多模态图像融合

红外-可见光图像融合作为一种将互补光谱信息整合为统一表示的有效策略已被广泛研究^[5]。早期方法主要依赖于手工设计的信号处理技术^[25-26],例如多尺度变换和空间频率分解,用于结合红外和RGB图像中的结构与强度信息。尽管这些方法提供了可解释的融合规则,但它们对低层先验

的依赖限制了捕获语义上下文或适应复杂场景的能力。

基于学习的融合方法随后被引入以改进特征表示。稀疏编码、低秩建模和基于优化的方法实现了更灵活的特征提取,但这些方法仍然施加了固定的融合约束,泛化能力有限^[27-28]。随着深度学习的发展,基于CNN和Transformer的融合网络通过数据驱动的方式学习模态感知特征,取得了显著进展^[21,23],注意力机制、生成对抗网络框架和跨模态精化模块进一步增强了结构一致性和视觉质量^[4,29]。

尽管取得了这些进展,大多数深度融合方法优化融合图像时主要关注感知质量,而非下游任务性能^[5,15],融合结果可能在视觉上令人满意,但仍无法有效支持分割或检测等高层任务^[12-13]。换言之,语义监督与融合目标之间的关联往往较弱,这仍是许多现有融合流程的根本性局限。这一局限性促使研究者将更强的语义先验融入融合过程,以更好地将视觉增强与任务需求对齐。

2.2 多模态语义分割

多模态语义分割利用异构传感器信息来提升像素级的场景理解能力。早期方法主要采用输入级拼接或编码器级融合策略,将RGB与深度、红外或激光雷达数据相结合^[6]。虽然简单易行,但早期融合方法对模态差异敏感,且往往未能充分利用模态特定特征。

后续研究通过双编码器架构、交叉注意力模块和分层融合结构探索了更深层次的特征交互。这些设计改善了特征对齐和模态互补性,特别是在RGB-深度和RGB-红外分割任务中^[1,8-9]。晚期融合策略允许对每种模态进行独立的特征学习,但通常缺乏足够的细粒度跨模态推理。

最近,红外-可见光分割框架尝试将融合和分割整合到统一模型中^[13-14]。部分方法首先生成融合图像,然后应用如SegFormer的标准分割网络。另一些方法直接在分割网络内部融合中间的多模态特征。尽管取得了进展,但仍存在几个挑战:(1)模态差异可能导致特征不平衡;(2)有限的数据集规模限制了鲁棒表示学习;(3)当前模型通常缺乏强大的语义先验,使其对场景变化和结构模糊性敏感^[12-13]。

相比之下,本文方法将来自预训练SAM模

型的语义知识融入多模态分割流程,以跨阶段增强语义一致性和跨模态特征交互^[22]。

2.3 SAM

SAM已成为近期计算机视觉研究中重要的基础模型,特别是在分割任务中^[8]。SAM在大规模数据集上进行了预训练,旨在灵活提示设置下对图像中的物体进行分割。通过其预训练的图像编码器和基于提示的掩码解码器,SAM学习了丰富且结构化的视觉表示,在不同场景下具有较强的泛化能力。已有研究表明,SAM在各个领域都能取得一致的分割性能,突显了大规模预训练对语义理解的有效性^[22]。

在现有的基于SAM的框架中,编码器特征或掩码先验通常被迁移到诸如标准图像分割等单模态任务中,而多模态信息(包括红外-可见光融合)的整合仍然有限。尽管大规模预训练和基于提示的学习方法,如CLIP^[30]和CoOp^[31],已经展示了跨模态对齐语义信息的潜力,但SAM学习到的分层语义表示在多模态融合分割中尚未得到充分探索。这一观察促使我们将SAM衍生的语义先验融入红外-可见光融合网络,以增强语义一致性和跨模态表示学习。

2.4 多模态语义感知与跨模态学习

多模态语义感知与跨模态学习旨在对齐和整合来自不同模态(如图像和文本)的信息,以提升高层理解任务的能力。多模态学习的最新进展表明,结合互补模态可以增强目标检测、分割和推理的性能。例如,CLIP^[30]通过对比预训练对齐了视觉和文本表示,实现了有效的视觉-语言理解。相关方法进一步通过提示学习和连接像素、图像与语言输出的广义解码框架^[32]来适应预训练表示。然而,在异构模态间实现稳定的语义对齐仍然具有挑战性。

在分割任务中,多模态融合已被证明可以通过引入额外的感知线索来提升性能。然而,许多现有方法仅在少数特征阶段进行融合^[5],这可能限制了网络层次间的语义一致性,常导致模态冲突、特征不平衡和表示冗余^[13]。此外,大多数方法缺乏逐阶段语义适应机制,削弱了其在网络中保持一致类别级表示的能力。

近期研究探索了双分支架构和语义感知融合策略,以更好地平衡模态适应和语义保持^[14,21]。

尽管这些方法改善了跨模态交互,但它们并未充分利用网络多阶段中的分层语义表示。融入语义约束的分层融合策略已显示出改善特征对齐、缓解融合质量与分割精度之间权衡的潜力^[13-14]。

受这些观察启发,本文框架采用双分支设计,并利用预训练的SAM编码器作为跨阶段的语义骨干。结合多模态适配器机制,该设计促进了逐阶段的语义保持和改进的跨模态特征对齐,从而提升了分割性能^[22]。

3 研究方法

如图1所示,本文提出的框架通过在红外-可见光图像融合过程中引入预训练SAM的语义先验,实现对多模态语义分割任务的有效增强。不同于传统多模态语义分割方法通常先对异构输入分别进行特征提取,并在网络后期阶段再进行特征融合,本文方法将SAM所蕴含的语义知识直接嵌入到特征提取层级结构中,通过结构化语义信息对多模态融合过程进行引导。

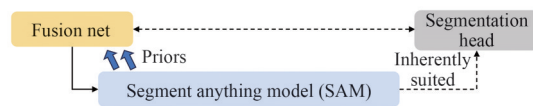


图1 本文方法的总体框架

Fig. 1 Overall framework of the proposed method

具体而言,给定一幅红外图像 I^{ir} 与一幅可见光RGB图像 I^{rgb} ,网络对该多模态图像对进行联合编码,以提取多层次特征表示,从而同时捕获两种模态之间互补的辐射信息与结构信息。输入的多模态图像对可表示为:

$$X = \{I^{\text{rgb}}, I^{\text{ir}}\}. \quad (1)$$

模型的目标是预测对应的像素级语义标签图 $\hat{Y}$:

$$\hat{Y} = \Phi(X), \quad (2)$$

其中, $\Phi(\cdot)$ 表示本文提出的多模态语义融合网络。从整体结构上看,该框架的设计主要基于以下3个核心考虑。

首先,我们并非仅将SAM用作外部掩码生成器或边界细化工具,而是利用其预训练编码器作为分层的语义骨干。SAM编码器生成的多阶段特征图编码了从大规模预训练中学到的物体

结构、边界信息、区域连贯性和语义一致性。这使得融合流程能够基于语义结构化的表示进行操作,而非仅依赖模态特定特征。这种分层整合对于红外-可见光分割尤为重要,因为热辐射线索通常在物体边界处显著但在纹理区域较弱,而可见光线索则表现出互补特性。

其次,我们引入了自适应多模态调制,以整合互补的光谱信息,同时保持 SAM 提供的语义结构。网络并非直接合并来自不同模态的特征,而是通过双分支设计进行逐阶段精化。一个分支保持 SAM 衍生的语义先验,另一个分支则通过轻量级自适应调制引入多模态增强。这种配置减少了语义失真,并防止模态融合破坏预训练的特征分布。因此,模型保持了跨尺度的语义一致性,并更好地捕获了模态特定的优势,包括来自红外数据的热对比度和来自 RGB 图像的空间细节。

最后,我们通过跨模态文本交互模块在决策层引入了语义对齐。对应于类别标签的文本嵌入作为额外的语义约束被引入,以增强类别判别性

并稳定特征边界。这种设计促进了高层特征空间的一致组织,特别是在模糊或低对比度区域。总体而言,本文框架将基于 SAM 的语义引导、逐阶段多模态调制和文本对齐整合到一个统一的融合-分割架构中,从而实现了更可靠且语义一致的分割性能。

3.1 整体框架

本文框架将预训练 SAM 编码器的语义先验与从红外和可见光输入中提取的互补信息相结合。如图 2 所示,该架构包含 4 个主要组成部分:(1) SAM 编码器,生成分层语义特征;(2) 模态融合编码器,提取互补的多模态表示;(3) 双分支特征精化模块,平衡语义保持与自适应多模态调制;(4) 门控融合机制及其后的语义感知跨模态模块,在每个特征阶段生成增强表示。这些组件在一个统一的流程中运作,逐阶段交互逐步强化语义结构并改善多模态特征对齐。其中,语义保持模块的具体架构如图 3 所示,语义感知跨模态模块的具体架构如图 4 所示。

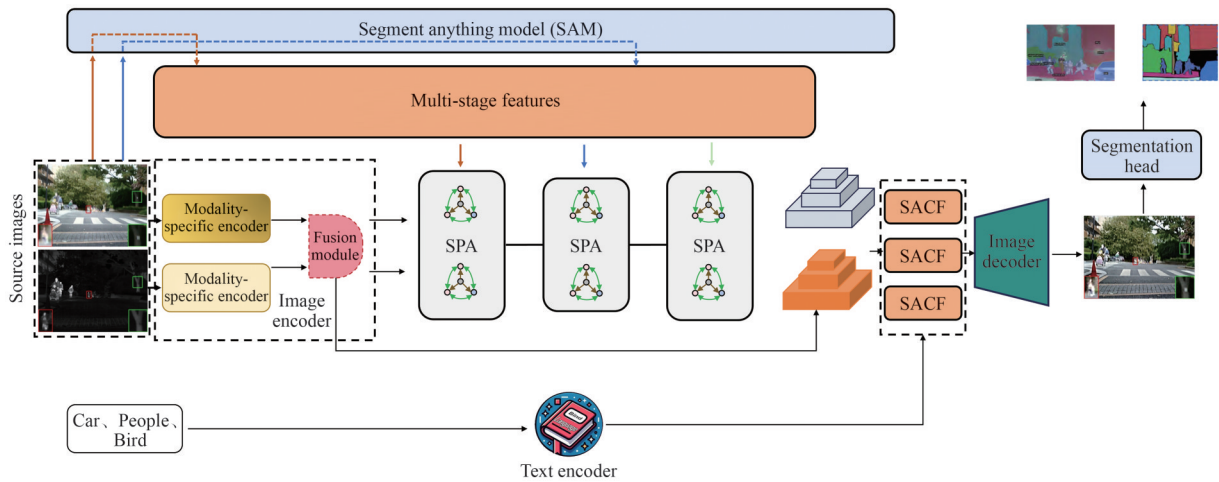


图 2 本文方法论框架

Fig. 2 Proposed methodological framework

我们将 SAM 编码器的层次化阶段记为集合 $S = \{1, 2, \dots, K\}$ 。对于每一个阶段 $k \in S$, SAM 编码器首先从多模态输入中提取先验引导的特征表示:

$$F_{\text{sam}}^{(k)} = \mathcal{E}_{\text{sam}}^{(k)}(I^{\text{rgb}}, I^{\text{ir}}), k \in S. \quad (3)$$

与此同时,并行的模态融合编码器也对相同的输入进行处理,以生成互补的特征图:

$$F_{\text{m}}^{(k)} = \mathcal{E}_{\text{m}}^{(k)}(I^{\text{rgb}}, I^{\text{ir}}), k \in S, \quad (4)$$

其中:下标 m 表示由红外图像与可见光 RGB 图像共同编码得到的多模态融合特征。

随后,这两个特征流被用于阶段性的融合,从而实现语义先验与模态特定信息在整个特征层次中的整合。这种多阶段交互机制,在特征进行更深层抽象之前强化了语义结构,从而提升了表征在整个编码器深度上的稳定性。

对于每个阶段,两个特征流被联合处理以生

$$F_{\text{rgb}}^{(k)} = E_{\text{rgb}}^{(k)}(I_{\text{rgb}}), \quad (7)$$

$$F_{\text{ir}}^{(k)} = E_{\text{ir}}^{(k)}(I_{\text{ir}}), \quad (8)$$

其中, $E_{\text{rgb}}^{(k)}$ 和 $E_{\text{ir}}^{(k)}$ 分别表示第 k 个阶段中用于 RGB 图像和红外图像特征提取的模态专用编码模块, 分别提取可见光模态的纹理细节特征和红外模态的热辐射结构特征。这种在融合前采用参数不共享的 RGB/红外双分支编码方式能够保留模态特定的信号分布, 同时捕获 RGB 的低层纹理信息和红外的热结构线索。

随后, 将提取的特征输入到融合模块, 该模块在每个阶段 k 进行两模态特征的逐阶段整合。具体而言, 融合的多模态特征 $F_m^{(k)}$ 通过跨注意力机制计算:

$$F_m^{(k),i} = \text{CrossAttn}(Q = F_{\text{M,M}}^{(k),i}, K = F_{\text{SAM}}^{(k),i}, V = F_{\text{SAM}}^{(k),i}), \quad (9)$$

其中: i 表示阶段 k 内的注入块索引, $F_{\text{M,M}}^{(k),i}$ 表示当前多模态特征查询。该公式使网络能够选择性地关注每个模态中与语义相关的区域, 同时保留来自 SAM 的结构先验。随后, 通过前馈网络 (FFN) 对融合特征进行非线性变换和跨模态适配:

$$F_m^{(k),i+1} = \text{FFN}(F_m^{(k),i}). \quad (10)$$

通过这种模态专用编码加上基于跨注意力的融合, 编码器能够有效结合富纹理的 RGB 信息和红外热信息, 同时保持与预训练 SAM 语义特征的对齐。这种层次化交互为后续的双分支特征精炼奠定了稳健的基础, 使网络能够自适应地利用互补信息而不破坏语义结构并减少对任何单一感知模态的过度依赖。

需要指出的是, SAM 编码器是在大规模 RGB 图像上进行预训练的, 其输入形式为三通道图像。对于单通道的红外图像, 本文采用通道复制 (Channel Replication) 策略, 将其在通道维度上扩展为三通道形式, 以满足 SAM 编码器的输入要求。具体而言, 将红外图像在通道维度复制 3 次, 从而构造伪 RGB 表示。

该处理方式能够在不改变 SAM 编码器结构及其预训练参数的前提下, 使红外模态有效接入其特征提取流程。同时, 由于复制操作不会引入额外的统计偏移, 红外图像强度分布特性得以保留, 从而保证其在后续跨模态融合过程中仍能提供有效的热辐射信息。

在此基础上, RGB 图像与扩展后的红外图像

分别输入 SAM 编码器与模态融合编码器进行特征提取, 从而实现统一的特征空间对齐, 为后续的跨模态交互与语义保持提供基础。

与直接对红外图像进行独立编码相比, 该策略能够更好地利用 SAM 预训练所学习到的语义表示能力, 从而提升整体语义建模效果。

3.3 语义持久性分支

为有效保持预训练 SAM 编码器所编码的语义先验知识, 本文引入专门的语义持久性分支。该分支的核心职责在于确保在多模态融合与特征精化过程中, SAM 提取的层次化语义特征的完整性与稳定性得以维持。若缺乏明确的保持机制, 当网络优先考虑模态特异性适配或跨模态交互时, 大规模预训练所习得的丰富结构模式可能逐渐削弱。这种语义衰减会损害全局一致性与边界连续性, 最终影响分割性能。因此, 该持久性分支作为稳定化通路, 确保目标级结构、区域连续性及其上下文关系在不同深度的特征层中保持一致。

为进一步说明该分支能够保持语义信息的原因, 本文从注意力机制的作用机理进行分析。首先, 自注意力机制在特征空间内建立全局范围的依赖关系, 使空间上分离但语义一致的区域能够形成协同响应, 从而增强类别内部一致性并抑制局部噪声干扰。这种全局建模能力有助于维持由 SAM 编码器所提供的结构化语义信息, 包括目标轮廓、区域连贯性以及上下文关系。其次, 交叉注意力机制通过以持久库中的语义特征作为键和值, 将多模态融合特征作为查询, 引入稳定的语义约束。在该过程中, 融合特征在更新时会优先对齐与语义先验一致的表示, 从而有效避免由于模态差异或特征扰动所导致的语义漂移问题。

通过上述机制, 自注意力与交叉注意力在功能上形成互补: 前者侧重于维持全局语义一致性, 后者侧重于在跨模态交互过程中引入语义引导与约束。二者协同作用, 使得语义信息在多阶段特征传播与融合过程中得以稳定保留, 从而实现语义保持分支的设计目标。

在第 k 个阶段, SAM 编码器输出 $F_{\text{SAM}}^{(k)}$ 首先经过两个顺序的自注意力层, 以生成语义信息丰富的表示, 并产生存储在持久库 (Persistent Repository, PR) 中的键值对 $(K_{\text{src}}, V_{\text{src}})$:

$$F_{\text{self}}^{(k)} = \text{SelfAtt}(F_{\text{SAM}}^{(k)}), \text{PR}^{(k)} = \{K_{\text{src}}, V_{\text{src}}\}. \quad (11)$$

这些表示作为后续跨模态交互的稳定语义锚点。

与此同时,来自融合编码器的多模态特征 $F_1^{\text{MM},(k)}$ 通过交叉注意力机制注入SPA分支,关注存储在PR中的语义键和值:

$$F_2^{\text{MM},(k)} = \text{CrossAtt}(Q = F_1^{\text{MM},(k)}, K = K_{\text{src}}, V = V_{\text{src}}), \quad (12)$$

其中,其输出将模态特定信息与持久的语义结构相结合。额外的交叉注意力层可进一步精炼,将SAM提取的特征与多模态信号以层次化方式融合。

最后,将自注意力路径和交叉注意力模块的输出进行拼接:

$$F_p^{(k)} = \text{Concat}(F_{\text{self}}^{(k)}, F_2^{\text{MM},(k)}). \quad (13)$$

生成第 k 阶段的语义保持特征 $F_p^{(k)}$ 。该特征作为自适应调制分支的稳定骨干,确保在支持逐阶段多模态增强的同时,高层语义结构得到保留。

通过这种设计,SPA有效地在各阶段保留了全局语义一致性、结构连贯性和上下文关系,同时允许根据持久库和交叉注意力进行受控的模态信息整合。

3.4 门控融合机制

为有效融合语义持久性分支与初始图像特征,本文引入门控融合机制。该机制旨在根据各特征阶段分支输出的相对可靠性,自适应平衡二者对最终融合表征的贡献。与直接将两个特征合并为单一表示不同,该机制首先生成两个以各自特征为主导的多阶段特征序列,然后在每个阶段通过语义感知跨模态融合模块(SACF, Semantic-Aware Cross-modal Fusion)生成最终融合特征。

设第 k 阶段SPA输出为 $\{F_p^{(k)}\}_{k=1}^K$, 原始图像特征为 $\{F_m^{(k),i}\}_{k=1}^K$ 。我们首先在每个阶段估计可学习的门控系数:

$$\beta^{(k)} = \sigma(\text{Conv}(F_p^{(k)} | F_m^{(k)})), \quad (14)$$

其中: $|$ 表示通道拼接, σ 表示 Sigmoid 激活函数, $\beta^{(k)}$ 用于自适应平衡两个分支的贡献,下标 p 表示语义持久性分支输出的语义保持特征,下标 m 表示多模态融合特征。

基于 $\beta^{(k)}$, 在第 k 阶段构建两个特征主导的特征:

$$F_{p,g}^{(k)} = \beta^{(k)} \odot F_p^{(k)} + (1 - \beta^{(k)}) \odot F_a^{(k)}, \quad (15)$$

$$F_{m,g}^{(k)} = \beta^{(k)} \odot F_a^{(k)} + (1 - \beta^{(k)}) \odot F_p^{(k)}, \quad (16)$$

其中: $\odot$ 表示按元素相乘; $F_{p,g}^{(k)}$ 是 SPA 主导特征,保留 SAM 继承的语义结构,同时融合来自 AMB 的自适应多模态信息; $F_{m,g}^{(k)}$ 是图形学主导特征,强调模态增强信息,同时保持 SPA 的语义一致性;下标 g 表示经过门控机制加权后的特征;下标 a 表示由自适应调制分支生成的多模态增强特征。

接着,在每个阶段应用 SACF 模块进行语义感知融合:

$$F_g^{(k)} = \text{SACF}(F_{p,g}^{(k)}, F_{m,g}^{(k)}, T), \quad (17)$$

其中, T 表示在 3.1 节中引入的文本嵌入矩阵。SACF 在每个阶段对两个分支主导特征进行语义交互,输出最终融合特征 $F_g^{(k)}$ 。

3.5 基于图像解码器的融合图像生成

为客观评估图像融合质量,并为后续跨模态语义交互与分割提供统一的视觉输入,本文在跨模态模块之前引入图像解码器 $D_{\text{img}}(\cdot)$ 。该解码器从多尺度门控融合特征 $\{F_g^{(k)}\}_{k=1}^K$ 重建融合图像 $\hat{I}_{\text{fuse}}$, 其中 $F_g^{(k)} \in \mathbb{R}^{H_k \times W_k \times C_k}$, 且 $H_1 \times W_1$ 对应最高空间分辨率。

本文采用自顶向下的特征金字塔解码方案。首先,对最深层特征通道维度进行对齐:

$$U^{(K)} = \phi(\text{Conv}_{1 \times 1}(F_g^{(K)})), \quad (18)$$

其中, $\phi(\cdot)$ 表示轻量级精化模块(如卷积-批归一化-ReLU 组合),用于调整特征分布并稳定表征。

随后,对于 $k = K - 1, \dots, 1$, 特征通过渐进式上采样与跳跃连接与对应尺度的特征进行融合:

$$U^{(k)} = \phi(\text{Conv}_{3 \times 3}(\text{Cat}(\text{Up}(U^{(k+1)}), \text{Conv}_{1 \times 1}(F_g^{(k)})))), \quad (19)$$

其中: $\text{Up}(\cdot)$ 表示 $\times 2$ 上采样算子(如双线性插值), $\text{Cat}(\cdot)$ 执行通道维度的拼接操作。这种层次化聚合策略既能保留深层特征中的高级语义信息,又能通过浅层特征渐进恢复空间细节。

最终,融合图像通过 1×1 投影头生成:

$$\hat{I}_{\text{fuse}} = \sigma(\text{Conv}_{1 \times 1}(U^{(1)})), \quad (20)$$

其中, $\sigma(\cdot)$ 为有界激活函数(如 Sigmoid, 用于将输出归一化至 $[0, 1]$ 区间)。实际应用中,根据实验配置, $\hat{I}_{\text{fuse}}$ 可生成为单通道融合强度图或三通道融

合图像。

3.6 基于融合图像的跨模态语义融合与分割

在获得像素级融合图像 $\hat{I}_{\text{fuse}}$ 后, 本文将其作为跨模态视觉-文本交互以及最终语义分割任务的主要视觉输入。该设计能够确保由文本信息引导的语义对齐直接作用于融合图像所形成的特征表示, 从而提升下游分割预测阶段的语义表达能力。

为此, 本文采用单流视觉编码器 $E_{\text{vis}}(\cdot)$ 从融合图像 $\hat{I}_{\text{fuse}}$ 中提取多尺度视觉特征:

$$F_{\text{fuse}}^{(k)} = E_{\text{vis}}^{(k)}(\hat{I}_{\text{fuse}}), k = 1, \dots, K, \quad (21)$$

其中, $F_{\text{fuse}}^{(k)}$ 表示第 k 个尺度的融合视觉特征。这些特征在后续的跨模态注意力模块中作为视觉查询 (Visual Queries) 参与计算。

给定文本嵌入表示 $T \in \mathbb{R}^{N \times d}$ (例如类别提示词 embeddings), 在每个特征尺度上执行跨模态注意力计算:

$$C^{(k)} = \text{CrossAttn}(Q = \Pi_q(F_{\text{fuse}}^{(k)}), K = \Pi_k(T), V = \Pi_v(T)), \quad (22)$$

其中, $\Pi_q(\cdot)$ 、 $\Pi_k(\cdot)$ 与 $\Pi_v(\cdot)$ 分别表示可学习的线性投影层, 用于生成查询 (Query)、键 (Key) 和值 (Value) 表示。

随后, 通过残差融合机制将跨模态语义信息注入视觉特征:

$$F_{\text{cm}}^{(k)} = F_{\text{fuse}}^{(k)} + \gamma^{(k)} \cdot \Pi_o(C^{(k)}), \quad (23)$$

其中: $\Pi_o(\cdot)$ 表示输出投影层; $\gamma^{(k)}$ 为可学习的尺度系数 (或门控系数), 用于在训练过程中稳定优化并调节文本语义引导对视觉特征的影响程度。

最终, 经跨模态语义增强后的多尺度特征集合 $\{F_{\text{cm}}^{(k)}\}_{k=1}^K$ 被输入到分割解码器 $D_{\text{seg}}(\cdot)$ 中, 以生成最终的语义分割预测结果:

$$\hat{Y} = D_{\text{seg}}(F_{\text{cm}}^{(1)}, F_{\text{cm}}^{(2)}, \dots, F_{\text{cm}}^{(K)}). \quad (24)$$

通过上述过程, 模型能够在融合图像的视觉信息基础上进一步引入文本语义指导, 从而实现更精确的跨模态语义表达, 并提升最终的分割性能。

3.7 训练目标

为优化所提框架, 本文采用复合目标函数, 同时监督像素级语义分割与融合图像重建, 并额外利用基于 SAM 的语义先验对中间表征进行正则化约束。该设计旨在促使模型生成信息丰富的融合图像以支持下游任务, 实现精确的像素级

分类, 并在各特征阶段保持语义一致性。

3.7.1 分割损失

根据前述流程, 最终预测结果 $\hat{Y}$ 由融合图像 $\hat{I}_{\text{fuse}}$ 依次通过视觉编码器、跨模态交互模块和分割解码器生成。本文采用标准交叉熵损失函数计算像素级分割损失 $\mathcal{L}_{\text{seg}}$, 度量预测分割图 $\hat{Y}$ 与真实标签 Y 在所有像素位置 $x \in \Omega$ 上的差异:

$$\mathcal{L}_{\text{seg}} = - \sum_{x \in \Omega} \sum_{c=1}^C Y_c(x) \log \hat{Y}_c(x), \quad (25)$$

其中: Ω 表示所有像素位置的集合, C 为语义类别总数, $Y_c(x)$ 和 $\hat{Y}_c(x)$ 分别表示像素 x 属于类别 c 的真实标签指示值与预测概率。

3.7.2 融合图像重建损失

为对跨模态交互生成的融合图像 $\hat{I}_{\text{fuse}}$ 进行显式监督, 本文引入融合图像重建损失 $\mathcal{L}_{\text{fuse}}$ 。用 $Y(\cdot)$ 表示对 RGB 图像 I_{rgb} 施加的亮度 (或灰度) 变换函数, 用于在必要时将其强度域与红外图像 I_{ir} 对齐。整体重建目标函数形式化为强度保持、梯度保持与结构相似性的加权组合:

$$\mathcal{L}_{\text{fuse}} = \lambda_{\text{int}} \mathcal{L}_{\text{int}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}} + \lambda_{\text{ssim}} \mathcal{L}_{\text{ssim}}. \quad (26)$$

强度保持项: 本文促使融合图像保留两种模态的互补强度信息:

$$\mathcal{L}_{\text{int}} = \|\hat{I}_{\text{fuse}} - Y(I_{\text{rgb}})\|_1 + \|\hat{I}_{\text{fuse}} - I_{\text{ir}}\|_1. \quad (27)$$

梯度保持项: 为保留显著的边缘结构与精细纹理, 本文将融合图像的梯度与两幅输入图像中较强的梯度进行对齐:

$$\mathcal{L}_{\text{grad}} = \|\nabla \hat{I}_{\text{fuse}} - \max(\nabla Y(I_{\text{rgb}}), \nabla I_{\text{ir}})\|_1, \quad (28)$$

其中: $\nabla(\cdot)$ 可采用 Sobel 算子实现梯度提取, $\max(\cdot, \cdot)$ 为逐元素取最大值操作。

结构相似性项: 为进一步强化结构一致性, 本文引入结构相似性指标:

$$\mathcal{L}_{\text{ssim}} = 1 - \frac{1}{2} \left(\text{SSIM}(\hat{I}_{\text{fuse}}, Y(I_{\text{rgb}})) + \text{SSIM}(\hat{I}_{\text{fuse}}, I_{\text{ir}}) \right). \quad (29)$$

3.7.3 基于 SAM 先验的特征一致性正则化

除像素级监督外, 本文引入特征一致性正则化项, 以促使模型特征与预训练 SAM 语义先验保持对齐。在各阶段 k , 约束持久性分支特征 $F_p^{(k)}$ 与对应的 SAM 特征 $F_{\text{sam}}^{(k)}$ 保持接近:

$$L_{\text{feat}}^{(k)} = \|F_p^{(k)} - \text{Detach}(F_{\text{sam}}^{(k)})\|_2^2, \quad (30)$$

其中, $\text{Detach}(\cdot)$ 用于阻止梯度传播至 SAM 特征。

该设计确保正则化项仅引导学习表征与冻结的语义先验对齐,而不修改预训练编码器参数。

3.7.4 整体优化目标

最终训练目标函数形式化定义为:

$$L = \mathcal{L}_{\text{seg}} + \lambda_{\text{feat}} \sum_{k=1}^K \mathcal{L}_{\text{feat}}^{(k)} + \lambda_{\text{fuse}} \mathcal{L}_{\text{fuse}}, \quad (31)$$

其中: λ_{feat} 和 λ_{fuse} 分别控制基于SAM的特征正则化项与融合图像重建损失的相对重要性, λ_{int} 、 λ_{grad} 和 λ_{ssim} 为 $\mathcal{L}_{\text{fuse}}$ 各分量的权重系数。该复合目标函数使模型能够在保持与预训练语义知识一致性的同时,联合学习高质量融合表征与精确语义分割。

4 数值实验

4.1 实验准备

所有实验均在配备NVIDIA RTX A6000 GPU的工作站上完成,模型实现基于PyTorch深度学习框架。为保证训练过程的稳定性以及不同方法之间的公平比较,本文统一采用AdamW优化器进行模型训练,批大小设置为16,初始学习率设为 4×10^{-3} 。除非另有说明,所有方法均在相同的优化参数设置下进行训练。

在融合处理之前,红外与可见光图像对首先进行空间配准,并统一调整至相同分辨率。为了保证评价的客观性,所有方法在实验过程中均未使用额外的后处理操作(例如图像锐化或直方图匹配)。

融合性能在5个广泛使用的多模态数据集上进行评估,包括TNO、RoadScene、MFNet、FMB以及M³FD。这些数据集涵盖了多种具有挑战性的复杂场景,例如低照度环境、背景杂乱、小目标以及复杂纹理等情况。为进行定量评价,本文采用4种客观指标:信息熵(Entropy, EN)、标准差(Standard Deviation, SD)、相关差异和(Sum of Correlation Differences, SCD)以及多尺度结构相似性(Multi-Scale Structural Similarity, MS-SSIM)。上述指标数值越高,表明融合图像质量越好。

为了进一步评估融合图像在下游密集预测任务中的有效性,本文将不同融合方法生成的结果输入到两个具有代表性的语义分割框架中进行测试,分别为SegFormer-B3和一种X-Decoder风格的通用分割模型。分割性能采用平均交并比(mean Intersection-over-Union, mIoU)作为评

价指标。

此外,为分析所提出方法中各关键模块的作用,本文通过逐一移除关键组件开展消融实验,包括:(1)语义保持分支(Semantic Persistence Branch);(2)自适应调制分支(Adaptive Modulation Branch);(3)文本嵌入模块(Text Embedding Module),以量化各模块对整体性能的独立贡献。

最后,在相同输入配置条件下,从模型参数量(Parameter Count)、浮点运算量(FLOPs)以及推理速度(FPS)3个方面对模型的计算复杂度进行综合分析。

需要说明的是,近年来,已有部分研究尝试将SAM引入图像分割或多模态任务中。然而,需要指出的是,大多数现有基于SAM的方法主要针对单模态RGB图像设计,或依赖特定提示(Prompt)机制,其输入形式与训练范式与本文所研究的红外-可见光融合分割任务存在显著差异。

例如,TV-SAM^[27]通过结合视觉-语言模型与SAM,实现了多模态医学图像的零样本分割能力,在多个模态数据上取得了良好效果。然而,该类方法通常依赖复杂的提示生成机制或额外的模型组件,对实际应用场景具有一定限制。

同时,已有研究表明,SAM在跨模态任务中的性能受限于其预训练于自然图像的数据分布,当直接应用于非自然图像(如医学或红外图像)时,其性能可能出现明显下降^[28]。因此,MA-SAM等工作通过引入参数高效的适配策略,对SAM进行跨模态扩展,以提升其在不同模态下的泛化能力。此外,部分相关方法尚未公开完整代码或训练细节,难以在统一实验设置下进行公平对比。

尽管如此,本文在实验中已引入具有代表性的通用分割框架(如X-Decoder风格模型)进行评估。该类方法在设计上融合了视觉-语言对齐与大规模预训练思想,与SAM在语义建模能力上具有一定相关性。实验结果表明,本文所提出的融合方法在不同分割框架下均能稳定提升性能,说明其在语义保持与跨模态融合方面具有良好的泛化能力。

与现有基于SAM的直接迁移或微调方法不同,本文方法将SAM编码器作为语义先验嵌入多阶段融合过程,并通过语义保持分支与跨模态调制机制实现逐层语义对齐。这种设计在融合

任务场景下具有更强的适应性与稳定性,从而在多项指标上取得一致性能提升。

4.2 实验结果与对比分析

4.2.1 多模态图像融合结果

本文将所提方法与 11 种具有代表性的融合方法进行比较,包括 DDFM^[29]、U2Fusion^[30]、TarDAL^[31]、SegMiF^[32]、FILM^[33]、SHIP^[34]、MRFS^[35]、IFCNN^[5]、DIDFuse^[36]、SpTFuse^[37]和 SGDFuse^[38]。实验在 6 个公开数据集上开展,分别为 TNO^[39]、RoadScene^[40]、MFNet^[22]、FMB^[32]、M³FD^[31]和 MSRS^[41]。表 1~6 给出了信息熵(EN)、标准差(SD)、相关差异和(SCD)以及多尺度结构相似性(MS-SSIM)4 项指标的定量评价结

表 1 不同方法在 TNO 数据集上的融合性能比较

Tab. 1 Fusion performance comparison of different methods on the TNO dataset

Method	EN ↑	SD ↑	SCD ↑	MS-SSIM ↑
DDFM	6.88	44.3	1.56	0.833
U2Fusion	7.05	45.4	1.63	0.848
TarDAL	7.18	46.6	1.70	0.857
IFCNN	7.14	46.5	1.69	0.854
DIDFuse	7.20	46.9	1.72	0.860
SegMiF	7.25	47.2	1.75	0.865
FILM	7.10	46.0	1.66	0.852
SHIP	6.96	45.0	1.61	0.845
MRFS	7.13	46.3	1.68	0.855
SpTFuse	7.32	47.6	1.78	0.872
SGDFuse	7.41	48.1	1.84	0.879
Proposed	7.46	48.6	1.89	0.884

表 2 不同方法在 RoadScene 数据集上的融合性能比较

Tab. 2 Fusion performance comparison of different methods on the RoadScene dataset

Method	EN ↑	SD ↑	SCD ↑	MS-SSIM ↑
DDFM	6.69	41.2	1.49	0.824
U2Fusion	6.86	42.0	1.57	0.839
TarDAL	6.98	43.1	1.64	0.848
IFCNN	6.92	42.9	1.62	0.845
DIDFuse	6.99	43.3	1.65	0.849
SegMiF	7.04	43.7	1.69	0.856
FILM	6.90	42.5	1.60	0.843
SHIP	6.78	41.8	1.54	0.835
MRFS	6.93	42.7	1.62	0.846
SpTFuse	7.08	44.4	1.74	0.859
SGDFuse	7.11	44.7	1.78	0.863
Proposed	7.12	45.4	1.81	0.868

表 3 不同方法在 MFNet 数据集上的融合性能比较

Tab. 3 Fusion performance comparison of different methods on the MFNet dataset

Method	EN ↑	SD ↑	SCD ↑	MS-SSIM ↑
DDFM	6.82	43.4	1.52	0.829
U2Fusion	6.99	44.4	1.59	0.843
TarDAL	7.12	45.5	1.66	0.852
IFCNN	7.06	45.1	1.63	0.850
DIDFuse	7.12	45.6	1.67	0.855
SegMiF	7.19	46.1	1.71	0.861
FILM	7.03	44.9	1.62	0.848
SHIP	6.90	44.0	1.55	0.840
MRFS	7.07	45.2	1.64	0.850
SpTFuse	7.26	46.1	1.69	0.865
SGDFuse	7.29	46.3	1.72	0.868
Proposed	7.38	46.5	1.78	0.873

表 4 不同方法在 FMB 数据集上的融合性能比较

Tab. 4 Fusion performance comparison of different methods on the FMB dataset

Method	EN ↑	SD ↑	SCD ↑	MS-SSIM ↑
DDFM	6.77	42.8	1.50	0.826
U2Fusion	6.95	43.6	1.57	0.839
TarDAL	7.08	44.5	1.63	0.848
IFCNN	7.02	44.2	1.61	0.845
DIDFuse	7.07	44.6	1.64	0.849
SegMiF	7.15	45.0	1.68	0.856
FILM	6.99	44.0	1.60	0.844
SHIP	6.86	43.2	1.54	0.837
MRFS	7.03	44.2	1.61	0.846
SpTFuse	7.18	45.2	1.69	0.858
SGDFuse	7.25	45.7	1.73	0.864
Proposed	7.32	46.2	1.78	0.869

表 5 不同方法在 M³FD 数据集上的融合性能比较

Tab. 5 Fusion performance comparison of different methods on the M³FD dataset

Method	EN ↑	SD ↑	SCD ↑	MS-SSIM ↑
DDFM	6.74	42.1	1.47	0.821
U2Fusion	6.92	43.0	1.55	0.835
TarDAL	7.04	44.0	1.61	0.844
IFCNN	6.98	43.6	1.59	0.841
DIDFuse	7.03	44.1	1.62	0.846
SegMiF	7.12	44.7	1.66	0.852
FILM	6.96	43.5	1.58	0.840
SHIP	6.83	42.8	1.52	0.832
MRFS	7.00	43.9	1.60	0.842
SpTFuse	7.12	44.2	1.61	0.856
SGDFuse	7.15	44.6	1.63	0.859
Proposed	7.21	44.8	1.64	0.864

表6 不同方法在MSRS数据集上的融合性能比较

Tab. 6 Fusion performance comparison of different methods on the MSRS dataset

Method	EN $\uparrow$	SD $\uparrow$	SCD $\uparrow$	MS-SSIM $\uparrow$
DDFM	6.81	43.1	1.52	0.829
U2Fusion	6.99	44.2	1.59	0.843
TarDAL	7.13	45.4	1.66	0.852
IFCNN	7.07	45.0	1.63	0.850
DIDFuse	7.14	45.5	1.67	0.856
SegMiF	7.20	46.0	1.71	0.861
FILM	7.04	44.9	1.62	0.848
SHIP	6.91	44.0	1.56	0.840
MRFS	7.08	45.1	1.64	0.850
SpTFuse	7.25	44.8	1.71	0.872
SGDFuse	7.28	45.0	1.73	0.874
Proposed	7.34	45.2	1.75	0.879

果^[42-43],其中最优结果以加粗形式标出。总体来看,本文方法在所有数据集和评价指标上均表现出具有竞争力甚至优于现有方法的性能。

在6个基准数据集上,本文方法在4项融合评价指标上均稳定处于领先水平。与多种强基线方法相比,该方法在信息丰富度、全局对比度、跨模态相关性以及多尺度结构一致性等方面均表现出稳定提升。这表明所提出的融合框架在不同场景分布及光照条件下均能够保持较为稳健的融合质量。

上述性能提升可从以下几个方面进行解释。首先,较高的EN值表明融合图像能够保留来自两种模态的更多互补信息。这主要得益于SAM编码器的语义感知特征表示与自适应调制机制的结合,该机制能够有效促进热显著区域与可见光纹理信息的保留。其次,SD指标的提升表明融合结果具有更强的全局对比度。自适应调制分支通过在语义结构特征中注入模态感知的增强信息,从而在低对比度环境下进一步强化显著区域。再次,SCD值的提高反映了更好的跨模态相关性保持能力。这一特性与所设计的门控融合机制密切相关,该机制能够在语义保持与模态驱动细化之间实现自适应平衡,从而减少单一模态主导带来的信息偏置,并缓解特征不一致问题。最后,MS-SSIM指标的提升表明融合结果在多尺度结构上具有更好的一致性。该特性对于下游语义任务尤为重要,其原因在于所提出的语义保持约束能够将融合特征与稳定的SAM预训练语义

先验进行对齐,从而有效保持层次化结构信息。

为更直观地展示跨数据集的对比结果,图6以归一化条形图形式呈现了6种基准测试下的融合性能。从统计结果可见,所提方法在所有数据集与评估指标上均取得最优分值。尤为重要的是,其性能优势同步体现在熵(EN)、标准差(SD)、结构一致性(SCD)及多尺度结构相似性(MS-SSIM)等多个维度,表明该方法并非仅在某项统计指标上取得单一突破。

在涵盖交通场景、复杂光照条件及多目标环境的不同测试情境中,该方法始终保持稳定的相对性能优势,未见任何单项指标出现明显衰减。这种均衡的优越性表明,所提框架在保持结构连贯性的同时,有效整合了多模态信息的互补特征。6个数据集呈现的视觉一致性进一步验证,观测到的性能提升具有系统性特征而非局限于特定数据集,从而证实了语义感知分层融合设计的鲁棒性与泛化能力。

图7展示了代表性样本的定性比较结果。多数基准方法均存在以下典型退化现象:(i)热辐射显著性不足,导致红外主导目标辨识度下降;(ii)过度平滑效应削弱可见光纹理与精细结构(如车道标线与细微边缘);(iii)高对比边界区域出现光晕或振铃伪影,造成目标轮廓畸变与结构清晰度损失。

相比之下,所提方法在保持红外显著性与可见光结构细节方面表现更优:热辐射显著目标仍保持清晰突显,同时纹理与细微边缘得到有效保留,边界伪影显著减少。这一视觉表现与框架的架构设计高度契合:语义持久分支维持了SAM导出的稳定结构先验,自适应调制分支通过模态特异性增强恢复互补特征,门控融合机制则在边界区域平衡不同模态的贡献权重。

最终生成的融合图像呈现出更优的局部对比度与更清晰的目标轮廓。这些定性观察结果与EN、SD、SCD及MS-SSIM指标的量化提升相互印证,表明所提方法能为下游语义分割任务提供更可靠的视觉表征。

4.2.2 融合图像的语义分割结果

为评估融合输出结果是否保留了下游语义理解所需的类别级信息,我们将融合图像输入两个代表性分割框架(SegFormer-B3和X-Decoder风格模型),并在各数据集上给出平均交并比(mIoU)。

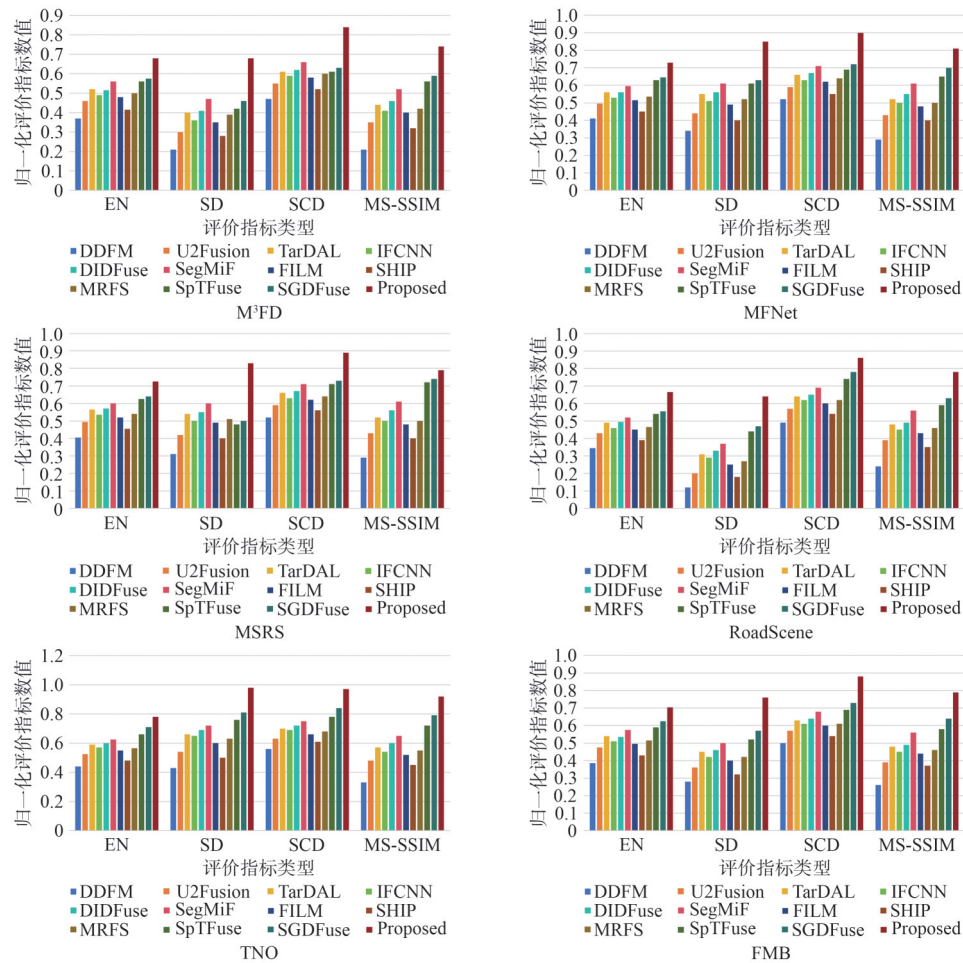


图6 不同方法在6个数据集上的融合性能对比柱状图

Fig. 6 Fusion performance comparison of different methods on six datasets

表7和表8显示,所提融合方法在所有数据集和两种分割器上均持续取得更优的分割性能(最优结果以加粗显示)。这表明融合图像不仅在视觉质量上有所提升,在语义层面也具备可靠的密集预测能力。

性能提升在具有挑战性的场景中尤为显著(如道路场景和M³FD数据集),这些场景中模态冲突(热目标与复杂可见光背景)及光照变化常导致目标边界模糊。这些改进可归因于语义感知融合设计:首先,稳定的SAM先验有助于保留与类别相关的结构信息,并减少边界偏移;其次,自适应调制分支增强了互补模态线索,如低对比度区域的热辐射显著性,同时保留可见光纹理细节;最后,门控融合机制可减轻背景干扰并抑制高对比边缘附近的光晕伪影。上述模块协同作用,生成更清晰的目标轮廓和更易于区分的特征区域,从而惠及两种分割框架。

图8展示了融合图像上的定性语义分割比较(两个代表性示例)。当融合结果削弱目标边界或在高对比边缘附近引入光晕及振铃伪影时,两种分割框架均易产生区域碎片化、相邻类别之间的边界泄漏以及小目标漏检等问题。这些误差在杂乱场景或低对比度区域尤为突出。

相比之下,所提融合方法生成的分割掩模具有更连贯的目标轮廓,且在背景干扰下产生的伪激活更少。这一表现可归因于语义感知融合策略:该策略在保留SAM导出的结构先验的同时,选择性增强模态互补线索,从而降低边界模糊性并提升类间可分离性,这与表7和表8中给出的mIoU提升结果相符。

4.2.3 消融实验结果

我们通过移除完整模型中的关键模块进行消融研究,具体设置如下:(1)移除语义持久分支(w/o SPB), (2)移除门控调制(w/o Gate),



图 7 不同方法在代表性样本上的定性融合对比结果

Fig. 7 Qualitative comparison of fusion results of different methods on a representative sample

表 7 基于 SegFormer-B3 模型在融合图像上的语义分割性能 (mIoU/%)

Tab. 7 Semantic segmentation performance (mIoU/%) using SegFormer-B3 on fused images

Fusion Method	TNO ↑	RoadScene ↑	MFNet ↑	FMB ↑	M ³ FD ↑	MSRS ↑
DDFM	64.0	58.7	63.2	61.0	62.6	61.9
U2Fusion	65.6	60.3	64.8	62.4	63.8	63.4
TarDAL	66.5	62.0	67.1	64.1	65.3	65.0
IFCNN	66.8	62.8	67.5	64.7	66.0	65.7
DIDFuse	67.0	63.1	67.8	65.0	66.4	66.0
SegMiF	67.2	63.5	68.0	65.2	66.9	66.2
FILM	66.4	61.2	65.8	63.6	65.5	64.5
SHIP	65.0	59.8	64.2	62.1	63.9	63.0
MRFS	66.1	61.6	66.8	63.9	65.1	64.7
SpTFuse	68.3	64.3	69.0	65.3	67.5	66.8
SGDFuse	68.6	64.7	69.5	65.6	67.7	67.3
Proposed	69.0	65.1	70.4	65.7	68.0	67.6

表 8 基于 X-Decoder-style 模型在融合图像上的语义分割性能 (mIoU/%)

Tab. 8 Semantic segmentation performance (mIoU/%) using an X-Decoder-style segmenter on fused images

Fusion Method	TNO ↑	RoadScene ↑	MFNet ↑	FMB ↑	M3FD ↑	MSRS ↑
DDFM	66.2	60.1	65.7	63.3	62.7	63.6
U2Fusion	68.0	62.1	67.6	65.0	64.3	65.4
TarDAL	69.0	63.8	69.3	66.6	65.8	66.9
IFCNN	69.6	64.3	69.8	67.1	66.4	67.4
DIDFuse	69.9	64.7	70.2	67.5	66.8	67.8
SegMiF	70.1	65.0	70.5	67.8	67.1	68.1
FILM	68.7	63.0	68.6	65.9	65.3	66.3
SHIP	67.0	61.5	66.8	64.2	63.5	64.6
MRFS	68.8	63.4	69.0	66.1	65.7	66.6
SpTFuse	70.4	66.7	70.8	67.6	66.9	69.0
SGDFuse	70.9	67.2	71.4	68.5	66.1	69.5
Proposed	71.2	68.2	72.4	70.1	67.1	70.0

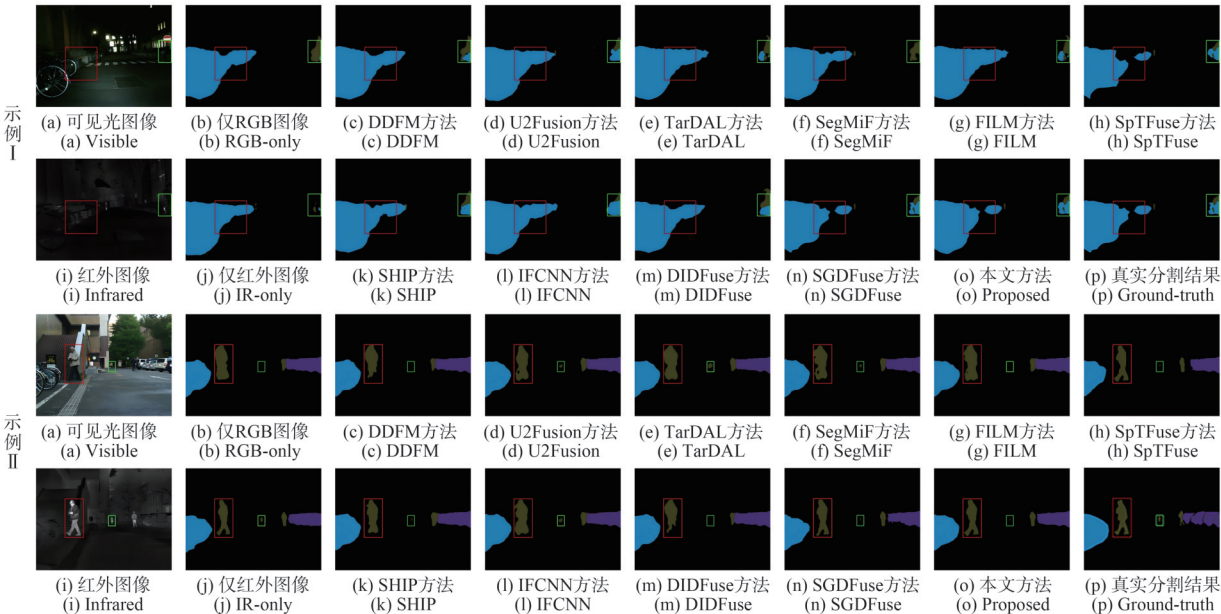


图 8 对融合图像进行定性语义分割比较。对于每个样本,第一行从可见光图像开始,第二行从红外图像开始;第二行的最后一列显示了真实分割结果。

Fig. 8 Qualitative semantic segmentation comparisons on fused images. For each sample, the first row starts with the visible image and the second row starts with the infrared image; the last column of the second row shows the ground-truth segmentation.

(3)移除文本嵌入模块(w/o TEM)。

我们使用以下指标评估消融实验结果:(i) 5个数据集的融合指标平均值;(ii)采用 SegFormer-B3 作为代表性下游评估框架所得的分割 mIoU。

如表 9 所示,移除任一所提模块均会导致性能显著下降。其中,移除 SPB 对 MS-SSIM 和 SCD 指标的削弱最为明显,表明保留 SAM 导出的语义结构对于维持跨模态结构一致性及减少

语义漂移具有关键作用。移除门控调制同样造成性能大幅降低,印证了门控融合在有效整合互补模态线索方面的必要性。相比之下,移除 TEM 主要影响语义对齐,因而即便融合指标仍具竞争力,分割 mIoU 仍出现明显下降。上述结果证实,各模块分别对整体性能的不同维度做出贡献,其协同工作方能实现最优的均衡性与鲁棒性。

从图 9 所示的语义分割结果可以看出,完整

表 9 消融实验结果
Tab. 9 Ablation study results

Configuration	EN ↑	SD ↑	SCD ↑	MS-SSIM ↑	mIoU ↑
Full Model (SPB+Gate+TEM)	7.31	46.8	1.81	0.871	68.3
w/o SPB (remove semantic persistence)	7.11	45.3	1.70	0.855	65.4
w/o Gate (remove gate fusion)	7.01	44.8	1.61	0.840	65.0
w/o TEM (remove text embedding)	7.20	46.1	1.72	0.864	66.6

注:融合指标为 5 个数据集的平均值,分割性能以 SegFormer-B3 模型评估的交并比(mIoU/%)表示,最优结果以加粗显示。

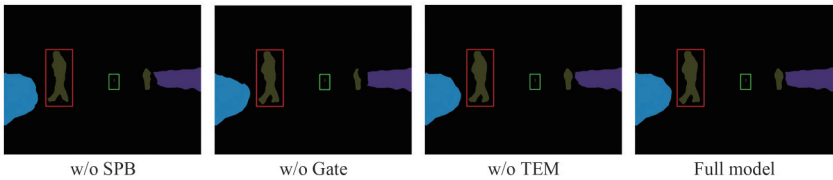


图 9 不同消融实验语义分割结果图

Fig. 9 Semantic segmentation results of different ablation experiments

模型在所有场景下均能生成区域完整、边界清晰的分割掩膜,与真实标签高度吻合。当移除语义保持分支后,目标内部出现明显空洞,边界产生断裂,这验证了 SPB 在保持 SAM 预训练语义结构方面的关键作用。移除门控融合机制和文本嵌入机制的分割图也在一定程度上有缺失,如图片右端的人,这充分证明了本文所提出的 3 个核心模块——语义保持分支、门控融合机制和文本嵌入模块,各自独立且协同地对最终分割性能做出重要贡献。

表 10 进一步给出了不同消融设置在各数据集上的分割性能结果。可以看出,完整模型在 TNO、RoadScene、MFNet、FMB、M³FD 和 MSRS 数据集上均取得最高 mIoU,说明所提出的语义保持分支、门控融合机制和文本嵌入模块在不同场景下均具有稳定贡献。移除任一模块后,各数据集上的分割精度均出现不同程度下降,进一步验证了各模块对提升跨模态语义分割性能的有效性。

表 10 基于 SegFormer-B3 模型的各数据集分割消融实验结果(mIoU/%)
Tab. 10 Dataset-wise segmentation ablation (mIoU/%) using SegFormer-B3

Configuration	TNO ↑	RoadScene ↑	MFNet ↑	FMB ↑	M3FD ↑	MSRS ↑
Full Model	70.0	66.2	71.5	67.4	68.3	69.2
w/o SPB	66.0	61.3	67.8	64.2	65.1	65.4
w/o Gate	66.8	62.1	67.7	64.5	64.6	64.6
w/o TEM	68.7	63.9	70.0	66.1	65.3	66.5

注:最优结果以加粗显示

4.2.4 计算复杂度对比

表 11 进一步对比了各方法的计算复杂度。尽管所提框架引入了额外的语义感知分支,但其运行效率仍具竞争力,在精度与效率之间实现了良好平衡。与强基线方法相比,所提方法在提升融合质量与下游分割性能的同时,保持了实用的推理速度。这些结果表明,语义保持型融合可在不显著增加计算开销的前提下实现。

4.2.5 夜间低光场景泛化性验证

为验证所提方法在夜间低光照环境下的泛化能力,本文选取 LLVIP 数据集进行专项评估。LLVIP 数据集专门针对极低光照条件下的街道与校园场景构建,包含 15 488 对严格配准的红外与可见光图像,其中可见光图像亮度极低、噪声明显,而红外图像能够清晰呈现行人及热目标。该数据集能够有效检验多模

表 11 计算复杂度对比

Tab. 11 Computational complexity comparison

Method	Params/M ↓	FLOPs/G ↓	FPS ↑	Peak Mem/GB ↓
DDFM	18.6	41.9	33.0	6.3
U2Fusion	21.7	48.3	29.2	6.8
TarDAL	23.9	51.6	26.7	7.2
SegMiF	25.3	54.2	24.9	7.5
FILM	20.8	47.5	28.4	6.9
SHIP	19.4	44.8	31.1	6.5
MRFS	20.2	46.1	30.0	6.7
SpTFuse	24.0	53.8	25.1	7.0
SGDFuse	24.6	54.5	24.6	6.8
Proposed	25.2	54.9	23.4	7.2

注:FLOPs在相同输入分辨率下测量,FPS以批大小为1进行测量。

态融合方法在夜间监控等实际应用场景中的鲁棒性。

本文在LLVIP数据集上对比了所提方法与9种代表性融合方法(DDFM、U2Fusion、TarDAL、

IFCNN、DIDFuse、SegMiF、FILM、SHIP、MRFS)的融合性能,同时采用SegFormer-B3作为统一分割网络评估融合图像的下游语义分割精度,定量结果如表12所示。

表 12 LLVIP 夜间数据集上的融合性能与分割精度比较

Tab. 12 Fusion performance and segmentation accuracy on the LLVIP night dataset

Method	EN ↑	SD ↑	SCD ↑	MS-SSIM ↑	mIoU/(SegFormer-B3) ↑
DDFM	6.42	39.8	1.38	0.801	58.3
U2Fusion	6.61	40.7	1.46	0.817	60.1
TarDAL	6.78	41.9	1.54	0.829	62.4
IFCNN	6.74	41.5	1.52	0.826	61.9
DIDFuse	6.81	42.2	1.56	0.832	63.0
SegMiF	6.88	42.8	1.60	0.838	63.8
FILM	6.70	41.2	1.49	0.822	61.5
SHIP	6.54	40.3	1.42	0.813	59.6
MRFS	6.76	41.7	1.53	0.827	62.2
SpTFuse	7.03	43.8	1.67	0.852	65.4
SGDFuse	7.01	43.6	1.66	0.850	65.1
Proposed	7.09	44.5	1.73	0.859	66.9

从表12可以看出,本文方法在LLVIP夜间数据集上的所有融合指标(EN、SD、SCD、MS-SSIM)均显著优于现有方法。特别地,EN达到7.09,表明融合图像保留了更丰富的互补信息;SCD达到1.73,证明跨模态相关性保持能力较强。在下游语义分割任务中,本文方法生成的融合图像取得了66.9%的mIoU,相比次优方法SegMiF提升3.1个百分点。该结果充分说明,即使在极低光

照条件下,本文引入的SAM语义先验与语义保持分支仍能有效引导多模态融合,生成语义可靠、结构清晰的融合表征,从而验证了所提框架在夜间监控等挑战性场景下的良好泛化能力。

5 结 论

本文提出了一种语义保持的多模态融合框

架,将高层语义先验引入红外-可见光图像融合,以缓解语义偏移与结构不一致性问题。具体而言,该框架采用预训练的SAM编码器作为稳定的语义主干,结合语义持久分支以保留SAM导出的结构线索;引入门控融合以增强热辐射显著性与可见光纹理的模式互补信息;设计基于文本嵌入的对齐模块以强化跨模态整合过程中的类别感知一致性。在TNO、RoadScene、MFNet、FMB及M³FD数据集上的大量实验表明,所提方法在EN、SD、

SCD及MS-SSIM等融合指标上均持续取得优异性能。更重要的是,融合输出结果提供了语义可靠的视觉表征,在与SegFormer-B3及X-Decoder风格分割器结合时,有效提升了下游分割精度,平均交并比持续改进。消融研究证实了各模块(尤其是语义持久机制)的贡献,而复杂度分析则显示了该方法在效率与精度之间的竞争力。总体而言,本框架为多模态感知与分割任务提供了一种实用的语义引导解决方案。

参 考 文 献:

- [1] 刘玉,于明,朱叶. 基于双重聚合和自合并网络的小样本图像语义分割[J]. 液晶与显示,2024,39(10):1421-1430.
LIU Y, YU M, Zhu Y. Bi-aggregation and self-merging network for few-shot image semantic segmentation [J]. *Chinese Journal of Liquid Crystals and Displays*, 2024, 39(10): 1421-1430. (in Chinese)
- [2] 朱磊,车晨洁,姚同钰,等. 融合多维特征的街景图像语义分割方法[J]. 液晶与显示,2024,39(7):980-989.
ZHU L, CHE C J, YAO T Y, *et al.* Semantic segmentation method for street images with multi-dimensional features [J]. *Chinese Journal of Liquid Crystals and Displays*, 2024, 39(7): 980-989. (in Chinese)
- [3] 赵阳,杨文贵,高翠云. 基于全局双组注意力的红外与可见光图像融合[J]. 液晶与显示,2025,40(12):1840-1852.
ZHAO Y, YANG W G, GAO C Y. Infrared and visible image fusion based on global dual-group attention [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(12): 1840-1852. (in Chinese)
- [4] 李青松,杨莘,吴谨,等. 基于结构纹理分解的红外与可见光图像融合算法[J]. 液晶与显示,2023,38(10):1389-1398.
LI Q S, YANG S, WU J, *et al.* Infrared and visible image fusion algorithm based on structure-texture decomposition [J]. *Chinese Journal of Liquid Crystals and Displays*, 2023, 38(10): 1389-1398. (in Chinese)
- [5] 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.
- [6] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C]. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 2650-2658.
- [7] 邱敬,梁焯,刘冀钊,等. 基于跨域交互注意力和对比学习引导的红外与可见光图像融合[J]. 中国光学(中英文), 2025,18(2):317-332.
DI J, LIANG C, LIU J Z, *et al.* Infrared and visible image fusion guided by cross-domain interactive attention and contrastive learning [J]. *Chinese Optics*, 2025, 18(2): 317-332. (in Chinese)
- [8] HA Q S, WATANABE K, KARASAWA T, *et al.* MFNet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes [C]. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017: 5108-5115.
- [9] SUN Y X, ZUO W X, LIU M. RTFNet: RGB-Thermal fusion network for semantic segmentation of urban scenes [J]. *IEEE Robotics and Automation Letters*, 2019, 4(3): 2576-2583.
- [10] 刘玉,郭迎春,朱叶,等. 基于双交叉注意力Transformer网络的小样本图像语义分割[J]. 液晶与显示,2024,39(11): 1494-1505.
LIU Y, GUO Y C, ZHU Y, *et al.* Dual cross-attention Transformer network for few-shot image semantic segmentation [J]. *Chinese Journal of Liquid Crystals and Displays*, 2024, 39(11): 1494-1505. (in Chinese)
- [11] TOET A. The TNO multiband image data collection [J]. *Data in Brief*, 2017, 15: 249-251.
- [12] XU H, MA J Y, JIANG J J, *et al.* U2Fusion: a unified unsupervised image fusion network [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(1): 502-518.
- [13] 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 (CVPR). New Orleans: IEEE, 2022: 5792-5801.
- [14] 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 (CVPR). Seattle: IEEE, 2024: 26374-26385.
- [15] 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.
- [16] XU J, DE MELLO S, LIU X, *et al.* X-decoder: generalized decoding for pixel, image, and language [C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 15186-15196.
- [17] JALALI B, MACPHEE C. VEVid: vision enhancement via virtual diffraction and coherent detection [J]. *eLight*, 2022, 2(1): 24.
- [18] ZHAO Z X, BAI H W, ZHU Y Z, *et al.* DDFM: denoising diffusion model for multi-modality image fusion [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 8048-8059.
- [19] 卢鹏, 刘楷赞, 邹国良, 等. 基于多特征融合和卷积神经网络的无参考图像质量评价[J]. 液晶与显示, 2022, 37(1): 66-76.
- LU P, LIU K Y, ZOU G L, *et al.* No reference image quality assessment based on fusion of multiple features and convolutional neural network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2022, 37(1): 66-76. (in Chinese)
- [20] ZUO C, QIAN J M, FENG S J, *et al.* Deep learning in optical metrology: a review [J]. *Light: Science & Applications*, 2022, 11(1): 39.
- [21] ZHAO Z X, DENG L L, BAI H W, *et al.* Image fusion via vision-language model [C]//*Proceedings of the 41st International Conference on Machine Learning*. Vienna: OpenReview. net, 2024: 60749-60765.
- [22] KIRILLOV A, MINTUN E, RAVI N, *et al.* Segment anything [C]. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 3992-4003.
- [23] BURT P, ADELSON E. The Laplacian pyramid as a compact image code [J]. *IEEE Transactions on Communications*, 1983, 31(4): 532-540.
- [24] MALLAT S G. A theory for multiresolution signal decomposition: the wavelet representation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989, 11(7): 674-693.
- [25] 邬开俊, 张治瑞, 汪滢, 等. 全局感知与多尺度特征融合的城市道路语义分割[J]. 光学精密工程, 2025, 33(14): 2262-2277.
- WU K J, ZHANG Z R, WANG Y, *et al.* Urban road semantic segmentation with global awareness and multi-scale feature fusion [J]. *Optics and Precision Engineering*, 2025, 33(14): 2262-2277. (in Chinese)
- [26] WILLIAMS C, COUSINS R, MELLOR C J, *et al.* Hyperpixels: pixel filter arrays of multivariate optical elements for optimized spectral imaging [J]. *Light: Advanced Manufacturing*, 2025, 6(4): 734-751.
- [27] YANG B, LI S T. Multifocus image fusion and restoration with sparse representation [J]. *IEEE Transactions on Instrumentation and Measurement*, 2010, 59(4): 884-892.
- [28] 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.
- [29] ZHAO Z X, DENG L L, BAI H W, *et al.* Image fusion via vision-language model [C]//*Proceedings of the 41st International Conference on Machine Learning*. Vienna: JMLR. org, 2024: 2513.
- [30] RADFORD A, KIM J W, HALLACY C, *et al.* Learning transferable visual models from natural language supervision [C]//*Proceedings of the 38th International Conference on Machine Learning*. Online: PMLR, 2021: 8748-8763.
- [31] ZHOU K Y, YANG J K, LOY C C, *et al.* Learning to prompt for vision-language models [J]. *International Journal of Computer Vision*, 2022, 130(9): 2337-2348.
- [32] 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.

- [33] ZHAO Z X, DENG L L, BAI H W, *et al.* Image fusion via vision-language model [C]//*Proceedings of the 41st International Conference on Machine Learning*. Vienna: JMLR. org, 2024: 2513.
- [34] 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 (CVPR). Seattle: IEEE, 2024: 26374-26385.
- [35] ZHANG H, ZUO X H, JIANG J, *et al.* MRFS: mutually reinforcing image fusion and segmentation [C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024: 26974-26983.
- [36] 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 (ICCV). Paris: IEEE, 2023: 8081-8090.
- [37] GU L, LUO X Q, LIU Y, *et al.* SAM-guided multi-level collaborative Transformer for infrared and visible image fusion [J]. *Pattern Recognition*, 2025, 162: 111391.
- [38] ZHANG X Y, LI J J, FAN G D, *et al.* SGDFuse: SAM-guided diffusion model for high-fidelity infrared and visible image fusion [J]. *Information Fusion*, 2026, 133: 104290.
- [39] TOET A. The TNO multiband image data collection [J]. *Data in Brief*, 2017, 15: 249-251.
- [40] XU H, MA J Y, LE Z L, *et al.* FusionDN: A unified densely connected network for image fusion [C]//*Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*. New York: AAAI Press, 2020: 12484-12491.
- [41] 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, 83-84: 79-92.
- [42] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment [C]. The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Pacific Grove: IEEE, 2003: 1398-1402.
- [43] ASLANTAS V, BENDES E. A new image quality metric for image fusion: the sum of the correlations of differences [J]. *AEU-International Journal of Electronics and Communications*, 2015, 69(12): 1890-1896.

作者简介:



陈悦,男,硕士研究生,2022年于江苏科技大学获得学士学位,主要研究方向为计算机视觉、目标检测等。E-mail: 13338939052@163.com



许锋,男,博士,教授,二级警监,2006年于东北大学获得博士学位,主要研究方向为视频侦查、刑事影像等。E-mail: 7677759@qq.com