文章编号　1004-924X(2026)12-1954-15

语义可控的红外与可见光图像融合

柳长源*，程兵云，吴海滨
(哈尔滨理工大学 测控技术与通信工程学院，黑龙江 哈尔滨 150080)

摘要：针对现有红外与可见光图像融合任务对开放式文本指令驱动的融合需求，本文提出一种基于可学习频带分解与文本引导门控的红外-可见光图像融合框架。通过语言－图像预训练模型(Contrastive Language-Image Pre-training，CLIP)与基于文本提示的通用分割模型(Grounded Segment Anything Model，Grounded SAM)协同构建文本-图像对齐模块，精准关联文本语义先验与图像局部区域，为双模态特征提取设计专属可学习频带分解模块，自适应分离表征整体结构的低频基底与刻画精细纹理的高频细节。随后设计文本感知门控融合模块，实现空间自适应调控红外与可见光模态的特征贡献权重，构建多约束损失函数，涵盖梯度保真、门控图正则损失等核心约束项，显式保障融合结果的结构完整性与文本语义一致性。实验结果表明：在LLVIP数据集上，本文网络相对于Textfusion的文本相关指标：边缘保持增强指标(Qabf+)和结构相似性增强指标(SSIM+)分别提升0.070 5和0.036 4，常见图像融合指标互信息(Mutual Information，MI)和峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)分别提升1.516 4和1.756 6 dB，在目标检测任务中交并比阈值范围为0.5～0.95的平均精度均值(mean Average Precision，mAP@50-95)提升0.9%。在实际融合中文本能够有效引导图像融合的生成过程，基本实现语义可控的图像融合目标。

关　键　词：图像融合；文本引导；红外图像；可见光图像；频带分解
中图分类号：TN911.73　　**文献标识码**：A
doi：10.37188/OPE.20263412.1954　　**CSTR**：32169.14.OPE.20263412.1954

Semantically controllable fusion of infrared and visible images

LIU Changyuan*，CHENG Bingyun，WU Haibin

(*College of Measurement-Control Technology and Communication Engineering*，
Harbin University of Science and Technology，*Harbin* 150080，*China*)
* *Corresponding author*，*E-mail*：*liuchangyuan@hrbust.edu.cn*

Abstract：In order to address the demand for open-ended text-instruction-guided fusion in infrared－visible image fusion tasks，this paper proposed an infrared－visible image fusion framework based on learnable frequency-band decomposition and text-guided gating. Specifically，the proposed method constructed a text－image alignment module by jointly exploiting Contrastive Language－Image Pre-training (CLIP) and the Grounded Segment Anything Model (Grounded SAM)，which associated textual semantic priors with local image regions. Then，a dedicated learnable frequency-band decomposition module was introduced for bimodal feature extraction，which adaptively separated the low-frequency base representing global structures from the high-frequency details describing fine textures. Based on this decomposition，a text-aware gated fusion module spatially regulated the feature contribution weights of the infrared and visible modali-

收稿日期：2026-04-14；**修订日期**：2026-05-06.
基金项目：黑龙江省交通运输厅科技项目(No. HJK2024B002)

ties. In addition, a multi-constraint loss function was formulated to incorporate gradient fidelity, and gating-map regularization, thereby preserving structural integrity and textual semantic consistency in the fused results. Experimental results on the Low-Light Visible-Infrared Paired (LLVIP) dataset demonstrate that, compared with TextFusion, the proposed method improves the text-related metrics edge preservation index plus (Qabf+) and Structural Similarity Index Measure plus (SSIM+) by 0.070 5 and 0.036 4, respectively. For common image fusion metrics, Mutual Information (MI) and Peak Signal-to-Noise Ratio (PSNR) are improved by 1.516 4 and 1.756 6 dB, respectively. Moreover, the proposed method achieves a 0.9% improvement in mean Average Precision over Intersection over Union thresholds from 0.5 to 0.95 (mAP@0.5∶0.95) in the downstream object detection task. These results indicate that textual instructions can effectively guide the image fusion process in practical scenarios, thereby achieving semantically controllable infrared - visible image fusion.

Key words: image fusion; text guidance; infrared image; visible image; frequency-band decomposition

1　引　言

面向复杂感知场景的多模态信息融合，旨在整合来自不同传感器、不同谱段或不同语义来源的互补信息，以获得更加完整、鲁棒且任务相关的场景表征[1]。红外与可见光图像融合是跨谱段多模态融合中的典型任务，其目标是在同一幅图像中同时保留红外模态对热辐射目标的显著响应，以及可见光模态所包含的纹理、边缘和背景细节信息[2]。

现有红外与可见光图像融合方法主要包括传统方法和基于深度学习的方法[3]。传统方法主要基于手工设计的特征与融合规则，如多尺度分解、稀疏表示和低秩建模等。近年来，也有研究继续围绕模型驱动或规则驱动范式进行改进。例如，龙志亮等人[4]提出改进多尺度结构化融合方法，以增强多尺度结构信息表达；任鹏百等人[5]基于HMSD与改进PCNN开展融合研究，提升了对结构信息和显著区域的建模能力。这些方法表明，模型驱动范式在结构保持和可解释性方面仍具有一定价值，但其融合策略通常依赖固定先验和预设规则，在复杂场景下的特征表达、语义感知和自适应调节能力仍然有限。

随着深度学习的发展，基于卷积神经网络(Convolutional Neural Network, CNN)[6]、注意力机制和视觉Transformer(Vision Transformer, ViT)[7]的端到端融合模型显著提升了融合性能，能够自动学习跨模态特征表示并优化融合策略。例如，Liu等人[8]提出的TarDAL在目标感知双对抗学习框架下联合建模红外-可见光图像融合与目标检测，使融合结果兼顾红外目标显著性与可见光纹理细节；Cheng等人[9]提出的MUFusion构建通用无监督端到端融合网络，并结合自适应统一损失提升融合质量；Zhao等人[10]提出DDFM，将融合任务建模为扩散概率模型下的条件生成问题，利用生成先验获得更自然的融合结果；Yi等人[11]提出的LUT-Fuse则将多模态融合能力蒸馏到可学习查找表中，在保证融合性能的同时显著提升推理效率。此外，赵阳等人[12]提出基于全局双组注意力的红外与可见光图像融合方法，通过注意力机制增强跨模态特征交互与全局信息建模能力。陶俣卓等人[13]提出目标语义分层驱动的红外和可见光图像融合方法，通过引入目标语义信息并进行分层建模，增强融合结果对目标区域和场景语义的表达能力。上述方法分别从检测协同、无监督学习、生成建模、注意力交互、语义增强和高效部署等角度推动了红外与可见光图像融合研究的发展。然而，目前多数基于深度学习的图像融合方法仍存在融合输出可控性不足的问题。模型在训练阶段通常通过预设损失函数、固定网络结构和训练数据分布学习一种相对稳定的跨模态融合策略，而在测试阶段难以根据具体任务目标或场景语义动态调整信息保留重点。

为突破上述“不可控”的瓶颈，近期研究开始尝试将高层语义先验，尤其是文本语义，引入融合过程。Wang等人[14]基于语言-图像预训练模型(Contrastive Language-Image Pre-training, CLIP)嵌入空间提出语言驱动损失，通过将融合目标表示为自然语言描述，并在视觉—语言共享嵌入空间中构建

语言驱动的融合关系，使融合结果能够在全局语义层面与文本目标保持一致；随后，Yi 等人[15]提出 Text-IF 将文本引导进一步扩展到退化感知和交互式融合场景；Cheng 等人[16]提出的 TextFusion 则通过视觉-语言模型建立从粗到细的文本-视觉关联机制，并在 Transformer 融合主干中嵌入仿射融合模块，从而将文本条件更深层次地注入特征融合过程。进一步地，Wang 等人[17]提出 RIS-Fuse，通过弱监督的实例级可控融合机制，利用文本指定目标并对目标区域与非目标区域施加差异化融合策略，从而实现对用户关注实例的自适应突出。总体来看，文本语义为红外与可见光融合提供了新的可控思路，但现有方法多侧重粗粒度关联建模，在开放式文本输入下仍难以实现精细的区域语义对齐、明确的目标和背景差异化融合，以及对红外-可见光频率差异的充分建模。

针对文本引导的红外-可见光图像融合的问题，本文围绕频域建模、文本门控与区域监督提出了一套新的解决方案。具体而言，首先构建了文本指令引导的跨模态融合框架，通过建立文本描述与图像中语义相关区域的对齐机制，实现融合过程的语境感知与任务导向；其次设计了可学习自适应频带分解模块，分离表征全局结构的低频基底与刻画精细纹理的高频细节，为红外显著性信息与可见光细节信息的互补融合提供更清晰的特征表示；进而提出文本感知门控融合模块，通过空间自适应调控双模态贡献权重，并设计多约束损失函数实现结构保持与文本语义一致性的显式统一。实验结果表明，所提方法能够在保持图像结构与细节信息的同时，实现更加稳定的文本语义引导融合。

2　原　理

2.1　网络总体架构

本文以红外 I_{ir}、可见光图像 I_{vis} 与文本描述 T 为输入：首先将 T 输入 CLIP 文本编码器以提取文本语义特征，并结合视觉定位模型（Grounding-DINO）[18]与分割一切模型（Segment Anything Model，SAM）[19]，在图像中定位并分割与文本描述相关的目标区域，生成语义对齐掩码。该掩码用于模型训练阶段，使模型能够重点关注文本所描述的显著目标区域。

同时将红外与可见光图像分别送入频率感知浅层编码器（Frequency Shallow Encoder，FSE），通过可学习二抽头（Learnable Two-Tap QMF，LQMF）分解为低频基底与高频细节特征。低频特征主要保留图像的整体结构与亮度分布，高频特征则侧重刻画边缘、纹理和目标细节。再在语义门控对称融合模块（Text-Aware Gated Fusion，TAGF）中引入文本语义特征，对红外与可见光的低频特征和高频特征分别进行像素级门控与仿射调制。将两个频段的特征进行拼接，并通过 1×1 卷积完成通道压缩与特征投影，实现跨频带信息重组。最后，将投影后的融合特征输入图像重建头（PatchUnEmbed），映射回图像域并生成最终的红外与可见光融合图像。如图 1 所示。

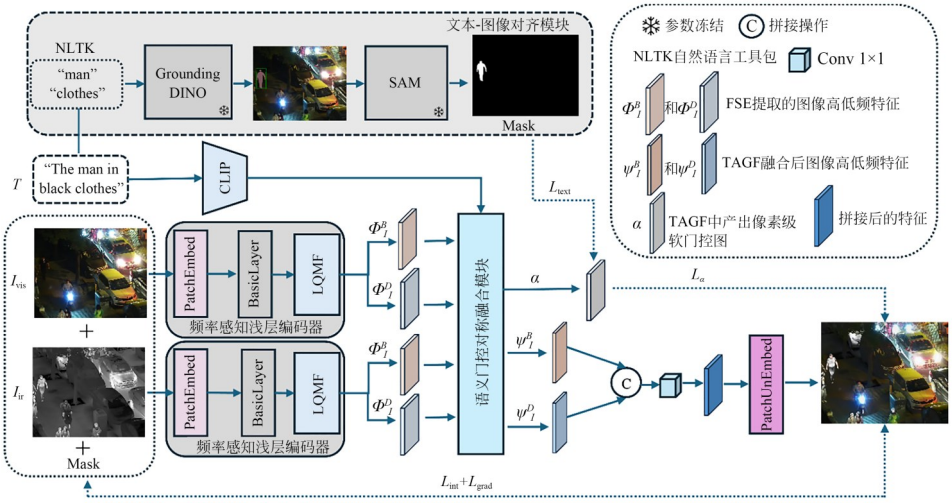


图 1　网络结构图

Fig. 1　Architecture of the proposed network

训练阶段，CLIP，Grounding DINO 和 SAM 参数保持冻结。模型仅优化FSE、TAGF 与解码器，并采用 L_{int}，L_{grad}，L_{text} 与 L_a 联合约束以提升目标显著性与细节保真。

2.2　文本-图像对齐模块

如图 1 的图文对齐模块，在文本引导的图像分割流程中，首先读取输入的文本描述信息，借助自然语言工具包（Natural Language Toolkit，NLTK）对文本进行关键语义信息提取。随后将提取的关键信息输入 GroundingDINO，由该模型在图像中检测并生成与文本语义匹配的目标检测区域 B，最后将上述检测区域作为提示信息输入 SAM，由 SAM 完成对文本描述目标的像素级分割，最终得到与文本语义对应的图像区域掩码 M 如公式（1）：

$$M = f_s(I, B), \tag{1}$$

其中：f_s 是 SAM 的核心分割映射函数，I 是原始图像。

2.3　频率感知浅层编码器

为在浅层阶段充分保留红外与可见光图像的空间细节和结构信息，本文针对两种模态的图像构建了频率感知浅层编码器（Frequency-aware Shallow Encoder，FSE），如图 2。

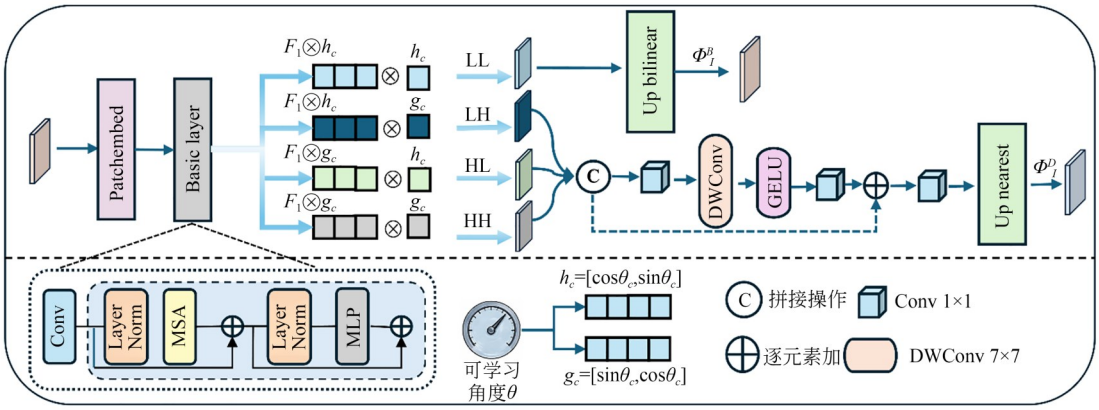


图 2　频带分解模块图

Fig. 2　Architecture of the frequency band decomposition module.

在空间编码阶段，FSE 首先采用带反射填充的卷积层完成 PatchEmbed，将输入单通道图像映射到 C 维特征空间。这里卷积步长设为 1，不在浅层做下采样，以避免过早降低分辨率，从而尽可能保留目标轮廓和纹理等细节信息。随后，引入由单层 TransformerBlock 构成的 BasicLayer。该模块采用 8×8 非移位局部窗口注意力，在不改变特征空间分辨率的条件下对局部邻域依赖关系进行建模，并利用残差连接增强跨通道信息交互能力，得到更具判别性的浅层特征图 F1。在频率建模阶段，FSE 通过 LQMF 对 F1 进行显式频带分解。与传统固定 Haar 小波不同，LQMF 为每个通道引入一个可学习角度参数 θ_c，并将一维分析滤波器参数化，如式（2）：

$$\begin{cases} h_c = [\cos\theta_c, \sin\theta_c] \\ g_c = [\sin\theta_c, -\cos\theta_c] \end{cases}, \tag{2}$$

其中：h_c 和 g_c 分别代表第 c 个通道不同的滤波器。

LQMF 为每个通道学习一对长度为 2 的正交镜像滤波器，通过可分二维卷积在空间上构造出四个子带响应，分别对应一个低频分量和三个方向的高频分量，低频子带 H_1^{LL} 主要携带场景的整体亮度与结构轮廓，高频子带 H_1^{LH}，H_1^{HL}，H_1^{HH} 则聚焦于物体边缘和细节纹理。可记为公式（3）：

$$\begin{cases} H_1^{LL} = F_1 \otimes h_c \otimes h_c \\ H_1^{LH} = F_1 \otimes h_c \otimes g_c \\ H_1^{HL} = F_1 \otimes g_c \otimes h_c \\ H_1^{HH} = F_1 \otimes g_c \otimes g_c \end{cases}, \tag{3}$$

其中，$\otimes$ 表示可分离卷积。

为了与解码端在原始分辨率下的重建操作相匹配，本文将低频子带通过双线性插值上采样（UP bilinear）到输入特征的空间尺寸，作为当前模态的基底特征；三个高频子带在通道方向拼接后，送入由 1×1 卷积、深度卷积、GELU 和 1×1 卷积构成的高频增强块进行局部细节建模。并且在残差连接时，增强结果进一步与输入高频特征相加，以提升训练稳定性并减少对原始纹理的

过度扰动。最后，通过1×1卷积将增强后的高频特征压缩回原通道数，并经最近邻插值(UP nearest)得到高频细节分支。通过上述设计，FSE在保持信息完整性的同时，为后续语义门控对称融合模块在不同频带上实施差异化的语义调制提供了良好的输入基础。

2.4　语义门控对称融合模块

为了充分利用FSE生成的低频(Base, B)和高频信息(Detail, D)，本文设计了文本引导的跨模态融合模块(Text-Aware Gated Fusion, TAGF)如图3所示：具体实现如下：设文本编码器输出的文本向量为ψ_T。首先，将同频的红外特征Φ_{ir}^{s}与可见光特征Φ_{vis}^{s}拼接，$s\in\{B,D\}$。得到用于注意力计算的多模态特征表示X，并通过共享1×1卷积配合归一化得到键(Key, K)和值(Value, V)，并由文本特征经线性层映射得到查询(Query, Q)。为了显式建模“文本语义－空间位置”的对应关系，TAGF采用以文本为查询的跨模态交叉注意力。将空间维度展平为$N=HW$后，对每个位置计算相关性分数与门控响应，通过多头交叉注意力计算空间门控响应如公式(4)：

$$g_j=\sigma\left(\frac{\boldsymbol{Q}\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d}\,\tau}\right),j=1,2,\cdots,N, \tag{4}$$

其中：d为单头维度，τ为温度系数，$\sigma(\ )$为Sigmoid函数。

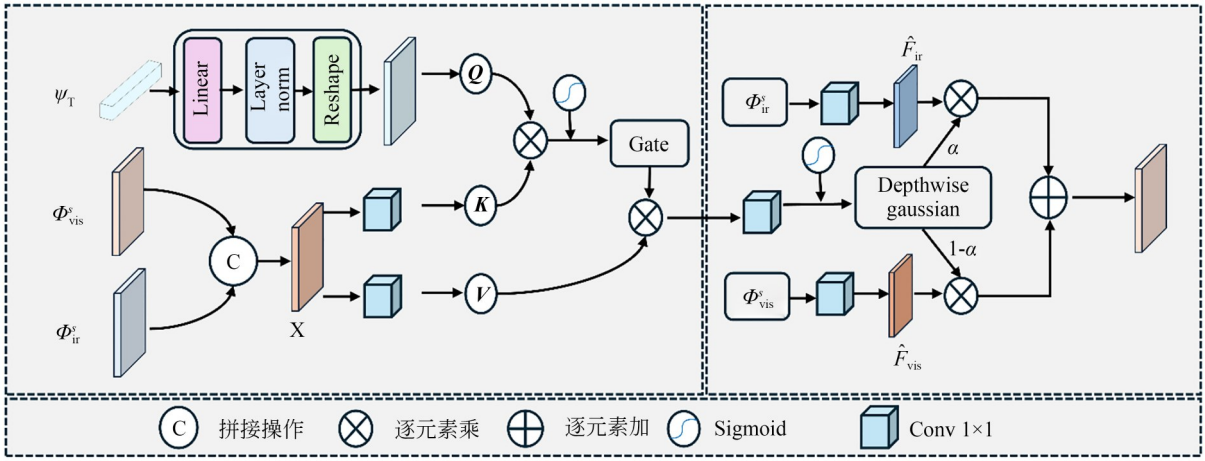


图3　语义门控对称融合模块图

Fig. 3　Structure of the semantic gating symmetric fusion module.

为避免门控在整幅图像上过度激活，引入局部归一化约束，并对g_j与值特征$\boldsymbol{V}$逐点重标定，聚合得到语义感知的上下文特征$\boldsymbol{C}$，如公式(5)：

$$\boldsymbol{C}=\sum_{j=1}^{N}\left(\frac{g_j}{\sum_m g_m+\varepsilon}\cdot(\rho N)\right)\boldsymbol{V}_j, \tag{5}$$

其中，ε为防止除零的微小常数，是超参数。

在得到上下文特征$\boldsymbol{C}$后，TAGF将其映射为像素级软门控图。首先通过1×1卷积与Sigmoid得到初始门控图，随后采用固定高斯核进行通道独立平滑，以抑制噪声并获得自然连续的融合边界，如公式(6)：

$$\alpha=\mathrm{clip}\left(\mathrm{Blur}\left(\sigma(\boldsymbol{W}_g*\boldsymbol{C})\right),0,1\right), \tag{6}$$

其中：$\boldsymbol{W}_g$为1×1卷积核，Blur(·)表示高斯平滑算子，clip(·)将门控值裁剪到[0,1]。

在融合阶段，为避免简单加权平均导致的模态偏置，并赋予门控清晰的物理含义，TAGF先将同频的两模态特征投影到统一子空间得到$\hat{F}_{\mathrm{vis}}$和$\hat{F}_{\mathrm{ir}}$，再采用对称仿射形式完成逐点融合，如公式(7)：

$$F_{\mathrm{fuse}}=(1-\alpha)\odot\hat{F}_{\mathrm{vis}}+\alpha\odot\hat{F}_{\mathrm{ir}}, \tag{7}$$

其中，$\odot$表示逐点乘。

2.5　TAGF与LQMF的理论分析

为进一步说明TAGF门控机制与LQMF频带分解设计的合理性，对公式(7)求梯度可得公式(8)：

$$\nabla F_{\mathrm{fuse}}=(1-\alpha)\odot\nabla\hat{F}_{\mathrm{vis}}+\alpha\odot\nabla\hat{F}_{\mathrm{ir}}+\left(\hat{F}_{\mathrm{ir}}-\hat{F}_{\mathrm{vis}}\right)\odot\nabla\alpha. \tag{8}$$

该式表明，融合特征的梯度不仅由红外和可见光特征共同决定，还受到文本引导门控图 α 及其空间变化的影响。因此，TAGF 能够根据文本语义自适应调节不同区域中红外与可见光特征的贡献。

对于 LQMF 频带分解模块，本文采用可学习二抽头正交镜像滤波器构造低通滤波器 h_c 和高通滤波器 g_c，由公式(2)可知公式(9)：

$$\| \boldsymbol{h}_c \|_2^2 = 1, \| \boldsymbol{g}_c \|_2^2 = 1, \boldsymbol{h}_c \boldsymbol{g}_c^{\mathrm{T}} = 0. \tag{9}$$

可以看出低通滤波器和高通滤波器具有归一化与正交性，有助于降低低频结构信息与高频细节信息之间的冗余耦合。由公式(3)可知二维滤波器由一维正交滤波器外积构造，四个子带在局部空间上构成一组正交基，因此可以将输入特征分解为互补的结构成分与细节成分。

此外，当 $\theta = \pi/4$ 时，LQMF 退化为固定 Haar 分解：相比固定 Haar 分解，LQMF 中的 θ 可在训练过程中自适应更新，使不同通道能够根据红外与可见光特征分布学习更合适的频带划分方式。因此，LQMF 既保留了小波分解的结构先验，又具备神经网络的自适应建模能力，更适合红外-可见光图像融合任务。

2.6　损失函数

与常规监督学习任务不同，红外－可见光图像融合缺乏对应的真实标签，网络只能通过源图像间的相对关系来约束训练。因此，本文从梯度、强度、文本引导的区域一致性损失和门控图正则损失四个方面联合设计损失函数。且本文根据数据集中的文本描述构建语义类别 s：对于红外响应更显著的目标，如行人，赋予较高的红外偏好；对于可见光纹理和结构信息更丰富的目标，如车辆，则赋予较高的可见光偏好。对于多目标或语义指向不明确的文本描述，采用中性权重设置，以避免错误语义偏置对融合结果造成过度影响：

梯度损失：用于在文本描述区域保留两模态中更显著的边缘结构，非文本描述区域贴近可见光结构。首先用 Sobel 算子计算梯度幅值 $G(\cdot) = \| \nabla \cdot \|$ 并构造目标梯度，如公式(10)：

$$\begin{cases} G_{\max} = \max(G_{\mathrm{ir}}, G_{\mathrm{vi}}) \\ G = M \odot G_{\max} + (1 - M) \odot G_{\mathrm{vi}}, \\ L_{\mathrm{grad}} = \| G_f - G \|_1 \end{cases} \tag{10}$$

其中：∇ 为梯度算子，$\| \ \|_1$ 为 L1 范数，M 为文本描述区域掩码。

强度损失：用于将文本描述区域根据文本语义自适应偏向某模态、在非文本描述区域保持可见光强度一致性。本文采用线性混合构造强度目标，如公式(11)：

$$\begin{cases} I_{\mathrm{mix}} = \lambda_s I_{\mathrm{ir}} + (1 - \lambda_s) I_{\mathrm{vi}} \\ I_{\mathrm{tar}} = M \odot I_{\mathrm{mix}} + (1 - M) \odot I_{\mathrm{vi}}, \\ \mathcal{L}_{\mathrm{int}} = \| F - I_{\mathrm{tar}} \|_1 \end{cases} \tag{11}$$

其中：$\lambda_s \in (0,1)$ 为红外与可见光强度的权重系数。红外偏好目标取较大 λ_s，可见光偏好目标取较小 λ_s，多目标或语义不明确文本采用默认权重。I 表示图像强度，M 为文本描述区域掩码。

文本引导的区域一致性损失：用于将文本描述区域及其模态偏好显式注入像素空间。利用 M 构造伪标签，如公式(12)：

$$\begin{cases} I_{\mathrm{text}} = \omega_s \odot I_{\mathrm{ir}} + (1 - \omega_s) \odot I_{\mathrm{vi}} \\ \omega_s = \alpha_s M + \beta (1 - M) \\ \mathcal{L}_{\mathrm{text}} = \| F - I_{\mathrm{text}} \|_2^2 \end{cases}, \tag{12}$$

其中：ω_s 为语义类别 s 对应的区域级模态偏好权重，α_s 和 β 分别表示文本描述区域和非文本描述区域的模态偏好权重，$\| \ \|_2$ 为 L2 范数，M 为文本描述区域掩码。

门控图正则损失：作用于 TAGF 模块输出的门控图 $\alpha \in [0,1]$，提升其与文本描述区域先验的一致性和可解释性。若 α 为多通道，则先在通道维度求平均得到 $\bar{\alpha}$。由于 TAGF 中 α 越大表示融合过程越偏向红外，α 越小表示越偏向可见光，本文根据文本语义类别 s 设置门控目标 ρ_s，并定义公式(13)：

$$L_\alpha = w_{\mathrm{fore}} \| \bar{\alpha} M - \rho_s M \|_1 + w_{\mathrm{back}} \| \bar{\alpha} (1 - M) \|_1, \tag{13}$$

其中：W_{fore}，W_{back} 为前景与背景门控正则的权重系数，M 为文本描述区域掩码，ρ_s 表示语义类别 s。对应的门控目标较大的 ρ_s 表示文本区域更偏向红外，较小的 ρ_s 表示更偏向可见光，非兴趣区域则约束 $\bar{\alpha} \to 0$，以保持可见光背景结构。

最终，总损失函数为各子损失的加权和，如公式(14)：

$$\mathcal{L}_{\mathrm{total}} = \lambda_g \mathcal{L}_{\mathrm{grad}} + \lambda_i \mathcal{L}_{\mathrm{int}} + \lambda_t \mathcal{L}_{\mathrm{text}} + \lambda_\alpha \mathcal{L}_\alpha, \tag{14}$$

其中：$\lambda_g, \lambda_i, \lambda_t, \lambda_\alpha$ 为超参数。

3　实验结果

3.1　数据集和实验细节

在本文在 LLVIP[20]，TNO[21]和 MSRS[22]数据集开展系统性实验。其中，LLVIP 数据集主要针对低照度及复杂城市场景；TNO 数据集聚焦于军事与安防目标的融合质量评估；MSRS 数据集则包含较丰富的城市道路与多目标场景。这 3 个数据集在光照条件、场景复杂度和目标类型上各具特色，能够全面验证本文方法的通用性与鲁棒性。文本来自 Textfusion 提供的 IVT 数据集，每组红外-可见光图像对应 5 条不同文本描述。

由于本文方法以文本作为融合过程的语义控制条件，文本描述的选择会影响模型对语义区域的关注，因此本文进一步明确文本引导的选择原则：文本应尽量描述图像中真实存在且与任务相关的显著目标，避免使用与图像内容无关、语义过于宽泛或指向不明确的描述；对于目标检测等任务，优先选择包含行人、车辆等关键目标的文本。实验基于 PyTorch 框架，在 Windows10 操作系统下进行。硬件环境包括 Intel i5-12600KF CPU (@2.50 GHz) 和 NVIDIA RTX 4070 Ti SUPER 显卡。训练初始学习率为 1e-4，优化器采用 Adam，共训练 10 个 Epoch，Batch size 设为 1，设置 $\lambda_g=1$，$\lambda_i=0.7$，$\lambda_t=1.5$，$\lambda_a=1$。

3.2　对比方法和评价指标

为全面验证所提出方法的有效性，本文选取了 9 种当前最具代表性的红外与可见光图像融合方法进行了系统性对比实验。所选对比方法涵盖了基于扩散模型的方法 DDFM、基于元学习的方法 MetaFusion[23]、基于对抗学习的方法 TarDAL、基于多模态特征解耦的方法 MMDRFuse[24]和 CDDFuse[25]、基于查找表快速推理的方法 LUT-Fuse 以及基于文本驱动的方法 RIS-Fuse，Text-IF 和 TextFusion。这些方法代表了当前多模态图像融合领域的最高水平。在客观评价方面，本文选取了 7 项公认的量化指标对融合性能进行评估，包括：文本注意力增强的边缘保持度（Textual Attention-enhanced Gradient-based Edge Preservation Metric，Qabf+）、文本注意力增强的结构相似性（Textual Attention-enhanced Structural Similarity，SSIM+）、视觉信息保真度（Visual Information Fidelity，VIF）、标准差（Standard Deviation，SD）、互信息（Mutual Information，MI）、峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）以及视觉感知度量（Perceptual Quality Metric，Q^{CB}）。这些指标从边缘细节保留、结构一致性、信息丰富度及人眼视觉感知等多个维度，对融合图像的综合性能进行了全面量化分析[26]。

对于检测任务，采用精确率（Precision，P）、召回率（Recall，R）、交并比阈值为 0.5 时的平均精度（mAP50）及交并比阈值从 0.5 到 0.95 平均精度均值（mAP@50-95）四项指标进行量化评价[27]。

3.2.1　融合结果客观评价

从表 1～表 3 可以看出，本文方法在三个数据集上均表现出较稳定的综合性能，尤其在 SSIM+，VIF，MI，PSNR 和 Q^{CB} 等指标上相对于大多数所列先进算法具有优势。

表 1　TNO 数据集上的融合结果的评价指标结果

Tab. 1　Quantitative evaluation metrics of fusion results on the TNO dataset.

方法	Qabf+	SSIM+	VIF	SD	MI	PSNR	Q^{CB}
DDFM	0.180 7	0.697 3	0.624 7	27.356 3	2.449 3	25.992 5	0.511 8
Metafusion	0.344 7	0.601 3	0.798 5	47.634 5	2.568 2	28.085 4	0.509 0
Tardal	0.389 7	0.698 0	0.824 4	41.075 3	2.885 2	28.111 4	0.436 6
MMDRFuse	0.373 7	0.592 6	**1.019 3**	**58.558 4**	2.577 5	28.057 6	0.452 2
CDDFuse	0.496 1	0.694 8	0.931 0	44.154 7	3.079 0	28.315 5	0.503 7
LUT-Fuse	0.407 4	0.679 1	0.853 0	38.106 8	3.618 2	29.249 8	0.400 6
Text-if	0.553 7	0.694 3	0.987 9	46.524 0	3.243 6	28.611 6	0.502 5
RIS-FUSE	**0.581 5**	0.715 7	0.891 8	41.396 7	3.507 1	28.963 4	0.531 4
Textfusion	0.528 9	0.725 7	0.920 6	40.227 5	3.339 9	27.489 8	0.526 5
Ours-o	0.522 8	0.709 8	0.943 1	39.802 5	3.946 8	28.555 8	0.592 1
Ours	0.539 0	**0.733 3**	1.011 5	43.221 2	**4.276 1**	**29.327 3**	**0.614 0**

表 2　LLVIP 数据集的融合结果的评价指标结果

Tab. 2　Quantitative evaluation metrics of fusion results on the LLVIP dataset.

方法	Qabf+	SSIM+	VIF	SD	MI	PSNR	Q^{CB}
DDFM	0.069 9	0.554 0	0.627 6	37.998 1	3.001 2	26.214 1	0.490 1
Metafusion	0.318 7	0.605 6	0.754 2	50.792 6	1.990 5	28.245 8	0.494 5
Tardal	0.368 3	0.594 6	0.717 8	45.722 6	3.133 3	28.567 8	0.321 1
MMDRFuse	0.547 1	0.570 6	1.012 9	**65.391 6**	3.601 2	28.097 8	0.388 3
CDDFuse	0.610 2	0.645 0	0.939 6	51.063 4	4.217 4	28.682 5	0.425 0
LUT-Fuse	0.468 9	0.617 7	0.900 7	51.259 1	4.152 0	29.741 1	0.397 7
Text-if	0.703 3	0.656 9	1.036 1	54.358 5	3.096 8	28.225 7	0.520 4
RIS-FUSE	**0.733 2**	0.687 2	0.961 3	44.529 6	2.970 9	28.569 1	0.543 0
Textfusion	0.557 9	0.687 6	0.816 9	48.296 9	2.944 0	28.000 8	0.469 0
Ours-o	0.601 9	0.709 8	0.991 8	44.329 2	3.936 4	29.481 8	0.596 1
Ours	0.628 4	**0.724 0**	**1.064 3**	49.652 2	**4.460 4**	**29.757 4**	**0.629 9**

表 3　MSRS 数据集融合结果的评价指标结果

Tab. 3　Quantitative evaluation metrics of fusion results on the MSRS dataset.

方法	Qabf+	SSIM+	VIF	SD	MI	PSNR	Q^{CB}
DDFM	0.171 2	0.442 5	0.557 8	24.364 9	2.403 7	26.219 9	0.486 8
Metafusion	0.504 5	0.725 6	0.875 8	36.960 1	1.616 5	29.487 9	0.446 0
Tardal	0.440 8	0.599 7	0.720 4	35.472 6	2.608 7	27.677 6	0.299 2
MMDRFuse	0.563 6	0.786 2	1.054 7	**55.697 4**	3.283 0	29.079 9	0.471 2
CDDFuse	0.744 0	0.814 2	1.042 5	43.010 7	**5.079 5**	31.141 1	0.562 3
LUT-Fuse	0.636 3	0.786 3	0.970 1	42.198 2	3.592 1	30.593 9	0.505 9
Text-if	0.742 7	0.820 7	1.053 0	44.609 8	4.026 9	30.887 1	0.563 0
RIS-FUSE	**0.748 7**	0.820 2	1.011 1	41.516 9	4.640 6	**31.222 9**	0.571 9
Textfusion	0.536 8	0.786 2	0.853 8	41.882 6	3.420 0	29.178 3	0.494 1
Ours-o	0.595 7	0.834 6	0.992 5	43.830 9	4.022 1	30.590 9	0.560 3
Ours	0.722 1	**0.849 5**	**1.060 0**	44.734 0	4.909 8	30.994 3	**0.588 6**

具体而言：部分方法如 LUT-Fuse 具有较高的 PSNR，但 MI，VIF 和 Q^{CB} 表现不够稳定，表明其更偏向像素级相似性，而跨模态信息保留和感知质量仍有不足。MMDRFuse 具有较高的 SD，但其 PSNR 并未同步提升，说明过高的灰度离散度可能伴随背景噪声。CDDFuse 和 RIS-FUSE 在 Qabf+等指标上表现较强，说明其对边缘信息具有较好的保留能力，但在 VIF 和 SSIM+上并不占优。与文本驱动方法 Text-IF 和 TextFusion 相比，本文方法在三个数据集上表现更稳定。尤其与 TextFusion 相比，Ours 在 MSRS，LLVIP 和 TNO 上的 VIF，MI，PSNR 和 Q^{CB} 均明显更高，说明本文并非仅依赖文本输入，而是通过 LQMF 频带分解和 TAGF 门控融合更有效地利用文本语义、红外和可见光图像的互补信息。

Ours-o 表示去除文本引导后的模型结果。由表 1～表 3 可知，引入文本信息后，Ours 在三个数据集的各项指标上均优于 Ours-o，说明文本引导不仅能够提升 Qabf+和 SSIM+等文本相关指标，也能改善信息保留、失真控制和感知质量，从而验证了文本引导机制的有效性。同理可以看出 Ours-o 在无文本条件下相较于多数无文本融合算法具有更稳定的综合性能。在 TNO 数据集上，Ours-o 的 Qabf+，SSIM+，MI 和 Q^{CB} 均优于 CDDFuse 和 LUT-Fuse 等方法，说明其在边缘保持、结构一致性和信息保留方面具有明显优势；

在LLVIP数据集上，Ours-o的SSIM+和Q^{CB}优于所有无文本方法，；在MSRS数据集上，Ours-o的SSIM+优于所有无文本方法，Q^{CB}接近CDDFuse并优于LUT-Fuse等方法。总体而言，Ours-o在不依赖文本引导的情况下，也能取得较好融合质量，验证了本文网络的有效性。

3.2.2　融合结果主观评价

图4～图6分别展示了在TNO，LLVIP和MSRS三种数据集上融合实验结果的视觉比较，其中左下角不同颜色方框中的图为该图中对应颜色框选区域放大后的结果（彩图见期刊电子版）。如图4所示，在TNO数据集中，文本描述人和房子时，无文本方法(e)以及有文本方法(i)和(k)对人物的融合中红外占比偏少，人物亮度明显偏暗。(d)和(e)均存在背景模糊，但本文方法达到了较好的人物和房屋细节的融合。

如图5所示，在LLVIP数据集中，文本描述车辆和人时，无文本方法(c)，(f)，(h)和有文本方法(k)在车轮高亮区域存在发白和细节被抹平的情况，本文方法在融合后能更好地保留可见光的纹理信息并兼顾红外图像的显著性：尤其在车辆局部细节上，车轮胎边缘轮廓更清晰、纹理更完整，同时行人等目标也保持了较好的对比度与可辨识度，整体呈现更自然的融合效果。

图6所示，在MSRS数据集中，文本描述人的时候，无文本方法(c)存在亮度偏高和细节雾化现象，(d)和(e)方法整体偏暗，有文本方法(i)，(j)和(k)均存在人物显著性不足的缺点，而本文方法能够在行人区域保持更高的可辨识度。

综上，本方法在各种数据中都表现良好，且完整地保留了红外和可见光图像中的互补信息，表明本方法具有较好的融合能力和对不同数据集的泛化能力。

图4　TNO数据集上本方法与各种先进算法对比的融合结果

Fig. 4　Qualitative comparison of our method with state-of-the-art algorithms on the TNO dataset

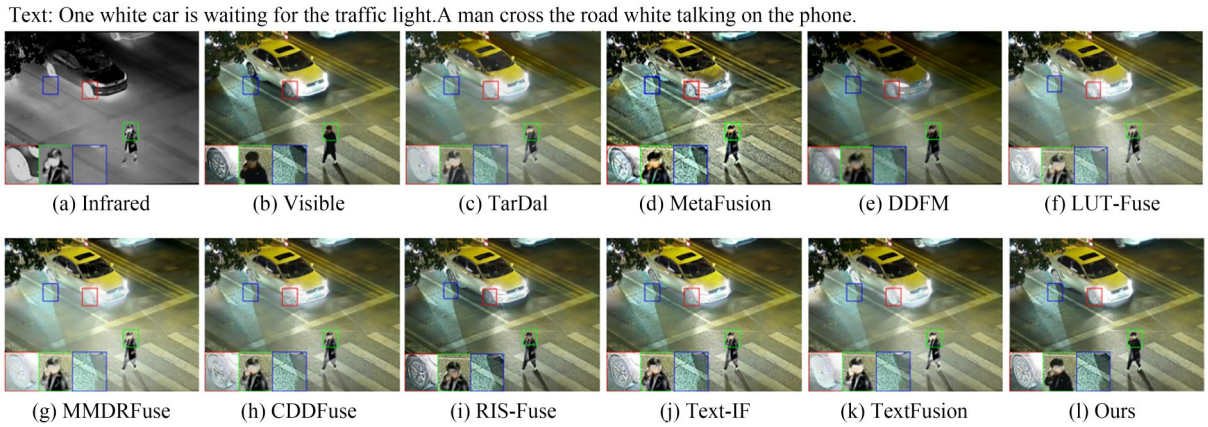


图5　LLVIP数据集上本方法与各种先进算法对比的融合结果

Fig. 5　Qualitative comparison of the proposed method with state-of-the-art algorithms on the LLVIP dataset

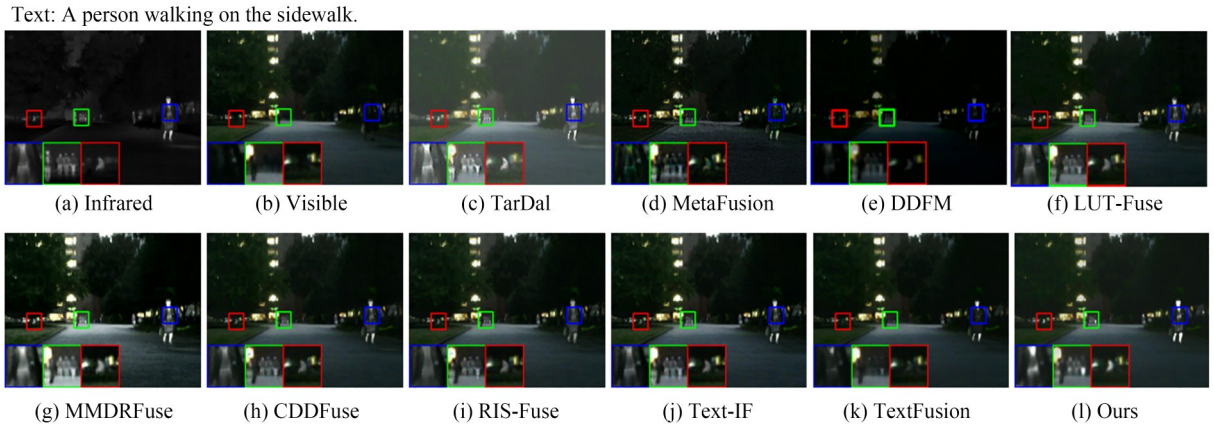


图 6　MSRS 数据集上本方法与各种先进算法对比的融合结果

Fig. 6　Qualitative comparison of the proposed method with state-of-the-art algorithms on the MSRS dataset

3.2.3　文本差异热图分析

为验证不同文本对图像融合的影响，在 LLVIP 数据集上选取一组图片进行热度分析，分别展示不同文本、空文本、错误文本情况下，TAGF 模块输出的低频 Base 分支热图、高频 Detail 分支热图、平均门控响应图以及最终融合响应热图。其中，Base 分支主要反映目标主体、亮度分布和整体结构响应，Detail 分支主要反映目标边缘、局部纹理和细节响应，Mean 表示高低频门控响应的综合结果，Fused image 表示最终融合结果，Fusion Heatmap 表示融合后的热图。

图 7 中可以看出，当输入“There are some

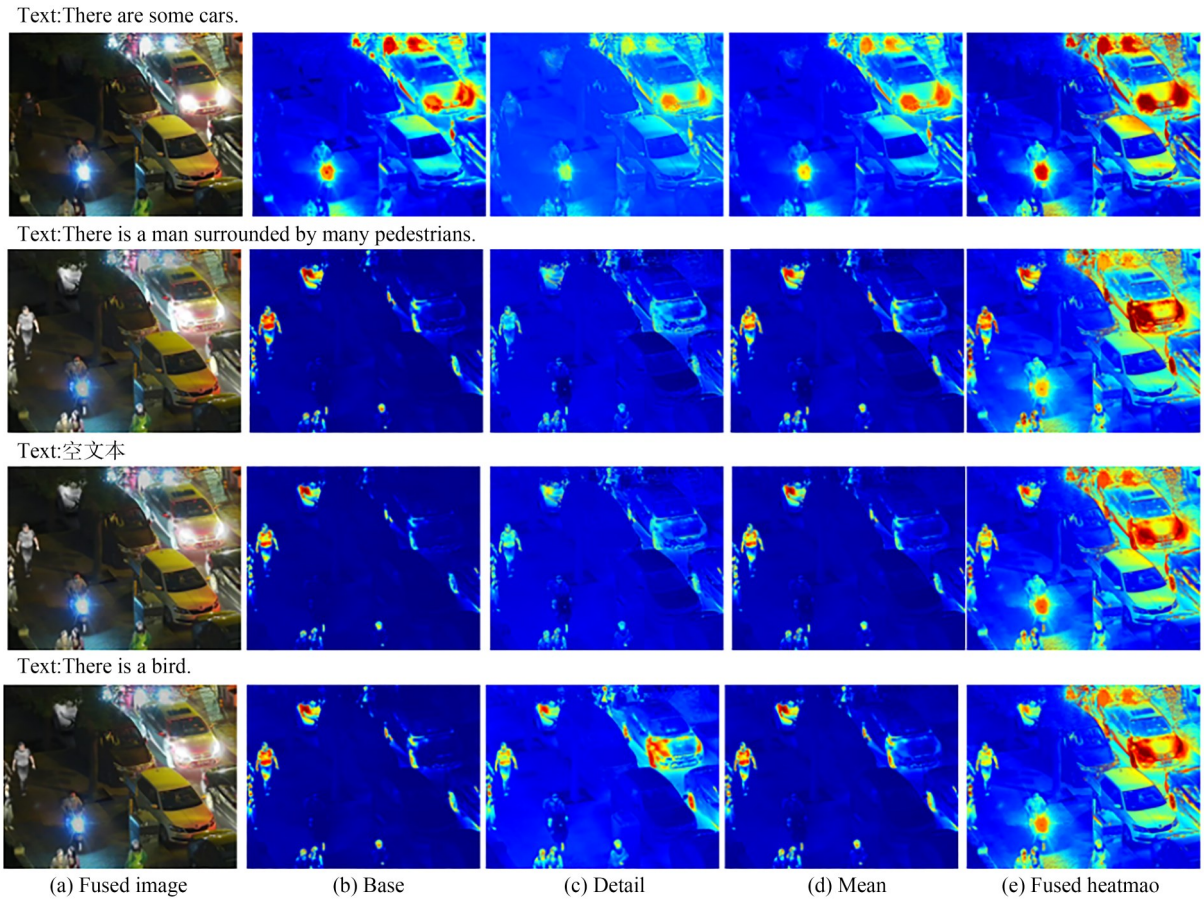


图 7　LLVIP 数据集上文本差异热图

Fig. 7　Heatmap of textual differences on the LLVIP dataset

cars.”时，门控响应主要集中在车辆区域；当输入“There is a man surrounded by many pedestrians.”时，响应明显转移到行人区域，说明 TAGF 能够根据文本语义动态调整红外与可见光特征的融合权重。相比之下，空文本条件下响应较为分散，主要依赖图像自身的强度和纹理信息。错误文本“There is a bird.”的响应分布与空文本较为相似，并未在图像中形成与“bird”对应的明显目标响应，而是仍集中在车辆、行人和强光等真实视觉显著区域。这说明当文本语义与图像内容不匹配时，文本引导作用会减弱，模型更倾向于依赖已有视觉证据进行融合，而不会凭空生成源图像中不存在的目标纹理。该可视化结果表明，正确文本能够有效调节 TAGF 的门控响应，而空文本和错误文本则更接近无有效语义约束的融合状态。

3.2.4　在目标检测上的结果及模型参数

表 4 给出了 LLVIP 数据集上不同融合方法的检测性能对比。

结果表明，本文方法在 mAP@50 与更严格的 mAP@50-95 上均取得最优，同时 Recall 达到 0.887 说明在保持较高精度的同时进一步降低了漏检率，并提升了整体检测性能与定位鲁棒性。进一步结合图 8 的定性结果可见，(c)，(d)，(e)，(f)，(g)和(k)均存在左下角行人漏检的情况，且本文方法使得人物的检测效果最好。总体而言，本文方法能够有效根据文本描述对行人区域表征进行融合，使得融合图像更加清晰、背景干扰更少。

表 4　LLVIP 数据集的目标检测结果的评价指标结果

Tab. 4　Objective detection metrics and parameter counts on the LLVIP dataset.

方法	P	R	F1	mAP@50	mAP@50-95
DDFM	0.848	0.787	0.816	0.850	0.492
Metafusion	0.863	0.812	0.837	0.872	0.507
Tardal	0.920	0.828	0.872	0.881	0.582
MMDRFuse	0.908	0.798	0.850	0.856	0.576
CDDFuse	0.927	0.873	0.899	0.926	0.587
LUT-Fuse	0.916	**0.887**	0.901	0.924	0.592
Text-if	0.946	0.870	**0.906**	0.927	0.590
RIS-FUSE	**0.950**	0.850	0.897	0.922	0.580
Textfusion	0.923	0.871	0.896	0.928	0.585
Ours	0.925	**0.887**	0.905	**0.932**	**0.594**

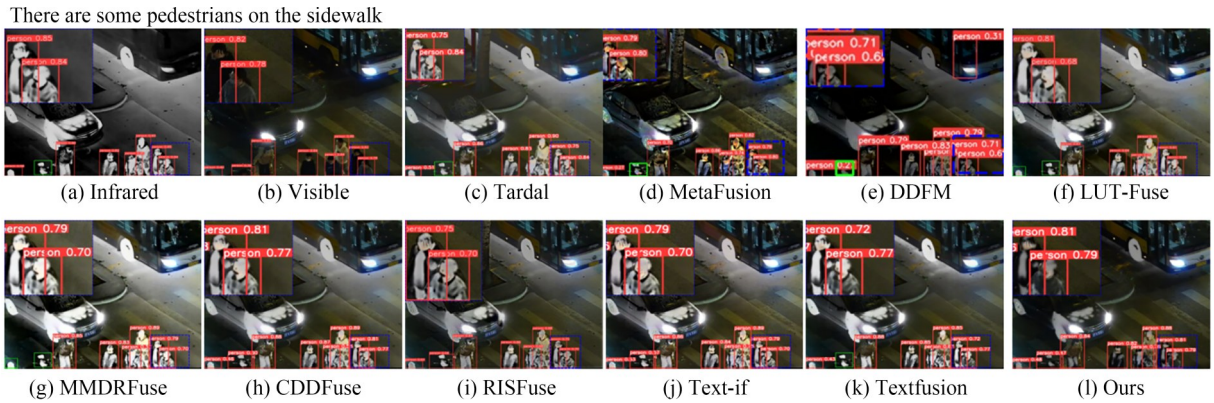


图 8　LLVIP 数据集上本方法与各种先进算法对比的目标检测可视化结果

Fig. 8　Visual comparison of object detection results on the LLVIP dataset

3.2.5　模型复杂度与部署代价分析

为更全面地评估模型复杂度与部署代价，本文在表 5 中补充报告了各方法的参数量、FLOPs 和平均推理时间。所有方法均在相同测试环境下进行评估，除 DDFM 采用 192×256 输入外，其余方法均统一采用 1 024×1 280 分辨率输入。对

于涉及文本输入的方法，参数量采用“核心融合网络参数量＋文本编码器参数量”的形式表示，其中“＋”后的部分为CLIP文本编码器参数量。核心融合网络参数量不包含预训练模块。

表5　LLVIP数据集上的模型复杂度与推理效率对比

Tab. 5　Model complexity and computational efficiency comparison on the LLVIP dataset

Method	Size	Parameters/M	FLOPs/G	Time/ms
DDFM	192×256	552.81	167.066K	8 562.196
MetaFusion	1 024×1 280	2.253	2.126K	324.735
TarDAL	1 024×1 280	0.297	776.409	943.396
MMDRFuse	1 024×1 280	**0.000 43**	**0.283**	83.267
CDDFuse	1 024×1 280	1.188	4.674K	1 212.533
LUT-Fuse	1 024×1 280	0.007 8	20.678	**58.871**
Text-IF	1 024×1 280	63.8＋63.4	12.948K	1 428.753
RIS-Fuse	1 024×1 280	0.476＋63.7	1.355K	8 274.26
TextFusion	1 024×1 280	0.074＋63.4	177.417	787.947
Ours	1 024×1 280	0.227＋63.4	119.504	268.215

从结果可以看出，本文网络在1 024×1 280分辨率下的FLOPs为119.504 G，平均推理时间为268.215 ms。相比Text-IF，RIS-Fuse和TextFusion这些相关方法，本文方法具有更低的计算量和更快的推理速度。这说明本文方法在引入文本语义信息后，仍保持了较低的部署开销。与无文本方法相比，本文方法的推理时间也低于MetaFusion，TarDAL和CDDFuse。虽然LUT-Fuse和MMDRFuse在参数量、FLOPs或推理速度方面更加轻量，但它们未显式建模文本语义。

综上所述，本文方法在轻量化和多模态语义扩展之间取得了较好的折中。但其局限性在于增加了文本编码器，推理时会带来额外计算和显存开销。

3.3　消融实验

3.3.1　模块有效性分析

为验证所提核心模块的有效性，本文设计了模块消融实验，分别构建去除LMQF中的可学习角度（θ）、去除LMQF中的高频分支（H）、去除LMQF中的低频分支（L）、去除TAGF和无文本输入的消融模型，并与完整模型进行对比。本节重点选取Qabf＋，VIF，PSNR，MI和mAP@50-95作为关键指标进行分析实验，结果如表6所示。结果表明，完整模型在Qabf＋，VIF，MI和mAP@50-95上取得最优结果，说明各模块之间具有良好的协同作用。具体来看，去除可学习角度θ后，Qabf＋，VIF，MI和mAP@50-95均低于完整模型，说明度会削弱频带分解的自适应建模能力。去除高频分支时，PSNR高于完整模型，减少了局部细节变化带来的像素级误差，但其Qabf＋，VIF和MI明显低于完整模型，表明边缘细节和跨模态信息保留能力下降。缺少低频分支时，mAP@50-95低于完整模型，说明仅依赖高频细节信息难以获得

表6　LLVIP数据集上的模块消融实验结果

Tab. 6　Results of ablation studies on the LLVIP dataset

方法	Qabf＋	SSIM＋	VIF	SD	MI	PSNR	Q^{CB}	P	R	mAP50	mAP@50-95
w/o LQMF(θ)	0.596 3	0.728 1	0.923 1	46.416 5	3.759 2	28.330 6	0.596 7	0.923 0	0.873 0	0.932 0	0.562 0
w/o LQMF(L)	0.618 6	**0.735 1**	0.998 7	39.465 6	4.072 7	29.800 8	0.619 9	0.940 0	0.873 0	0.933 0	0.552 0
w/o LQMF(H)	0.509 8	0.723 0	0.913 4	37.225 8	3.449 2	**29.981 2**	0.623 3	0.923 0	0.900 0	0.931 0	0.555 0
w/o TAGF	0.614 1	0.721 0	0.909 6	48.203 1	3.779 3	29.901 1	0.576 6	0.909 0	**0.909 0**	0.931 0	0.549 0
w/o text	0.601 9	0.709 8	0.991 8	44.329 2	3.936 4	29.481 8	0.596 1	**0.959 0**	0.873 0	**0.941 0**	0.569 0
Ours	**0.628 4**	0.724 0	**1.064 3**	**49.652 2**	**4.460 4**	29.757 4	**0.629 9**	0.925 0	0.887 0	0.932 0	**0.594 0**

稳定的信息表达和任务性能。去除 TAGF 后，VIF，MI 和 mAP@50-95 均下降，表明缺少门控机制会削弱红外与可见光信息的自适应选择能力。不输入文本时，虽然 P 和 mAP50 较高，但 Qabf+，VIF，PSNR，MI 和 mAP@50-95 均低于完整模型，说明文本引导有助于提升语义相关区域的融合质量和更严格检测条件下的任务表征能力。

3.3.2　损失函数有效性分析

为验证损失函数中各组成项的有效性，本文设计了损失函数消融实验，分别移除强度约束损失 L_{int}、梯度损失 L_{grad}、文本引导的区域一致性损失 L_{text} 和门控图正则损失 L_a，并与完整模型进行对比，结果如表 7 所示。考虑到不同评价指标关注重点不同，本文选取 Qabf+，VIF，MI，PSNR 和 mAP@50-95 作为代表性指标进行分析。从表 7 可以看出，完整模型在 Qabf+，VIF 和 mAP@50-95 上取得较优结果，说明完整损失函数有助于提升边缘保持、视觉信息保真和严格检测性能。去除 L_{int} 后，VIF，MI，PSNR 和 mAP@50-95 均低于完整模型，说明强度约束对维持融合图像的亮度分布和跨模态信息保留具有重要作用。去除 L_{grad} 后，Qabf+ 和 VIF 明显下降，表明梯度损失能够有效增强边缘结构和细节纹理的保持能力。去除 L_{text} 后，MI 和 PSNR 反而达到最高，说明缺少文本区域约束时，模型更倾向于保留低层统计信息和像素级相似性。但是 Qabf+ 和 mAP@50-95 均低于完整模型，表明仅提高互信息或像素级相似性并不一定带来更好的任务相关融合效果。去除 L_a 后，Qabf+ 与完整模型较为接近，但 VIF，MI 和 mAP@50-95 均下降，说明门控图正则有助于稳定模态选择过程，提高融合结果的视觉保真和下游任务适应性。

表 7　LLVIP 数据集上的损失函数验证实验结果

Tab. 7　Ablation study of different loss functions on the LLVIP dataset

方法	Qabf+	SSIM+	VIF	SD	MI	PSNR	Q^{CB}	P	R	mAP50	mAP@50-95
w/o L_{int}	0.615 9	0.722 8	0.810 6	35.986 7	2.740 5	28.869 9	0.539 5	0.931 0	**0.906 0**	**0.944 0**	0.586 0
w/o L_{grad}	0.596 8	0.727 1	0.851 8	38.604 1	3.534 6	29.568 8	0.562 6	0.903 0	0.889 0	0.929 0	0.562 0
w/o L_{text}	0.615 3	0.719 5	0.995 5	38.436 8	**4.675 0**	**30.441 9**	**0.676 7**	**0.934 0**	0.877 0	0.931 0	0.553 0
w/o L_a	0.626 1	**0.727 8**	1.015 3	43.007 7	3.983 6	29.974 3	0.626 5	0.914 0	0.879 0	0.923 0	0.554 0
Ours	**0.628 4**	0.724 0	**1.064 3**	**49.652 2**	4.460 4	29.757 4	**0.629 9**	0.925 0	0.887 0	0.932 0	**0.594 0**

4　结　论

本文针对开放式文本指令驱动的红外-可见光图像融合需求，提出了一种基于可学习频带分解与文本引导门控的语义可控红外与可见光图像融合框架，旨在充分挖掘文本语义先验并提升融合结果的目标辨识度与细节保真性。本文将文本-图像对齐与融合生成统一建模：通过 CLIP 与 Grounded SAM 协同构建对齐模块，将文本语义精准映射到图像局部区域，实现可解释的语义定位与约束。为增强跨模态信息互补性，设计专属频带分解模块，自适应分离表征全局结构的低频基底与刻画精细纹理的高频细节。进一步提出文本感知门控融合模块，在空间维度自适应调控红外与可见光特征贡献权重，实现面向不同语义目标的可控融合。与此同时，结合多约束损失函数，显式保障结构完整性与文本语义一致性。实验结果表明，本文方法在 LLVIP 数据集上 mAP50 和 mAP@50-95 分别达到 0.932 0 和 0.594 0，且根据不同文本产生不同的融合效果图像证明文本能够有效引导融合生成过程，实现真正的语义可控融合效果。

作者贡献声明：

柳长源：论文方法构思和撰写，论文审核；

程兵云：实验设计及数据分析，论文编辑；

吴海滨：为项目研究成果成功发表而去争取并获得资助。

参考文献:

[1] 杨艳春,李佳龙,李毅,等. 夜间红外与可见光多尺度信息注入式图像融合[J]. 光学 精密工程, 2025, 33(2): 282-297.
YANG Y C, LI J L, LI Y, *et al*. Multiscale semantic injective fusion of nighttime infrared and visible [J]. *Opt. Precision Eng.*, 2025, 33(2): 282-297. (in Chinese)

[2] 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.

[3] 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*, 2025, 59(1): 18.

[4] 龙志亮,邓月明,谢竞,等. 改进多尺度结构化融合的红外与可见光图像融合[J]. 光学 精密工程, 2024, 32(7): 1101-1110.
LONG Z L, DENG Y M, XIE J, *et al*. Infrared and visible images fusion based on improved multi-scale structural fusion [J]. *Opt. Precision Eng.*, 2024, 32(7): 1101-1110. (in Chinese)

[5] 任鹏百,雷慧云,党建武,等. 基于HMSD与改进PCNN的红外与可见光图像融合[J]. 光学 精密工程, 2025, 33(9): 1481-1495.
REN P B, LEI H Y, DANG J W, *et al*. Infrared and visible image fusion based on HMSD and improved PCNN[J]. *Opt. Precision Eng.*, 2025, 33(9): 1481-1495. (in Chinese)

[6] 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.

[7] 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.

[8] 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*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 5792-5801.

[9] CHENG C Y, XU T Y, WU X J. MUFusion: a general unsupervised image fusion network based on memory unit [J]. *Information Fusion*, 2023, 92: 80-92.

[10] 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*). *October* 1-6, 2023, *Paris*, *France*. IEEE, 2024: 8048-8059.

[11] YI X P, ZHANG Y B, XIANG X Y, *et al*. LUT-Fuse: towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables[C]. 2025 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 19-25, 2025, *Honolulu*, *HI*, *USA*. IEEE, 2026: 14559-14568.

[12] 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)

[13] 陶侯卓,罗玉婷,赵凡. 目标语义分层驱动的红外和可见光图像融合[J/OL]. 液晶与显示, 2025,1-12.
TAO Y, LUO Y, ZHAO F. Target semantic hierarchy driven infrared and visible image fusion[J/OL]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 1-12.

[14] WANG Y H, MIAO L J, ZHOU Z Q, *et al*. Infrared and visible image fusion with language-driven loss in CLIP embedding space [C]. *Proceedings of the 33rd ACM International Conference on Multimedia*. *Dublin Ireland*. ACM, 2025: 1443-1451.

[15] 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* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*, *USA*. IEEE, 2024: 27016-27025.

[16] 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.

[17] WANG Z Y, ZHANG J Z, SONG H Y, *et al*. Highlight what you want: weakly-supervised instance-level controllable infrared-visible image fu-

sion[C]. 2025 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 19-25, 2025, *Honolulu*, *HI*, *USA*. IEEE, 2026: 12637-12647.

[18] LIU S L, ZENG Z Y, REN T H, *et al*. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection[C]. *Computer Vision – ECCV* 2024. *Cham*: Springer, 2025: 38-55.

[19] KIRILLOV A, MINTUN E, RAVI N, *et al*. Segment anything[C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 1-6, 2023, *Paris*, *France*. IEEE, 2024: 3992-4003.

[20] 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* (*ICCVW*). *October* 11-17, 2021, *Montreal*, *BC*, *Canada*. IEEE, 2021: 3489-3497.

[21] TOET A. The TNO Multiband Image Data Collection[J]. *Data in Brief*, 2017, 15: 249-251.

[22] 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.

[23] 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* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 13955-13965.

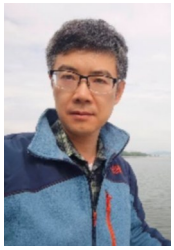
[24] DENG Y L, XU T Y, CHENG C Y, *et al*. MMDRFuse: distilled mini-model with dynamic refresh for multi-modality image fusion[C]. *Proceedings of the* 32*nd ACM International Conference on Multimedia*. *Melbourne VIC Australia*. ACM, 2024: 7326-7335.

[25] 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* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 5906-5916.

[26] 唐霖峰, 张浩, 徐涵, 等. 基于深度学习的图像融合方法综述[J]. 中国图象图形学报, 2023, 28(1): 3-36.
TANG L F, ZHANG H, XU H, *et al*. Deep learning-based image fusion: a survey[J]. *Journal of Image and Graphics*, 2023, 28(1): 3-36. (in Chinese)

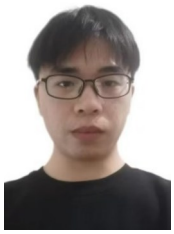
[27] LI X, WANG J, CHEN W W, *et al*. Infrared and visible image fusion based on text-image core-semantic alignment and interaction[J]. *Digital Signal Processing*, 2025, 163: 105203.

通讯作者：



柳长源(1970—),男,黑龙江哈尔滨人,博士,副教授,硕士生导师,2013年于哈尔滨工程大学获得博士学位,2016-2017年普渡大学计算机与电子工程系访问学者,主要从事模式识别、人工智能与机器学习、数字图像处理领域研究。E-mail: liuchangyuan@hrbust.edu.cn

作者简介：



程兵云(2001—),男,江西上饶人,硕士研究生,2024年于江西科技师范大学获得学士学位,现就读于哈尔滨理工大学,主要从事图像融合。E-mail: 2420600038@stu.hrbust.edu