arXiv:2602.04405v1 [cs.CV] 4 Feb 2026

面向多模态图像融合的交互式空间-频率融合 Mamba

Yixin Zhu、Long Lv、Pin gping Zhang、Xuehu Liu、Tongdan Tang、Feng Tian、Weibing Sun 和 Huchuan Lu

摘要—多模态图像融合（MMIF）旨在结合来自不同模态的图像，生成融合图像，同时保留纹理细节和重要信息。近年来，一些 MMIF 方法引入频域信息以增强空间特征。然而，这些方法通常仅采用简单的串行或并行空频融合，缺乏交互。本文提出了一种面向 MMIF 的新型交互式空间-频率融合 Mamba（ISFM）框架。具体而言，我们首先利用模态专属提取器（MSE）从不同模态中提取特征。该提取器以线性计算复杂度对图像中的长程依赖进行建模。为了有效利用频率信息，我们进一步提出多尺度频率融合（MFF）。该方法自适应地融合多个尺度上的低频和高频分量，从而获得稳健的频率特征表示。更重要的是，我们进一步提出交互式空间-频率融合（ISF）。该方法利用频率特征引导跨模态空间特征，增强互补表示。我们在六个 MMIF 数据集上开展了大量实验。实验结果表明，与其他先进方法相比，我们的 ISFM 能够取得更优的性能。源代码可在 https://github.com/Namn23/ISFM 获取。

索引词——**多模态图像融合、空频交互、视觉状态空间模型、视觉 Mamba。**

I. 引言

多模态（MULTI-MODAL）图像融合旨在整合来自多个模态的图像，生成信息量更丰富、更全面的图像。红外-可见光图像融合和医学图像融合是两种典型的多模态图像融合任务。具体而言，红外-可见光图像融合旨在保留红外图像中的热辐射信息以及可见光图像中的精细纹理细节，并克服可见光图像对各种光照的敏感性，以及红外图像对噪声的敏感性。医学图像融合能够整合磁共振成像（MRI）模态中的软组织信息与计算机断层扫描（CT）模态中的骨结构信息。

Yixin Zhu 和 Long Lv 贡献相同，并列第一作者。（通讯作者：Pingping Zhang 和 Tongdan Tang。）Yixin Zhu 和 Pingping Zhang 隶属于大连理工大学未来技术学院及教育部数据科学与智能教育重点实验室（海南师范大学）。（电子邮箱：lcabert1@mail.dlut.edu.cn；zhpp@dlut.edu.cn）Long Lv、Feng Tian 和 Weibing Sun 隶属于大连大学附属中山医院。（电子邮箱：lvlong113@126.com；tian-feng73@163.com；massurm@163.com）Xuehu Liu 隶属于武汉理工大学计算机科学与人工智能学院。（电子邮箱：liuxuehu@whut.edu.cn）Tongdan Tang 隶属于大连理工大学中心医院。（电子邮箱：tangtongdan2002@sina.com）Huchuan Lu 隶属于大连理工大学信息与通信工程学院。（电子邮箱：lhchuan@dlut.edu.cn）

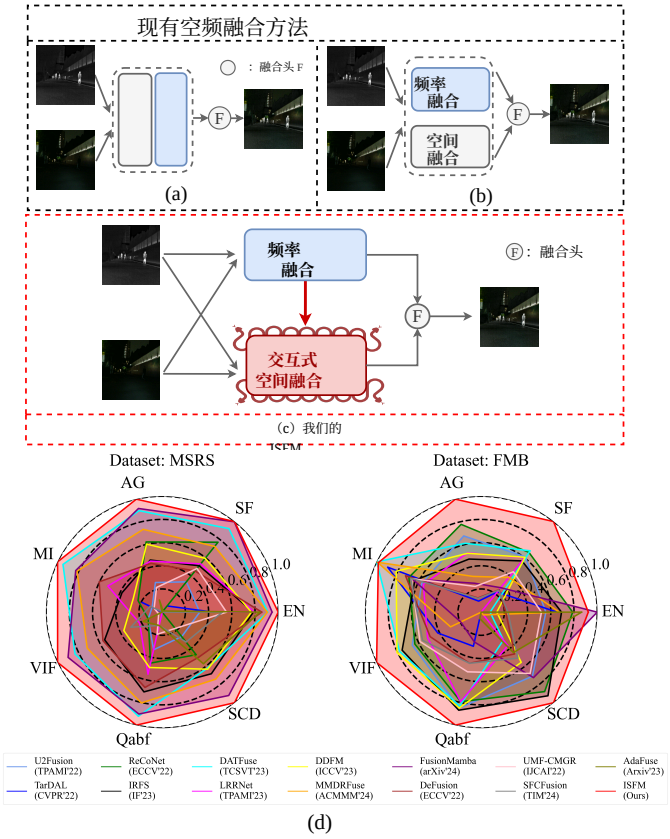


图1. 我们提出的ISFM与现有多模态图像融合方法的范式及性能比较。(a) 顺序空频融合方法；(b) 并行空频融合方法；(c) 我们提出的ISFM；(d) 在MSRS [1] 和FMB [2] 上七项指标的性能比较。

它提供互补见解并提升医学诊断能力。凭借这些优势，多模态图像融合能够提升许多下游任务的性能，例如，目标检测 [3]，语义分割 [2]，[4]。

多模态图像融合的主要挑战在于模态之间固有的差距，以及互补信息的有效提取。近年来，一些方法 [5]–[10] 开始联合利用空间和频率信息来应对这些挑战，并取得了显著进展。如图1(a)所示，一些方法 [5]，[7] 提出按顺序融合空间和频率信息。同时，如图1(b)所示，一些方法 [6]，[8]–[10] 采用并行空频融合。二者均未能探索空间域与频域之间的交互。这导致特征提取不完整，以及对领域特定特征的利用效率低下。事实上，空间域

能够捕获细粒度纹理细节，而频域则揭示全局模式。两个域之间的交互确保了更全面的信息整合，从而提升融合图像的整体质量。

此外，现有方法 [11]–[17] 主要采用卷积神经网络（CNN）或 Transformer 进行特征提取。卷积神经网络在建模长程依赖方面存在固有局限。Transformer 则需要巨大的计算开销。近年来，Mamba [18]凭借以线性计算复杂度建模长程依赖的能力，在视觉任务中展现出卓越性能。因此，如图 1(c) 所示，我们提出一种基于 Mamba 的交互式空间-频率融合框架，以充分利用领域特定特征的互补性。

本文提出一种用于多模态图像融合（MMIF）的交互式空间-频率融合 Mamba 框架，称为 ISFM。该框架将频率信息融入空间融合过程，并利用 Mamba 捕获长程依赖。具体而言，我们首先采用模态专属提取器（MSE）提取模态专属特征。随后，我们提出多尺度频率融合（MFF），以有效利用频率信息，并实现稳健的频率特征表示。MFF 在多个尺度上自适应地整合不同模态的低频和高频分量。为充分挖掘领域特定特征的互补性，我们提出交互式空间-频率融合（ISF）。ISF 由两部分组成：频率引导 Mamba（FGM）和频率引导门（FGG）。FGM 利用 Mamba 的能力增强互补表示。FGG 融入频率特征，以引导跨模态空间特征的融合。通过结合 MFF 和 ISF，我们的 ISFM 能够全面整合空间域与频域中的互补信息。在六个多模态图像融合数据集上进行的大量实验表明，与其他最先进的方法相比，我们的方法能够取得更优的性能。

我们的贡献总结如下：

- 我们提出了一种新颖的交互式空频融合 Mamba（ISFM）框架，用于多模态图像融合（MMIF）。该框架为空频融合提供了独特的视角。
- 我们提出了多尺度频率融合（MFF），以有效融合多个尺度上的频率信息。此外，我们提出了交互式空频融合（ISF），以充分利用空频信息的互补性。
- 在红外与可见光图像融合（IVIF）和医学图像融合（MIF）任务上的大量实验验证了我们方法的有效性。我们还验证了该方法对高层次计算机视觉任务的帮助作用。

II. 相关工作

A. 多模态图像融合

多模态图像融合旨在整合源图像中的重要信息，生成信息丰富的融合图像。现有的多模态图像融合方法大致可分为三类：基于卷积神经网络的方法 [11]–[14]、基于生成对抗网络的方法 [3], [19], [20], 以及基于 Transformer 的方法 [15]–[17], [21]。基于卷积神经网络的方法利用卷积神经网络实现图像融合，通常通过精心设计的损失函数和网络架构进行优化。例如，Ma 等人 [22] 设计了一种掩码损失，以实现目标区域与背景区域的选择性融合。Zhang 等人 [23] 设计了一种统一形式的损失函数，用于生成最终的融合图像。然而，由于局部感受野的限制，基于卷积神经网络的方法难以建模长程依赖。基于生成对抗网络的方法利用 GAN [24]确保融合图像与输入图像的分布相似。Ma等人 [19] 首次将 GAN 应用于图像融合领域。他们利用判别器引导生成器生成纹理丰富的融合图像。此外，Ma等人 [25] 提出了一种双判别器条件生成对抗网络，以实现均衡融合，使红外图像和可见光图像都参与对抗过程。然而，基于生成对抗网络的方法存在模型不稳定性和崩溃问题。此外，研究人员还探索了基于 Transformer 的方法。例如，Zhao 等人 [16]将卷积神经网络与 Transformer 相结合，对图像特征进行分解和重构。Tang 等人 [21] 利用 SwinTransformer [26] 克服基于卷积神经网络的融合方法的不足。近期研究[27]–[29]进一步提升了鲁棒性和语义一致性，为改进多模态图像融合提供了新的思路。然而，这些方法面临显著的计算挑战。与这些方法不同，我们提出了一种面向多模态图像融合的基于Mamba的空间-频率融合框架。该框架将频率信息融入空间融合过程，并利用 Mamba 捕获长程依赖。

B. 图像融合中的频域表示

频域表示能够突出图像中潜在的全局模式和结构特征 [30]。因此，一些研究 [7]–[10], [31]–[34] 利用频域信息进行多模态图像融合。早期工作只是在融合过程中引入频率信息。例如，Balasubramanian 等 [33] 使用小波包变换，并融合高频子带。Bhat 等 [32] 将主成分分析应用于频域变换，以实现医学图像融合。Wang等 [31]设计了频域自编码器来实现图像融合。近年来，一些研究者开始将空间域与频域信息结合起来，用于多模态图像融合。例如，Zheng 等 [5] 提出按顺序融合空间域和频域特征。Gu 等 [6] 提出交叉注意力模块，以并行融合跨域信息。Hu 等 [8] 构建了并行空间与频率融合网络，以引导融合过程。Chen 等 [10] 在频域中分解图像，并采用并行融合过程。Wu 等 [7] 设计了用于图像融合的级联空间与频率增强模块。然而，这些方法通常只是整合空间域和频域的信息，缺乏交互。与它们不同，我们提出交互式空间-频率融合（ISF），将频率信息融入空间融合过程。该方法利用 Mamba 从空间域和频域中提取互补特征。

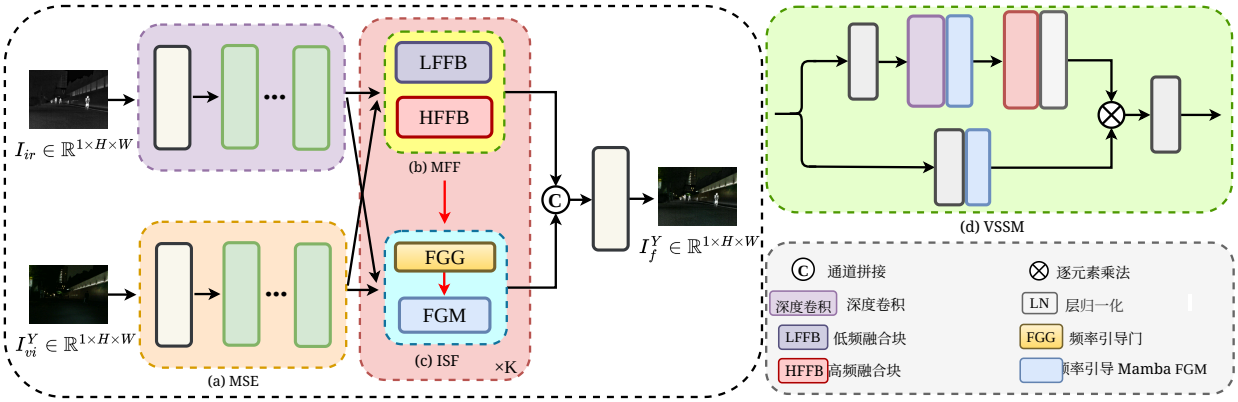


图 2. 我们提出的 ISFM 框架概览。（a）模态专属提取器（MSE）提取模态特定特征；（b）多尺度频率融合（MFF）在不同尺度上采用频域融合；（c）交互式空间-频率融合（ISF）将频率信息融入空间融合；（d）视觉状态空间模块结构（VSSM）。

C. 计算机视觉中的 *Mamba*

近年来，Mamba [18] 引入了一种选择机制，能够以线性计算复杂度对长序列进行建模，因此在计算机视觉领域受到了广泛关注。例如，Vision Mamba [35]提出了一种双向 Mamba 方法，实现了全局感知。Liu 等 [36] 进一步提出 VMamba，通过四向扫描策略实现更多空间连接。Guo等 [37] 设计了 MambaIR，其中引入了用于图像复原的残差状态空间块。He 等 [38] 提出 Pan-Mamba，以应对全色锐化的挑战。针对多模态图像融合，Xie 等 [39] 提出结合 Mamba 的动态特征增强方法，以实现多模态图像融合。Zhu等 [40] 提出了一种由 Transformer 和 Mamba 构成的双分支网络。Cao 等 [41] 提出随机扫描策略，以解决固定序列扫描的问题。然而，这些方法仅限于空间域融合。本文提出多尺度频率融合，以有效利用频率信息。此外，我们还提出一种基于 Mamba 的交互式空间-频率融合方法，以整合空间域和频域中的互补信息。

III. 方法

A. 总体架构

如图 2 所示，我们的框架包含两个分支，用于实现空间-频率融合。需要注意的是，这是一个通用的多模态图像融合框架。在不失一般性的前提下，我们以红外与可见光图像融合任务为例展示其工作流程。给定一对红外图像 $I_{ir} \in \mathbb{R}^{1\times H\times W}$ 和可见光图像 $I_{vi} \in \mathbb{R}^{3\times H\times W}$，我们采用 RGB 到 YCbCr 颜色转换 [14] 来解决通道不匹配问题。首先，我们提取可见光图像的 Y 通道，记为 $I_{vi}^{Y} \in \mathbb{R}^{1\times H\times W}$。其中，$H$ 和 W 分别表示输入图像的高度和宽度。随后，我们采用基于 Mamba 的模态专属提取器（MSE）提取模态特定特征。之后，通过多尺度频率融合（MFF）和交互式空间-频率融合（ISF）处理这些特征。MFF 旨在对低频和高频特征进行自适应融合。ISF 将频率特征融入空间融合过程，以引导空间融合过程，从而全面提取互补特征。最后，沿通道维度拼接来自空间域和频域的特征。我们采用两个卷积层生成最终融合图像的 Y 通道$I_f^Y \in \mathbb{R}^{1\times H\times W}$。

B. 模态特定提取器

为了从不同模态中提取模态特定特征，我们采用基于 Mamba 的模态专属提取器（MSE）。具体而言，给定输入 I_{ir} 和 I_{vi}^{Y}，我们首先应用两个 3×3 卷积层，生成浅层特征 $F_{ir}^{S} \in \mathbb{R}^{C\times H\times W}$ 和 $F_{vi}^{S} \in \mathbb{R}^{C\times H\times W}$。其中，$C$ 表示特征的通道维度。该过程表示如下：

$$\begin{aligned} F_{ir}^{S} &= SiLU(Conv_{3\times 3}(I_{ir})), \\ F_{vi}^{S} &= SiLU(Conv_{3\times 3}(I_{vi}^{Y})), \end{aligned} \tag{1}$$

其中，$SiLU$ 是 SiLU 激活函数 [42]。

随后，我们采用视觉状态空间模块（VSSM）[36] 提取高层特征。它能够有效捕获图像特征的长距离依赖关系。需要注意的是，两个模态的权重不共享。通过结合卷积神经网络和 Mamba，MSE 生成模态特定特征 F_{ir}^{M}，$F_{vi}^{M} \in \mathbb{R}^{C\times H\times W}$。该过程表示为：

$$\begin{aligned} F_{ir}^{M} &= \Psi_{ir}(F_{ir}^{S}), \\ F_{vi}^{M} &= \Psi_{vi}(F_{vi}^{S}), \end{aligned} \tag{2}$$

其中，Ψ 表示 VSSM 模块。

C. 多尺度频率融合

频域表示能够捕捉图像的全局特征，并揭示其内在模式 [30]。这使我们能够以更系统的方式分解图像细节，从而实现更精确的特征提取与分析。因此，我们提出多尺度频率融合（MFF），以有效整合频率信息。如图 3 所示，我们应用离散小波变换（DWT）[43]，获得低频分量 LL 和高频分量 LH、HL 和 HH。随后，我们提出低频融合块（LFFB）和高频融合

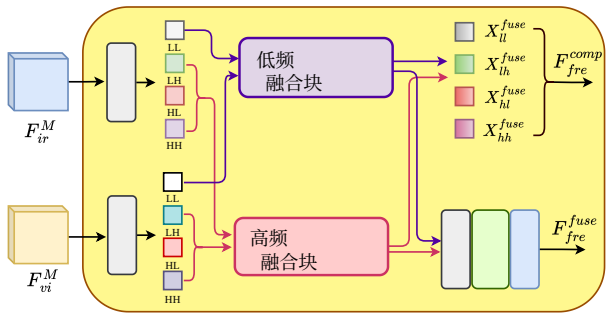


图 3. 所提出的 MFF 示意图。

块（HFFB），分别用于实现低频分量和高频分量的自适应融合。

1）低频融合块：如图 4(a) 所示，我们首先对低频分量执行逐元素加法。随后，采用一个 1×1 卷积层生成 X_{lf}，表示如下：

$$\begin{aligned} X_{lf}^{ir} &= DWT(F_{ir}^{M}), X_{lf}^{vi} = DWT(F_{vi}^{M}), \\ X_{lf} &= SiLU(Conv_{1\times1}(X_{lf}^{ir} + X_{lf}^{vi})). \end{aligned} \tag{3}$$

随后，我们通过两条支路处理合并后的分量。第一条支路应用最大池化层和平均池化层，自适应地捕获 X_{lf} 的全局信息。然后，我们通过一个 3×3 卷积层和一个 Sigmoid 层计算空间注意力权重σ_s。该过程表示为：

$$\sigma_s = Sigmoid(Conv_{3\times3}([\mathcal{M}(X_{lf}); \mathcal{A}(X_{lf})])), \tag{4}$$

其中，$\mathcal{M}(\cdot)$ 和 $\mathcal{A}(\cdot)$ 分别表示最大池化层和平均池化层。

第二条支路利用具有不同卷积核尺寸的深度卷积层处理 X_{lf}。随后，我们采用一个 1×1 卷积层生成不同尺度的特征。该过程表示为：

$$\begin{aligned} U' &= SiLU(Conv_{1\times1}(DWConv_{3\times3}(X_{lf}))), \\ U'' &= SiLU(Conv_{1\times1}(DWConv_{5\times5}(X_{lf}))), \end{aligned} \tag{5}$$

其中，$DWConv_{3\times3}(\cdot)$ 和 $DWConv_{5\times5}(\cdot)$ 分别表示一个 3×3和 5×5 深度卷积层。

针对不同尺度的上下文信息，我们利用注意力权重 σ_s 通过逐元素乘法提取重要信息。随后，我们对 X_{lf}应用残差连接。最后，将聚合后的特征输入 1×1 卷积层，得到融合后的低频特征 X_{lf}^{fuse}。该过程表示为：

$$\begin{aligned} U &= (U' + U'') \otimes \sigma_s + X_{lf}, \\ X_{lf}^{fuse} &= SiLU(Conv_{1\times1}(U)), \end{aligned} \tag{6}$$

其中，$\otimes$ 表示逐元素乘法。

2）高频融合块：如图 4(b) 所示，我们对高频分量进行逐元素加法。随后，应用一个 1×1 卷积层来生成 X_{hf}，其表达式如下：

$$\begin{aligned} X_{hf}^{ir} &= DWT(F_{ir}^{M}), X_{hf}^{vi} = DWT(F_{vi}^{M}), \\ X_{hf} &= SiLU(Conv_{1\times1}(X_{hf}^{ir} + X_{hf}^{vi})). \end{aligned} \tag{7}$$

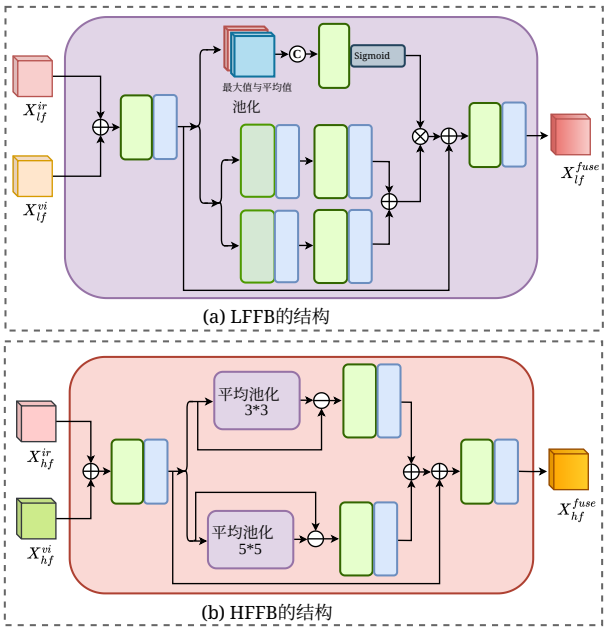


图 4. 所提出的 LFFB 和 HFFB 示意图。

为增强 X_{hf} 的边缘细节，我们应用两个尺寸不同的平均池化层。该过程表示如下：

$$X'_{hf} = \mathcal{A}_{3\times3}(X_{hf}), X''_{hf} = \mathcal{A}_{5\times5}(X_{hf}), \tag{8}$$

其中，$\mathcal{A}_{3\times3}(\cdot)$ 和 $\mathcal{A}_{5\times5}(\cdot)$ 分别表示一个平均池化层和一个平均池化层。

随后，我们从原始高频特征中减去池化结果。通过强调原始高频信息与平滑版本 [44] 之间的差异，该操作突出了更精细、更显著的边缘细节。接着，我们采用 1×1 卷积层生成 S_1 和 S_2。该过程表示为：

$$\begin{aligned} S_1 &= SiLU(Conv_{1\times1}(X_{hf} - X'_{hf})), \\ S_2 &= SiLU(Conv_{1\times1}(X_{hf} - X''_{hf})). \end{aligned} \tag{9}$$

在高频特征中，噪声通常呈现随机模式，而边缘细节则具有结构性。将平均池化特征从原始高频特征中减去，可以增强边缘细节并抑制噪声。最后，我们对特征求和，并对 X_{hf}应用残差连接。随后，将其输入 1×1 卷积层，得到融合后的高频特征 X_{hf}^{fuse}。该过程表示如下：

$$\begin{aligned} S &= S_1 + S_2, \\ X_{hf}^{fuse} &= SiLU(Conv_{1\times1}(S + X_{hf})). \end{aligned} \tag{10}$$

通过结合 LFFB 和 HFFB，融合频率特征经过两种操作。首先，将其用作 ISF 的频域引导信息，记为 F_{fre}^{comp}。这使得交互式融合能够充分利用空间域与频域中的互补信息，相关内容将在第 III-D 节详细介绍。其次，融合频率特征通过逆离散小波变换和卷积层转换至空间域，记为

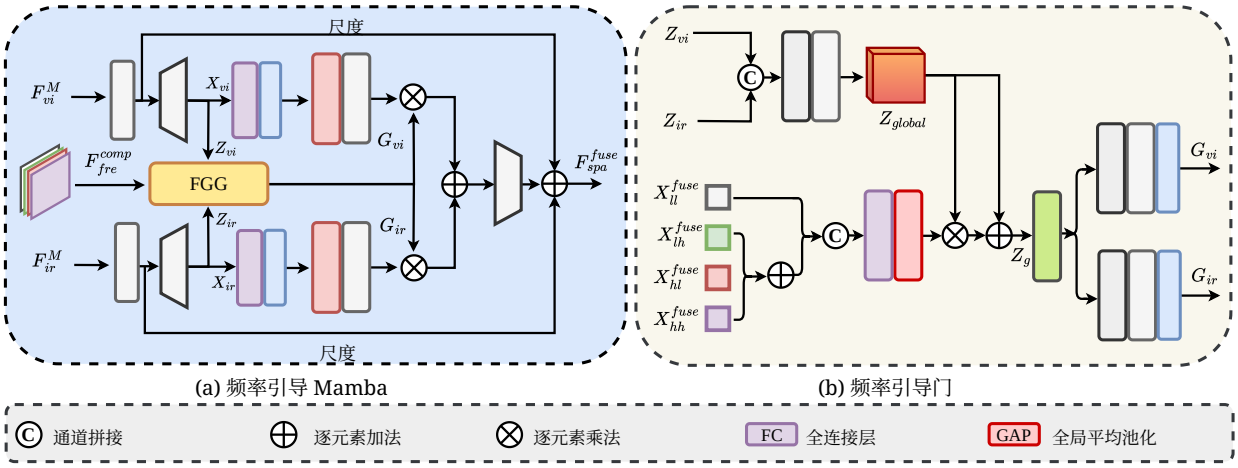


图 5. 所提出的 ISF 的架构。(a) 频率引导 Mamba（FGM）利用 Mamba 增强互补表示；(b) 频率引导门（FGG）融入频率特征，以引导跨模态的空间特征。

为 F_{fre}^{fuse}。随后，它将与融合空间特征结合，生成最终的融合图像。

D. 交互式空间-频率融合

交互式空间-频率融合（ISF）旨在利用频率特征，全面提取不同模态之间的互补特征。它通过频率引导 Mamba（FGM）和频率引导门（FGG）实现。

1）频率引导 *Mamba*：如图 5(a)所示，输入特征 F_{ir}^{M}和 F_{vi}^{M}首先经过归一化层。以红外流为例，随后通过线性层处理特征，并沿通道维度进行拆分，从而生成两个不同的特征序列，即 $X_{ir} \in \mathbb{R}^{\eta C \times H \times W}$ 和 $Z_{ir} \in \mathbb{R}^{\eta C \times H \times W}$。其中，$\eta$ 为通道扩展因子。该过程可表示为：

$$\begin{aligned} X_{ir} = Linear(LN(F_{ir}^{M})), Z_{ir} = Linear(LN(F_{ir}^{M})), \\ X_{vi} = Linear(LN(F_{vi}^{M})), Z_{vi} = Linear(LN(F_{vi}^{M})), \end{aligned} \tag{11}$$

其中，$Linear(\cdot)$ 和 $LN(\cdot)$ 分别表示线性层和层归一化 [45], 。

随后，对 X_{ir} 进行深度卷积处理。然后，将其输入二维状态空间模型块和归一化层，以生成隐藏状态特征 H_{ir}。该过程表示如下：

$$\begin{aligned} H_{ir} = LN(SSM(SiLU(DWConv_{3\times3}(X_{ir})))), \\ H_{vi} = LN(SSM(SiLU(DWConv_{3\times3}(X_{vi})))). \end{aligned} \tag{12}$$

同时，初始特征 Z_{ir} 通过 FGG 与频率特征交互。它生成门控特征，以引导更稳健的空间融合，具体内容将在第 III-D2 节中详述。该过程表示如下：

$$G_{ir}, G_{vi} = \mathcal{G}(Z_{ir}, Z_{vi}, F_{freq}^{comp}), \tag{13}$$

其中，$\mathcal{G}(\cdot)$ 表示 FGG。

随后，我们利用 G_{ir} 调制 H_{ir}，以实现空间和频率信息的交互融合。最后，我们将可调尺度因子应用于输入特征。它们用作残差连接，以生成融合空间特征 F_{spa}^{fuse}。该过程可表示为：

$$\begin{aligned} F_h = H_{ir} \otimes G_{ir} + H_{vi} \otimes G_{vi}, \\ F_{spa}^{fuse} = Linear(F_h) + s_1 \cdot F_{ir}^{M} + s_2 \cdot F_{vi}^{M}, \end{aligned} \tag{14}$$

其中，s_1 和 s_2 表示可调尺度因子。

2）频率引导门：为充分利用频率互补性，我们提出了频率引导门（FGG）。它增强了 Mamba 跨模态提取互补信息的能力。如图 5(b) 所示，我们首先对特征 Z_{ir} 和 Z_{vi}进行通道拼接。随后采用线性层生成 Z_{global}，其表示如下：

$$Z_{global} = LN(Linear[Z_{ir}; Z_{vi}]). \tag{15}$$

给定频率分量后，我们首先将高频分量相加，并与低频分量拼接。然后，使用全连接层和全局平均池化层聚合通道信息。该操作对频率分量的全局统计信息进行编码，表示如下：

$$\begin{aligned} F_f = [X_{ll}^{fuse}; X_{lh}^{fuse} + X_{hl}^{fuse} + X_{hh}^{fuse}], \\ Z_f = GAP(FC(F_f)), \end{aligned} \tag{16}$$

其中，$FC(\cdot)$ 表示全连接层，$GAP(\cdot)$ 表示全局平均池化层。

随后，利用编码后的特征引导空间特征的选择性融合，其表达式为：

$$Z_g = Z_f \otimes Z_{global} + Z_{global}. \tag{17}$$

随后，我们沿通道维度拆分 Z_g。每个部分依次经过线性层、归一化层和SiLU激活函数。这有助于实现更精确的自适应门控，同时增强互补特征。该过程表示为

$$\begin{aligned} Z_g', Z_g'' = Split(Z_g), \\ G_{ir} = SiLU(LN(Linear(Z_g'))), \\ G_{vi} = SiLU(LN(Linear(Z_g''))). \end{aligned} \tag{18}$$

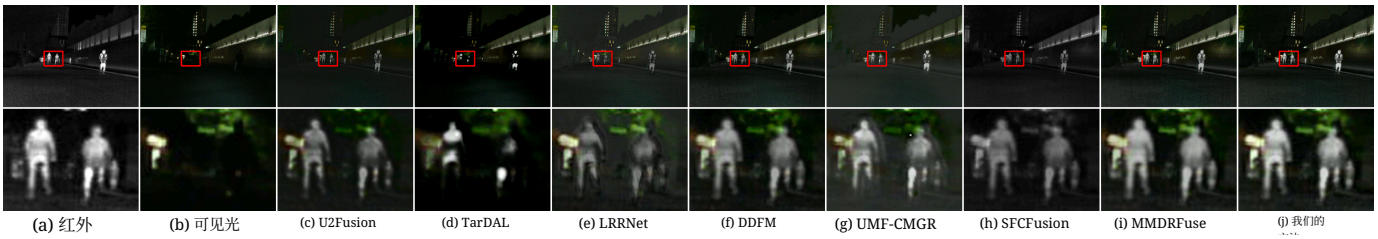


Fig. 6. MSRS数据集上的融合结果比较。第二行显示第一行中的放大区域

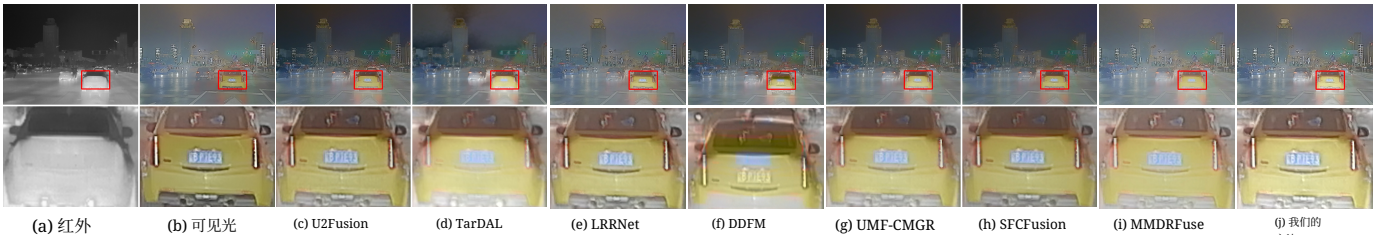


Fig. 7. FMB数据集上的融合结果比较。第二行显示第一行中的放大区域

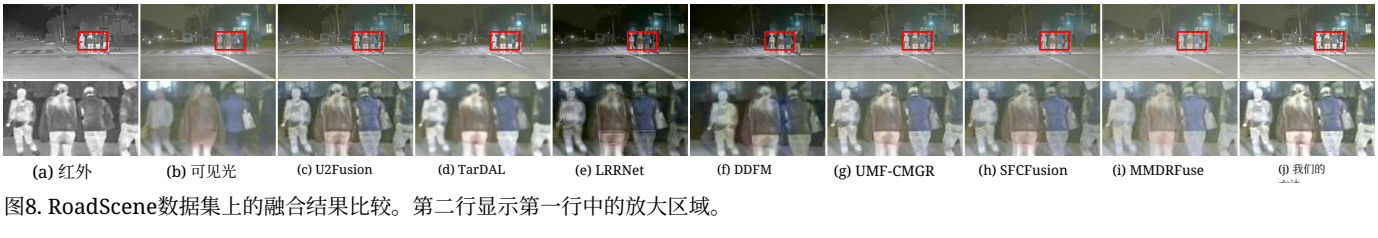


图8. RoadScene数据集上的融合结果比较。第二行显示第一行中的放大区域。

其中， G_{ir} 和 G_{vi} 是生成的门控权重，分别源自不同的模态和频率分量。

E. 损失函数

损失函数在平衡保留不同图像模态信息方面起着关键作用。本文通过内容损失 $\mathcal{L}_{cont}$、像素强度损失 $\mathcal{L}_{int}$、梯度损失$\mathcal{L}_{grad}$ 和结构相似性损失 $\mathcal{L}_{ssim}$的组合来优化训练过程。总损失表示如下：

$$\mathcal{L}_{total} = \mathcal{L}_{cont} + \alpha\mathcal{L}_{int} + \lambda\mathcal{L}_{grad} + \gamma\mathcal{L}_{ssim}, \quad (19)$$

其中， α、 λ 和 γ 是用于控制各个子损失项权衡的超参数。

我们将 $\mathcal{L}_{cont}$ 计算为融合图像 I_f^Y与源图像 I_{ir} 和 I_{vi}^Y 之间差值的总和：

$$\mathcal{L}_{cont} = \frac{1}{HW}(\|I_f^Y - I_{ir}\|_1 + \|I_f^Y - I_{vi}^Y\|_1), \quad (20)$$

其中， H 和 W 分别表示图像的高度和宽度。 $\|\cdot\|_1$ 表示 l_1-范数。

我们调整 $\mathcal{L}_{int}$ 和 $\mathcal{L}_{grad}$ ，以最大限度地保留融合图像中源模态的显著目标和细节纹理信息。可表示为：

$$\begin{aligned}\mathcal{L}_{int} &= \frac{1}{HW}(\|I_f^Y - max(I_{ir}, I_{vi}^Y)\|_1),\\ \mathcal{L}_{grad} &= \frac{1}{HW}(\||\nabla I_f^Y| - max(|\nabla I_{ir}|, |\nabla I_{vi}^Y|)),\end{aligned} \quad (21)$$

其中， $|\cdot|$ 和 ∇ 分别表示绝对值运算和 Sobel 梯度算子。

为全面评估融合图像的结构质量，我们采用 $\mathcal{L}_{ssim}$，其表达式为：

$$\mathcal{L}_{ssim} = \frac{1 - SSIM(I_f^Y, I_{ir})}{2} + \frac{1 - SSIM(I_f^Y, I_{vi}^Y)}{2}, \quad (22)$$

其中， $SSIM(\cdot)$ 是结构相似性指数 [46]，用于衡量两幅图像的结构相似性。

IV. 实验

A. 实验设置

1）数据集和指标：我们在六个广泛使用的多模态图像融合数据集上进行了大量实验，以评估我们的方法。对于红外与可见光图像融合，我们选用了三个数据集：MSRS [1], FMB[2] 和 RoadScene [47]。我们在 MSRS 数据集上训练和测试框架，其中包含 1083 对训练图像和 361 对测试图像。我们还在 FMB 和 RoadScene 数据集上进行测试，分别包含 280 对和 221 对测试图像。这些数据集涵盖了丰富多样的图像，图像采集时间包括白天和夜晚，且包含人、汽车和自行车等各种物体。对于医学图像融合，我们采用哈佛医学数据集[1]。哈佛医学数据集是一个公开的医学图像网站，我们从中选取了 150 对图像进行测试，包括 50 对 MRI-CT 图像、50 对 MRI-PET 图像和 50 对 MRI-SPECT 图像。请注意，为了充分验证泛化性能，我们未在 FMB、RoadScene 和哈佛医学数据集上进行微调。

[1]http://www.med.harvard. edu/AANLIB/home.html

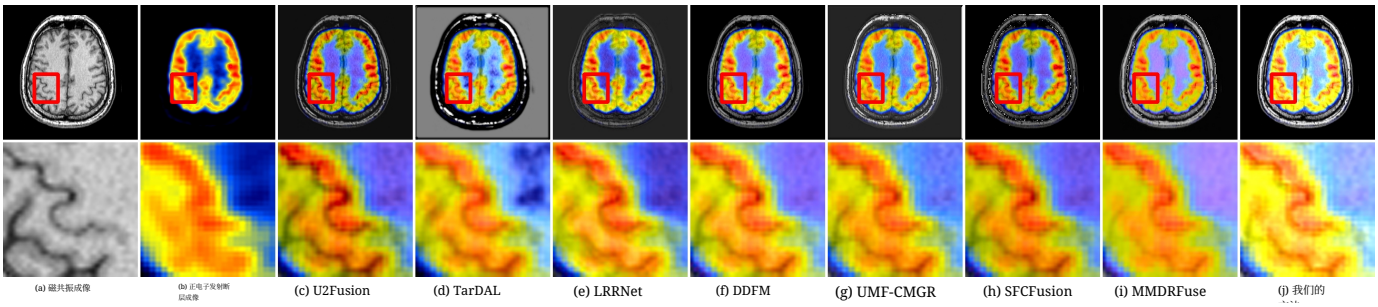


图 9. 融合结果在哈佛医学 MRI-PET 数据集上的比较。第二行显示第一行中放大后的区域。

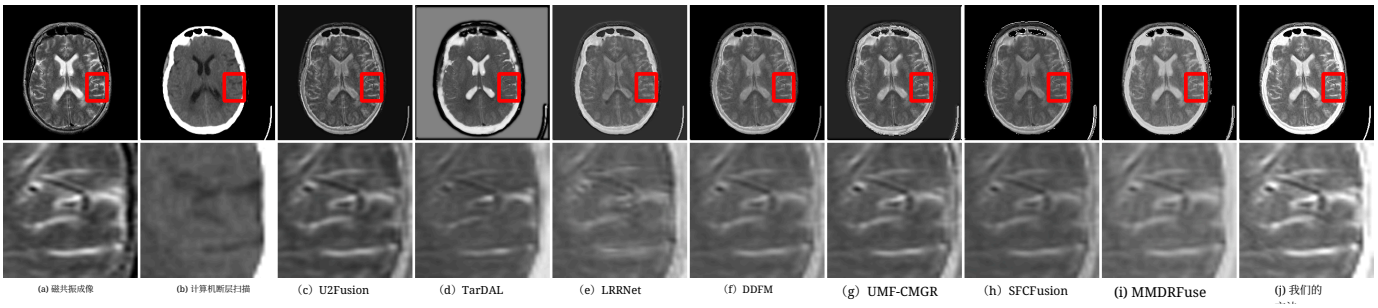


图 10. 融合结果在哈佛医学 MRI-CT 数据集上的比较。第二行显示第一行中放大后的区域。

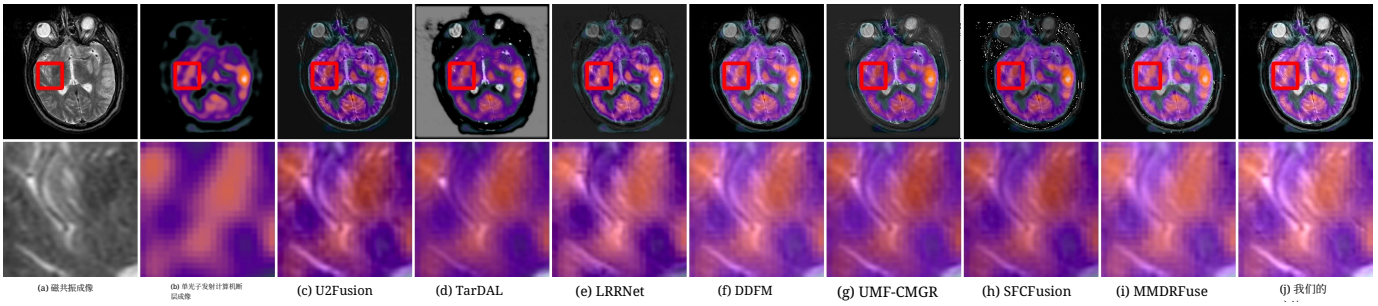


图 11. 融合结果在哈佛医学 MRI-SPECT 数据集上的比较。第二行显示第一行中放大后的区域。

我们使用八项指标衡量融合结果：熵（信息熵）、空间频率、平均梯度、视觉信息保真度、互信息、基于边缘的相似性度量（$Q^{AB/F}$）、差异相关性之和以及平均排名（平均排名）。这些指标的详细信息见 [54], [55]。

2) 实现细节： 在本研究中，所有实验均使用 PyTorch 工具箱和一块 GeForce RTX 3090 图形处理器完成。预处理阶段将训练样本随机裁剪为 128×128 个图像块。训练轮数设为 100，批次大小设为 4。为优化模型，我们采用 Adam[56] 优化器，初始学习率设为 10^{-4}。学习率采用预热策略并进行余弦衰减。超参数方面，VSSM 的数量和 K 分别设为 2 和 4。VSSM 中的通道维度设为 128。扩展因子 η 设为 2。 α、 λ 和 γ分别设为 5、5 和 0.5。

B. 与其他方法的比较

在本小节中，我们将所提方法与 15 种最先进的方法进行比较，包括 ReCoNet [13], DeFusion [11], U2Fusion [47], TarDAL [3], IRFS [48],AdaFuse [6], DAT- 融合 [49], LRRNet [12], DDFM [50],FusionMamba [39], SFCFusion [10], MMDRFuse [51],UMF-CMGR [52], FIS- C Net [5]和 RPFNet [53]。

1) 定性分析： 对于红外与可见光图像融合，如图 6、图 7 和图 8 所示，我们分别比较了多种方法在 MSRS、FMB 和 RoadScene数据集上的融合结果。可以看出，我们的 ISFM 能有效保留红外图像中的热辐射信息，同时保留目标的纹理细节。例如，在图 6 中，红色框内的行人等前景目标清晰突出，细节锐利。然而，其他方法（如 U2Fusion、TarDAL、LRRNet、等等。）生成的融合图像在视觉上并不理想。例如，图 7 绿色框内的车牌细节保留较差。我们的方法生成的融合图像呈现出丰富的纹理细节，视觉质量也更出色。

对于医学图像融合，我们在图9、图10和图11中展示了定性比较。可以观察到，U2Fusion和MMDRFuse方法未能很好地保留正电子发射断层成像、计算机断层扫描和单光子发射计算机断层成像图像中的功能信息。LRRNet和DDFM方法存在明显的颜色失真。该

表 I 红外与可见光图像融合任务的定量结果。最佳性能以**粗体** 标示，次佳性能以下划线标示。

数据集	方法	信息熵(↑)	空间频率(↑)	平均梯度(↑)	VIF(↑)	互信息(↑)	问$^{AB/F}$(↑)	差异相关和(↑)	平均排名(↓)
MSRS	ReCoNet [13]	4.23	9.98	2.99	0.49	1.58	0.40	1.26	10.71
	DeFusion [11]	6.35	7.98	2.60	0.75	2.16	0.51	1.29	8.71
	U2Fusion [47]	5.03	7.22	2.25	0.49	1.37	0.34	1.06	13.14
	TarDAL [3]	5.32	5.54	1.71	0.41	1.52	0.17	0.79	14.71
	IRFS [48]	6.17	8.34	2.66	0.74	1.82	0.53	1.47	8.29
	AdaFuse [6]	6.31	5.19	1.93	0.52	1.22	0.23	1.36	12.57
	DATFuse [49]	6.48	10.93	3.56	0.91	2.70	0.64	1.41	4.57
	LRRNet [12]	6.19	8.47	2.64	0.54	2.03	0.45	0.79	10.00
	DDFM [50]	6.13	8.86	2.95	0.63	1.65	0.42	1.42	9.00
	FusionMamba [39]	6.57	11.40	3.60	0.95	2.51	0.63	1.71	3.00
	SFCFusion [10]	5.51	5.11	1.72	0.59	1.40	0.23	1.12	13.29
	MMDRFuse [51]	6.47	9.69	3.22	0.84	2.49	0.58	1.53	5.57
	UMF-CMGR [52]	5.60	8.15	2.16	0.43	1.34	0.27	0.97	13.29
	FISCNet [5]	6.67	11.36	3.69	0.92	2.44	0.65	1.55	3.43
	RPFNet [53]	**7.26**	11.39	3.75	0.83	1.75	0.59	1.61	4.14
	ISFM(我们的方法)	6.70	**11.42**	**3.77**	**1.01**	**2.78**	**0.68**	**1.79**	**1.14**
FMB	ReCoNet [13]	6.69	10.70	3.62	0.65	2.25	0.55	1.60	6.00
	DeFusion [11]	6.39	7.13	2.22	0.59	2.29	0.37	1.33	11.43
	U2Fusion [47]	6.57	10.27	3.40	0.66	2.10	0.58	1.49	7.29
	TarDAL [3]	6.62	6.91	2.16	0.56	2.50	0.29	1.03	12.29
	IRFS [48]	6.64	9.88	2.96	0.70	2.18	0.59	1.63	6.86
	AdaFuse [6]	6.73	4.98	1.95	0.39	1.40	0.16	1.32	13.57
	DATFuse [49]	6.32	10.85	3.26	0.71	2.93	0.57	1.23	7.43
	LRRNet [12]	6.28	10.18	3.08	0.60	2.08	0.55	1.27	10.86
	DDFM [50]	6.35	10.23	3.07	0.72	2.40	0.58	1.39	7.29
	FusionMamba [39]	6.84	7.32	2.50	0.47	2.21	0.27	1.50	10.00
	SFCFusion [10]	6.40	7.08	2.21	0.62	2.23	0.37	1.36	11.29
	MMDRFuse [51]	6.63	8.30	2.63	0.51	2.91	0.13	1.05	11.00
	UMF-CMGR [52]	6.55	8.84	2.44	0.63	2.22	0.41	1.47	9.86
	FISCNet [5]	6.72	13.61	3.97	0.77	**2.94**	0.64	1.34	3.71
	RPFNet [53]	**7.38**	13.63	4.06	0.74	2.03	0.55	1.63	4.57
	ISFM(我们的方法)	6.76	**13.65**	**4.10**	**0.80**	2.61	**0.66**	**1.68**	**1.71**
RoadScene	ReCoNet [13]	7.07	9.31	3.90	0.53	2.15	0.38	1.57	8.57
	DeFusion [11]	6.93	8.99	3.59	0.50	2.06	0.37	1.35	11.86
	U2Fusion [47]	6.78	12.38	5.00	0.52	1.82	0.49	1.27	9.43
	TarDAL [3]	7.24	10.65	4.22	0.54	2.13	0.40	1.47	7.00
	IRFS [48]	7.01	10.79	4.08	0.57	2.01	0.46	1.58	7.43
	AdaFuse [6]	7.09	7.64	3.44	0.41	1.52	0.25	1.33	13.43
	DATFuse [49]	6.77	11.81	4.22	0.58	**2.49**	0.47	1.16	7.43
	LRRNet [12]	7.14	12.79	4.80	0.49	1.95	0.35	1.57	8.14
	DDFM [50]	6.56	7.71	2.94	0.54	2.20	0.32	1.12	12.00
	FusionMamba [39]	7.20	13.85	5.41	0.53	2.10	0.40	1.53	5.57
	SFCFusion [10]	6.87	9.65	3.76	0.52	1.92	0.37	1.35	11.86
	MMDRFuse [51]	7.09	10.38	4.26	0.49	2.15	0.14	1.04	11.00
	UMF-CMGR [52]	7.06	10.82	4.18	0.57	2.13	0.48	1.48	7.00
	FISCNet [5]	7.09	14.62	5.31	0.57	2.20	0.52	1.26	4.43
	RPFNet [53]	**7.40**	13.83	5.05	0.53	1.94	0.46	1.54	5.86
	ISFM（我们的方法）	7.11	**14.87**	**5.50**	**0.59**	2.16	**0.53**	**1.59**	**2.00**

所提方法能够更好地同时保留源图像中的结构细节和功能信息，生成具有较高视觉质量的融合图像。

2）定量分析：对于红外与可见光图像融合，定量结果如表 I 所示。所提方法在三个数据集上均取得了最佳平均排名，表明我们的方法优于其他方法。与仅依赖空间融合的方法（如 ReCoNet、DeFusion 和 U2Fusion）相比，我们的 ISFM 在所有指标上都具有明显优势。与其他空频融合方法（如 SFCFusion、FISCNet 和 RPFNet）相比，我们的方法提升了整体性能，并在平均排名中位居第一。

对于医学图像融合，定量结果如表 II 所示。请注意，我们的 ISFM 未在医学图像融合数据集上进行微调，而 U2Fusion、SFCFusion 和 FusionMamba 等方法则在该数据集上进行了训练。尽管信息熵和空间频率等部分指标未取得最佳结果，我们的 ISFM 在平均排名方面仍表现领先。实验结果表明，我们的方法能够实现稳健的性能。它进一步证明——

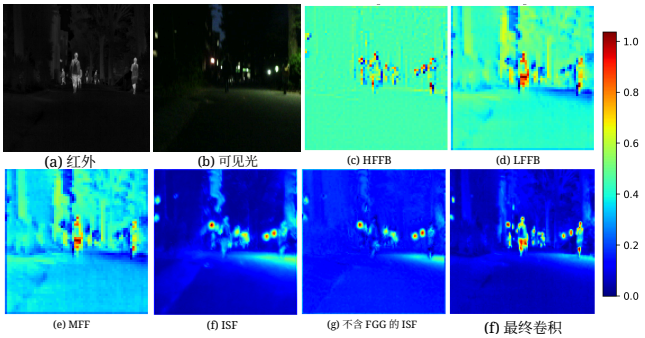


图 12. 不同模块获取的特征图可视化结果。(a) 红外图像；(b) 可见光图像；(c) HFFB；(d) LFFB；(e) MFF；(f) ISF；(e) 不含 FGG 的 ISF；(f) 最终卷积层。

这表明我们的方法在不同数据集上具有很强的泛化能力。

3）空间与频率融合特征的可视化：为验证所提出模块的有效性，我们进一步可视化了不同模块提取的特征。

表 II MIF 任务的定量结果。最佳性能以**粗体** 标示，次佳性能以下划线标示。

数据集	方法	信息熵(↑)	空间频率(↑)	平均梯度(↑)	VIF(↑)	互信息(↑)	问$^{AB/F}$(↑)	差异相关和(↑)	平均排名(↓)
磁共振成像-正电子发射断层成像	ReCoNet [13]	4.36	16.5	5.70	0.42	1.51	0.30	1.14	13.29
	DeFusion [11]	5.58	29.17	9.13	0.58	2.04	0.62	0.95	6.43
	U2Fusion [47]	5.37	19.86	6.66	0.46	1.80	0.41	0.52	12.14
	TarDAL [3]	5.72	27.54	9.11	0.31	1.56	0.30	0.77	10.86
	IRFS [48]	5.60	23.80	7.46	0.62	2.27	0.54	1.47	6.29
	AdaFuse [6]	6.18	19.91	6.60	0.68	2.23	0.47	1.18	7.29
	DATFuse [49]	6.00	**45.91**	11.19	0.42	1.93	0.54	0.84	6.57
	LRRNet [12]	5.57	15.88	5.32	0.33	1.65	0.18	0.25	14.71
	DDFM [50]	5.26	21.10	6.93	0.68	2.35	0.55	1.28	7.00
	FusionMamba [39]	4.95	42.05	**12.20**	0.59	2.02	0.65	1.62	4.86
	SFCFusion [10]	5.37	40.47	7.89	0.43	2.33	0.36	0.75	9.14
	MMDRFuse [51]	5.56	33.03	8.14	0.50	2.01	0.51	1.06	8.14
	UMF-CMGR [52]	6.20	30.72	8.44	0.41	1.72	0.43	0.35	9.86
	FISCNet [5]	6.22	23.72	9.04	0.57	1.81	0.59	1.10	6.71
	RPFNet [53]	**7.04**	17.97	5.59	0.52	1.83	0.51	1.04	9.29
	ISFM（我们的方法）	5.61	33.13	10.30	**0.73**	**2.40**	**0.70**	**1.66**	**2.57**
磁共振成像-计算机断层扫描	ReCoNet [13]	4.42	21.45	5.71	0.42	1.91	0.40	1.11	9.14
	DeFusion [11]	4.67	19.76	5.09	0.42	1.94	0.37	1.07	9.57
	U2Fusion [47]	4.69	20.05	5.74	0.34	1.66	0.40	0.55	11.29
	TarDAL [3]	5.79	22.72	7.45	0.23	1.49	0.28	0.44	11.14
	IRFS [48]	5.15	26.93	6.53	0.41	1.97	0.52	1.45	5.86
	AdaFuse [6]	5.42	18.92	4.74	0.41	**2.26**	0.33	0.94	9.71
	DATFuse [49]	5.05	27.61	6.29	0.46	2.09	0.48	1.20	5.86
	LRRNet [12]	4.90	20.20	4.80	0.37	1.86	0.33	0.18	12.29
	DDFM [50]	4.31	19.64	4.81	0.40	2.18	0.37	1.03	10.57
	FusionMamba [39]	4.40	34.80	**7.67**	0.50	1.99	0.52	1.57	4.86
	SFCFusion [10]	4.49	**37.86**	6.48	0.29	2.16	0.34	0.70	9.14
	MMDRFuse [51]	4.58	35.11	6.02	0.39	2.01	0.37	1.25	7.57
	UMF-CMGR [52]	5.56	34.27	7.40	0.32	1.78	0.38	0.39	9.29
	FISCNet [5]	6.56	33.49	8.54	0.50	1.90	0.53	1.28	**3.86**
	RPFNet [53]	**6.78**	18.10	4.57	0.37	1.80	0.45	1.04	10.29
	ISFM（我们的方法）	4.61	33.10	7.51	**0.53**	2.02	**0.54**	**1.72**	4.14
磁共振成像-单光子发射计算机断层成像	ReCoNet [13]	3.71	10.61	3.40	0.35	1.55	0.28	0.53	15.14
	DeFusion [11]	4.86	14.81	4.55	0.57	2.10	0.54	0.65	9.43
	U2Fusion [47]	4.61	13.53	4.22	0.44	1.84	0.44	0.90	12.14
	TarDAL [3]	5.59	16.67	5.63	0.32	1.55	0.23	0.88	10.29
	IRFS [48]	5.39	14.49	4.63	0.55	2.11	0.53	1.34	7.57
	AdaFuse [6]	5.42	14.22	4.61	0.56	2.17	0.55	1.47	6.86
	DATFuse [49]	5.23	23.40	6.27	0.51	2.24	0.60	0.88	5.43
	LRRNet [12]	5.12	11.62	3.78	0.36	1.80	0.20	0.52	14.00
	DDFM [50]	4.79	15.20	4.77	0.55	2.22	0.58	1.53	6.57
	FusionMamba [39]	4.79	24.89	**7.18**	0.64	2.16	0.68	1.53	3.86
	SFCFusion [10]	4.70	**27.37**	4.54	0.42	2.28	0.31	0.50	10.00
	MMDRFuse [51]	5.15	21.85	4.88	0.49	2.08	0.49	1.18	7.71
	UMF-CMGR [52]	5.52	22.92	5.25	0.38	1.81	0.44	0.78	8.86
	FISCNet [5]	6.21	17.18	6.00	0.53	1.96	0.61	1.16	5.71
	RPFNet [53]	**6.96**	15.06	4.57	0.45	1.81	0.59	0.96	8.14
	ISFM（我们的方法）	5.02	20.74	6.44	**0.65**	**2.28**	**0.70**	**1.75**	**3.14**

可视化结果如图 12 所示。可以观察到，低频特征主要捕捉图像的整体结构，而高频特征则突出边缘细节。所提出的 MFF 能够有效整合这些频域特征。接下来，我们展示所提出 ISF 的特征图。可以看出，这些特征突出了显著的目标细节，同时保留了必要的背景信息。此外，若缺少 FGG 提供的频域交互，特征便无法增强纹理细节，从而导致表征不够全面。最后，最终卷积层的特征保留了空间域与频域各自的鲜明特性，显著提升了融合图像的整体质量。

表III MSRS数据集上关键模块的消融研究。

	配置	EN	SF	AG	SCD	VIF	$Q^{AB/F}$
I	基线	6.46	10.69	3.39	1.68	0.87	0.59
II	基线+MFF	6.63	11.24	3.7	1.70	0.99	0.66
III	基线+FGM	6.63	11.17	3.58	1.73	0.95	0.65
IV	基线+ISF	6.65	11.25	3.66	1.74	0.96	0.66
V	基线+MFF+FGM	6.67	11.32	3.69	1.77	1.00	0.68
	Ours	**6.70**	**11.42**	**3.77**	**1.79**	**1.01**	**0.68**

表 IV FGM 在 MSRS数据集上的分析。

	配置	EN	SF	AG	SCD	VIF	$Q^{AB/F}$
I	相加	6.65	11.11	3.55	1.68	0.89	0.65
II	拼接	6.66	11.13	3.56	1.68	0.91	0.65
III	交叉注意力	6.65	11.22	3.68	1.69	0.98	0.66
IV	FGM	**6.70**	**11.42**	**3.77**	**1.79**	**1.01**	**0.68**

C. 消融实验

我们在MSRS数据集上进行了一系列消融实验，以评估所提方法的有效性。

1）关键模块的作用：为验证所提模块的有效性，我们在以下条件下进行消融实验不同配置。如表 III 所示，基线是不含附加模块的模型。我们将模态特定特征拼接起来，生成融合图像。在

表V MSRS数据集上FGG的分析。

	配置	EN	SF	AG	SCD	VIF	$Q^{AB/F}$
I	无交互	6.67	11.32	3.69	1.77	1.00	0.68
II	仅频率	6.67	11.34	3.70	1.76	1.01	0.68
III	FGG	**6.70**	**11.42**	**3.77**	**1.79**	**1.01**	**0.68**

表VI MSRS数据集上的频域融合消融研究。

	LFFB	HFFB	FGG	EN	SF	AG	SCD	VIF	$Q^{AB/F}$
I				6.63	11.17	3.58	1.73	0.95	0.65
II	✓			6.66	11.13	3.55	1.76	0.99	0.67
III		✓		6.64	11.23	3.57	1.70	0.95	0.65
IV	✓	✓		6.67	11.32	3.69	1.77	1.00	0.68
V	✓	✓	✓	**6.70**	**11.42**	**3.77**	**1.79**	**1.01**	**0.68**

配置 II 中，我们将 MFF 融入基线，结果显示性能显著提升。在配置 III 中，我们在基线中引入 FGM，与基线相比，这增强了融合结果。配置 IV 在配置 III 的基础上加入 FGG，所有指标均显著提升。结果表明，空间域与频域之间的交互能够有效提升融合质量。在配置 V 中，我们将 MFF 与 FGM 结合，但不进行空间域与频域的交互。与最终模型相比，缺少交互会导致融合效果不理想。这些实验结果证明了所提模块的有效性。

2）交互式空间-频率融合分析：在所提出的 ISF 中，FGM 在提取空间域中的互补信息方面发挥着关键作用。为了评估该模块的影响，我们移除 FGM，改用逐元素加法来提取互补信息，如表 IV 的配置 I 所示。在配置 II 中，我们用按通道拼接取代 FGM，并在其后接入一个全连接层。在配置 III 中，我们使用交叉注意力来实现相同目的。结果表明，所提出的 FGM 显著优于其他方法。这些发现凸显了 FGM 在增强互补信息方面的有效性。

此外，我们还研究了 FGG 的影响，如表 V 所示。第一种配置未在空间域与频域之间引入任何交互。在配置 II 中，我们将频率信息引入交互过程，但门控权重仅由频率特征驱动。结果表明，这仍未能实现最佳交互。最后，在配置 III 中，我们采用了所提出的 FGG。它取得了最佳的行人

性能。这些结果表明，ISF 能够实现空间域与频域之间更稳健的融合。

3）频域融合分析：我们进一步开展消融实验，以验证所提出的频域融合方法的有效性，如表 VI 所示。在实验 I 中，由于未引入频域信息，融合方法取得了最差的结果。在实验 II 中，使用 LFFB 融合低频信息，同时直接对高频特征求和。在实验 III 中，使用 HFFB 进行高频融合，并对低频特征

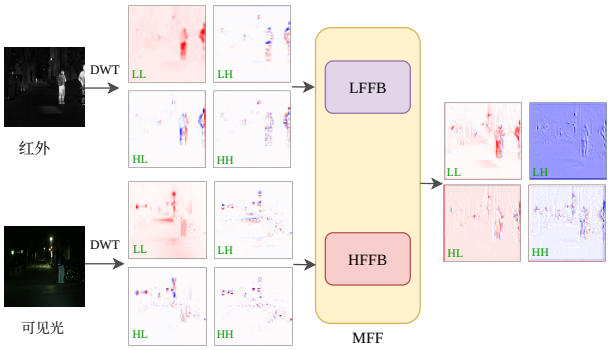


图 13. 可见光和红外输入的离散小波变换分解（低-低、低-高、高-低、高-高）以及我们提出的 MFF 生成的融合特征的可视化。

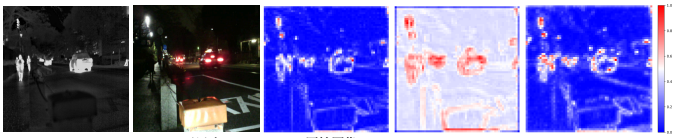


图 14. 所提出的高频增强操作的可视化。

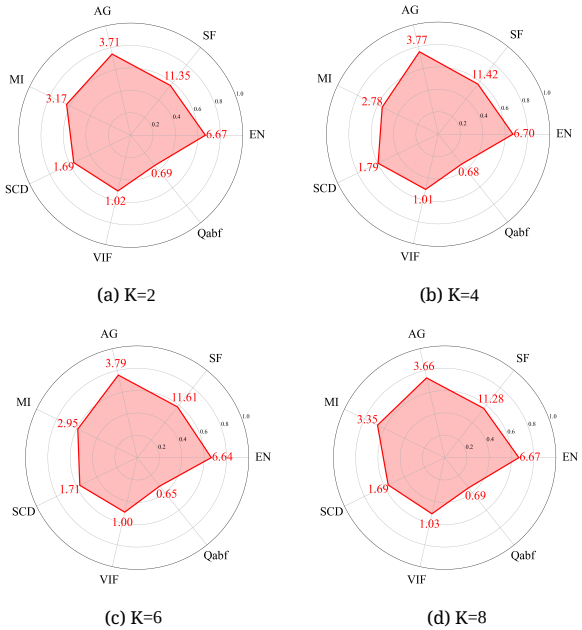
求和。在实验 IV 中，我们使用 FFB 整合频域信息。最后，通过 FGG 实现空频交互。整合所有模块后，我们的方法取得了最佳性能。这些结果表明，我们的方法在频域融合方面有效。

为了直观验证我们的频域融合机制的有效性，我们进行了两类可视化实验。首先，我们展示源图像的离散小波变换分解，以及通过所提出的 MFF 融合得到的相应特征。如图 13 所示，低-低、低-高、高-低和高-高子带揭示了不同模态之间的互补特性。我们的 MFF 能够有效整合这些特性，生成结构丰富且包含丰富高频细节的融合特征。其次，我们对高频增强操作的效果进行可视化。从概念上讲，该操作类似于信号处理中的反锐化掩模。事实上，它通过从原始信号中减去平均池化（模糊）后的特征来增强高频分量。如图 14 所示，可视化结果证实了这一效果：物体边界处的响应更强，而平滑区域的响应则受到抑制。

4) 超参数的影响 K：为研究超参数 K的影响，我们在MSRS数据集上进行了实验。如图15所示，当 K 增大至4时，整体性能显著提升。超过该值后，空间频率、平均梯度、互信息上的性能有所提升，但其他指标下降。随着进一步增大，大多数指标出现显著下降。为平衡性能与计算效率，我们将 $K = 4$ 设为默认值。

D. 计算效率分析

在本小节中，我们对计算效率进行了全面分析。表 VII 展示了各关键模块的参数分布。此外，我们从推理时间、参数量和 GFLOPs 方面，将我们的方法与多种最先进的方法进行了比较。如表所示。



(a) K=2 (b) K=4

(c) K=6 (d) K=8

图 15. MSRS数据集上采用不同 K 的消融研究。

表VII 所提出框架的参数量。

模块	参数（M）
MSE	0.9752
MFF	4.4186
ISF	3.1635
其他层	0.5905
总计	9.148

VIII，我们的方法保持了适中的 GFLOPs 和参数量，在模型复杂度与性能之间实现了良好的平衡。

表 VIII MSRS数据集上的计算效率比较。

方法	时间（秒）	GFLOPs（G）	参数（M）
IRFS [48]	1.45	74	0.242
FusionMamba [39]	0.17	125	164.539
DeFusion [11]	0.33	143	7.874
UMF-CMGR [52]	0.03	193	0.629
DDFM [50]	63.46	139k	552.814
AdaFuse [6]	1.65	431	538.529
SFCFusion [10]	0.21	68	0.074
RPFNet [5]	0.03	47	0.072
FISCNet [53]	0.07	101	0.324
Ours	0.28	371	9.148

表 IX MSRS数据集上的目标检测性能。

	精确率	召回率	mAP@0.50
ReCoNet [13]	0.972	0.853	0.935
DeFusion [11]	0.982	0.947	0.976
U2Fusion [47]	0.962	0.961	0.980
TarDAL [3]	0.931	0.835	0.922
IRFS [48]	0.981	0.961	0.978
AdaFuse [6]	0.966	0.923	0.976
DATFuse [49]	0.977	0.934	0.969
LRRNet [12]	0.980	0.949	0.983
DDFM [50]	0.978	0.956	0.981
FusionMamba [39]	0.956	0.928	0.966
SFCFusion [10]	0.962	0.928	0.977
MMDRFuse [51]	0.982	0.946	0.983
UMF-CMGR [52]	0.964	0.876	0.953
FISCNet [5]	0.967	0.965	0.971
RPFNet [53]	0.973	0.958	0.976
Ours	**0.984**	**0.962**	**0.985**

表 X MSRS数据集上的语义分割性能。

	Unl	Car	Per	Bik	Cur	CS	GD	CC	公交车平均交并比
可见光	96.92	79.89	34.53	61.98	18.99	50.48	41.47	47.78	60.73	54.75
红外	96.57	73.58	46.99	52.63	29.94	24.51	52.81	22.98	47.79	49.76
ReN [13]	97.33	81.48	57.48	59.09	36.11	49.48	55.48	36.13	57.77	58.93
DeF [11]	97.79	86.21	64.91	63.90	28.33	64.58	58.39	46.51	60.67	63.48
U2F [47]	97.84	86.20	65.51	65.74	33.68	61.29	35.29	49.49	61.33	61.82
TarD [3]	97.67	84.70	63.10	61.96	28.41	60.56	38.98	48.56	57.62	60.17
IRFS [48]	97.69	85.41	59.43	67.39	42.45	53.71	63.82	52.68	65.78	65.37
AdaF [6]	97.30	83.39	57.18	63.11	34.60	47.25	55.24	49.13	44.04	59.03
DAT [49]	97.64	84.07	60.51	63.62	40.15	56.21	61.05	48.22	61.18	63.63
LRR [12]	97.51	82.65	63.16	65.22	36.46	52.53	52.71	49.48	59.57	62.14
DDFM [50]	97.64	84.84	60.31	65.09	41.37	52.00	62.52	51.13	63.67	64.29
FuMa [39]	98.02	87.93	65.88	66.17	43.58	65.45	58.01	52.75	68.13	67.32
SFC [10]	97.84	86.20	65.51	65.74	33.68	61.29	35.29	49.49	61.33	61.82
MDR [51]	98.05	87.91	65.68	67.20	46.12	68.01	65.05	53.66	69.51	69.02
UMF [52]	97.79	86.21	64.91	63.90	28.33	64.58	58.39	46.51	60.67	63.48
FISC [5]	97.67	84.29	59.89	64.55	41.80	56.98	66.17	49.05	63.97	64.93
RPF [53]	97.15	82.23	28.43	64.93	36.06	55.41	**66.86**	49.19	64.18	60.49
Ours	**98.15**	**88.66**	**66.30**	**68.69**	**49.88**	**69.00**	66.22	**54.29**	**73.60**	**70.53**

E. 下游任务评估

在本小节中，我们进一步评估了我们的方法在两项下游任务，即目标检测和语义分割中的有效性。

1）目标检测评估：我们首先在 MSRS 数据集上进行目标检测。采用 YOLOv5 [57] 评估检测性能。图 16 展示了我们的 ISFM 在提升目标检测性能方面的优势。在“00002D”图像中，由于场景复杂且光照退化，检测器无法从其他方法的融合结果中识别出行人，而我们的 ISFM 则有效融合了红外与可见光图像的互补信息，确保所有行人都能成功检测出来。我们进一步使用定量指标衡量目标检测性能：精确率、召回率和平均精度均值（mAP）。表 IX 所示结果表明，所提出的 ISFM 能够提升目标检测性能，从而促进恶劣环境条件下的计算机视觉任务。

2）语义分割评估：我们还在MSRS数据集上进行了语义分割。该数据集包含九类目标的标注（即背景、凸起、彩色锥桶、护栏、弯道、自行车、行人、停车标志和汽车）。我们在红外图像、可见光图像以及各方法生成的融合图像上重新训练DeeplabV3+ [58]，以评估其性能。如图17所示，我们的方法生成的分割图具有更好的区域完整性。我们进一步在表X中报告了平均交并比（mIoU）。结果表明，我们的方法能够更好地保留源模态中的边缘信息，从而提高分割准确率。



图 16。C MSRS数据集上的目标检测结果对比。第二行展示了第一行中放大后的区域 .



图17. MSRS数据集上的语义分割结果对比。

V. 结论

本文提出了一种新颖的交互式空间-频率融合Mamba框架，用于多模态图像融合。具体而言，我们提出了多尺度频率融合（MFF），以整合多个尺度上的频域信息。为充分利用域特定特征之间的互补性，我们提出了交互式空间-频率融合（ISF）。该方法引入频率特征来指导跨模态空间融合，并利用Mamba增强空间-频率域的互补表示。在六个MMIF数据集上的大量实验验证了所提方法的有效性。

参考文献

[1] L. Tang、J. Yuan、H. Zhang、X. Jiang 和 J. Ma，“Piafusion：一种基于光照感知的渐进式红外与可见光图像融合网络”，《信息融合》，第83卷，第79–92页，2022年。[2] J. Liu、Z. Liu、G. Wu、L. Ma、R. Liu、W. Zhong、Z. Luo 和 X. Fan，“用于图像融合与分割的多交互特征学习及全天候多模态基准”，载于《*IEEE/CVF*国际计算机视觉会议论文集》，2023年，第8115–8124页。[3] J. Liu、X. Fan、Z. Huang、G. Wu、R. Liu、W. Zhong 和 Z. Luo，“面向目标检测的红外与可见光融合：目标感知双对抗学习及多场景多模态基准”，载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》，2022年，第5802–5811页。[4] A. Kirillov、E. Mintun、N. Ravi、H. Mao、C. Rolland、L. Gustafson、T. Xiao、S. Whitehead、A. C. Berg、W.-Y. Lo 等，“万物分割”，载于《*IEEE/CVF*国际计算机视觉会议论文集》，2023年，第4015–4026页。[5] N. Zheng、M. Zhou、J. Huang 和 F. Zhao，“用于红外与可见光图像融合的频率融合与空间补偿网络”，《信息融合》，第109卷，文章编号102359，2024年。[6] X. Gu、L. Wang、Z. Deng、Y. Cao、X. Huang 和 Y.-m. Zhu，“Ada-fuse：基于空间-频率交叉注意力的自适应医学图像融合”，*arXiv:2310.05462*，2023年。[7] D. Wu、M. Han、Y. Yang、S. Zhao、Y. Rao、H. Li、X. Lin、C. Zhou 和 H. Bai，“Dcfusion：用于红外与可见光图像融合的双频交叉增强融合网络”，《*IEEE*仪器与测量汇刊》，第72卷，第1–15页，2023年。[8] K. Hu、Q. Zhang、M. Yuan 和 Y. Zhang，“Sfdfusion：一种高效的红外与可见光图像空间-频域融合网络”，载于欧洲人工智能会议论文集。IOS Press，2024年，第482–489页。[9] X. Hou、Y. Xue 和 S. Yang，“一种基于频域信息提取的红外与可见光图像融合网络”，载于*IEEE*系统仿真技术与应用会议论文集IEEE，2024年，第298–304页。

[10] H. Chen, L. Deng, Z. Chen, C. Liu, L. Zhu, M. Dong, X. Lu, 和 C. Guo, 《SFCFusion：空频协同红外与可见光图像融合》，*IEEE*仪器与测量汇刊，第73卷，第1–15页，2024年。[11] P. Liang, J. Jiang, X. Liu, 和 J. Ma, 《基于分解的融合：一种用于图像融合的自监督分解方法》，载于欧洲计算机视觉会议论文集，Springer，2022年，第719–735页。[12] H. Li, T. Xu, X.-J. Wu, J. Lu, 和 J. Kittler, 《LRRNet：一种由表征学习引导的红外与可见光图像新型融合网络》，*IEEE* 模式分析与机器智能汇刊，第45卷，第9期，第11 040–11 052页，2023年。[13] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, 和 Z. Luo, 《ReCoNet：用于快速高效多模态图像融合的循环校正网络》，载于欧洲计算机视觉会议论文集，Springer，2022年，第539–555页。[14] L. Tang, X. Xiang, H. Zhang, M. Gong, 和 J. Ma, 《Divfusion：无暗区红外与可见光图像融合》，《信息融合》，第91卷，第477–493页，2023年3月。[15] W. Tang, F. He, Y. Liu, 和 Y. Duan, 《Matr：通过多尺度自适应变换器实现多模态医学图像融合》，*IEEE* 图像处理汇刊，第31卷，第5134–5149页，2022年。[16] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, 和 L. Van Gool, 《Cddfuse：用于多模态图像融合的相关性驱动双分支特征分解》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第5906–5916页。[17] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, 和 L. Van Gool, 《等变多模态图像融合》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第25 912– 25 921页。[18] A. Gu 和 T. Dao, 《Mamba：利用选择性状态空间进行线性时间序列建模》，*arXiv:2312.00752*，2023年。[19] J. Ma, W. Yu, P. Liang, C. Li, 和 J. Jiang, 《Fusiongan：用于红外与可见光图像融合的生成对抗网络》，《信息融合》，第48卷，第11–26页，2019年。[20] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, 和 J. Jiang, 《通过细节保持对抗学习实现红外与可见光图像融合》，《信息融合》，第54卷，第85–98页，2020年。[21] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, 和 Y. Ma, 《Swinfusion：通过 Swin Transformer 实现通用图像融合的跨域长程学习》，*IEEE/CAA*自动化学报，第9卷，第7期，第1200–1217页，2022年7月。[22] J. Ma, L. Tang, M. Xu, H. Zhang, 和 G. Xiao, 《Stdfusionnet：一种基于显著目标检测的红外与可见光图像融合网络》，*IEEE*仪器与测量汇刊，第70卷，第1–13页，2021年。[23] H. Zhang, H. Xu, Y. Xiao, X. Guo, 和 J. Ma, 《重新思考图像融合：一种基于梯度与强度比例保持的快速统一图像融合网络》，载于*AAAI*人工智能会议论文集，第34卷，第07期，2020年，第12 797–12 804页。[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, 和 Y. Bengio, 《生成对抗网络》，神经信息处理系统进展，第27卷，2014年。

[25] J. Ma, H. Xu, J. Jiang, X. Mei, 和 X.-P. Zhang, 《Ddcgan：用于多分辨率图像融合的双判别器条件生成对抗网络》，IEEE 图像处理汇刊，第29卷，第4980– 4995页，2020年。 [26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, 和 B. Guo, 《Swin Transformer：利用移位窗口的分层视觉 Transformer》，载于 IEEE/CVF国际计算机视觉会议论文集，2021年，第10 012– 10 022页。[27] H. Zhang, L. Cao, X. Zuo, Z. Shao, 和 J. Ma, 《Omnifuse：语言驱动语义的复合退化鲁棒图像融合》，IEEE 模式分析与机器智能汇刊，2025年。 [28] X. Yi, Y. Ma, Y. Li, H. Xu, 和 J. Ma, 《人工智能促进复杂环境中的感知信息融合》，The Innovation，第6卷，第4期，2025年。 [29] Z.-H. Cao, Y.-J. Liang, L.-J. Deng, 和 G. Vivone, 《一种利用统一语言与掩码引导的高效图像融合网络》，IEEE 模式分析与机器智能汇刊，2025年。 [30] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, 和 R. Ng, 《傅里叶特征使网络能够在低维域中学习高频函数》，神经信息处理进展，第33卷，第7537–7547页，2020年。 [31] C. Wang, J. Wu, A. Fang, Z. Zhu, P. Wang, 和 H. Chen, 《一种高效的红外与可见光图像频域融合网络》，工程应用中的人工智能，第133卷，文章编号108013，2024年。 [32] M. Bhat, M. V. Karki 等, 《基于主成分分析和粒子群优化、利用双树复小波变换进行多模态医学图像融合的特征选择》，arXiv:1701 08918，2017年。 .[33] R. Balasubramanian 和 G. Bhatnagar, 《一种新的基于对比度并利用小波包的图像融合方法》，arXiv:0812.0759，2008年。 [34] Q. Li, J. Du, F. Song, C. Wang, H. Liu, 和 C. Lu, 《利用局部空间频率的基于区域的多焦点图像融合》，载于 IEEE 控制与决策会议论文集，IEEE，2013年，第3792–3796页。 [35] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, 和 X. Wang, 《Vision Mamba：利用双向状态空间模型实现高效视觉表征学习》，国际机器学习会议论文集，2024年。 [36] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, 和 Y. Liu, 《VMamba：视觉状态空间模型》，神经信息处理进展，2024年。[37] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, 和 S.-T. Xia, 《 MambaIR：一种基于状态空间模型的简易图像复原基线》，载于欧洲计算机视觉会议论文集，Springer，2025年，第222–241页。 [38] X. He, K. Cao, J. Zhang, K. Yan, Y. Wang, R. Li, C. Xie, D. Hong, 和 M. Zhou, 《Pan-Mamba：利用状态空间模型实现有效的全色锐化》，《信息融合》，第115卷，文章编号102779，2025年。 [39] X. Xie, Y. Cui, C.-I. Ieong, T. Tan, X. Zhang, X. Zheng, 和 Z. Yu, 《FusionMamba：利用 Mamba 实现多模态图像融合的动态特征增强》，视觉智能，第2卷，第1期，文章编号37，2024年。 [40] C. Zhu, S. Gao, H. Chen, G. Guo, C. Wang, Y. Wang, C. S. Lei, 和 Q. Fan, 《为何 Mamba 有效？探索用于多模态图像融合的线性 Transformer-Mamba 网络》，CoRR，2024年。 [41] K. Cao, X. He, T. Hu, C. Xie, J. Zhang, M. Zhou, 和 D. Hong, 《Shuffle Mamba：用于多模态图像融合的随机打乱状态空间模型》，arXiv:2409.01728，2024年。 [42] N. Shazeer, 《Glu 变体改进 Transformer》，arXiv:2002 05202，2020年。 .[43] S. G. Mallat, 《多分辨率信号分解理论：小波表示》，IEEE 模式分析与机器智能汇刊，第11卷，第7期，第674–693页，1989年。 [44] T. Yan, Z. Wan, P. Zhang, G. Cheng, 和 H. Lu, 《Transy-net：学习全 Transformer 网络以检测遥感图像中的变化》，IEEE 地球科学与遥感汇刊，第61卷，第1–12页，2023年。 [45] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, 和 T. Liu, 《关于 Transformer 架构中的层归一化》，载于国际机器学习会议论文集，2020年，第10 524–10 533页。 [46] Z. Wang, A. C. Bovik, H. R. Sheikh, 和 E. P. Simoncelli, 《图像质量评估：从误差可见性到结构相似性》，IEEE 图像处理汇刊，第13卷，第4期，第600–612页，2004年。 [47] H. Xu, J. Ma, J. Jiang, X. Guo, 和 H. Ling, 《U2Fusion：统一的无监督图像融合网络》，IEEE 模式分析与机器智能汇刊，第44卷，第1期，第502–518页，2020年。 [48] D. Wang, J. Liu, R. Liu, 和 X. Fan, 《用于红外与可见光图像联合融合及显著性目标检测的交互强化范式》，《信息融合》，第98卷，文章编号101828，2023年。 [49] W. Tang, F. He, Y. Liu, Y. Duan, 和 T. Si, 《DATFuse：通过双重注意力 Transformer 实现红外与可见光图像融合》，IEEE 电路与系统视频技术汇刊，第33卷，第7期，第3159–3172页，2023年。 [50] Z. Zhao、H. Bai、Y. Zhu、J. Zhang、S. Xu、Y. Zhang、K. Zhang、D. Meng、R. Timofte 和 L. Van Gool，“DDFM：用于多模态图像融合的去噪扩散模型”，载于*2023*年*IEEE/CVF*国际计算机视觉会议论文集，2023年，第8082–8093页。[51] Y. Deng、T. Xu、C. Cheng、X.-J. Wu 和 J. Kittler，“MMDRFuse：用于多模态图像融合的具有动态刷新的蒸馏微型模型”，载于*2024*年*ACM*国际多媒体会议论文集，2024年，第7326–7335页。[52] D. Wang、J. Liu、X. Fan 和 R. Liu，“通过跨模态图像生成与配准实现无监督错位红外与可见光图像融合”，国际人工智能联合会议论文集，第3508–3515页，2022年。[53] Z. Guan、X. Wang、W. Qian、P. Liu 和 R. Ma，“用于图像融合的残差先验驱动频率感知网络”，载于*2025*年*ACM*国际多媒体会议论文集，2025年，第1082–1091页。[54] J. Ma、Y. Ma 和 C. Li，“红外与可见光图像融合方法及应用：综述”，《信息融合》，第45卷，第153–178页，2019年1月。[55] S. Park、A. G. Vien 和 C. Lee，“用于红外与可见光图像融合的跨模态变换器”，*IEEE* 电路与系统视频技术汇刊，第34卷，第2期，第770–785页，2023年。[56] D. P. Kingma，“Adam：一种随机优化方法”，国际学习表征会议论文集，第1–15页，2014年。[57] J. Redmon、S. Divvala、R. Girshick 和 A. Farhadi，“你只看一次：统一的实时目标检测”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2016年，第779–788页。[58] L.-C. Chen、Y. Zhu、G. Papandreou、F. Schroff 和 H. Adam，“采用空洞可分离卷积的编码器-解码器语义图像分割网络”，载于欧洲计算机视觉会议论文集，2018年，第801–818页。