基于文本语义引导的红外与可见光图像融合方法

朱明瑞,陈希茹,卫 鑫,王楠楠*,高新波
(西安电子科技大学空天地一体化综合业务网全国重点实验室,陕西西安 710071)

摘 要: 红外与可见光图像融合(Infrared and Visible image Fusion,IVF)旨在结合两种图像模态中的互补信息,将红外图像中的显著目标与可见光图像的丰富纹理细节进行有效整合,从而生成在信息量与视觉质量方面均优于单一模态的融合图像。现有研究证实,基于深度学习的融合方法已在提升融合图像质量方面取得了显著进展,但这类方法大多仍局限于低层视觉特征层面的建模,对于高层语义信息与视觉特征之间的深层语义关联挖掘仍不充分。近年来,随着大规模视觉-语言模型(Vision-Language Models,VLMs)的快速发展,文本引导的图像融合方法因其灵活性与多样性而展现出巨大潜力。然而,文本语义信息的有效整合与利用仍有待深入研究。针对上述问题,本文提出了一种用于红外与可见光图像融合的文本语义引导方法(Textual Semantic Guidance,TeSG),该方法以下游目标检测与语义分割等视觉任务为目标,通过在融合过程中显式引入由 VLMs 生成的高层语义信息,实现对融合过程的精准调控。TeSG 从两个层级引入文本语义信息:一是由 VLMs 自动生成文本描述,作为全局文本语义级引导,为融合过程提供高层语义约束;二是基于文本描述生成关键目标区域的掩码语义,实现对前背景区域的定位与差异化建模。基于此,本文设计了三个核心模块:语义信息生成(Semantic Information Generator,SIG)模块基于自动生成的文本描述生成掩码语义与文本语义;掩码引导交叉注意力(Mask-Guided Cross-Attention,MGCA)模块在掩码语义的指导下,对红外与可见光图像的视觉特征进行基于注意力的初步融合,实现掩码级别跨模态特征的交互;文本驱动注意力融合(Text-Driven Attentional Fusion,TDAF)模块通过文本引导注意力和门控机制实现语义级的融合与动态加权。实验结果表明,所提 TeSG 方法通过双层语义引导的融合范式,在保持多模态图像纹理和对比度方面均优于现有先进方法(State Of The Art,SOTA),并在下游目标检测与语义分割任务中也取得了更优的性能,相较于当前最优的图像融合方法平均提升了 1.4%,验证了其竞争力与有效性。本文方法有效解决了现有图像融合算法文本与视觉特征的深层关联探索不充分的问题,实现了融合质量与下游任务性能的双重提升。

关键词: 图像融合;红外与可见光图像;文本语义引导;深度学习;视觉-语言模型;注意力
基金项目: 国家自然科学基金(No.62576261,No.U22A2096)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0086-16
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250906

Textual Semantic Guidance for Infrared and Visible Image Fusion

ZHU Mingrui, CHEN Xiru, WEI Xin, WANG Nannan*, GAO Xinbo
(*State Key Laboratory of Integrated Services Networks*, *Xidian University*, *Xi'an*, *Shaanxi* 710071, *China*)

Abstract: Infrared and visible image fusion (IVF) aims to integrate the complementary information contained in both image modalities by effectively combining the salient targets in infrared images with the rich texture details present in visible images. Through this integration, IVF produces more informative and comprehensive fused images that surpass single-modality inputs. Existing research has demonstrated that deep learning-based fusion methods have achieved remarkable progress in improving fused image quality. However, most of these approaches focus mainly on low-level visual features, and the deep semantic associations between high-level semantic information and visual features have not yet been sufficiently explored. In recent years, with the rapid development of large vision-language models (VLMs), text-guided image fusion methods have exhibited great potential due to their flexibility and versatility. However, the effective integration and utilization of textual semantic information in the image fusion process remain insufficiently studied. To tackle these challenges, this paper proposes a textual semantic guidance method for infrared and visible image fusion, termed textual semantic guidanc (TeSG), which guides the image synthesis process in a way that is optimized for downstream tasks such as object detection and semantic segmentation. By explicitly introducing high-level semantic information generated by VLMs into the fusion pipeline, TeSG achieves precise regulation of the fusion process and enhances the semantic consistency of the fused results. TeSG introduces textual semantics at two levels: the mask semantic level and the text semantic level. First, automati-

收稿日期:2025-12-21;录用日期:2026-01-06;责任编委:丛润民
*通信作者:王楠楠

cally generated textual descriptions from VLMs are employed as global text-level semantic guidance, providing high-level semantic constraints for the fusion process. Second, based on these textual descriptions, mask semantics corresponding to key target regions are constructed, enabling accurate localization and differentiated modeling of foreground and background regions. Building on this, three core modules are designed to implement the proposed framework. The semantic information generator (SIG) module generates both mask semantics and text semantics from automatically produced textual descriptions. The mask-guided cross-attention (MGCA) module performs preliminary attention-based fusion of visual features from both infrared and visible images under the guidance of mask semantics, thereby realizing mask-level cross-modal feature interaction. Finally, the text-driven attentional fusion (TDAF) module achieves text-level fusion and dynamic weighting through text-guided attention and a gating mechanism, allowing semantic cues to modulate the contribution of different modalities in an adaptive manner. Experimental results demonstrate that the proposed TeSG method, through its dual-level textual semantic guidance paradigm, performs favorably against existing state of the art (SOTA) methods in preserving multimodal texture information and enhancing contrast in the fused images. In addition, TeSG yields superior performance in downstream tasks such as object detection and semantic segmentation, highlighting its task-oriented fusion capability. Compared with current SOTA image fusion approaches, the proposed TeSG achieves an average improvement of 1.4% on downstream tasks, validating its competitiveness and effectiveness while also exhibiting strong generalization ability across different datasets and scene conditions. The proposed method effectively addresses the insufficient exploration of deep correlations between textual and visual features in existing image fusion algorithms, achieving simultaneous improvements in fusion quality and downstream task performance.

Keywords: image fusion; infrared and visible images; textual semantic guidance; deep learning; vision-language models; attention

Foundation Item(s): National Natural Science Foundation of China (No.62576261, No.U22A2096)

0 引言

由于硬件条件和成像技术的限制,单一模态传感器往往难以在复杂场景中捕获全部有效信息。红外图像在低光照或复杂环境下表现优异,能够有效检测热辐射信息,但缺乏对场景的细节描述;而可见光图像能够提供丰富的纹理细节,但在低光照或低对比度条件下表现不佳。因此,仅依赖单一模态难以对场景进行全面而准确的表征[1-2]。红外与可见光图像融合(Infrared and Visible image Fusion, IVF)[3-4]作为一种关键的图像融合技术,旨在充分利用两种模态的互补优势。IVF的主要目标是同时保留红外图像中的热辐射信息与可见光图像中的纹理细节,从而生成高质量的融合图像。融合后的图像能够更好地表征场景,在目标检测[5-6]、场景理解[7]、遥感[8]等诸多重要领域均具有广泛的应用价值。

在过去的几十年中,已有大量研究致力于解决红外与可见光图像融合任务中的挑战。传统方法大多依赖多尺度变换[9-11]、稀疏表示[12-13]、子空间分析[14-15]以及显著性检测[16-17]等方法,直接对多模态信息进行融合。这些方法尽管在一定程度上具备有效性,但其依赖固定的特征提取方式与人工设计的融合规则,这限制了它们对复杂多样场景的处理能力。近年来,基于深度学习的方法迅速推动了图像融合的发展,例如基于自编码器(AutoEncoder, AE)的图像融合框架[11,18-20]、基于卷积神经网络(Convolutional Neural Network, CNN)的图像融合框架[21-23]、基于生成对抗网络(Generative Adversarial Network, GAN)的图像融合框架[24-28]以及基于Transformer的图像融合框架[29-30]等,通过端到端的学习显著提升了融合图像的质量。然而,这些方法多聚焦于视觉特征,往往忽略了高层语义信息与场景上下文。在复杂场景下,这种语义引导的缺失会导致语义一致性与可理解性下降,进而使融合性能欠佳。

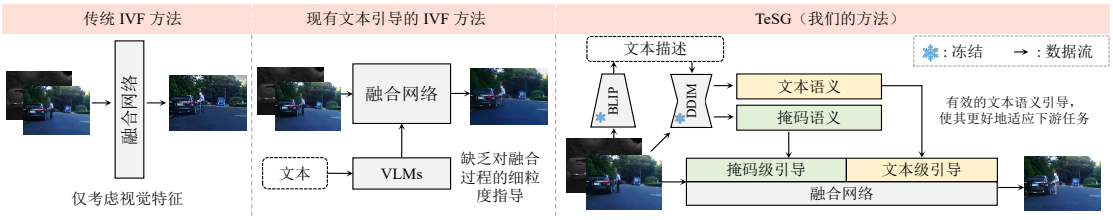
近年来,随着大型视觉-语言模型(Vision-Language Models, VLMs)的发展,文本引导的红外与可见光图像融合(Text-guided IVF)[31-35]受到了广泛关注。TextIF(Text guidance Image Fusion)[31]提出了一种语义交互引导模块,通过文本注明图像退化类型,实现退化感知的融合。Zhao等人[32]利用视觉-语言模型对源图像生成详细的语义描述,如图像标题与密集描述等,以引导并增强视觉特征的融合。尽管文本信息的引入提升了现有方法的灵活性与融合性能,但这些方法仍存在显著局限。首先,文本与视觉特征之间的深层语义关联尚未得到充分探索,缺乏全面的跨模态交互与融合;其次,文本语义在有效增强跨模态语义一致性方面仍待深入研究。

为解决上述问题,本文提出了一种文本语义引导方法(Textual Semantic Guidance, TeSG),以更有效地引导图像融合过程,使其更好地服务于下游任务。TeSG采用预训练的BLIP(Bootstrapping Language-

Image Pre-training)模型[36]自动提取图像的文本描述，从而避免了人工标注文本的成本。在图像融合过程中，这些文本描述的引导方式主要通过两个关键途径实现：一是基于文本描述生成关键目标的掩码语义，为关键区域视觉特征的融合提供精准的掩码级引导；二是通过利用文本语义与视觉特征建立语义关联，实现文本语义级的门控滤波，进而引导视觉特征融合。

为实现上述目标，TeSG包含三个功能模块：语义信息生成(Semantic Information Generator，SIG)模块、掩码引导交叉注意力模块(Mask-Guided Cross-Attention，MGCA)以及文本驱动注意力融合模块(Text-Driven Attentional Fusion，TDAF)。通过这种方式，该方法能够充分利用文本信息，相较于现有的文本驱动图像融合方法，提供了更有效的文本语义引导方案，具体如图1所示。

本文的主要贡献归纳如下：(1)探索了一种新颖的双层语义引导机制，通过掩码级和语义级双重引导实现图像融合，且无需手动标注文本，为文本引导的图像融合框架提供新的结构创新。(2)设计了三个紧密关联的模块以实现文本语义引导，包括语义信息生成模块、掩码引导交叉注意力模块以及文本驱动注意力融合模块。(3)定性和定量实验表明，所提方法能够生成高质量的融合图像，融合图像边缘信息保留更充分，充分验证了其在下游任务应用上的有效性。



注：传统方法仅关注低层次的视觉特征，而现有的文本引导融合方法往往在融合过程中忽略了深层次的细节信息。相比之下，本文的方法结合了掩码级、文本级和图像级特征，有效地引导了融合过程。

图1 现有红外与可见光图像融合方法与本文提出的TeSG方法的对比

Figure 1 Comparison of existing IVF methods with our TeSG

1 相关工作

1.1 红外与可见光图像融合

红外与可见光图像融合方法可大致分为传统方法和基于深度学习的方法。

传统红外与可见光图像融合方法主要包括以下五类。基于多尺度变换(Multi-Scale Transform，MST)的方法[9-10,37]主要利用小波变换[9-10]和非下采样轮廓波变换[37]等多尺度分解工具提取特征，能够在不同尺度上有效捕捉输入图像的局部与全局信息。基于子空间的方法[14-15,38]通过将高维数据投影到低维子空间，能够有效提取原图像的固有特征，同时降低冗余并保留关键信息。基于显著性检测的方法[16-17,39-40]通过突出最显著区域，充分利用红外与可见光图像的优势以实现有效融合。基于显著性检测的融合方法可大致分为两类：一类通过计算显著性权重进行图像融合[16,39]，另一类则提取显著目标进行融合[17,40]。基于稀疏表示(Sparse Representation，SR)的方法[12-13,41-42]通过构建能够有效表征图像特征的字典，将图像分解为稀疏系数，从而在融合过程中提取并组合红外与可见光图像的关键特征。每个传统IVF方法均具有自身优势与局限性，为克服单一方法的不足，研究者提出了混合方法[16,43]，例如Liu等人[16]提出了结合MST与SR的新型图像融合框架，旨在同时解决MST与SR方法各自的固有缺陷。然而，以上传统方法通常依赖于人工设计的特征提取策略或固定的融合规则，在复杂场景下缺乏灵活性与鲁棒性。

随着深度学习的快速发展，基于数据驱动的融合方法已逐步成为该领域的研究热点。近年来，深度学习技术显著推动了红外与可见光图像融合的发展进程。当前基于深度学习的融合方法可进一步分为以下四类。基于AE的方法[11,18,20,44]利用编码器-解码器框架完成特征提取与图像重建。例如，DenseFuse[18]将编码器-解码器网络与密集卷积块相结合，用于提取图像特征，并采用人工设计的融合策略实现特征融合；Li等人[44]提出端到端残差融合网络，采用两阶段训练策略以实现更有效的图像融合。基于CNN的方法[4,21-23,29]通过精心设计的网络架构和损失函数实现图像特征的提取、融合与重建。以RFNet(image Registration and Fusion Network)[23]为例，其提出一种适用于多模态图像配准与融合的无监督网络，通过引入互相增强框架有效提升了融合结果的性能。基于GAN的方法[3,24-25,28]因其在无监督分布学习中的强大能力而受到广泛关注，且GAN[45-46]在红外与可见光图像融合任务中表现优异，例如，FusionGAN[24]利用生成器-判别器框架，在实现高质量图像融合的同时，能够有效保留源图像的细节信息。基于扩散模型

的方法也成为近期图像融合领域涌现的突破性技术。DDFM(Denoising Diffusion image Fusion Model)[47]采用去噪扩散概率模型进行无条件图像生成,在跨模态信息保留方面表现出色;类似地,Yue 等人[48]利用扩散模型直接对多通道输入数据的分布进行建模,增强了多源信息的聚合并提升了融合图像色彩保真度。

1.2　文本引导的红外与可见光图像融合

随着图像生成与自然语言处理(Natural Language Processing,NLP)领域的快速发展,文本信息在指导图像生成[49-51]、图像编辑[52-53]以及图像翻译[54]等视觉任务中展现出强大潜力。Ramesh 等人[49]借助大规模文本-图像配对数据集及自回归生成模型实现了文本到图像的转换,使自然语言文本能够直接驱动高质量图像的生成。Qi 等人[55]通过文本描述指定生成图像的风格与内容,实现了可控的风格迁移。Nichol 等人[56]针对条件文本图像合成问题展开研究,提出基于扩散模型的文本引导图像生成与编辑方法。在上述视觉任务中,文本不仅为图像生成提供语义指导,还能直接控制图像的风格、内容或细节,实现文本描述与视觉内容的深度耦合。

然而,尽管文本信息在图像生成中取得了显著进展,其在图像融合任务中的潜力尚未得到充分挖掘。现有的 IVF 方法多聚焦于提升融合图像的视觉质量,对于语义信息的引入缺乏足够重视。一些研究尝试通过引入下游任务来增强融合效果,例如目标检测[28,57]、语义分割[58-59]等,但这些方法往往具有任务特定性,无法实现多模态特征的全面交互或精细控制,限制了其在更广泛应用场景中的适用性。相反,文本引导图像融合方法则直接将文本信息引入 IVF 任务。TextIF[31]提出了一种将图像融合管道与语义交互引导模块结合的方法,可针对退化图像执行感知式和交互式图像融合任务。Zhao 等人[32]将 VLMs 引入图像融合领域,通过多层次文本提示(如图像标题和密集描述)引导并增强视觉特征的融合。Wang 等人[33]提出了一个灵活的图像融合框架,利用多个 VLMs 解析文本指令,实现文本驱动与区域感知的图像融合。TextFusion(Textual semantics image Fusion)[34]引入文本-视觉关联映射及仿射融合单元,实现了对图像融合过程的精细控制。尽管现有方法在融合质量与模型灵活性方面有所突破,但仍面临着高计算复杂度以及文本与视觉特征之间深层语义关联探索不足等问题。为解决这些局限性,本文提出了一种有效的文本引导融合方法,通过掩码层级和语义层级的双重引导实现多模态特征的深度融合,且无需人工手动输入文本。

2　方法

2.1　方法概述

给定输入图像对 $\{(\boldsymbol{I}_{\mathrm{v}},\boldsymbol{I}_{\mathrm{i}})_i\}_{i=1}^{N}$,其中 $\boldsymbol{I}_{\mathrm{v}}\in\mathbf{R}^{H\times W\times 3}$ 表示可见光图像域的样本,$\boldsymbol{I}_{\mathrm{i}}\in\mathbf{R}^{H\times W\times 1}$ 表示红外图像域的样本。本文的目标是学习一个参数化映射 $\boldsymbol{I}_{\mathrm{f}}=\mathcal{G}(\boldsymbol{I}_{\mathrm{v}},\boldsymbol{I}_{\mathrm{i}}|\theta)$,能够生成高保真度的融合图像,有效整合两种模态的互补优势。为实现这一目标,本文提出了文本语义引导方法 TeSG,以更有效地引导图像融合过程,使其更好地应用于下游任务。TeSG 由三个核心模块组成:SIG(详见 2.2 节)、MGCA(详见 2.3 节)以及 TDAF(详见 2.4 节)。

如图 2 所示,对齐后的可见光图像 $\boldsymbol{I}_{\mathrm{v}}$ 和红外图像 $\boldsymbol{I}_{\mathrm{i}}$ 首先由 SIG 模块处理,该模块通过冻结参数的 BLIP[36]编码器从 $\boldsymbol{I}_{\mathrm{v}}$ 中提取文本描述。这些文本描述与输入图像一起被输入到预训练的扩散模型 Stable Diffusion[60]中,用于生成掩码语义 $\boldsymbol{M}$ 与文本语义 $\boldsymbol{F}_{\mathrm{t}}$。随后,输入图像及其对应的掩码语义 $\boldsymbol{M}$ 被送入编码器以生成特征表示,包括全局特征(即 $\boldsymbol{F}_{\mathrm{v}}$、$\boldsymbol{F}_{\mathrm{i}}$)和局部特征(即 $\boldsymbol{F}_{\mathrm{v}}^{\mathrm{m}}$、$\boldsymbol{F}_{\mathrm{v}}^{\bar{\mathrm{m}}}$、$\boldsymbol{F}_{\mathrm{i}}^{\mathrm{m}}$、$\boldsymbol{F}_{\mathrm{i}}^{\bar{\mathrm{m}}}$)。编码器与解码器均采用基于 Transformer 的模块结构。

这些编码后的特征随后进入 MGCA 模块,通过掩码引导的注意力机制实现跨模态特征融合。初步融合后的特征被送入 TDAF 模块,其中文本语义 $\boldsymbol{F}_{\mathrm{t}}$ 在引导图像融合过程中发挥关键作用。此外,TDAF 模块还引入了动态门控机制,通过调整融合权重有效融合可见光与红外图像特征。最终,解码器对融合后的特征进行重建,输出融合图像 $\boldsymbol{I}_{\mathrm{f}}$。

2.2　语义信息生成模块

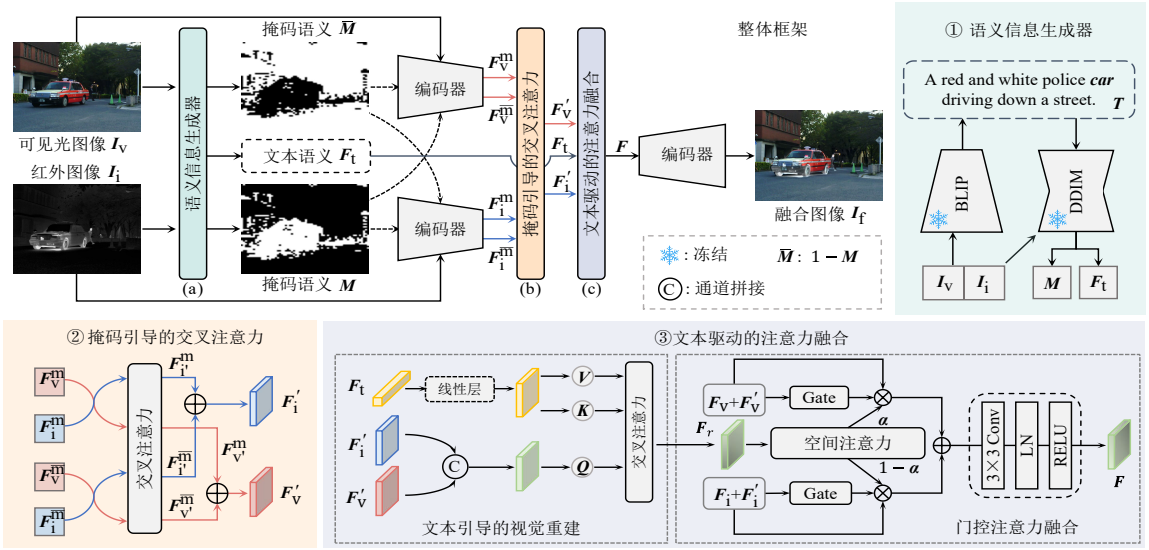
SIG 模块的详细工作流程以及生成的文本语义和掩码语义的可视化示例如图 3 所示。

2.2.1　文本描述生成

为了生成输入图像的语境准确的文本描述,本文首先利用预训练的 BLIP 模型[36]。BLIP 模型基于大规模自然图像进行预训练,确保了在大多数场景中生成描述的总体稳定性,足以满足语义引导的需求。因此,将可见光图像输入冻结的 BLIP 编码器,以生成对应的文本描述 $\boldsymbol{T}$。该过程可表示为

$$\boldsymbol{T}=\mathrm{BLIP}(\boldsymbol{I}_{\mathrm{v}}) \tag{1}$$

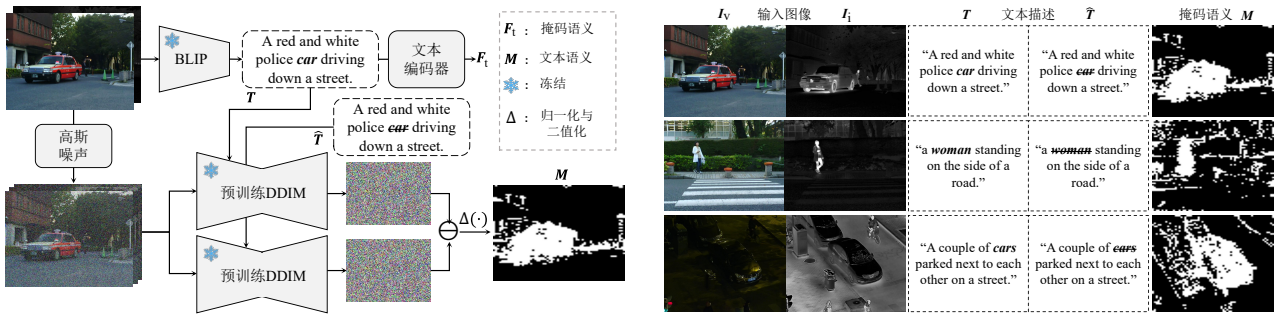
其中,$\boldsymbol{I}_{\mathrm{v}}$ 表示输入的可见光图像;$\boldsymbol{T}$ 表示 BLIP 模型生成的文本描述。此外,本文对 $\boldsymbol{T}$ 进行修改,去除其中关键词 V^{*},得到修改后的文本描述 $\hat{\boldsymbol{T}}$。具体来说,关键词 V^{*}(例如图 3(a)中的“car”)对应下游任务中的关键目标,如目标检测或语义分割中的对象。在生成及后续融合过程中,可见光图像与红外图像共享同一组文本描述 $\boldsymbol{T}$ 和 $\hat{\boldsymbol{T}}$。



注：TeSG的整体框架由三个模块组成：① 语义信息生成模块，用于生成掩码语义与文本语义；② 掩码引导的交叉注意力模块，通过掩码引导的注意力机制实现跨模态特征的有效融合；③ 文本驱动的注意力融合模块，利用提取到的文本语义，通过门控注意力机制进一步优化融合过程。

图2 TeSG网络架构

Figure 2 The overall framework of TeSG



(a) 语义信息生成模块的详细流程

(a) The process of the semantic information generator module

(b) 不同场景下文本描述与对应掩码语义的可视化示例

(b) Example visualizations of text descriptions and corresponding mask semantics in different scenarios

图3 语义信息的可视化与生成过程

Figure 3 The visualization and generation process of semantic information

2.2.2 掩码语义生成

在IVF任务中，一个长期存在的挑战是如何在融合过程中管理信息冗余并防止关键细节丢失。为解决这一问题，本文引入掩码语义，采用预训练Stable Diffusion[60]作为噪声估计器，通过建模文本语义变化对图像局部区域的影响，来强调红外与可见光模态之间的互补特征。以可见光图像$\boldsymbol{I}_v$为例，首先，将输入可见光图像$\boldsymbol{I}_v$编码为初始潜变量，并在DDIM（Denoising Diffusion Implicit Model）反演框架下，于固定的中间时间步向潜变量注入高斯噪声，以获得噪声扰动后的潜表示。随后，在相同的噪声与相同的时间步下，分别引入参考文本描述$\boldsymbol{T}$及其对照文本$\hat{\boldsymbol{T}}$，由扩散模型预测对应的噪声分量。不同文本条件将导致与语义相关区域产生显著差异的噪声估计，因此，通过计算两种文本条件下噪声预测结果的差异，并对其进行归一化与二值化处理，最终得到掩码，其过程可表示为

$$\boldsymbol{M}_v=\Delta\left(\mathrm{D}_\theta(\boldsymbol{I}_v,\boldsymbol{T})-\mathrm{D}_\theta(\boldsymbol{I}_v,\hat{\boldsymbol{T}})\right) \tag{2}$$

其中，D_θ表示扩散模型；Δ表示对计算得到的噪声差进行归一化和二值化。为进一步缓解单一模态信息提取不足的问题，将该方法同时应用于可见光和红外图像。具体而言，分别为可见光图像$\boldsymbol{I}_v$和红外图像$\boldsymbol{I}_i$生成掩码$\boldsymbol{M}_v$与$\boldsymbol{M}_i$，然后取并集得到最终的掩码语义$\boldsymbol{M}$：

$$\boldsymbol{M}=\boldsymbol{M}_v\cup\boldsymbol{M}_i \tag{3}$$

2.2.3 文本语义生成

本文采用Stable Diffusion的文本编码器结构(即基于CLIP(Contrastive Language-Image Pretraining)的文本编码器)将文本描述$\boldsymbol{T}$转换为嵌入向量$\boldsymbol{F}_{\mathrm{t}}\in\mathbf{R}^{\mathrm{B}\times 77\times 768}$。随后经线性映射将$\boldsymbol{F}_{\mathrm{t}}$投影至与视觉特征相匹配的维度,使其可直接用于TDAF模块中的文本引导的视觉重建模块。文本语义向量在后续模块中作为跨模态交互的语义先验,用于增强跨模态语义一致性。

综上所述,语义信息生成模块同时生成掩码级和文本级语义信息,以指导后续的图像融合过程。

2.3 掩码引导交叉注意力模块

MGCA模块首先利用掩码语义将图像分解为关键目标前景图像和背景图像。给定可见光图像$\boldsymbol{I}_{\mathrm{v}}$、红外图像$\boldsymbol{I}_{\mathrm{i}}$以及对应的掩码语义$\boldsymbol{M}$,该模块将前景语义$\boldsymbol{M}$与背景语义$\bar{\boldsymbol{M}}$应用于每个输入图像,其中,$\bar{\boldsymbol{M}}$定义为前景掩码的补集区域。由此得到四幅不同的掩码图像,能够同时捕捉每个模态的前景与背景信息。

随后,这些掩码图像被输入编码器,以提取初始特征表示,分别记为$\boldsymbol{F}_{\mathrm{v}}^{\mathrm{m}}$、$\boldsymbol{F}_{\mathrm{v}}^{\bar{\mathrm{m}}}$、$\boldsymbol{F}_{\mathrm{i}}^{\mathrm{m}}$和$\boldsymbol{F}_{\mathrm{i}}^{\bar{\mathrm{m}}}$,其中上标m与$\bar{\mathrm{m}}$分别表示前景特征和背景特征。

接下来,利用交叉注意力机制对可见光与红外图像的前景和背景特征进行特征重构。红外图像的前景和背景特征$\boldsymbol{F}_{\mathrm{i}}^{\mathrm{m}}$与$\boldsymbol{F}_{\mathrm{i}}^{\bar{\mathrm{m}}}$用于重构可见光图像的前景和背景特征$\boldsymbol{F}_{\mathrm{v}}^{\mathrm{m}}$与$\boldsymbol{F}_{\mathrm{v}}^{\bar{\mathrm{m}}}$,得到$\boldsymbol{F}_{\mathrm{v}'}^{\mathrm{m}}$与$\boldsymbol{F}_{\mathrm{v}'}^{\bar{\mathrm{m}}}$:

$$\begin{aligned}\boldsymbol{F}_{\mathrm{v}'}^{\mathrm{m}}&=\mathrm{CA}(\boldsymbol{F}_{\mathrm{v}}^{\mathrm{m}},\boldsymbol{F}_{\mathrm{i}}^{\mathrm{m}})=\mathrm{CA}(\boldsymbol{Q}_{\mathrm{v}}^{\mathrm{m}},\boldsymbol{K}_{\mathrm{i}}^{\mathrm{m}},\boldsymbol{V}_{\mathrm{i}}^{\mathrm{m}})\\\boldsymbol{F}_{\mathrm{v}'}^{\bar{\mathrm{m}}}&=\mathrm{CA}(\boldsymbol{F}_{\mathrm{v}}^{\bar{\mathrm{m}}},\boldsymbol{F}_{\mathrm{i}}^{\bar{\mathrm{m}}})=\mathrm{CA}(\boldsymbol{Q}_{\mathrm{v}}^{\bar{\mathrm{m}}},\boldsymbol{K}_{\mathrm{i}}^{\bar{\mathrm{m}}},\boldsymbol{V}_{\mathrm{i}}^{\bar{\mathrm{m}}})\end{aligned}\tag{4}$$

反之,可见光图像的前景和背景特征$\boldsymbol{F}_{\mathrm{v}}^{\mathrm{m}}$与$\boldsymbol{F}_{\mathrm{v}}^{\bar{\mathrm{m}}}$用于重构红外图像的前景和背景特征$\boldsymbol{F}_{\mathrm{i}}^{\mathrm{m}}$与$\boldsymbol{F}_{\mathrm{i}}^{\bar{\mathrm{m}}}$,得到$\boldsymbol{F}_{\mathrm{i}'}^{\mathrm{m}}$与$\boldsymbol{F}_{\mathrm{i}'}^{\bar{\mathrm{m}}}$:

$$\begin{aligned}\boldsymbol{F}_{\mathrm{i}'}^{\mathrm{m}}&=\mathrm{CA}(\boldsymbol{F}_{\mathrm{i}}^{\mathrm{m}},\boldsymbol{F}_{\mathrm{v}}^{\mathrm{m}})=\mathrm{CA}(\boldsymbol{Q}_{\mathrm{i}}^{\mathrm{m}},\boldsymbol{K}_{\mathrm{v}}^{\mathrm{m}},\boldsymbol{V}_{\mathrm{v}}^{\mathrm{m}})\\\boldsymbol{F}_{\mathrm{i}'}^{\bar{\mathrm{m}}}&=\mathrm{CA}(\boldsymbol{F}_{\mathrm{i}}^{\bar{\mathrm{m}}},\boldsymbol{F}_{\mathrm{v}}^{\bar{\mathrm{m}}})=\mathrm{CA}(\boldsymbol{Q}_{\mathrm{i}}^{\bar{\mathrm{m}}},\boldsymbol{K}_{\mathrm{v}}^{\bar{\mathrm{m}}},\boldsymbol{V}_{\mathrm{v}}^{\bar{\mathrm{m}}})\end{aligned}\tag{5}$$

其中,CA表示交叉注意力机制。

最后,将对应特征逐元素相加,得到完整的重构可见光与红外特征:

$$\boldsymbol{F}_{\mathrm{v}}'=\boldsymbol{F}_{\mathrm{v}'}^{\mathrm{m}}+\boldsymbol{F}_{\mathrm{v}'}^{\bar{\mathrm{m}}},\boldsymbol{F}_{\mathrm{i}}'=\boldsymbol{F}_{\mathrm{i}'}^{\bar{\mathrm{m}}}+\boldsymbol{F}_{\mathrm{i}'}^{\bar{\mathrm{m}}}\tag{6}$$

2.4 文本驱动注意力融合模块

2.4.1 文本引导的视觉重建

给定重构后的可见光和红外特征$\boldsymbol{F}_{\mathrm{v}}'$与$\boldsymbol{F}_{\mathrm{i}}'$,以及文本语义$\boldsymbol{F}_{\mathrm{t}}$,首先将$\boldsymbol{F}_{\mathrm{v}}'$与$\boldsymbol{F}_{\mathrm{i}}'$拼接,生成多模态视觉特征。随后,通过交叉注意力机制利用文本语义$\boldsymbol{F}_{\mathrm{t}}$对多模态视觉特征进行重建,实现文本与视觉特征的有效交互。该过程可形式化表示为

$$\boldsymbol{Q}_{\mathrm{r}}=\boldsymbol{W}_{\mathrm{q}}\left(\mathrm{Cat}(\boldsymbol{F}_{\mathrm{v}}',\boldsymbol{F}_{\mathrm{i}}')\right)\tag{7}$$

$$\boldsymbol{K}_{\mathrm{t}}=\boldsymbol{W}_{\mathrm{k}}(\boldsymbol{F}_{\mathrm{t}}),\boldsymbol{V}_{\mathrm{t}}=\boldsymbol{W}_{\mathrm{v}}(\boldsymbol{F}_{\mathrm{t}})\tag{8}$$

$$\boldsymbol{F}_{\mathrm{r}}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{r}}\boldsymbol{K}_{\mathrm{t}}^{\mathrm{T}}}{\sqrt{\boldsymbol{d}_{\mathrm{k}}}}\right)\boldsymbol{V}_{\mathrm{t}}\tag{9}$$

其中,$\boldsymbol{Q}_{\mathrm{r}}$、$\boldsymbol{K}_{\mathrm{t}}$和$\boldsymbol{V}_{\mathrm{t}}$分别表示查询矩阵、键矩阵和值矩阵;$\boldsymbol{d}_{\mathrm{k}}$表示键的维度;Cat(·)表示特征拼接。通过上述过程,TDAF模块生成了文本引导的视觉特征$\boldsymbol{F}_{\mathrm{r}}$,实现文本语义信息与视觉特征的深度融合。

2.4.2 门控注意力融合模块

为了充分利用红外与可见光特征的互补特性,该模块引入门控机制,对两种模态进行动态加权融合。首先,将重构后的可见光与红外特征$\boldsymbol{F}_{\mathrm{v}}'$和$\boldsymbol{F}_{\mathrm{i}}'$分别与初始图像特征$\boldsymbol{F}_{\mathrm{v}}$与$\boldsymbol{F}_{\mathrm{i}}$相加,得到融合后的特征,再通过卷积层生成门控权重$\boldsymbol{G}_{\mathrm{v}}$与$\boldsymbol{G}_{\mathrm{i}}$,其计算公式为

$$\begin{aligned}\boldsymbol{G}_{\mathrm{v}}&=\sigma\left(\boldsymbol{W}_{\mathrm{v}}\cdot(\boldsymbol{F}_{\mathrm{v}}+\boldsymbol{F}_{\mathrm{v}}')\right)\\\boldsymbol{G}_{\mathrm{i}}&=\sigma\left(\boldsymbol{W}_{\mathrm{i}}\cdot(\boldsymbol{F}_{\mathrm{i}}+\boldsymbol{F}_{\mathrm{i}}')\right)\end{aligned}\tag{10}$$

其中,σ表示S型(Sigmoid)激活函数;$\boldsymbol{W}_{\mathrm{v}}$与$\boldsymbol{W}_{\mathrm{i}}$分别表示可见光与红外特征卷积层的权重矩阵。门控权重用于控制各模态对最终融合结果的贡献。

随后,从文本引导的视觉特征$\boldsymbol{F}_{\mathrm{r}}$中计算空间注意力权重$\boldsymbol{\alpha}$,并通过多层感知机(MultiLayer Perceptron,MLP)对特征进行精炼:

$$\boldsymbol{\alpha}=\mathrm{SA}(\boldsymbol{F}_{\mathrm{r}})=\sigma\left(\Phi_{\mathrm{SA}}(\boldsymbol{F}_{\mathrm{r}})\right)\tag{11}$$

其中,Φ_{SA}是一个由两层卷积块和ReLU(Rectified Linear Unit)激活组成的MLP融合函数。

最后,将空间注意力权重$\boldsymbol{\alpha}$与门控权重$\boldsymbol{G}_{\mathrm{v}}$和$\boldsymbol{G}_{\mathrm{i}}$联合应用,对红外与可见光特征进行自适应加权融合。该融合过程表示为

$$\begin{aligned}\boldsymbol{F}=&\boldsymbol{\alpha}\cdot(1+\boldsymbol{G}_{\mathrm{v}})\cdot(\boldsymbol{F}_{\mathrm{v}}+\boldsymbol{F}_{\mathrm{v}}')\\&+(1-\boldsymbol{\alpha})\cdot(1+\boldsymbol{G}_{\mathrm{i}})\cdot(\boldsymbol{F}_{\mathrm{i}}+\boldsymbol{F}_{\mathrm{i}}')\end{aligned}\tag{12}$$

其中,$\boldsymbol{F}$表示最终融合特征,随后通过解码器生成高质量融合图像$\boldsymbol{I}_{\mathrm{f}}$。

2.5 目标函数

对于红外与可见光图像融合任务,本文目标是在融合过程中同时保留红外图像的高频信息和可见光图像的纹理细节。为此,本文采用了四种广泛用于图像融合任务的通用损失函数:结构相似性损失、梯度损失、强度损失以及颜色损失。

首先,结构相似性损失用于保证融合图像与输入图像在结构上的一致性,从而保留输入图像的纹理细节和全局结构信息,增强图像的结构感知能力,其定义如下:

$$\begin{aligned}L_{\mathrm{SSIM}}=&\delta_{\mathrm{v}}\left(1-\mathrm{SSIM}(\boldsymbol{I}_{\mathrm{f}},\boldsymbol{I}_{\mathrm{v}})\right)\\&+\delta_{\mathrm{i}}\left(1-\mathrm{SSIM}(\boldsymbol{I}_{\mathrm{f}},\boldsymbol{I}_{\mathrm{i}})\right)\end{aligned}\tag{13}$$

其中,SSIM表示结构相似性指数[61];δ_{v}与δ_{i}分别设为1和0.5。

梯度损失强调边缘信息的保留，确保融合图像的边缘细节与输入图像一致，有效防止融合过程中边缘信息的丢失，其定义为

$$L_{\text{grad}}=\frac{1}{\text{HW}}\left\|\nabla \boldsymbol{I}_{\text{f}}-\max(\nabla \boldsymbol{I}_{\text{v}},\nabla \boldsymbol{I}_{\text{i}})\right\|_1 \tag{14}$$

其中，∇ 表示 Sobel 梯度算子；max 为取最大值操作。

强度损失主要用于增强融合图像中显著区域的强度信息，其定义为

$$L_{\text{int}}=\frac{1}{\text{HW}}\left\|\boldsymbol{I}_{\text{f}}-\max(\boldsymbol{I}_{\text{v}},\boldsymbol{I}_{\text{i}})\right\|_1 \tag{15}$$

颜色损失用于约束融合图像的颜色一致性。具体做法是将 RGB 图像转换为 YCbCr 颜色空间，比较融合图像与可见光图像在 Cb 和 Cr 通道上的差异，其定义为

$$\begin{aligned}L_{\text{color}}=&\left\|F_{\text{Cb}}(\boldsymbol{I}_{\text{f}})-F_{\text{Cb}}(\boldsymbol{I}_{\text{v}})\right\|_1\\&+\left\|F_{\text{Cr}}(\boldsymbol{I}_{\text{f}})-F_{\text{Cr}}(\boldsymbol{I}_{\text{v}})\right\|_1\end{aligned} \tag{16}$$

最终，总损失函数定义为

$$L_{\text{total}}=\lambda_1 L_{\text{SSIM}}+\lambda_2 L_{\text{grad}}+\lambda_3 L_{\text{int}}+\lambda_4 L_{\text{color}} \tag{17}$$

其中，$\lambda_i(i=1,2,3,4)$ 为平衡系数，用于调节各个损失项的相对贡献，以生成更高质量的融合结果。

3 实验

3.1 实验设置

3.1.1 实现细节

本文的实验基于 PyTorch 实现，并在拥有 4 张 24 GB NVIDIA GeForce RTX 3090 GPU 的服务器上进行训练。训练过程采用 AdamW 优化器，初始学习率设为 0.000 1，共训练 140 个轮次（epoch），批次大小为 8。训练样本随机裁剪为 96×96 的图像。模型架构包含两个编码器和一个解码器，每个模块由四个 Transformer 块组成。损失函数中各项权重参数 $\lambda_i(i=1,2,3,4)$（详见式（17））设置为 {0.4, 9, 1, 10}，以平衡各损失的贡献。关键词 V^*（详见 2.2 节）的获取基于人工定义的策略：首先使用 BLIP 生成文本描述，然后从中提取与检测和分割任务相关的名词作为关键目标类别。关键词列表包括“{car, people, person, woman, man, truck, van, traffic}”。

3.1.2 数据集介绍

实验使用三个公开可用的数据集：MSRS（Multi-Spectral Road Scenarios）[62]、RoadScene[22] 和 LLVIP（Low-Light Visible-Infrared Paired dataset）[63]。模型在 MSRS 数据集上进行训练，并在 RoadScene 与 LLVIP 数据集上进行评估。MSRS 数据集包含 1 083 对红外可见光图像训练图像对，以及 361 对测试图像对。为了评估模型在不同环境下的泛化能力，实验从 RoadScene 数据集中选取 221 对图像对、从 LLVIP 数据集中随机选取 347 对图像对构成测试集。

3.1.3 对比方法

本文将所提方法与十一种最先进的（State Of The Art，SOTA）方法进行比较，包括：TarDAL（Target-aware Dual Adversarial Learning）[28]、ReCoNet（Recurrent Correction Network）[64]、LRRNet（Low-Rank Representation-learning guided fusion Network）[65]、MetaFusion（image Fusion via Meta-feature embedding）[66]、CDDFuse（Correlation-Driven feature Decomposition Fusion）[67]、DDFM[47]、TextIF[31]、TextFusion[34]、MMDRFuse（Mini-Model with Dynamic Refresh for image Fusion）[68]、TeRF（Text-driven and Region-aware Flexible image fusion）[33] 以及 MulFS-CAP（Multimodal Fusion Supervised Cross-modality Alignment Perception）[69]。

3.1.4 评估指标

为了定量评估所提 TeSG 方法的性能，本文采用五个广泛使用的评价指标，从多个维度衡量融合图像质量，包括：信息熵（ENtropy，EN）[1]，用于衡量融合图像的信息量；标准差（Standard Deviation，SD），反映图像的信息丰富度和对比度；差异相关性总和（Sum of the Correlations of Differences，SCD）[70]，用于评估融合图像与源图像的相似性；视觉保真度（Visual Information Fidelity，VIF）[71]，衡量融合图像与源图像之间共享的视觉信息量，反映其与人类视觉感知的一致性；以及基于边缘信息的指标 $Q^{AB/F}$[1]，用于评价源图像边缘信息的保留情况。在上述指标中，数值越高表示融合方法的性能越优。

3.2 融合对比实验和分析

3.2.1 定量实验结果和分析

本文在 MSRS[62]、RoadScene[22] 和 LLVIP[63] 数据集上，对所提方法与十一种最先进的融合方法进行了定量比较，实验结果如表 1 所示。

从表 1 实验结果可见，所提 TeSG 方法在 MSRS 和 LLVIP 数据集上的结果均优于现有方法，且在 MSRS 数据集上所有指标均取得最优结果。其中，TeSG 在 SD、SCD 与 VIF 三项指标上的优异表现，充分证明该方法在保留来自红外和可见光两种模态的结构与纹理细节方面具有较强能力，从而显著提升了融合图像的颜色保留和视觉细节。同时，较高的 $Q^{AB/F}$ 分数进一步说明 TeSG 能够有效保留两种模态的互补特征，显著增强了融合图像的对比度与轮廓清晰度。此外，TeSG 在 EN 指标上位列第二，进一步验证了其在多源信息融合方面的有效性。对于 RoadScene 数据集，TeSG 的性能提升幅度相对有限，这一现象主要归因于该数据集部分输入图像存在质量退化，从而导致语义信息的精度下降，客观上限制了融合效果。即便如此，TeSG 在 SCD、VIF 以及 $Q^{AB/F}$ 等指标上仍保持领

表1　在MSRS、RoadScene和LLVIP数据集上,本文的方法与十一种最新SOTA方法的定量比较

Table 1　Quantitative comparison of our method against eleven SOTA methods on the MSRS, LLVIP, and RoadScene datasets

方法	MSRS					RoadScene					LLVIP				
	EN	SD	SCD	VIF	$Q^{AB/F}$	EN	SD	SCD	VIF	$Q^{AB/F}$	EN	SD	SCD	VIF	$Q^{AB/F}$
TarDAL	5.322	23.346	0.697	0.406	0.172	7.259	47.897	1.435	0.546	0.396	6.322	37.047	1.030	0.526	0.211
ReCoNet	4.234	41.714	1.262	0.490	0.404	7.054	41.275	1.536	0.545	0.380	5.800	46.898	1.461	0.559	0.405
LRRNet	6.192	31.758	0.791	0.541	0.454	7.132	42.436	1.569	0.494	0.352	6.381	29.189	0.869	0.547	0.420
MetaFusion	6.357	39.133	1.502	0.686	0.464	7.391	50.749	1.549	0.527	0.402	7.042	46.631	1.327	0.623	0.293
CDDFuse	6.701	43.374	1.621	1.051	0.693	**7.453**	**56.322**	1.712	0.624	0.483	7.342	49.971	1.580	0.88	0.642
DDFM	6.175	28.922	1.449	0.743	0.474	7.250	43.188	1.713	0.571	0.407	7.089	40.374	1.417	0.648	0.258
TextIF	6.665	43.190	1.689	1.049	0.716	7.319	49.871	1.522	0.686	**0.591**	**7.428**	50.250	1.444	1.033	0.747
TextFusion	6.029	38.022	1.432	0.722	0.521	7.004	39.752	1.537	0.677	0.397	6.552	38.248	1.278	0.689	0.493
MMDRFuse	6.475	37.323	1.532	0.851	0.591	6.505	27.631	1.102	0.545	0.320	7.245	44.390	1.475	0.812	0.586
TeRF	6.505	41.445	1.610	0.844	0.611	7.091	44.183	1.355	0.655	0.573	7.359	49.553	1.469	0.837	0.660
MulFS-CAP	5.898	31.765	0.937	0.269	0.388	7.138	39.857	1.383	0.507	0.543	6.846	35.290	0.974	0.414	0.428
Ours	**6.712**	**43.472**	**1.719**	**1.080**	**0.733**	7.180	53.678	**1.741**	**0.690**	0.577	7.396	**52.553**	**1.614**	**1.059**	**0.763**

注:加粗表示最优结果,下划线表示次优结果。

先,展现了其在次优条件下的稳健性。总之,上述定量结果充分体现了TeSG在不同数据集上的泛化能力和融合质量,验证了其在红外与可见光图像融合中对细节信息和语义一致性的有效保留优势。

3.2.2　可视化定性实验结果和分析

图4展示了在三个数据集上的可视化结果,进一步揭示了现有SOTA方法的局限性。TarDAL、ReCoNet、LRRNet和MulFS-CAP在融合过程中存在显著的多模态信息保留不足问题(例如,MSRS与LLVIP数据集的红框标注区域中,行人的热特征未得到清晰呈现)。此外,尽管LRRNet在一定程度上保留了局部纹理细节,TarDAL、ReCoNet和MulFS-CAP方法对细节保留效果较差(例如,RoadScene数据集红框标注的地面文字区域出现明显模糊)。TextFusion和MetaFusion虽在细节保留上优于前述方法,但仍存在局限:TextFusion无法清晰突出热目标(例如红框标注的热目标不够明显),MetaFusion则存在明显的色彩失真问题,且两种方法的融合结果均存在亮度偏低的现象。MMDRFuse和DDFM的融合结果对比度较低,且整体色彩接近单一模态图像,未能有效融合多模态信息,同时对热目标的突出效果不佳。CDDFuse、TextIF和TeRF生成的融合图像质量相对较高,能够较好地保留红外模态中的热目标信息,但在可见光纹理细节的保留上存在明显缺陷;TextIF还存在色彩失真问题(例如,MSRS数据集中蓝框标注的背景区域)。相比之下,所提TeSG方法通过双层语义引导和门控机制,显著增强了前景热目标,同时更好地保留了可见光纹理细节,更自然地保留了源图像的亮度信息。同时,该方法在低光照条件下仍能呈现较高对比度、良好的轮廓信息和场景细节,实现红外与可见光信息的均衡融合,可稳定生成适用于昼夜不同光照条件的高质量融合结果。

3.3　下游任务对比实验和分析

3.3.1　目标检测实验结果和分析

为确保定性与定量比较的公平性,本实验采用YOLOv8(You Only Look Once,https://github.com/ultralytics/ultralytics)作为目标检测网络。结果如图5和表2所示。

从图5的可视化检测结果可见,所提TeSG方法能够以高置信度准确检测场景中的所有目标,显著优于现有融合方法,在检测准确性上具有明显优势。对比其他方法,MulFS-CAP将“行人(person)”及“车辆(car)”错误分类为“自行车(bike)”及“行人(person)”,而TarDAL和TextFusion在识别“person”类别时存在问题,检测框存在重叠,凸显了它们在保留关键热目标方面的局限性。此外,ReCoNet、LRRNet等方法由于无法捕捉源图像中的关键细节,存在漏检问题,直接导致检测性能下降。相比之下,TeSG能够有效融合红外热目标信息与可见光纹理细节,充分保留两种模态的关键特征。同时依托文本语义引导增强了对场景目标的理解,即使在具有挑战性的复杂场景下,TeSG仍能提供清晰的目标轮廓与丰富的上下文信息,从而提升了检测准确性和稳定性。在定量评估方面,本文采用了五项指标,包括精度(Precision)、召回率(Recall)以及不同阈值下的平均精度均值(mean Average Precision,mAP)。如表2所示,TeSG在mAP指标上取得最高分,显著优于其他方法。即使在更严格的评估标准下(如mAP@0.75和mAP@0.5:0.95),TeSG仍表现出优越性能。上述定量结果表明,TeSG

图4 本文方法与十一种最新SOTA方法的定性对比(从上到下依次来自MSRS、RoadScene和LLVIP数据集)

Figure 4 Qualitative comparison of our method against eleven SOTA methods (The images, arranged from top to bottom, are sourced from the MSRS, RoadScene and LLVIP datasets)

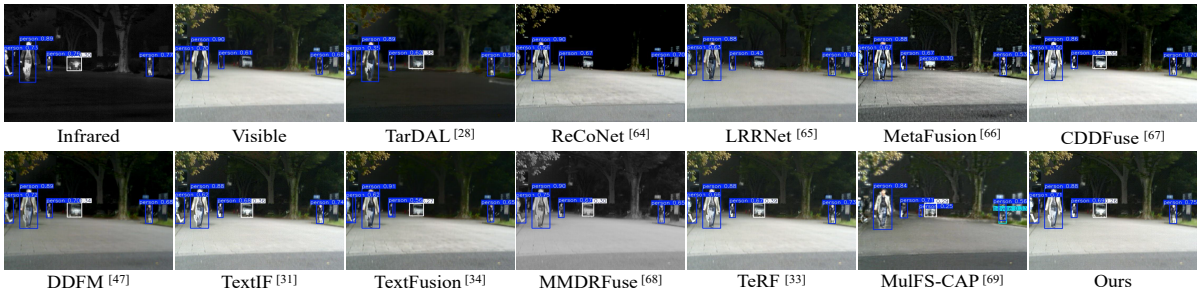


图5 各方法在MSRS数据集上目标检测的定性比较

Figure 5 Qualitative comparison of object detection performance on the MSRS dataset

不仅能提升目标检测的整体性能,还能在更严格的检测标准下保持出色的准确性。

3.3.2 语义分割实验结果和分析

为了进一步评估所提TeSG方法在语义分割任务中的表现,本文选择DeeplabV3+[72]作为语义分割网络,并在MSRS数据集上进行训练。该数据集包含九个类别:背景(Background)、汽车(Car)、行人(Person)、自行车(Bike)、弯道(Curve)、停车标识(Car stop)、防护栏(Guardrail)、彩色路锥(Color cone)和减速带(Bump)。在训练过程中,采用交叉熵损失作为优化目标,初始学习率设为0.01。总训练轮次(epoch)为370,批次大小为4。为确保各融合方法间的公平比较,本文对各方法生成的融合图像进行了定性和定量评估,结果分别如图6与表3所示。

图6中的可视化结果表明,所提TeSG能够准确分割场景中的所有语义目标,表现出优异的分割性

表2　在MSRS数据集上目标检测性能的定量比较

Table 2　Quantitative comparison of object detection performance on the MSRS dataset

方法	mAP			Precision	Recall
	@0.50	@0.75	@0.50:0.95		
TarDAL	0.850	0.597	0.553	0.891	0.744
ReCoNet	0.715	0.526	0.466	0.914	0.622
LRRNet	0.921	0.735	0.674	**0.971**	0.852
MetaFusion	0.905	0.728	0.647	0.905	0.855
CDDFuse	0.942	0.778	0.684	0.895	**0.881**
DDFM	0.923	0.818	0.699	0.940	0.835
TextIF	0.939	0.808	0.709	0.947	0.860
TextFusion	0.861	0.636	0.598	0.888	0.748
MMDRFuse	0.914	0.782	0.688	0.915	0.820
TeRF	0.938	0.808	0.711	0.899	0.885
MulFS-CAP	0.724	0.567	0.465	0.774	0.644
Ours	**0.945**	**0.820**	**0.717**	0.948	0.879

注:加粗表示最优结果,下划线表示次优结果。

能。相比之下,TarDAL、TeRF和MulFS-CAP对场景中远距离目标的分割能力不足(如图中红框标注的行人(person)和自行车(bike)),未能准确分割对象。ReCoNet、LRRNet、MetaFusion、CDDFuse、TextFusion、MMDRFuse、TeRF和MulFS-CAP等方法在融合图像的减速带(Bump)区域(如图中绿框标注)未能保留足够的语义信息,导致模型无法准确识别该类别,出现分割不完整等问题。DDFM和TextIF虽然能够成功识别并分割所有目标,但在边缘分割精度上存在明显缺陷,目标边界存在模糊现象。相比之下,TeSG在边缘梯度信息保留方面表现突出。通过增强融合图像的边缘细节,使融合图像的边缘特征与源图像高度一致,显著增强了目标轮廓的对比度和清晰度,最终提升了其在下游语义分割任务中的表现。

在表3的定量分析中,TeSG在平均交并比(mean Intersection over Union,mIoU)指标上取得最佳成绩且在多个类别上均优于其他对比方法,充分体现其在图像分割任务中的综合优势。综上所述,所提TeSG方

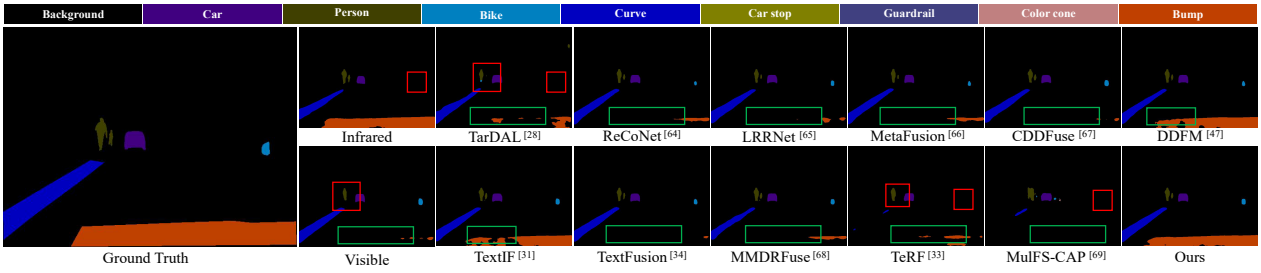


图6　各方法在MSRS数据集上语义分割的定性比较

Figure 6　Qualitative comparison of semantic segmentation performance on the MSRS dataset

表3　在MSRS数据集上语义分割性能的定量比较　　单位:%

Table 3　Quantitative comparison of semantic segmentation performance on the MSRS dataset　　unit: %

方法	IoU									mIoU
	Background	Car	Person	Bike	Curve	Car stop	Guardrail	Color cone	Bump	
Infrared	96.57	67.30	70.93	44.85	34.75	22.98	0.00	33.97	46.61	46.44
Visible	97.91	88.03	42.00	69.55	52.39	70.30	76.72	60.56	66.05	69.28
TarDAL	97.70	85.05	57.67	64.41	39.52	63.70	34.32	54.66	55.98	61.44
ReCoNet	97.55	83.95	56.54	58.69	35.54	59.52	69.24	45.33	46.03	61.38
LRRNet	98.29	89.60	68.34	69.06	49.64	71.14	76.51	61.17	64.82	72.06
MetaFusion	98.15	87.92	62.25	69.18	54.51	68.18	74.61	60.09	53.43	70.15
CDDFuse	98.47	90.10	74.13	72.00	59.85	71.72	76.16	61.58	65.54	74.40
DDFM	98.39	89.85	73.83	70.73	56.30	70.88	72.42	60.50	63.82	72.97
TextIF	98.49	90.16	74.60	72.12	57.89	71.85	**76.98**	62.59	**70.38**	75.01
TextFusion	98.27	89.17	65.97	70.18	54.92	71.04	75.04	59.61	64.28	72.05
MMDRFuse	98.43	89.97	73.83	71.50	58.05	72.04	74.96	61.87	64.14	73.87
TeRF	97.65	84.79	66.21	66.43	36.15	59.27	60.37	52.08	51.44	63.82
MulFS-CAP	96.67	73.45	45.46	47.83	21.07	51.51	55.49	40.10	34.57	51.79
Ours	**98.53**	**90.55**	**74.86**	**72.27**	**60.70**	**72.42**	76.60	**62.62**	69.86	**75.38**

注:加粗表示最优结果,下划线表示次优结果。

法能够有效突出前景目标，同时保留关键纹理细节，在融合质量和语义一致性方面均具有显著优势。

3.4 消融实验

本文提出的TeSG通过掩码级与语义级的双重引导，有效地指导了图像融合过程，使模型能够更加关注源图像中目标区域的融合，最终生成语义一致性更高的融合图像。为了评估各模块对模型整体性能的贡献，本节在LLVIP[63]数据集上设计并开展消融实验，定量结果如表4所示。

表4 模块消融实验结果

Table 4 Quantitative ablation results

模型设置	EN	SD	SCD	VIF	$Q^{AB/F}$
(a) w/o Mask	7.366	51.690	1.576	**1.065**	**0.765**
(b) w/o MGCA	7.359	51.057	1.523	1.063	0.765
(c) w/o TIVR	7.382	52.201	1.608	1.063	0.764
(d) w/o GAF	7.383	51.956	1.572	1.065	0.764
(e) Ours	**7.396**	**52.553**	**1.614**	1.059	0.763

注：加粗表示最优结果。

（1）掩码语义的有效性。为验证掩码语义在融合过程中的作用，本文设计了一个掩码引导交叉注意力模块的变体，该变体不再通过掩码语义将输入图像分解为关键目标前景和背景区域，而是直接从两种模态中提取特征以进行交叉注意力操作。如表4（a）所示，移除掩码语义后，模型的多数评价指标出现下降，尽管VIF有轻微提升，但该提升幅度极小，无法弥补整体性能的损失。这表明掩码语义在TeSG的融合过程中起着关键作用，尤其是在引导融合、保持图像对比度和边缘信息方面。SD和SCD指标的下降进一步表明，掩码语义有效减少了融合过程中的信息损失，提高了融合质量。

（2）掩码引导交叉注意力（MGCA）模块。为了评估MGCA模块的有效性，本文移除了整个模块，既不使用掩码语义，也不对红外和可见光特征进行跨模态注意力操作。如表4（b）所示，这导致多数评价指标出现显著下降，充分凸显了该模块的必要性。这表明MGCA模块对模型网络有着至关重要的贡献，在促进两种模态特征间的信息交互、减少信息损失并提升融合性能方面发挥了关键作用。

（3）文本引导视觉重构（Text-Informed Visual Reconstruction，TIVR）模块。为了验证文本语义的有效性，本文移除了TIVR模块，不使用文本特征引导，直接对拼接后的图像特征进行融合。如表4（c）所示，在缺乏文本语义指导时，融合图像的丰富信息和纹理细节呈现明显下降趋势。这表明在没有文本引导的情况下，模型难以有效识别并融合图像中的关键信息，导致融合结果存在细节缺失问题，尤其在纹理保持和语义一致性方面表现不佳。该实验充分证明文本语义引导在提升融合图像质量中的重要性。

（4）门控注意力融合（Gated Attentional Fusion，GAF）模块。为探究动态门控机制对融合性能的影响，本文移除了GAF模块，直接对初始融合特征进行解码并生成融合图像。如表4（d）所示，模型的多数评价指标显著下降，强调了GAF模块的重要性。GAF模块能够动态调整不同模态特征的融合权重，既有效保留两种源图像的丰富信息，又维持融合图像与源图像的相似性，实现融合质量的显著提升。

3.5 超参数敏感性分析

为了系统评估TeSG方法中超参数（如λ_1、λ_2和λ_4，其中λ_3默认设为1）的稳健性，本文采用固定其中两个参数，仅改变剩余一个参数的方式，观察其在指定取值范围内变化时模型性能指标的波动情况。图7定量展示了各参数变化下指标的稳定性，所有指标在测试范围内波动均控制在1%以内。

对于$\lambda_1 \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$（图7第一行），当$\lambda_1 > 0.4$时，融合结果的所有评价指标均低于$\lambda_1 = 0.4$的表现；当$\lambda_1 < 0.4$时，仅有EN和SCD指标略有提升，其余指标均下降。因此，综合各指标的整体表现，本文将λ_1设为0.4，以获得最佳融合效果。进一步观察发现，除$Q^{AB/F}$和VIF外，其余指标在测试范围内的波动均低于0.5%，证明该参数在测试范围对模型性能影响较小，参数鲁棒性良好。

对于$\lambda_2 \in \{7, 8, 9, 10, 11\}$（图7第二行），实验结果显示当$\lambda_2 = 9$时，EN、SD和$Q^{AB/F}$指标达到最大值。进一步比较发现，虽然其他$\lambda_2$值下部分指标略有波动，但均未超过$\lambda_2 = 9$时的性能表现。并且当$\lambda_2 > 9$时，所有评价指标均低于$\lambda_2 = 9$的表现。因此，综合各项性能指标，本文选择将$\lambda_2$的最优值确定为9。此外，几乎所有指标在测试范围内的波动均低于0.5%，进一步说明该参数的稳定性。

如图7第三行所示，λ_4的测试取值范围为$\{6, 8, 10, 12, 14\}$。结果显示，当$\lambda_4 > 10$时，所有评价指标均下降，明显低于$\lambda_4 = 10$时的表现；当$\lambda_4 < 10$时，虽然部分指标（如SD和SCD）相较于$\lambda_4 = 10$有波动，但整体性能仍未超过$\lambda_4 = 10$。例如，当$\lambda_4 = 6$时，尽管SCD指标相对较高，但VIF和$Q^{AB/F}$均下降。基于上述综合分析，本文将λ_4设为10，以有效提升融合图像的质量。综上所述，TeSG在超参数的测试范围内始终保持稳定波动，证明了其对超参数变化的强鲁棒性。

3.6 局限性分析

本文仅将可见光图像输入视觉语言模型，这可能导致红外图像中的关键信息被忽略。然而，BLIP模型的预训练数据为大规模自然图像文本对，确保了在

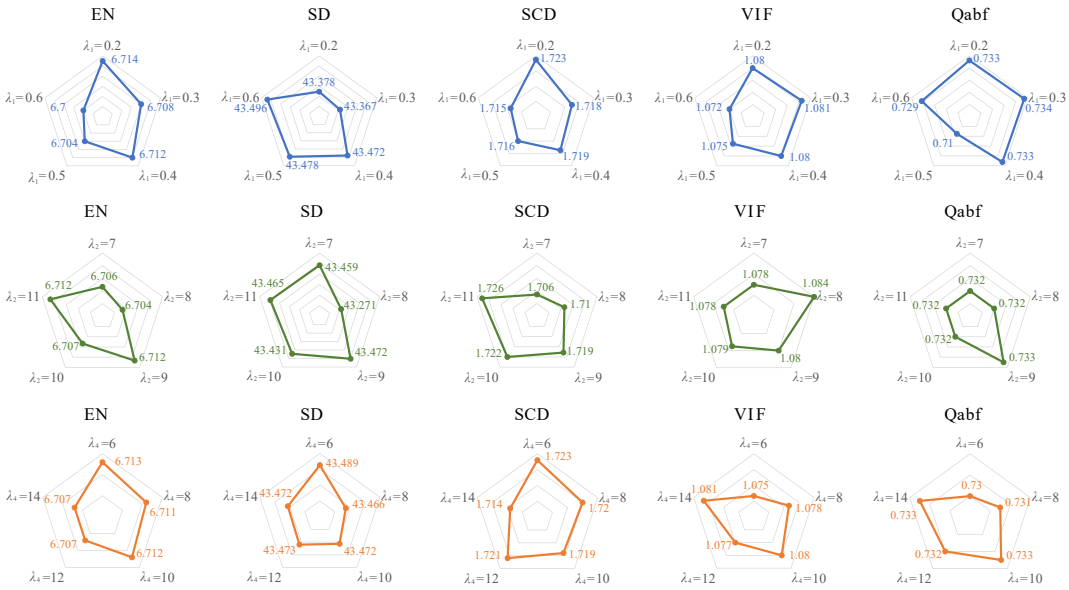


图7　在MSRS数据集上针对超参数λ_1、λ_2和λ_4的敏感性分析,采用五种评价指标进行评估(图中从上到下依次展示了λ_1、λ_2和λ_4的分析结果)

Figure 7　Sensitivity analysis of hyperparameters λ_1, λ_2 and λ_4 on the MSRS dataset evaluated using five metrics (From top to bottom, the figure shows the results for λ_1, λ_2 and λ_4, respectively)

大多数标准场景中生成描述的总体稳定性,在本文使用的数据集上语义描述生成的准确率约75%,仅在一些复杂场景下(如低光照、遮挡环境)存在较小误差,足以满足语义引导的需求。同时,本文使用的红外与可见光数据集是完全空间对齐的,因此BLIP基于可见光图像生成的全局文本语义可迁移至红外模态。

其次,本文使用的关键词V^*是基于数据集统计的目标类别,尚未实现完全自动化与语义自适应。这在一定程度上限制了方法在复杂语义或跨数据集条件下的泛化能力。未来工作将考虑引入大语言模型驱动的关键词生成机制,减少对数据集先验的依赖。

在实际部署与应用中,存在一些极端质量退化场景,BLIP生成的文本描述可能不够准确,进而导致掩码语义生成质量下降,语义引导作用略有减弱。但是在大多数实际测试场景中,BLIP生成的文本语义与扩散模型生成的区域掩码均稳定可靠,即使图像存在轻微模糊、遮挡或噪声,这完全适配多数实际应用场景的自动化需求。本文所用公开数据集虽未包含此类极端质量退化样本,但已在RoadScene数据集(含可见光图像曝光不均衡现象)上完成针对性实验,结果显示,即使图像存在轻微质量退化,模型的融合性能仍显著优于其他融合框架。

4　结束语

本文提出了一种基于文本语义信息引导的红外与可见光图像融合新方法TeSG。该方法包含三个核心模块:语义信息生成模块(SIG)、掩码引导的跨注意力模块(MGCA)以及文本驱动注意力融合模块(TDAF)。这些模块通过充分利用掩码级和文本级的双重语义信息,有效指导图像融合过程,最终生成高质量的融合图像。大量实验结果表明,TeSG在各项指标上均优于现有的最先进方法,尤其在目标检测、语义分割等下游任务中表现出显著性能提升。未来工作将探索更可控的融合机制,并拓展跨模态信息的应用场景。

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] Zhang H, Xu H, Tian X, et al. Image fusion meets deep learning: A survey and perspective[J]. Information Fusion, 2021, 76: 323-336.

[3] Yang Y, Liu J X, Huang S Y, et al. Infrared and visible image fusion via texture conditional generative adversarial network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(12): 4771-4783.

[4] Zhao Y Y, Zheng Q C, Zhu P H, et al. TUFusion: A transformer-based universal fusion algorithm for multimodal images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(3): 1712-1725.

[5] Davis J W, Sharma V. Background-subtraction using contour-based fusion of thermal and visible imagery[J]. Computer Vision and Image Understanding, 2007, 106(2/3):

162-182.

[6] Han J G, Pauwels E J, de Zeeuw P. Fast saliency-aware multi-modality image fusion[J]. Neurocomputing, 2013, 111: 70-80.

[7] Xu P, Davoine F, Bordes J B, et al. Multimodal information fusion for urban scene understanding[J]. Machine Vision and Applications, 2016, 27(3): 331-349.

[8] Li H G, Ding W R, Cao X B, et al. Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing[J]. Remote Sensing, 2017, 9(5): 441.

[9] 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.

[10] Ma J L, Zhou Z Q, Wang B, et al. Infrared and visible image fusion based on visual saliency map and weighted least square optimization[J]. Infrared Physics & Technology, 2017, 82: 8-17.

[11] Liu J Y, Fan X, Jiang J, et al. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 105-119.

[12] Li H, Liu L, Huang W, et al. An improved fusion algorithm for infrared and visible images based on multi-scale transform[J]. Infrared Physics & Technology, 2016, 74: 28-37.

[13] Li G F, Lin Y J, Qu X D. An infrared and visible image fusion method based on multi-scale transformation and norm optimization[J]. Information Fusion, 2021, 71: 109-129.

[14] Cvejic N, Bull D, Canagarajah N. Region-based multimodal image fusion using ICA bases[J]. IEEE Sensors Journal, 2007, 7(5): 743-751.

[15] Wang J, Peng J Y, Feng X Y, et al. Fusion method for infrared and visible images by using non-negative sparse representation[J]. Infrared Physics & Technology, 2014, 67: 477-489.

[16] Bavirisetti D P, Dhuli R. Two-scale image fusion of visible and infrared images using saliency detection[J]. Infrared Physics & Technology, 2016, 76: 52-64.

[17] Liu C H, Qi Y, Ding W R. Infrared and visible image fusion method based on saliency detection in sparse domain[J]. Infrared Physics & Technology, 2017, 83: 94-102.

[18] 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.

[19] Zhao Z X, Xu S, Zhang C X, et al. DIDFuse: Deep image decomposition for infrared and visible image fusion[C]// Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2020: 970-976.

[20] Zhang H, Ma J Y. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion[J]. International Journal of Computer Vision, 2021, 129(10): 2761-2785.

[21] Liu Y, Chen X, Cheng J, et al. Infrared and visible image fusion with convolutional neural networks[J]. International Journal of Wavelets, Multiresolution and Information Processing, 2018, 16(3): 1850018.

[22] 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.

[23] Xu H, Ma J Y, Yuan J T, et al. RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 19647-19656.

[24] Ma J Y, Yu W, Liang P W, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48: 11-26.

[25] Ma J Y, Xu H, Jiang J J, et al. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995.

[26] Ma J Y, Liang P W, Yu W, et al. Infrared and visible image fusion *via* detail preserving adversarial learning[J]. Information Fusion, 2020, 54: 85-98.

[27] Ma J Y, Zhang H, Shao Z F, et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 5005014.

[28] 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.

[29] Ma J Y, Tang L F, Fan F, et al. SwinFusion: Cross-do-

main long-range learning for general image fusion via swin transformer[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.

[30] Wang Z S, Chen Y L, Shao W Y, et al. SwinFuse: A residual swin transformer fusion network for infrared and visible images[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 5016412.

[31] Yi X P, Xu H, Zhang H, et al. Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 27016-27025.

[32] 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. New York: ACM, 2024: 60749-60765.

[33] Wang H, Zhang H, Yi X P, et al. TeRF: Text-driven and region-aware flexible visible and infrared image fusion[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 935-944.

[34] Cheng C Y, Xu T Y, Wu X J, et al. TextFusion: Unveiling the power of textual semantics for controllable image fusion[J]. Information Fusion, 2025, 117: 102790.

[35] Wang Z Y, Zhao L B, Zhang J Z, et al. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model[J]. International Journal of Computer Vision, 2025, 133(7): 4646-4668.

[36] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022: 12888-12900.

[37] Bhatnagar G, Jonathan Wu Q M, Liu Z. Directive contrast based multimodal medical image fusion in NSCT domain[J]. IEEE Transactions on Multimedia, 2013, 15(5): 1014-1024.

[38] Mitianoudis N, Stathaki T. Pixel-based and region-based image fusion schemes using ICA bases[J]. Information Fusion, 2007, 8(2): 131-142.

[39] Cui G M, Feng H J, Xu Z H, et al. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition[J]. Optics Communications, 2015, 341: 199-209.

[40] Zhang B H, Lu X Q, Pei H Q, et al. A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled Shearlet transform[J]. Infrared Physics & Technology, 2015, 73: 286-297.

[41] Kong W W, Lei Y, Zhao H X. Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization[J]. Infrared Physics & Technology, 2014, 67: 161-172.

[42] 方帅, 万旗, 曹洋 . 基于跨尺度相似先验的遥感图像时空融合算法[J]. 电子学报, 2024, 52(6): 2037-2052.
Fang Shuai, Wan Qi, Cao Yang. A spatiotemporal fusion algorithm of remote sensing images based on cross-scale similarity prior[J]. Acta Electronica Sinica, 2024, 52(6): 2037-2052. (in Chinese)

[43] Liu Y, Liu S P, Wang Z F. A general framework for image fusion based on multi-scale transform and sparse representation[J]. Information Fusion, 2015, 24: 147-164.

[44] 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.

[45] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2014: 2672-2680.

[46] Mirza M, Osindero S. Conditional generative adversarial nets[PP/OL]. V1. arXiv (2014-11-06) [2025-12-21]. https://doi.org/10.48550/arXiv.1411.1784.

[47] 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. Piscataway: IEEE, 2023: 8048-8059.

[48] Yue J, Fang L Y, Xia S B, et al. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models[J]. IEEE Transactions on Image Processing, 2023, 32: 5705-5720.

[49] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C]//International Conference on Machine Learning. PMLR, 2021: 8821-8831.

[50] Kim G, Kwon T, Ye J C. DiffusionCLIP: Text-guided diffusion models for robust image manipulation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 2416-2425.

[51] Ruiz N, Li Y Z, Jampani V, et al. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 22500-22510.

[52] Lin Y Z, Chen Y W, Tsai Y H, et al. Text-driven image editing via learnable regions[C]//2024 IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 7059-7068.

[53] Kawar B, Zada S, Lang O, et al. Imagic: Text-based real image editing with diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 6007-6017.

[54] Tumanyan N, Geyer M, Bagon S, et al. Plug-and-play diffusion features for text-driven image-to-image translation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 1921-1930.

[55] Qi T H, Fang S C, Wu Y Z, et al. DEADiff: An efficient stylization diffusion model with disentangled representations[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 8693-8702.

[56] Nichol A Q, Dhariwal P, Ramesh A, et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models[C]//International Conference on Machine Learning. PMLR, 2022: 16784-16804.

[57] Yang Z Y, Zhang Y F, Li H F, et al. Instruction-driven fusion of Infrared-visible images: Tailoring for diverse downstream tasks[J]. Information Fusion, 2025, 121: 103148.

[58] 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.

[59] 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, 2023: 8081-8090.

[60] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10674-10685.

[61] 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.

[62] 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.

[63] Jia X Y, Zhu C, Li M Z, et al. LLVIP: A visible-infrared paired dataset for low-light vision[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE, 2021: 3489-3497.

[64] Huang Z B, Liu J Y, Fan X, et al. ReCoNet: Recurrent correction network for fast and efficient multi-modality image fusion[M]//Computer Vision - ECCV 2022. Cham: Springer International Publishing, 2022: 539-555.

[65] Li H, Xu T Y, Wu X J, et al. LRRNet: A novel representation learning guided fusion network for infrared and visible images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 11040-11052.

[66] Zhao W D, Xie S G, Zhao F, et al. MetaFusion: Infrared and visible image fusion via meta-feature embedding from object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 13955-13965.

[67] 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.

[68] 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.

[69] Li H F, Yang Z Y, Zhang Y F, et al. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(5): 3673-3690.

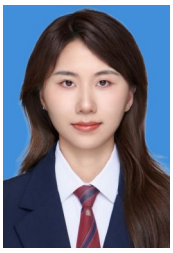
[70] Aslantas V, Bendes E. A new image quality metric for image fusion: The sum of the correlations of differences[J]. AEU - International Journal of Electronics and Communications, 2015, 69(12): 1890-1896.

[71] Han Y, Cai Y Z, Cao Y, et al. A new image fusion performance metric based on visual information fidelity[J]. Information Fusion, 2013, 14(2): 127-135.

[72] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 833-851.

作者简介

朱明瑞　男,1992年4月出生于山东省莱芜市。现为西安电子科技大学空天地一体化综合业务网全国重点实验室副教授、硕士生导师。主要研究方向为计算机视觉与机器学习。
E-mail: mrzhu@xidian.edu.cn



陈希茹　女,2001年10月出生于江苏省徐州市。现为西安电子科技大学通信工程学院研究生。主要研究方向为计算机视觉与机器学习。
E-mail: xrchen@stu.xidian.edu.cn

卫　鑫　男,1994年10月出生于陕西省铜川市。现为西安电子科技大学空天地一体化综合业务网全国重点实验室华山菁英副教授、硕士生导师。主要研究方向为三维计算机视觉。
E-mail: weixin@xidian.edu.cn

王楠楠　男,1986年11月出生于山东省潍坊市。现为西安电子科技大学空天地一体化综合业务网全国重点实验室教授、博士生导师。主要研究方向为计算机视觉与机器学习。中国电子学会会员编号:E190027116M。
E-mail: nnwang@xidian.edu.cn

高新波　男,1972年8月出生于山东省莱芜市。现为西安电子科技大学空天地一体化综合业务网全国重点实验室教授。主要研究方向为人工智能、机器学习、计算机视觉和模式识别。中国电子学会会员编号:E190004421F。
E-mail: xbgao@mail.xidian.edu.cn