动态时序扩散和多层级细化的红外与可见光图像融合网络

邸敬[1]，李涵[1]，石淑慧[1]，刘冀钊[2]，廉敬[1]
（1.兰州交通大学电子与信息工程学院，甘肃 兰州 730070；2.兰州大学信息科学与工程学院，甘肃 兰州 730070）

摘　要：针对现有扩散模型进行红外与可见光图像融合时出现的时序梯度失稳、图像细节保留不充分等问题，提出了一种动态时序扩散和多层级细化的红外与可见光图像融合网络。首先，在动态去噪时序扩散融合模块中，使用图神经网络中间模块精准捕捉局部特征并建模特征区域间的复杂关系。然后，使用动态去噪路径规划方法对反向去噪过程进行最优去噪路径划分。接着，根据最优动态去噪路径使用分段解码器进行针对性优化去噪。最后，构建了跨层级特征拼接的多层级细化模块，再次提取源图像特征与初步融合图像实现进一步的细化融合。与9种代表性融合方法进行主客观分析比较，在MSRS、M3FD和RoadScene数据集上的实验结果表明，所提方法在7个客观评价指标上都有一定的提升，在各种复杂的场景下表现优异，图像细节保留更加充分，符合人眼视觉特征。此外，目标检测下游实验也展现出所提方法的实际应用潜力。

关键词：图像融合；红外与可见光；动态时序扩散；图神经网络；多层级细化模块
中图分类号：TP391
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025177

Dynamic temporal diffusion and multi-level refinement network for infrared and visible image fusion

DI Jing[1], LI Han[1], SHI Shuhui[1], LIU Jizhao[2], LIAN Jing[1]
1. School of Electronic & Information Engineering, Lan Zhou Jiao Tong University, Lanzhou 730070, China
2. School of Information Science & Engineering, Lan Zhou University, Lanzhou 730070, China

Abstract: To address the problems of temporal gradient instability and inadequate image detail preservation in existing diffusion models for infrared and visible image fusion, a dynamic temporal diffused and multi-level refined network for infrared and visible image fusion was proposed. Firstly, in the diffusion fusion module, a graph neural network middle block was employed to precisely capture local features and model complex relationships between feature regions. Then, a dynamic denoising path planning method was developed to achieve optimal partitioning of the reverse denoising process. Subsequently, a segmented decoder was designed based on the optimal dynamic denoising path to perform targeted optimization denoising of the diffusion process. Finally, a multi-level refinement module with cross-hierarchical feature was constructed to perform further refined fusion through re-extraction of source image features and preliminary denoised fused images. An analysis and comparison with 9 representative fusion methods, both subjectively and objectively, on the MSRS, M3FD, and RoadScene datasets show that the proposed method achieves improvements in 7 objective evaluation metrics. The proposed method has excellent performance in complex scenes, enhances the preservation of image details, and can better conform to human visual characteristics. Furthermore, practical applicability is validated through downstream object detection experiments.

Keywords: image fusion, infrared and visible, dynamic temporal diffusion, graph neural network, multi-level refinement module

收稿日期：2025–07–10；**修回日期：**2025–10–04
通信作者：李涵，2438813120@qq.com
基金项目：国家自然科学基金资助项目（No.62565013）；甘肃省自然科学基金资助项目（No.24JRRA231）；甘肃省级科技计划重点研发计划基金资助项目（No.24YFFA024）
Foundation Items: The National Natural Science Foundation of China (No.62565013), The Natural Science Foundation of Gansu Province (No.24JRRA231), Gansu Provincial Science and Technology Program Key Research and Development Program Project (No.24YFFA024)

0 引言

多模态图像融合是指将不同传感器或模态的图像信息进行整合，以获得包含更多互补信息的图像。红外图像能够捕捉热辐射信息但缺乏细节，可见光图像细节丰富但受光照影响。红外与可见光图像融合作为多模态图像融合的重要分支，突破了单一模态的局限性，整合2种模态的细节信息，为监控安防[1]、目标检测[2]、道路交通[3]等关键领域提供可靠的技术支撑，具有重要的实际意义。

图像融合方法主要分为传统方法[4-6]和深度学习方法两大类，深度学习方法相较于传统方法有更强的特征提取能力，适应性更强。随着深度学习的发展，基于深度学习的融合方法已成为图像融合领域的主流方法。现有基于深度学习的图像融合算法包括基于自编码器[7-8]（AE, autoencoder）、基于卷积神经网络[9-13]（CNN, convolutional neural network）、基于Transformer[14-16]、基于生成对抗网络[17-19]（GAN, generative adversarial network）和基于去噪扩散概率模型[20-24]（DDPM, denoising diffusion probabilistic model）等。文献[7]提出了卷积层与密集块结合的编码器结构增强特征提取，但融合策略依赖手工设计的规则，无法充分挖掘特征的互补性，故融合结果不理想。文献[9]根据红外热辐射信息设计了显著目标掩模，为不同模态信息融合提供了空间引导，但模型过度依赖于掩模质量，影响融合的可靠性。文献[10]引入残差融合模块，缓解了融合过程中信息丢失和梯度消失问题。文献[14]构建了Swin Transformer架构，但基于L1范数活动级别图的融合策略导致信息保留不均衡。文献[15]提出了基于Transformer的紧凑型多模态图像融合方法，但该方法使用迭代混合注意力机制需要多阶段优化，导致计算复杂度升高。文献[16]提出了一种基于掩码注意力机制的通用图像融合框架，解决多模态图像融合过程中动态权重分配问题。文献[17]提出了基于GAN的融合框架FusionGAN，利用生成器和判别器的博弈保留红外图像的热辐射和可见光图像的纹理细节。文献[18]提出了基于纹理条件的生成对抗网络TC-GAN，该网络能够有效融合红外与可见光图像。文献[19]设计了双对抗融合网络，主要面向目标检测任务。虽然基于GAN的图像融合模型性能良好，但它们存在训练不稳定、缺乏可解释性等问题，严重影响生成样本质量。

为了应对这一挑战，保证图像生成过程的稳定性和可解释性，文献[20]提出了DDPM，该模型训练过程更加稳定，表征学习能力更强，在复杂数据上的生成质量更高。在DDPM采样框架下，文献[21]提出了多模态图像融合DDFM，该模型使用预训练权重，通过期望最大化算法生成融合图像，但其权重不适用于特定融合任务导致生成图像质量较差。文献[22]针对现有扩散模型缺乏基础真实值的问题，提出了基于融合知识先验的多模态图像融合框架Diff-IF，通过融合知识先验引导前向扩散，生成符合融合知识先验概率分布的融合图像。文献[23]提出了高效潜在特征引导的扩散模型，使用像素空间自动编码器和基于Transformer的紧凑扩散网络实现融合任务。文献[24]在使用扩散模型实现融合的基础上加入了文本动态调控融合过程。然而，扩散模型的反向去噪过程会因时间步梯度差异造成时序梯度失稳，进而影响图像细节保留。针对以上问题，本文首次尝试联合扩散模型和图神经网络（GNN, graph neural network）增强空间协调性，使用多层级细化模块，在加强训练稳定性的同时提升了图像融合质量。本文主要贡献如下。

1) 提出了一种动态时序扩散和多层级细化的红外与可见光图像融合网络，该网络结合动态去噪时序扩散和多层级细化模块实现了图像融合质量的提升。

2) 提出了一种图神经嵌入的动态时序扩散融合模块，该模块使用基于GNN的U-Net中间模块，并使用动态去噪路径规划方法实现最优去噪路径的选择，同时分段解码器针对不同去噪路径聚焦不同任务，提高阶段性去噪效果，缓解了反向去噪过程的时序梯度失稳问题，实现了红外与可见光图像的初步融合。

3) 提出了一种多层级细化模块，该模块通过双分支编码架构和跨层级特征拼接增强红外与可见光初始细节特征和修复特征的进一步融合，能够有效保留2种模态源图像的结构和细节信息，提高了网络的融合性能。

4) 使用MSRS、M3FD和RoadScene的红外与可见光图像数据集进行实验分析。结果表明，本文方法在图像细节保留、融合质量等方面优于其他9种代表性融合方法。

1　模型设计

1.1　总体框架

为了提高红外与可见光图像的融合质量，缓解扩散模型在红外与可见光图像融合任务反向去噪过程中出现的时序梯度失稳现象，提升训练稳定性，增强图像细节保留，本文提出了一种动态时序扩散和多层级细化的红外与可见光图像融合网络，该网络首次尝试联合扩散模型和图神经网络实现 t 步去噪融合，总体框架如图 1 所示。首先，利用共享编码器提取不同模态特征并学习不同模态间的对齐关系，通过图结构聚合全局上下文，增强空间协调性。然后，通过计算时间步损失的梯度相似性，将扩散过程动态划分最优去噪路径，再根据不同扩散阶段设计分段解码器，使模型能够动态适配不同扩散阶段。在特征重建过程中，利用分段解码器进行针对性优化去噪融合。最后，将分段解码器输出的初步去噪融合图像和红外与可见光图像同时送入多层级细化模块，处理初步去噪融合图像和原始多模态输入，增强跨模态特征交互，通过多层级处理恢复图像的清晰纹理，进一步细化初步融合结果，最终生成更高融合质量的红外与可见光图像。

1.2　图神经嵌入的动态时序扩散融合模块

扩散模型是通过定义一个前向扩散过程和一个反向去噪过程来实现数据生成。在前向扩散过程中，模型对真实数据逐步添加高斯噪声，该过程可以表示为一个马尔可夫链，每个时间步的状态由前一个时间步的状态添加噪声得到。在时间步 t 处，加噪图像 $f_t(i,v)$ 可以表示为

$$q(f_t|f_0)=N\left(f_t(i,v);\sqrt{\bar{\alpha}_t}\,f_0(i,v),(1-\bar{\alpha}_t)I\right) \tag{1}$$

其中，$f_t(i,v)$ 和 $f_0(i,v)$ 分别表示当前 t 步的加噪图像和原始图像，i 和 v 分别表示红外图像和可见光图像，$\bar{\alpha}_t$ 表示前 t 步所有噪声方差之积，I 表示标准正态分布，$N(f_t;\cdot,\cdot)$ 表示 f_t 服从高斯分布。

在反向去噪过程中，通过逐步去噪从噪声中恢复原始数据，在已知第 t 步噪声图像 f_t 和源图像 f_0 的条件下，可以推导出上一个时间步 $t-1$ 的噪声图像 f_{t-1} 为

$$p_\theta(f_{t-1}|f_t f_0)=N\left(f_{t-1};\mu_\theta(f_t,t),\frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)I}{(1-\bar{\alpha}_t)}\right) \tag{2}$$

其中，α_t 表示第 t 步的噪声方差，$\mu_\theta(f_t,t)$ 表示噪声图像 f_{t-1} 的条件分布均值。

在扩散模型中，通常使用 U-Net 结构实现反向去噪，噪声图像经过编码器进行特征提取，然后将提取到的特征送入中间层，对编码器输出的最深层次特征进行进一步处理，提取更复杂的抽象特征作为解码器的初始输入，最后进行特征上采样，逐步恢复空间分辨率，将抽象特征映射回原始图像尺寸。在反向去噪过程中，梯度在跨越多步传递时可

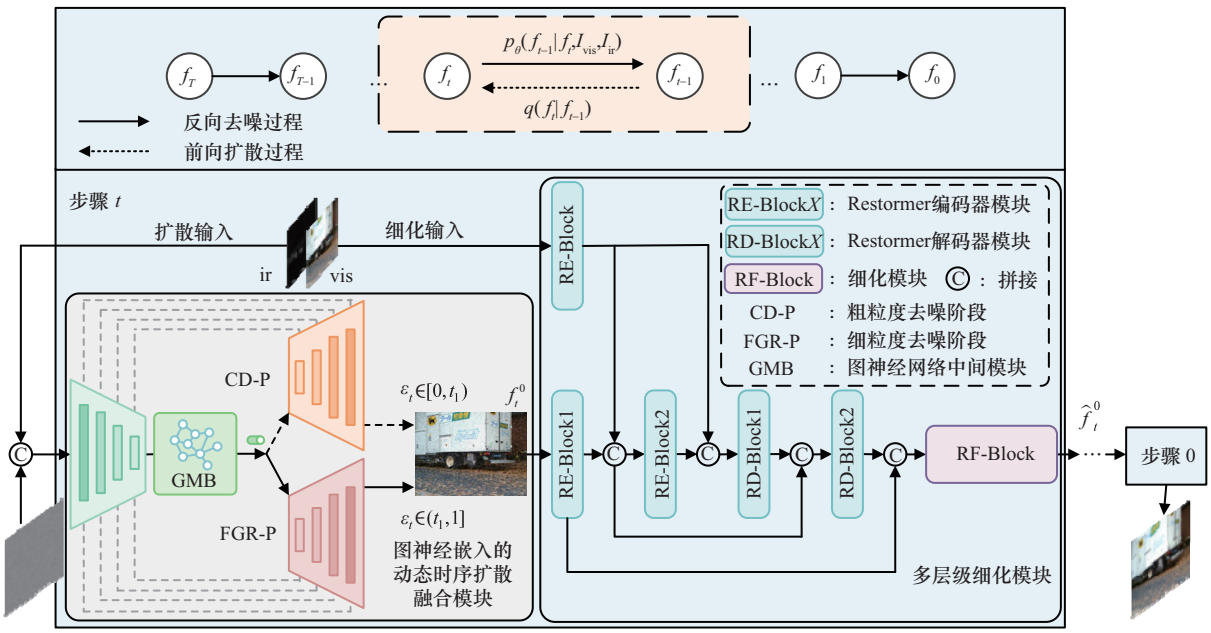


图 1　动态时序扩散和多层级细化的红外与可见光图像融合网络总体框架

能会出现梯度爆炸或消失的失稳问题。早期粗粒度去噪阶段（CD-P, coarse denoising-phase）与后期细粒度优化阶段（FGR-P, fine grained refinement-phase）损失梯度分布差异显著，U-Net 中的单一解码器难以兼顾，而传统的卷积中间层受限于局部感受野，难以灵活捕获不规则特征和复杂特征。针对以上问题，本文提出了基于图神经网络中间模块（GMB, GNN middle block）和分段解码器 U-Net 架构的时序扩散融合模块。首先使用共享编码器进行图像特征提取，再通过图神经网络中间模块，利用节点间消息传递聚合全局上下文，然后通过动态去噪路径规划划分反向去噪阶段，最后使用分段解码器使模型动态适配不同的扩散阶段，利用分段解码器进行针对性优化去噪，得到初步融合图像，图神经嵌入的动态时序扩散融合模块如图 2 所示。

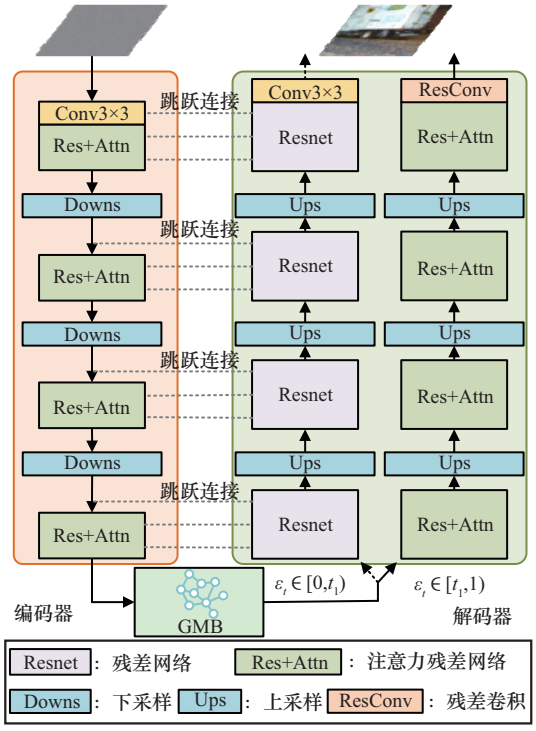


图 2　图神经嵌入的动态时序扩散融合模块

1.2.1　图神经网络中间模块

传统 U-Net 使用 CNN 进行特征提取和重建，但是 CNN 作为中间层时深层全局特征可能会被局部细节稀释，受限于局部感受野，难以捕捉长程依赖关系，本文使用 CNN 和 GNN 混合的 U-Net，CNN 提取丰富的局部特征表示，GNN 作为中间层建模特征图区域之间的复杂关系。GNN 对特征图中的不同特征建立连接，提供更好的灵活性，利用较大的感受野和灵活的图卷积实现图像中的远程信息交互。在红外与可见光图像融合任务中，通常需要保存 2 种模态的局部信息，尤其是红外图像中的热辐射信息和可见光图像中的纹理细节信息，GNN 的消息传递机制通过聚合邻域节点信息能够直接建模和传播 2 种模态之间的依赖关系。

噪声图像在经过编码器提取特征并进行下采样后，通过 GNN 中间层的动态图卷积有效捕捉特征关系，增强特征表示。中间层将输入特征输入到第一个全连接层提取高阶特征，为图卷积做准备，然后将提取到的高阶特征转换为图结构，将深层特征图中的特征点作为动态图中节点，通过计算节点特征相似度建立节点间的连接。

$$D_{i,j}=\left\| x_i - x_j \right\|_2^2 \tag{3}$$

其中，x_i 和 x_j 分别表示节点 i 和 j 的特征向量。在得到节点特征相似度后，每个节点选择特征相似度最高的 $K\times d$ 个候选邻居，从候选邻居中等间隔选取 K 个间隔为 d 的邻居，扩大感受野，也可以通过增大间隔 d 使节点关注更远距离的语义相关区域。邻接关系为

$$\mathcal{N}_i=\mathrm{DilatedKNN}\left(x_i;d,K\right) \tag{4}$$

其中，K 表示每个节点选择 K 个邻居，d 表示稀疏化因子，$\mathcal{N}_i$ 表示节点 i 特征相似度最高的 K 个邻居节点的集合。

通过以上方法构建好邻接关系矩阵后再执行图卷积，本文选用边卷积捕捉中心节点与其邻居节点之间的局部差异来聚合邻居特征，增强特征表示。首先拼接中心节点特征和邻居–中心差异特征，保留中心节点原始特征的同时捕捉局部结构差异，再通过多层感知机（MLP, multi-layer perceptron）对拼接后的特征进行非线性映射组合中心特征与差异信息，提取高阶局部模式，学习更复杂的节点间关系，构建邻接边特征。

$$m_{i,j}=\mathrm{MLP}\left(x_i,x_j-x_i\right) \tag{5}$$

其中，x_i 表示中心节点特征，x_j-x_i 表示邻居–中心差异特征，MLP 表示可学习的多层感知机。

然后通过聚合邻居特征更新节点信息，增强节点的特征表示。

$$x_i' = \sum_{j \in \mathcal{N}_i} m_{i,j} \tag{6}$$

最后经过全连接层和残差连接恢复图特征并还原图像形状。

1.2.2　动态去噪路径规划

当使用扩散模型实现图像融合任务时，反向去噪过程中时间步的损失梯度分布差异显著，同时梯度在长距离反向传播中会出现梯度爆炸或消失的失稳问题，时序梯度动态失衡将引发训练震荡和特征退化问题。针对这一问题，本文提出的动态去噪路径规划和分段解码器协同作用，动态去噪路径规划基于损失曲线的动态特性，通过梯度相似性和动态规划实现时间步的最优分割，将反向去噪过程划分为CD-P和FGR-P这2个阶段，动态选择最优去噪路径，提升训练稳定性。

扩散模型在不同时间步t上的损失值序列为$\{L_t\}_{t=1}^{T}$，使用Savitzky-Golay滤波器对损失值序列进行平滑处理得到$\tilde{L}_t$，再对平滑处理后的损失值序列计算梯度并进行梯度归一化得

$$g_t = \nabla \tilde{L}_t = \tilde{L}_{t+1} - \tilde{L}_t, t = 1,2,\cdots,T-1 \tag{7}$$

$$\hat{g}_t = \frac{g_t}{\left(\| g \|_2 + \varepsilon \right)}, \varepsilon = 10^{-8} \tag{8}$$

其中，T表示总时间步，g_t表示损失值梯度，$\hat{g}_t$表示归一化后时间步t处的梯度，ε表示平滑项。

对任意时间区间$[i,j]$，定义其代价为段内梯度与均值的偏离平方和，用以衡量区间内梯度的一致性，代价越小表示该区间越适合作为一个聚类。预计算所有可能的区间代价并构建代价矩阵$\boldsymbol{C} \in \mathbb{R}^{T \times T}$为

$$\boldsymbol{C}(i,j) = \sum_{m=i}^{j} \left(\hat{g}_m - \mu_{i,j} \right)^2 \tag{9}$$

$$\mu_{i,j} = \sum_{n=i}^{j} \frac{\hat{g}_n}{(j - i + 1)} \tag{10}$$

$$\boldsymbol{C}(i,j) = \boldsymbol{C}(i,j-1) + \left(\hat{g}_j - \mu_{i,j} \right)^2 \tag{11}$$

其中，$\boldsymbol{C}(i,j)$表示时间区间$[i,j]$的代价，$\mu_{i,j}$表示时间区间$[i,j]$的梯度均值。

得到代价矩阵后，通过动态规划递推寻找最优分割点，当聚类数$k=1$时，所有时间步属于一个聚类。

$$\mathrm{dp}(1,t) = \boldsymbol{C}(0,t), t = 0,1,\cdots,T-1 \tag{12}$$

当$k \geqslant 2$时，对每个时间步t寻找最优分割点s，即

$$\mathrm{dp}(k,t) = \min_{s \in [k-1, t-1]} \left[dp(k-1,s) + \boldsymbol{C}(s+1,t) \right] \tag{13}$$

并记录分割点，如式(14)所示。

$$\mathrm{partition}(k,t) = \arg\min_{s} \left[\mathrm{dp}(k-1,s) + \boldsymbol{C}(s+1,t) \right] \tag{14}$$

其中，$\mathrm{dp}(k,t)$表示前t个时间步划分为k个聚类的最小代价，s表示$[k-1,t-1]$时间步内最优分割点，$\mathrm{partition}(k,t)$用来存储总代价最小的分割点位置。

存储每一步的最佳决策后，进行反向追踪分割点，从最后一个时间步$t = T-1$和最大聚类数k开始，根据$\mathrm{partition}(k,t)$找到最优分割点s并将其加入分割点列表，随后更新$t = s$并减少k，直到$k=1$，最后在分割点列表中补充首尾时间步，排序去重，得到最终的聚类区间，即最优动态去噪路径。

1.2.3　分段解码器模块

经过GNN中间层提取到高阶特征后，使用分段解码器重构特征，在此之前，本文提出的动态去噪路径规划已将扩散去噪过程划分为CD-P和FGR-P这2个阶段，对于不同去噪阶段，使用分段解码器进行针对性优化去噪融合。分段解码器通过动态路径选择适配不同阶段需求，关注特定阶段的梯度，避免长距离依赖，缓解时序梯度失稳。在粗粒度去噪阶段主要消除全局噪声，使用简单解码器快速捕捉全局特征，建立数据的基本语义结构，而细粒度优化阶段需要精准修复局部细节，聚焦纹理和边缘细节，对此使用深度结构，在简单解码器中加入自注意力机制，捕捉全局特征的同时恢复图像细节，再通过残差输出保留输入图像的保真度。

1.3　多层级细化模块

经过动态时序扩散融合模块后得到初步去噪融合图像，由于该图像存在细节信息缺失和整体结构信息较差等问题，本文提出了一种多层级细化模块，将初步去噪融合图像，以及红外与可见光源图像三者共同送入多层级细化模块进一步对融合图像进行层级细化，通过在反向去噪过程中对图像初步融合结果逐步引入原始图像的语义信息，避免在融合图像中丢失细节和边缘信息或丢失全局语义内容。本文使用双分支编码–融合–解码架构和跨层级特征拼接实现多层级细化，编解码器使用由多头深度可分离卷积自注意力（MDTA, multi Dconv

head transposed attention）和门控深度可分离卷积前馈网络（GDFN, gated Dconv feed forward network）构成的 Restormer 编解码器模块（RE/D-Block, Restormer encoder/decoder block），该模块如图 3 所示，再通过一个自注意力细化模块（SARB, self-attention refinement block）进一步细化集成特征，自注意力细化模块如图 4 所示。

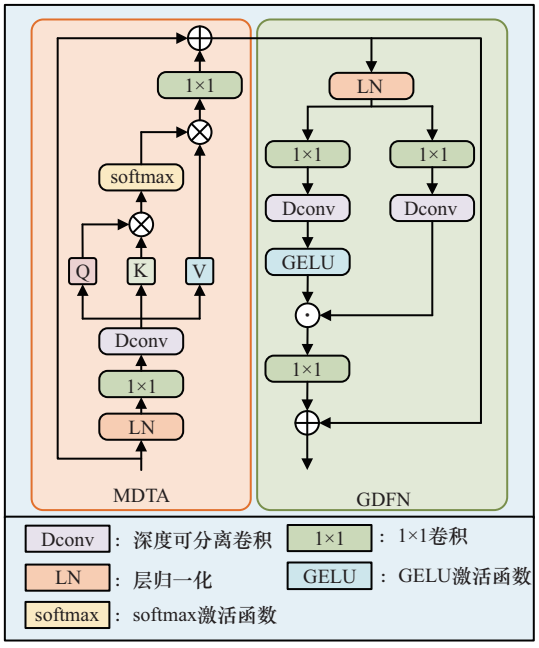


图 3　Restormer 编解码器模块

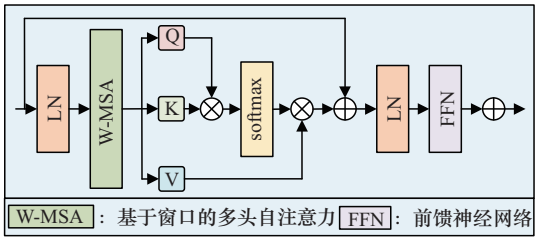


图 4　自注意力细化模块

在多层级细化模块中，将初步去噪融合图像 f_t^0 作为主分支输入，送入第一级主分支编码器 RE-Block1，同时将红外与可见光源图像作为细化分支输入，送入细化分支编码器 RE-Block，随后将 2 个分支的输出进行特征拼接，送入第二级主分支编码器 RE-Block2，对初步融合图像特征和多模态细节信息进行融合。在编码融合后，将编码器输出送入第一级解码器 RD-Block1 实现多层级特征重构，在该级解码器中实现主分支编码器输出的融合特征和细化分支编码器输出的多模态特征的拼接，结合高层语义信息和编码阶段保留的底层细节信息，实现全局结构引导局部修复的多层级协作，最后将第一级解码器输出与第一级编码器输出进行特征拼接，送入第二级解码器 RD-Block2 中整合修复特征和最原始的编码特征。通过双分支编码–融合–解码架构完成跨层级特征修复融合后，将该融合特征输入 SARB，通过建模长距离依赖关系增强全局信息整合，生成红外与可见光融合图像。

1.4　损失函数

本文使用扩散模型实现红外与可见光图像融合，损失函数包含 3 个主要组成部分：噪声预测损失、重建先验损失和一般损失，整体损失为

$$L = L_{\text{simple}} + \lambda_{\text{eps}} L_{\text{eps}} + \lambda_{x_0} L_{x_0} \tag{15}$$

其中，L_{simple} 为一般损失，L_{eps} 为噪声预测损失，L_{x_0} 为重建先验损失，λ_{eps} 和 λ_{x_0} 为调节参数。

在反向去噪过程中，预测噪声与真实噪声之间的差异为噪声预测损失，该损失是训练扩散模型的核心。噪声预测损失 L_{eps} 定义为

$$L_{\text{eps}} = \left\| \epsilon - \hat{\epsilon}_t \right\| \tag{16}$$

其中，ϵ 和 $\hat{\epsilon}_t$ 分别表示模型真实噪声和预测噪声。

模型从噪声中逐步重建出干净图像，重建先验损失代表预测去噪图像与参考图像之间的差异。重建先验损失可以表示为

$$L_{x_0} = \left\| x_{0,t}^f - x_0^{ts} \right\| \tag{17}$$

其中，$x_{0,t}^f$ 表示预测去噪图像，x_0^{ts} 表示参考图像。

在一般融合任务中，强度损失鼓励融合图像在亮度上与输入图像最大值保持一致，用于强化融合图像保留红外图像的显著热目标。梯度损失衡量图像梯度差异，保留清晰的边缘结构和细节结构。通过基于结构相似性（SSIM, structural similarity）指标设计的结构相似性损失来衡量图像在亮度、对比度和结构上的相似性，提升视觉质量。一般损失定义为

$$L_{\text{simple}} = \lambda_{\max} L_{\max} + \lambda_{\text{grad}} L_{\text{grad}} + \lambda_{\text{SSIM}} L_{\text{SSIM}} \tag{18}$$

$$L_{\max} = \left\| x_0^f - \max\left(I_{\text{vis}}, I_{\text{ir}}\right) \right\| \tag{19}$$

$$L_{\text{grad}} = \left\| \nabla x_0^f - \max\left(\nabla I_{\text{vis}}, \nabla I_{\text{ir}}\right) \right\| \tag{20}$$

$$L_{\text{SSIM}} = \left\| 1 - \text{SSIM}\left(x_0^f, I_{\text{ir}}\right) \right\| + \left\| 1 - \text{SSIM}\left(x_0^f, I_{\text{vis}}\right) \right\| \tag{21}$$

其中，$L_{\max}$ 表示强度损失，L_{grad} 表示梯度损失，L_{SSIM} 表示结构相似性损失，x_0^f、I_{ir} 和 I_{vis} 分别表示融合图像、红外源图像和可见光源图像，$\lambda_{\max}$、

λ_{grad}和λ_{SSIM}表示调节参数。

本文综合强度损失、梯度损失、结构相似性损失、噪声预测损失和重建先验损失同时优化多个目标，保证融合图像的保真度、清晰度和整体结构的一致性，提高融合效果。

2　实验结果和分析

本文模型包括动态时序扩散融合模块和多层级细化模块，其训练阶段和测试阶段均在Pytorch框架下实现。所有实验均在Windows 11（64位）系统上进行，硬件配置为16 vCPU AMD EPYC 9K84 96-Core处理器和H20-NVLink GPU。

实验中，选用公开数据集MSRS、M3FD和RoadScene完成红外与可见光图像融合任务。在训练阶段，使用MSRS数据集中1 083张图片进行训练；在测试阶段，选用MSRS数据集中361张、M3FD数据集中300张和RoadScene数据集中221张图片作为测试数据。

2.1　客观评价指标

为说明本文方法的有效性，选择7种客观评价指标来衡量融合结果，包括熵（EN, entropy）[25]、互信息（MI, mutual information）[26]、结构相似性（SSIM）[27]、峰值信噪比（PSNR, peak signal to noise ratio）[28]、标准差（STD, standard deviation）[29]、相关系数（CC, correlation coefficient）[30]和视觉保真度（VIFF, visual information fidelity for fusion）[31]。

EN是衡量图像包含信息量的重要指标，数值越高表示融合图像的信息量越丰富，细节保留越好。MI用来衡量融合图像和源图像之间的信息相关性，反映融合图像保留源图像信息的能力。SSIM用来评估融合图像与源图像的相似性，值越大，表示融合图像的结构特征与源图像越一致。PSNR通过均方误差衡量融合图像的保真度，数值越大，图像保真度越高。STD反映图像灰度分布的离散程度，图像对比度强，细节更突出。CC衡量融合图像与源图像的线性相关性，值越接近1，表明融合图像与源图像的一致性越高。VIFF基于人眼视觉系统评估融合图像的视觉信息损失，视觉信息保留越好，人眼感知质量越高。在进行评价指标分析时需要综合各个指标进行互补分析。

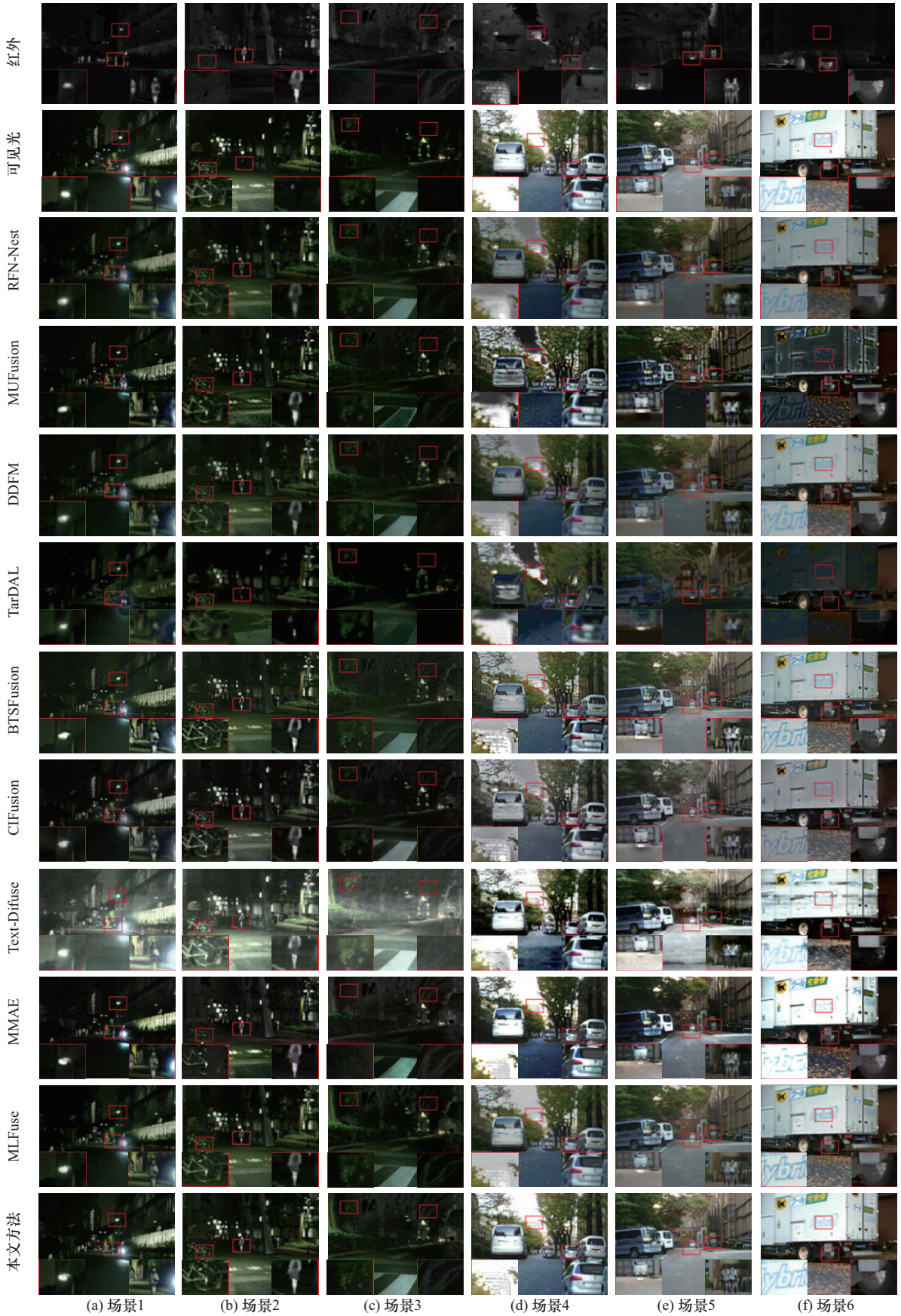
2.2　融合图像分析

为进一步验证本文方法的有效性和实用性，基于模型方法和前沿性选用9种对比方法，分别为RFN-Nest[10]、MUFusion[8]、DDFM[21]、TarDAL[19]、BTSFusion[12]、GIFusion[13]、Text-Difuse[24]、MMAE[16]和MLFuse[32]，并将这9种对比方法从主观视觉和客观评价指标两方面与本文方法进行对比分析。

2.2.1　MSRS数据集主客观评价

从MSRS数据集中选取3张白天和3张夜间具有代表性的红外与可见光图像进行主观对比分析，其中白天图像为01384D、01474D和01527D，夜间图像为00754N、01098N和01194N。图5展示了6组图像的融合结果。通过主观对比显示，RFN-Nest和DDFM模型在保留红外目标信息方面表现不佳，红外热辐射目标的显著特征信息丢失，同时融合结果中可见光图像色彩饱和度降低导致场景呈现灰暗视觉效果。MUFusion和TarDAL模型在红外与可见光图像融合时受红外色彩污染导致合成图像部分场景可见光色彩丢失且红外目标信息缺失严重，如这2个模型在场景4中的高楼内部细节信息表征不明显。BTSFusion和GIFusion模型能够有效保留可见光纹理和红外目标信息，但BTSFusion模型的融合图像存在局部曝光问题，高亮细节丢失，如场景1中路灯和场景5中井盖局部曝光且内部细节丢失，GIFusion模型的融合图像物体边缘不够清晰，如场景1和场景5中人物边缘模糊。Text-Difuse模型融合图像对比度适中，可见光纹理细节保留较好，但对于红外图像中较微弱的热辐射信息保留较差，如场景3中树木轮廓不清晰，场景6中货车轮胎内部细节模糊。MMAE模型能够较好地保留可见光纹理和红外目标信息，但其色彩饱和度增强导致融合图像色彩不自然，局部曝光严重导致部分细节缺失，如场景6中货车上字体因曝光严重而模糊不清，无法辨认。MLFuse模型融合图像质量较高，但存在部分细节缺失，如场景5井盖内部细节不清晰。本文方法在保证色彩保真度的同时，能够保留可见光图像的复杂纹理细节和红外图像的热辐射信息，并且在保留较弱的热特征信息方面表现出更好的效果，对比度和亮度自然，使融合图像展现出更优的视觉质量。

为了使融合结果更具说服力，使用MSRS数据集中361张融合图像进行融合性能和计算复杂度分析，结果如表1所示，其中，加粗数据为最优，加下划线数据为次优。本文方法在EN、MI、SSIM、



(a) 场景1　(b) 场景2　(c) 场景3　(d) 场景4　(e) 场景5　(f) 场景6

图5　MSRS数据集6组图像的融合结果

表 1　　MSRS 数据集 361 张融合图像的客观评价指标

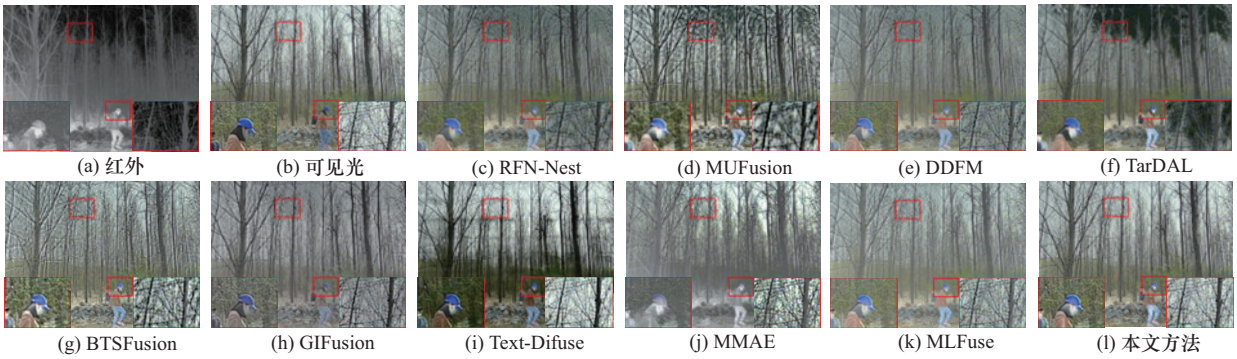
方法	EN	MI	SSIM	PSNR	STD	CC	VIFF	推理时间/s	FLOPs
RFN-Nest	6.304 6	2.729 9	0.650 7	17.545 0	5.519 2	0.936 4	0.539 8	0.231 3	520.81×10^9
MUFusion	5.948 5	1.533 4	0.594 4	16.890 3	5.352 4	0.629 0	0.672 2	0.908 9	$\mathbf{3.01}\times10^9$
DDFM	6.288 6	2.637 5	0.640 1	17.678 3	5.638 0	0.938 1	0.554 1	41.53	$1\,113.75\times10^9$
TarDAL	5.054 1	2.167 2	0.439 7	17.062 7	4.395 4	0.435 8	0.163 6	0.090 0	19.44×10^9
BTSFusion	6.514 0	2.493 8	0.603 7	17.606 9	6.044 4	0.928 6	0.719 4	0.185 8	$\underline{3.36\times10^9}$
GIFusion	6.376 8	2.334 4	0.624 4	17.844 3	5.836 2	0.947 0	0.661 5	0.416 0	37.70×10^9
Text-Difuse	**7.466 0**	2.229 6	0.381 3	10.326 3	**7.876 9**	0.919 5	**0.936 6**	13.63	113.92×10^9
MMAE	6.227 8	<u>3.499 6</u>	<u>0.655 9</u>	16.831 9	6.742 7	0.850 2	0.687 3	<u>0.078 0</u>	42.06×10^9
MLFuse	6.696 8	3.182 7	0.654 0	<u>18.631 1</u>	6.240 9	<u>0.956 5</u>	0.727 8	**0.027 3**	31.00×10^9
本文方法	<u>6.872 6</u>	**4.541 3**	**0.670 9**	**20.997 8**	<u>6.802 8</u>	**0.969 7**	<u>0.814 2</u>	1.414 7	85.31×10^9

PSNR、STD、CC 和 VIFF 等客观评价指标上均表现优异。Text-Difuse 模型的 EN 和 STD 指标较高，融合图像能够保留丰富的热辐射特征和纹理细节，并且增强图像对比度，但这种方法会破坏源图像的自然结构，增加高频噪声的影响，导致 MI、SSIM 和 PSNR 指标显著降低，保留源图像信息的能力下降，而本文方法的 7 个客观评价指标均为最优或次优，综合性能更好。在计算复杂度分析中，本文方法、DDFM 与 Text-Difuse 均属于计算复杂度较高的扩散融合模型，数据表明三者的推理时间和浮点运算量（FLOPs, floating point operations）均处于较高水平，但本文方法在保持融合性能的同时显著降低了计算开销，展现出更优的效率。通过与以上 9 种方法对比，在计算复杂度可接受的合理范围内，本文方法在主客观评价中取得了更优的表现，尤其是在色彩保真度、纹理细节、自然结构保留和融合质量方面，均体现出明显优势。

2.2.2　M3FD 数据集主客观评价

为了进一步验证本文方法的优越性和泛化能力，本文选取 M3FD 数据集进行泛化测试，并从测试数据中选取 2 组代表性图像进行不同方法的主客观对比分析，其中包括树林场景“01393”和道路场景“02846”，2 种场景的融合结果如图 6 和图 7 所示。

通过主观对比发现，RFN-Nest、GIFusion 和 TarDAL 模型保留了良好的可见光纹理细节，但在融合过程中受红外噪声影响，导致图片质量较差，如图 6 中树木背景灰暗且树木枝条轮廓模糊。MU-Fusion 和 BTSFusion 模型的融合图像局部曝光，高亮细节缺失，如图 6 中人物脸部产生局部曝光，脸部细节丢失，同时 MUFusion 模型对比度过高导致色彩不自然，边缘细节模糊，如图 7 中高楼边缘细节不清晰。Text-Difuse 模型在融合过程中引入过多噪声导致图像背景模糊，如图 6 和图 7 中背景噪点增多，图像融合质量不理想。MMAE 模型有效地融合了可见光信息和红外目标信息，但图像出现了严重的曝光导致图 6 中人物脸部信息和色彩丢失。DDFM 模型图像同时保留了可见光纹理和红外信



(a) 红外　(b) 可见光　(c) RFN-Nest　(d) MUFusion　(e) DDFM　(f) TarDAL

(g) BTSFusion　(h) GIFusion　(i) Text-Difuse　(j) MMAE　(k) MLFuse　(l) 本文方法

图 6　M3FD 数据集“01393”树林场景融合结果

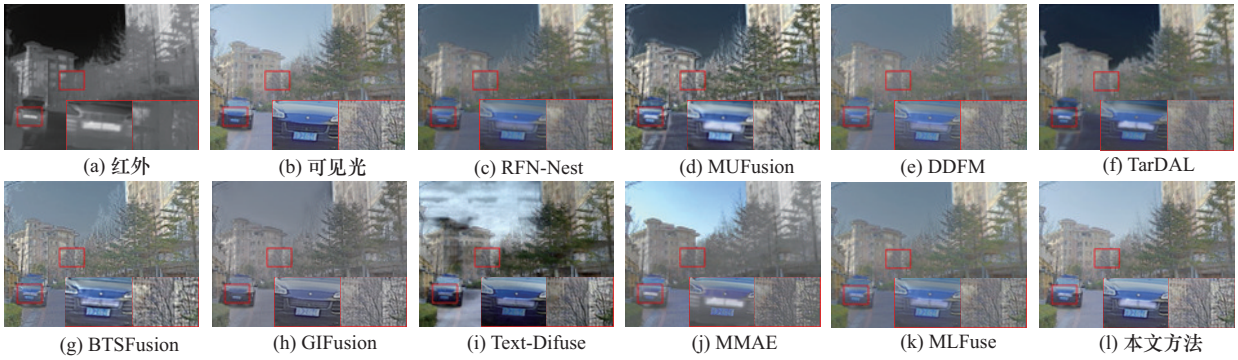


图7　M3FD数据集“02846”道路场景融合结果

息，但色彩保真度较低。MLFuse模型图像细节较清晰，但图像整体呈现雾感。通过对比可以看出，本文方法在有效保留可见光纹理细节和红外热辐射信息的同时，保证了图像的色彩保真度，具有较高的泛化能力，能够更好地完成各种复杂光照和环境下的红外与可见光图像融合任务。

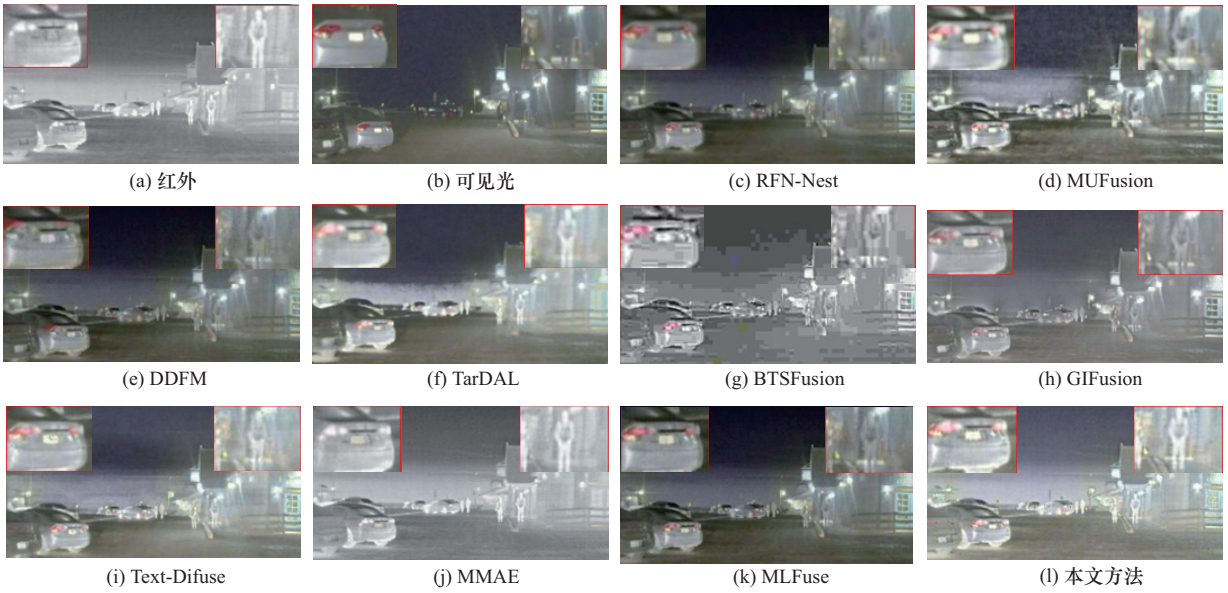
M3FD数据集300张融合图像的客观评价指标如表2所示，其中加粗数据为最优，加下划线数据为次优。数据表明，本文方法在EN、MI、SSIM、PSNR、STD和CC等客观评价指标上取得了较好的结果。MUFusion模型融合图像的VIFF指标较高，而SSIM、MI和CC指标较低，这是由于其对比度过高，符合人眼视觉特性，但这造成了图像局部结构变形和边缘模糊，丢失了源图像的细节信息，过度的视觉优化牺牲了结构一致性和信息完整性。Text-Difuse模型在融合过程中过度强调细节和对比度的增强，使其EN和STD指标显著提高，但该模型忽略了整体结构的保持，破坏了源图像的自然结构，导致SSIM指标显著降低。本文方法综合指标更为优异，仅在VIFF指标上略低于MUFusion和BTSFusion模型，这是因为模型在融合时更注重2种模态的交互，而非提取单一可见光图像的细节信息。本文方法在学习红外图像的热辐射信息和可见光图像丰富纹理细节的同时，能够充分保留整体结构信息，与主观视觉评价一致。此外，本文方法在色彩保真度和泛化性方面也具有较大优势，能够适应不同复杂场景下的红外与可见光图像融合任务。

2.2.3　RoadScene数据集主客观评价

为了进一步验证本文方法的泛化能力和有效性，本文选取RoadScene数据集道路场景进行测试，在该数据集中选取一组代表性图像“FLIR_05879”进行主客观对比分析，结果如图8所示。MUFusion、GIFusion和TarDAL模型融合图像局部曝光，高亮细节缺失。RFN-Nest、DDFM和Text-Difuse模型融合图像

表2　M3FD数据集300张融合图像的客观评价指标

方法	En	MI	SSIM	PSNR	STD	CC	VIFF
RFN-Nest	6.924 8	1.589 9	0.546 4	14.402 8	5.808 3	0.340 9	0.372 0
MUFusion	7.501 5	1.547 7	0.505 1	13.384 5	6.987 7	0.392 3	**0.670 1**
DDFM	6.778 4	1.671 3	0.564 5	14.466 7	5.481 9	0.489 9	0.374 9
TarDAL	7.260 0	2.566 2	0.549 1	14.981 7	6.534 5	0.152 8	0.283 6
BTSFusion	7.435 1	1.557 1	0.522 4	13.979 2	6.708 4	0.705 1	0.556 0
GIFusion	7.112 1	2.281 5	0.560 8	14.925 1	5.934 3	0.830 7	0.296 7
Text-Difuse	**7.851 9**	2.299 8	0.480 9	12.289 3	**8.157 6**	0.893 8	0.351 4
MMAE	7.300 4	3.468 3	0.554 9	13.291 5	6.856 5	0.720 0	0.198 5
MLFuse	7.018 9	1.957 1	0.561 7	14.150 1	5.822 1	0.696 3	0.373 3
本文方法	7.536 9	**4.057 7**	**0.571 5**	**16.570 4**	6.998 5	**0.941 7**	0.412 5



(a) 红外　(b) 可见光　(c) RFN-Nest　(d) MUFusion

(e) DDFM　(f) TarDAL　(g) BTSFusion　(h) GIFusion

(i) Text-Difuse　(j) MMAE　(k) MLFuse　(l) 本文方法

图 8　RoadScene 数据集“FLIR_05879”道路场景融合结果

更好地融合了红外与可见光信息，但边缘纹理细节模糊，图中人物边缘轮廓细节丢失。BTSFusion 和 MMAE 模型在融合过程中受红外噪声影响，引入过多噪声导致图像模糊。MLFuse 模型图像边缘细节不清晰。通过对比可以看出，本文方法能够更好地实现该数据集道路场景下的红外与可见光图像融合任务。

RoadScene 数据集 221 张融合图像的客观评价指标如表 3 所示，其中加粗数据为最优，加下划线数据为次优。数据表明，本文方法在 EN、MI、SSIM、PSNR、CC 和 VIFF 等客观评价指标上取得了较好的结果。MUFusion 模型融合图像的 VIFF 指标最高，符合人眼视觉特性，但是图像局部结构曝光，丢失了源图像的细节信息。TarDAL 模型在图像融合过程中出现明显的局部过曝和边缘细节缺失，导致像素值剧烈波动，STD 指标显著升高，而 SSIM 等指标表现不佳。相比之下，本文方法在各项指标上达到了最优或次优水平，在图像融合性能上表现出色，同时在泛化性方面也表现出较大优势，能够灵活应对不同复杂场景下的红外与可见光图像融合需求，具有良好的适应性。

2.3　消融实验

为了验证分段解码器、图神经网络中间模块和多层级细化模块的有效性，本文设计了 5 组消融实验，将本文方法降级为 4 种不同的网络结构，同时使用 MSRS 数据集进行测试。第 1 组实验不使用动态去噪路径规划，并将时序扩散融合模块中的分段解码器替换为单个简单解码器，其余网络结构不变，记为 GMB+MLRB。第 2 组实验直接去除多层

表 3　RoadScene 数据集 221 张融合图像的客观评价指标

方法	En	MI	SSIM	PSNR	STD	CC	VIFF
RFN-Nest	7.254 9	3.112 7	0.593 3	14.133 8	6.818 0	0.745 3	0.504 0
MUFusion	7.318 0	2.481 0	0.659 5	15.631 4	7.165 6	0.595 4	**0.870 2**
DDFM	7.227 7	2.167 9	0.400 7	13.510 8	6.717 0	0.629 5	0.391 6
TarDAL	7.091 1	3.289 8	0.699 4	15.551 2	**7.705 3**	0.548 4	0.574 6
BTSFusion	5.137 7	2.565 4	0.605 1	16.003 6	6.616 9	0.687 2	0.553 2
GIFusion	7.105 5	2.258 4	0.496 0	15.481 3	6.287 5	0.601 9	0.367 3
Text-Difuse	7.340 3	2.274 2	0.509 6	14.922 6	7.162 1	0.619 0	0.225 5
MMAE	7.395 2	4.748 6	**0.737 4**	18.850 6	7.245 2	0.518 6	0.443 3
MLFuse	7.293 3	3.369 4	0.670 4	14.753 6	6.879 1	0.684 0	0.641 6
本文方法	**7.432 8**	**5.074 7**	0.700 6	**19.246 7**	7.334 6	**0.773 4**	0.656 2

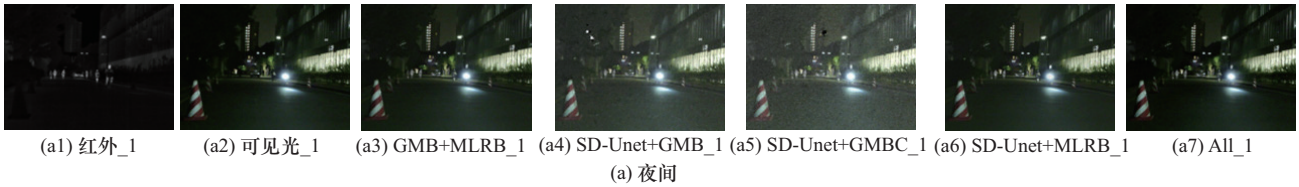
级细化模块，使用动态去噪路径规划且其余网络结构不变，记为SD-Unet+GMB。第3组实验将多层级细化模块简化为简单卷积输出，使用动态去噪路径规划且其余网络结构不变，记为SD-Unet+GMBC。第4组实验将时序扩散融合模块中的图神经网络中间模块替换为简单卷积中间层，其余网络结构不变，记为SD-Unet+MLRB。第5组实验使用动态去噪路径规划，并保留去噪扩散融合网络中的分段解码器和图神经网络中间模块，保留多层级细化模块，用来观察完整模型架构下的融合结果，记为All。

图9展示了5种不同网络结构下夜间和白天2种场景的消融实验融合结果。消融实验中5种不同网络结构的361张融合图像均值客观评价指标如表4所示，加粗数据为最优，加下划线数据为次优。观察图像可知，前4组实验的融合图像质量均低于完整模型架构下生成的融合图像，尤其是在细节信息、边缘轮廓和图像清晰度等方面，完整模型均优于其他4个降级网络结构。第5组实验除EN指标外，其余指标都获得了最高值。SD-Unet+GMB和SD-Unet+GMBC实验中的EN指标较高，但其他指标显著低于All实验，这是由于模型初步完成图像融合后保留了大量源图像的细节信息，但是图像还存在结构失真、噪声等影响，需要进一步对图像进行细化。综合所有指标，证明了分段解码器、图神经网络中间模块和多层级细化模块的重要性和有效性。

2.4　下游实验

下游任务是图像融合应用的一个重要方面，为了研究图像融合对于下游任务的作用，本文使用YOLOv8的目标检测方法进行实验。红外与可见光图像融合与YOLOv8目标检测结合的下游任务核心在于通过跨模态信息互补，显著提升目标检测系统在复杂场景下的鲁棒性、精度和实用性。

本文选用MSRS数据集中红外与可见光图像进行目标检测实验，并与单一模态图像和9种对比方法的融合图像进行对比，可视化结果如图10所示。同时，本文使用平均精确率均值（MAP, mean average precision）指标进行客观评价，结果如表5所示，加粗数据为最优，加下划线数据为次优。综合主客观评价可得，本文方法的融合图像进行目标检测的精度优于单一模态图像和其他方法的融合图像。可见光图像MAP0.5最高，但目标检测类别较少，同时存在目标识别错误的现象。相较于其他方法，本文方法的融合图像MAP0.5最高且未出现目标识别错误的现象。总体而言，本文方法在目标检测任务中表现最佳，有效提升了下游视觉任务的精度和实用性。



(a1) 红外_1　(a2) 可见光_1　(a3) GMB+MLRB_1　(a4) SD-Unet+GMB_1　(a5) SD-Unet+GMBC_1　(a6) SD-Unet+MLRB_1　(a7) All_1

(a) 夜间

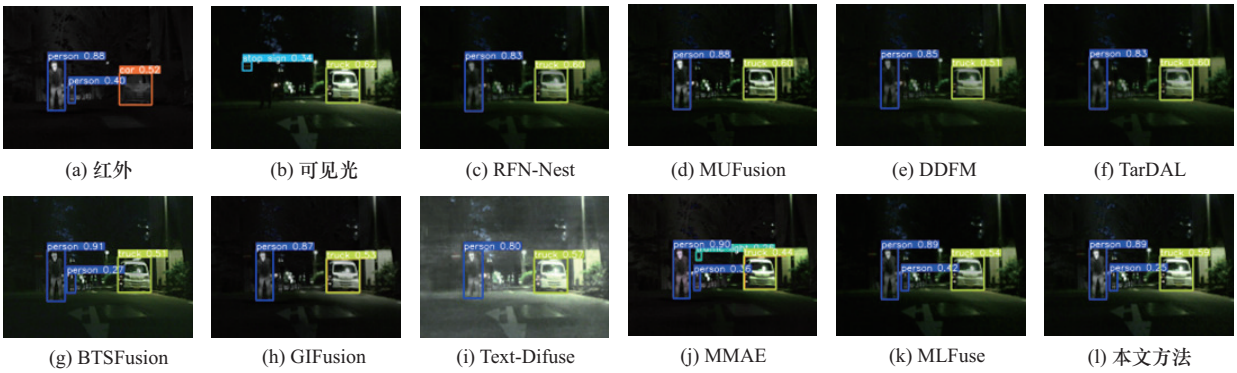
(b1) 红外_2　(b2) 可见光_2　(b3) GMB+MLRB_2　(b4) SD-Unet+GMB_2　(b5) SD-Unet+GMBC_2　(b6) SD-Unet+MLRB_2　(b7) All_2

(b) 白天

图9　夜间和白天2种场景的消融实验融合结果

表4　　消融实验中5种不同网络结构的361张融合图像均值客观评价指标

网络	EN	MI	SSIM	PSNR	STD	CC	VIFF
GMB+MLRB	6.776 8	4.072 4	0.631 7	19.901 6	6.655 1	0.964 3	0.772 6
SD-Unet+GMB	6.962 1	2.202 3	0.376 2	14.967 6	6.691 6	0.951 0	0.721 9
SD-Unet+GMBC	**7.126 8**	1.619 6	0.240 6	13.317 8	6.720 2	0.909 2	0.680 7
SD-Unet+MLRB	6.706 5	3.655 8	0.610 3	18.760 7	6.567 5	0.963 2	0.758 9
All	6.872 6	**4.541 3**	**0.670 9**	**20.997 8**	**6.802 8**	**0.969 7**	**0.814 2**



(a) 红外　(b) 可见光　(c) RFN-Nest　(d) MUFusion　(e) DDFM　(f) TarDAL

(g) BTSFusion　(h) GIFusion　(i) Text-Difuse　(j) MMAE　(k) MLFuse　(l) 本文方法

图10　目标检测任务场景可视化结果

表5　目标检测任务场景客观评价指标

方法	MAP0.5
红外	0.333 5
可见光	**0.833 5**
RFN-Nest	0.708 5
MUFusion	0.594 0
DDFM	0.708 5
TarDAL	0.373 5
BTSFusion	0.500 0
GIFusion	0.500 0
Text-Difuse	0.361 3
MMAE	0.625 0
MLFuse	0.778 0
本文方法	0.791 5

3　结束语

本文提出了一种动态时序扩散和多层级细化的红外与可见光图像融合网络。该网络为了增强U-Net结构编解码器间的消息传递，将GNN与U-Net结合为GMB，GMB通过聚合邻域节点信息，利用较大的感受野和灵活的图卷积实现图像中的远程信息交互。通过动态去噪路径规划并设计分段解码器实现扩散反向去噪过程的分阶段针对性去噪，有效避免了时间步梯度差异造成时序梯度失稳问题。同时，为了兼顾红外信息结构和可见光细节，设计了多层级细化模块融合源图像信息和恢复图像结构，提高了图像细节保留能力。实验结果表明，本文方法在MSRS、M3FD和RoadScene这3个红外与可见光图像数据集上展现出较强的融合性能和泛化能力，同时有效提升了下游视觉任务目标检测的精度和实用性。然而，本文方法将U-Net与图卷积结合，在实现动态去噪路径规划时会处理大量节点信息，导致计算开销较高，未来研究将在GMB模块内部设计更高效的图构建策略以提高计算效率，提升其在监控安防、目标检测、道路交通等关键领域的实时处理性能。

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] YANG B, JIANG Z H, PAN D, et al. Detail-aware near infrared and visible fusion with multi-order hyper-Laplacian priors[J]. Information Fusion, 2023, 99: 101851.

[3] TANG L F, XIANG X Y, ZHANG H, et al. DIVFusion: darkness-free infrared and visible image fusion[J]. Information Fusion, 2023, 91: 477-493.

[4] 连静, 王珂, 李光鑫. 基于边缘的小波图像融合算法[J]. 通信学报, 2007, 4: 18-23.
LIAN J, WANG K, LI G X. Edge-based wavelet image fusion algorithm[J]. Journal on Communications, 2007, 4: 18-23.

[5] LI S T, YANG B, HU J W. Performance comparison of different multi-resolution transforms for image fusion[J]. Information Fusion, 2011, 12(2): 74-84.

[6] 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.

[7] 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.

[8] 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.

[9] MA J Y, TANG L F, XU M L, et al. STDFusionNet: an infrared and visible image fusion network based on salient target detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.

[10] 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.

[11] XU H, MA J Y, JIANG J J, et al. U2Fusion: a unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Ma-

chine Intelligence, 2022, 44(1): 502-518.
[12] QIAN Y, LIU G, TANG H J, et al. BTSFusion: fusion of infrared and visible image via a mechanism of balancing texture and salience[J]. Optics and Lasers in Engineering, 2024, 173: 107925.
[13] WANG W, DENG L J, VIVONE G. A general image fusion framework using multi-task semi-supervised learning[J]. Information Fusion, 2024, 108: 102414.
[14] 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: 1-12.
[15] LI J F, SONG H, LIU L, et al. MixFuse: an iterative mix-attention transformer for multi-modal image fusion[J]. Expert Systems with Applications, 2025, 261: 125427.
[16] WANG X X, FANG L X, ZHAO J L, et al. MMAE: a universal image fusion method via mask attention mechanism[J]. Pattern Recognition, 2025, 158: 111041.
[17] 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.
[18] 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.
[19] 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]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 5802-5811.
[20] HO J, JAIN A N, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.
[21] ZHAO Z X, BAI H W, ZHU Y Z, et al. DDFM: denoising diffusion model for multi-modality image fusion[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 8082-8093.
[22] YI X P, TANG L F, ZHANG H, et al. Diff-IF: multi-modality image fusion via diffusion model with fusion knowledge prior[J]. Information Fusion, 2024, 110: 102450.
[23] YANG B, JIANG Z H, PAN D, et al. LFDT-Fusion: a latent feature-guided diffusion Transformer model for general image fusion[J]. Information Fusion, 2025, 113: 102639.
[24] ZHANG H, CAO L, MA J Y. Text-DiFuse: an interactive multi-modal image fusion framework based on text-modulated diffusion model[J]. Advances in Neural Information Processing Systems, 2024, 37: 39552-39572.
[25] ROBERTS J W, VARDT J A V, AHMED F B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification[J]. Journal of Applied Remote Sensing, 2008, 2(1): 023522.
[26] QIU D F, HU X Y, LIANG P W, et al. A deep progressive infrared and visible image fusion network[J]. Journal of Image and Graphics, 2023, 28(1): 156-165.
[27] SINGH S, SINGH H, BUENO G, et al. A review of image fusion: methods, applications and performance metrics[J]. Digital Signal Processing, 2023, 137: 104020.
[28] 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.
[29] ZHANG H, XU H, TIAN X, et al. Image fusion meets deep learning: a survey and perspective[J]. Information Fusion, 2021, 76: 323-336.
[30] DESHMUKH M, BHOSALE U. Image fusion and image quality assessment of fused images[J]. International Journal of Image Processing (IJIP), 2010, 4(5): 484.
[31] ZHANG X C, DEMIRIS Y. Visible and infrared image fusion using deep learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10535-10554.
[32] LEI J, LI J W, LIU J Y, et al. MLFuse: multi-scenario feature joint learning for multi-modality image fusion[J]. IEEE Transactions on Multimedia, 2025, 27: 3880-3894.

[作者简介]

邸敬（1979– ），女，甘肃兰州人，兰州交通大学副教授、硕士生导师，主要研究方向为人工智能、类脑计算、机器视觉、图像处理等。

李涵（2001– ），女，河北保定人，兰州交通大学硕士生，主要研究方向为多模态图像融合、图像处理。

石淑慧（2001– ），女，甘肃平凉人，兰州交通大学硕士生，主要研究方向为医学图像融合、图像处理。

刘冀钊（1989– ），男，甘肃兰州人，兰州大学副教授、硕士生导师，主要研究方向为类脑计算、混沌理论与应用技术等。

廉敬（1983– ），男，甘肃兰州人，兰州交通大学教授、博士生导师，主要研究方向为人工智能、类脑计算、模式识别、机器视觉等。