论文

面向跨模态图像生成与配准的无监督未对齐红外与可见光图像融合

王迪[1], 刘晋源[1], 樊鑫[2], 刘日升[2*]

1. 大连理工大学软件学院, 大连，116620
2. 大连理工大学国际信息与软件学院, 大连, 116620
* 通信作者. E-mail: rsliu@dlut.edu.cn

摘要 最近基于学习的图像融合方法在预配准多模态数据方面取得了许多进展, 但由于空间变形和难以缩小跨模态差异, 在处理未对齐多模态数据时出现了严重的重影. 为了克服这些障碍, 本文提出了一种鲁棒的无监督跨模态红外与可见光图像融合 (IVIF) 交叉模态生成-配准范式. 具体地, 本文提出了一种跨模态感知风格迁移网络 (CPSTN) 生成以可见光图像为输入的伪红外图像. 得益于 CPSTN 良好的几何保存能力, 生成的伪红外图像具有清晰的结构, 结合红外图像的结构敏感, 更有利于将跨模态图像对齐转化为单模态配准. 之后, 本文引入了多级细化配准网络 (MRRN) 来预测失真和伪红外图像之间的位移矢量场, 并在单模态设置下重建配准红外图像. 此外, 为了更好地融合配准的红外图像和可见图像, 本文提出了一个特征交互融合模块 (IFM), 以自适应地选择更有意义的特征经由双路径交互融合网络 (DIFN) 进行融合.

关键词 图像融合, 跨模态配准, 感知风格迁移, 自适应特征交互

1 引言

图像融合是从不同模态图像中提取互补信息, 并聚合它们以生成更丰富, 更有意义的特征表示. 通常, 红外和可见光图像融合 (IVIF) 具有这样的优势, 并有利于自动驾驶和视频监控等实际应用.

现有的大多数 IVIF 方法都是专门为手工制作的预配准图像设计的, 尽管取得了许多进展, 但这是非常耗费劳力时间的. 然而, 它们对未对齐红外和可见光图像的强度与分布差异很敏感, 一旦存在轻微的偏差和形变就会导致融合图像中出现严重的重影现象. 其内在原因是红外图像和可见光图像之间的跨模态变化较大, 在共享的特征空间中直接跨越二者的领域鸿沟是不现实的. 再加上缺乏跨模态图像的相似性约束, 很少有工作尝试融合未对齐的红外和可见光图像. 主要的障碍是跨模态的图像对齐.

通常, 现有的广泛使用的图像对齐方法 [2,6] 执行像素级和特征级对齐, 通过显式估计变形图像和其参考图像之间的变形场. 然而, 由于它们高度依赖于具有邻域参考的合成或真实数据的分布和外观相似性, 因此只能在单一模态设置下工作. 此外, 用于优化跨模态对齐过程的相似性度量的设计

已被证明是相当具有挑战性的. 这些障碍促成了跨模态医学图像翻译 [21] 和跨模态 Re-ID 任务的发展 [17,19]. 前者利用 cGAN [7] 执行 NeuroImage 到 NeuroImage 的转换, 而后者通过 RGB 到红外图像转换来学习跨不相交摄像机视图的行人图像中的跨模态匹配. 最近的一项研究 [1] 通过在两种输入模态上训练图像到图像的转换网络, 提出了一种多模态图像配准方法. 上述方法的基本思想是使用图像到图像的转换来实现跨模态变换, 以缩小不同模态之间的巨大差异.

启发于此, 我们提出了一种专门的跨模态生成配准范式, 用于无监督的未对齐红外和可见图像融合. 红外图像的一个固有特征是重结构, 轻纹理. 换句话说, 几何结构对于红外图像是必不可少的. 生成保留清晰结构信息的伪红外图像更有利于实现失真红外与生成的伪图像之间的单模态配准. 为了在从可见图像生成伪红外图像的过程中更好地保留几何结构, 我们提出了一种跨模态感知风格转移网络 CPSTN, 它继承了 CycleGAN [25] 的基本循环一致学习方式, 同时开发了一个感知风格迁移约束和跨双循环学习路径的交叉正则化, 以进一步指导 CPSTN 生成更清晰的结构. 没有它们, CPSTN 的生成能力将退化到方法 [1] 的水平. 这样做为红外图像的单模态配准奠定了基础. 本文利用多级细化配准网络 (MRRN) 来从粗到细地预测失真和伪红外图像之间的变形场, 并重建配准的红外图像. 之后, 本文进一步进行配准红外图像和可见图像的融合. 为了使融合网络能够更多地关注真实的纹理细节, 同时避免由一些不成熟的融合规则（例如, 连接、加权平均）引起的特征平滑, 本文开发了一个交互融合模块 (IFM) 来自适应地从红外和可见光图像中选择有意义的特征以实现具有清晰纹理的融合结果. 我们在人工处理过的未对齐数据集上评估所提出的方法, 并通过综合分析证明其优势. 本文的主要贡献总结如下：

- 提出了一种高度鲁棒的无监督红外和可见图像融合框架, 与专门用于预配准图像的基于学习的融合方法相比, 该框架更侧重于减轻由未对齐对图像融合引起的重影.
- 考虑到跨模态图像对齐的难度, 本文开发了一种专门的跨模态生成配准范式来弥合模态之间的巨大差异, 从而实现有效的红外和可见图像对齐.
- 开发了交互融合模块, 自适应融合多模态特征, 避免了融合规则不成熟导致的特征平滑, 强调了可信的纹理细节.

大量的实验结果表明, 所提出的方法在未对齐的跨模态图像融合方面表现出色.

2 本文方法

2.1 研究动机

传感器内部不同的成像流程和热量耗散导致观察到的红外和可见图像之间存在未对齐现象, 表现为移位和变形. 观察发现, 直接融合未对齐的红外和可见图像经常会出现严重的重影. 受到 [17,19] 的启发, 通过图像到图像的转换来降低跨模态差异. 为了迎合红外图像“重结构轻纹理”的固有特性, 本文提出了一种专门的跨模态生成配准范式, 以减轻未对齐的红外和可见图像融合过程中的重影。

2.2 跨模态感知风格迁移

本文提出的跨模态生成-配准范式 (CGRP) 的核心部分之一是跨模态图像生成. 考虑到红外图像容易因热辐射而失真, 提出了一种跨模态感知风格转移网络 CPSTN 将可见图像 I_{vis} 转换为对应

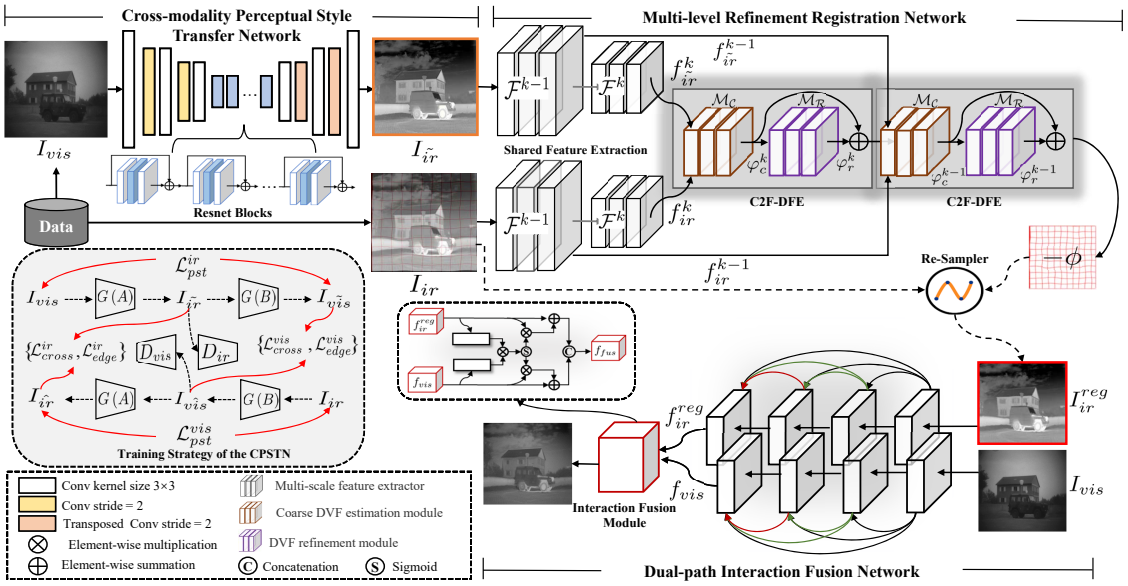


图 1 所提出的无监督跨模态融合网络的工作流程, 用于未对齐的红外和可见图像. 该网络主要由三个子网络组成, 分别是跨模态感知风格转移网络 (CPSTN)、多级细化注册网络 (MRRN) 和双路径交互融合网络 (DIFN). 该方法以未对齐的红外可见图像作为输入, 依次执行上述三个网络以获得最终的融合图像.

的伪红外图像 $I_{\tilde{ir}}$, 一起形成了一个伪红外图像对 $(I_{ir}, I_{\tilde{ir}})$, 以提供统一的表示. 正如图 1所, CPSTN 由两个生成器 $G(A)$ 和 $G(B)$ 以及两个判别器 D_{ir} 和 D_{vis} 构成. 生成器的网络是一个 U 形结构, 它的底部包含 9 个残差块. 不同于 CycleGAN [25], 我们倾向于提出一种由感知风格转移约束控制的特定学习策略, 并在两条循环生成路径之间建立路径间相关性, 以进一步优化伪红外图像中清晰结构的生成. 伪红外图像通过 $I_{\tilde{ir}} = \mathcal{T}_\theta(I_{ir})$ 生成, $\mathcal{T}_\theta$ 指的是参数为 θ 的 CPSTN. 这个优化过程在图 1的左侧虚线框中进行了说明, 感知风格转移约束和路径间相关性的正则化在第 2.5节中进行了解释.

2.3 多尺度细化配准

由于 CPSTN 减少了跨模态差异, 单模态设置下的红外图像配准成为 CGRP 的另一个核心部分. 正如图 1所示, 本文利用多尺度细化配准网络 MRRN 来预测失真和伪红外图像之间的变形场 (DF) 并重建配准的红外图像. MRRN 由一个共享的多尺度特征提取器 (SM-FE)$\mathcal{F}^k$, 两个由粗到细的变形场预测模块 (C2F-DFE), 以及一个重采样层构成. 在每一个 C2F-DFE 中, 包含一个粗的 DFE 模块 $\mathcal{M}_\mathcal{C}$ 和一个精细的 DFE 模块 $\mathcal{M}_\mathcal{R}$. 那么，粗变形场首先被预测为：

$$\varphi_c^k = \mathcal{M}_\mathcal{C}\left(\mathcal{F}^k(I_{\tilde{ir}}, I_{ir})\right), \tag{1}$$

细化变形场由下式估计：

$$\varphi_r^k = \mathcal{M}_\mathcal{R}\left(\varphi_c^k\right) \oplus \varphi_c^k, \tag{2}$$

这里的 $\mathcal{F}^k$ 指的是第 k 尺度的特征提取器. 假设 SM-FE 总共包含 K 个尺度, 当 $k = K$ 时, 最终的变形场 $-\phi = \varphi_r^K$ 就被估计到了. 最后, 我们使用类似于 STN [8] 的重采样层去重构配准后的红外图像, 如下式所示:

$$I_{ir}^{reg} = I_{ir} \circ (-\phi), \tag{3}$$

操作 $\circ$ 指的是用于配准的空间变换.

2.4 双路径交互融合

为了融合配准后的红外图像 I_{ir}^{reg} 和可见光图像 I_{vis}, 我们提出了一种双路径交互融合网络 (DIFN). 该网络双路径特征提取模块和特征交互融合模块组成. 这里, 双路径特征提取模块的结构继承了残差密集网络 [23], 提取的特征表示为

$$f_{ir}^{reg}, f_{vis} = \mathcal{M}_{\theta_E}\left(I_{ir}^{reg}, I_{vis}\right), \tag{4}$$

在这个公式中, $\mathcal{M}_{\theta_E}$ 是特征提取模块，并且 θ_E 是它的参数.

交互融合模块. 正如图 1中间的虚线框所示, 本文提出了一个特征交互融合模块 (IFM) 去自适应地从红外和可见光图像中选择特征进行融合. 为了关注更多重要的特征, 我们需要通过下式重新校正特征响应：

$$\mathcal{A}_{tt} = \mathcal{S}\left(Conv_{1\times1}\left(f_{ir}^{reg}\right) \otimes Conv_{1\times1}\left(f_{vis}\right)\right), \tag{5}$$

那么校正后的红外与可见光图像特征表示为:

$$\begin{aligned} f_{ir}^{Att} &= f_{ir}^{reg} \otimes \left(1 + \mathcal{A}_{tt}\right), \\ f_{vis}^{Att} &= f_{vis} \otimes \left(1 + \mathcal{A}_{tt}\right), \end{aligned} \tag{6}$$

我们获得最终的融合结果为：

$$I_{fus} = Conv_{3\times3}\left(Concat\left(f_{ir}^{Att}, f_{vi}^{Att}\right)\right), \tag{7}$$

式中的 $\mathcal{S}$ 是 Sigmoid 函数, $\otimes$ 指的是像素间的乘法操作.

2.5 损失函数

感知风格迁移损失. 为了生成更逼真的伪红外图像, 本文引入了感知风格迁移 (PST) 损失来控制 CPSTN 中的循环一致性. 这个 PST 损失函数由两项组成, 第一项是感知损失 $\mathcal{L}_{pcp}$, 第二项是风格损失 $\mathcal{L}_{sty}$. 首先, $\mathcal{L}_{pcp}$ 被定义为:

$$\begin{aligned} \mathcal{L}_{pcp}^{\psi_j} = &\left\|\psi_j(I_{vis}) - \psi_j\left(G_B\left(G_A(I_{vis})\right)\right)\right\|^2 \\ &+\left\|\psi_j(I_{ir}) - \psi_j\left(G_A\left(G_B(I_{ir})\right)\right)\right\|^2, \end{aligned} \tag{8}$$

式中的 ψ_j 是 VGG-19 [16] 模型中第 j 层的特征, 并且 $j \in [2, 7, 12, 21, 30]$, 对应的权重为 $\omega \in [\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, 1, 1]$. 这些特征也被用于计算 $\mathcal{L}_{sty}$, 这一项用来度量 $(I_{vis}, I_{\tilde{vis}})$ 和 $(I_{ir}, I_{\hat{ir}})$ 图像对之间的统计误差, 并且被定义为：

$$\begin{aligned} \mathcal{L}_{sty}^{\psi_j} = &\omega_j\left\|\mathcal{G}_{\psi_j}(I_{vis}) - \mathcal{G}_{\psi_j}\left(G_B\left(G_A(I_{vis})\right)\right)\right\|^2 \\ &+\omega_j\left\|\mathcal{G}_{\psi_j}(I_{ir}) - \mathcal{G}_{\psi_j}\left(G_A\left(G_B(I_{ir})\right)\right)\right\|^2, \end{aligned} \tag{9}$$

式中的 $\mathcal{G}$ 指的是 Gram [15] 矩阵. 它是一种有效的工具，用于抑制图像中的棋盘伪影. 总的感知风格迁移损失定义为:

$$\mathcal{L}_{pst} = \lambda_p\mathcal{L}_{pcp} + \lambda_s\mathcal{L}_{sty}, \tag{10}$$

交叉正则项损失. 我们创造性地提出 CPSTN 中两条循环路径之间的交叉正则化 $\mathcal{L}_{cross}$ 以建立路径间相关性. 这一项包含内容正则项 $\mathcal{L}_{con}$ 和边缘正则项 $\mathcal{L}_{edge}$, 被定义为

$$\begin{aligned}\mathcal{L}_{con} &= \|I_{\tilde{ir}} - I_{\hat{ir}}\|_1 + \|I_{\tilde{vis}} - I_{\hat{vis}}\|_1, \\ \mathcal{L}_{edge} &= \|\nabla I_{\tilde{ir}} - \nabla I_{\hat{ir}}\|_{char} + \|\nabla I_{\tilde{vis}} - \nabla I_{\hat{vis}}\|_{char},\end{aligned} \tag{11}$$

式中的 ∇ 指的是拉普拉斯梯度算子. 使用 Charbonnier Loss [9] 计算 $\mathcal{L}_{edge}$. 总的交叉正则可以通过下式计算得到：

$$\mathcal{L}_{cross} = \lambda_c \mathcal{L}_{con} + \lambda_e \mathcal{L}_{edge}, \tag{12}$$

配准损失. 本文采用双向相似性损失来约束特征空间中失真和伪红外图像之间的对齐, 其定义为:

$$\mathcal{L}_{sim}^{bi} = \|\psi_j(I_{ir}^{reg}) - \psi_j(I_{\tilde{ir}})\| + \lambda_{rev}\|\psi_j(\phi \circ I_{\tilde{ir}}) - \psi_j(I_{ir})\|_1, \tag{13}$$

第一项为前向变形, 第二项为反向变形, 权重 $\lambda_{rev} = 0.2$. 其中反向变形场 ϕ 用于扭曲伪红外图像 $I_{\tilde{ir}}$ 并使其接近扭曲的输入 I_{ir}. 为了确保预测到一个平滑的变形场，我们定义了一个平滑损失函数如下：

$$\mathcal{L}_{\text{smooth}} = \|\nabla\phi\|_1, \tag{14}$$

那么，配准部分的总的损失函数通过下式来计算:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{sim} + \lambda_{sm}\mathcal{L}_{smooth}, \tag{15}$$

其中, 在本文的工作中 λ_{sm} 被设置为 10.0.

融合损失. 在融合阶段, 我们利用 MS-SSIM 损失函数 $\mathcal{L}_{ssim}^{ms}$ 来保持融合图像清晰的强度分布. 该损失函数表示如下:

$$\mathcal{L}_{ssim}^{ms} = (1 - SSIM(I_{fus}, I_{ir}^{reg})) + (1 - SSIM(I_{fus}, I_{vis})), \tag{16}$$

为了鼓励纹理细节的恢复, 我们对梯度分布进行建模并开发联合梯度损失, 表示为:

$$\mathcal{L}_{JGrad} = \|max(\nabla I_{ir}^{reg}, \nabla I_{vis}), \nabla I_{fus}\|_1, \tag{17}$$

为了获得更清晰的纹理, 融合图像的梯度被迫接近红外和可见图像梯度之间的最大值. 此外, 为了保留 IR-VIS 图像中的显着性目标, 我们利用由 [4] 启发的自视觉显着性图来构建另一个逐像素约束, 表示为:

$$\begin{aligned}\boldsymbol{\omega}_{ir} &= \boldsymbol{S}_{f_{ir}} / (\boldsymbol{S}_{f_{ir}} - \boldsymbol{S}_{f_{vis}}), \boldsymbol{\omega}_{vis} = 1 - \boldsymbol{\omega}_{ir}, \\ \mathcal{L}_{svs} &= \| (\boldsymbol{\omega}_{ir} \otimes I_{ir}^{reg} + \boldsymbol{\omega}_{vis} \otimes I_{vis}), I_{fus}\|_1,\end{aligned} \tag{18}$$

式中的 $\boldsymbol{S}$ 指的是显著性矩阵, $\boldsymbol{\omega}_{vis}$ 和 $\boldsymbol{\omega}_{ir}$ 分别指的是红外和可见光图像的加权图. 融合阶段总的损失函数可以通过下式来计算：

$$\mathcal{L}_{fus} = \lambda_{ssim}\mathcal{L}_{ssim}^{ms} + \lambda_{JG}\mathcal{L}_{JGrad} + \lambda_{svs}\mathcal{L}_{svs}, \tag{19}$$

在这里, λ_{ssim}, λ_{JG} 和 λ_{svs} 分别被设置为 1.0, 20.0 和 5.0.

整体损失. 我们通过最小化以下整体损失函数来训练我们的网络:

$$\mathcal{L}_{total} = \mathcal{L}_{pst} + \mathcal{L}_{cross} + \mathcal{L}_{GAN} + \mathcal{L}_{reg} + \mathcal{L}_{fus}. \tag{20}$$

需要说明的是 $\mathcal{L}_{GAN}$ 继承于 CycleGAN [25] 用于判别伪红外图像的真假性。

3 实验及分析

3.1 数据集与执行细节

数据集. 我们的跨模态图像配准和融合是在两个广泛使用的数据集上进行的: TNO[1)]和 RoadScene[2)], 它们的数据都是预先配准好的. 为了满足未对齐图像对的要求, 我们首先通过执行不同程度的仿射和弹性变换来生成几个变形场 (DF), 然后将它们应用于红外图像以获得失真图像. 我们使用来自 RoadScene 数据集的所有 221 张图像作为训练样本. 由于 TNO 数据集的数量有限，我们仅将其用作测试集。此外, 我们从 RoadScene 数据集中随机选择 55%（121 张图像）图像进行测试, 因为我们的方法是完全无监督的.

执行细节. 我们方法的代码是使用带有 NVIDIA 2080Ti GPU 的 PyTorch 框架实现的. 我们随机选择 8 付大小为 256×256 的图像构成一个 batch, 使用 Adam($\beta_1 = 0.9$, $\beta_2 = 0.999$) 优化器优化我们的模型. 初始的学习率为 1×10^{-3} 并且在经过 300 epoch 的整个训练阶段保持不变.

3.2 配准性能评估

我们使用三个常用指标评估中间注册结果, 包括 MSE、互信息 (MI) [14] 和归一化互相关 (NCC) [3]. 考虑到多模态图像对齐方法比较缺乏, 为了与我们方法进行公平的比较, 我们选择了两种典型的基于变形场 DF 的图像对齐算法, 分别是 FlowNet [6] 和 VoxelMorph [2]. 需要注意的是, FlowNet 仅预测输入图像对的 DF, 因此使用空间变换层 (STN) [8] 来生成配准后的图像. 正如表 1所示, 直接使用这两种算法来执行 IR-VIS 图像的跨模态对齐产生的改进可以忽略不计, 甚至比未对齐输入更差. 相比之下, 本文提出的跨模态生成配准范式在 TNO 和 RoadScene 数据集上的 MI 指标分别获得了大约 **0.10** 和 **0.23** 的提升. 事实证明, 所提出的 CGRP 比现有流行的图像对齐方法更有效.

表 1 TNO 和 Roadscene 数据集上跨模态图像对齐的定量比较. 最好的结果用 红色粗体显示.

Methods	TNO			RoadScene		
	MSE↓	NCC↑	MI↑	MSE↓	NCC↑	MI↑
Misaligned Input	0.007	0.876	1.558	0.01	0.894	1.602
FlowNet + STN	0.029	0.636	1.095	0.009	0.910	1.549
FlowNet + STN + CPST	0.007	0.893	1.465	0.005	0.949	1.744
VoxelMorph	0.007	0.880	1.545	0.008	0.914	1.589
VoxelMorph + CPST	0.005	0.919	1.573	0.006	0.941	1.689
Ours CGRP	**0.004**	**0.926**	**1.648**	**0.004**	**0.963**	**1.833**

3.3 融合性能评估

我们使用几种最先进的 IVIF 方法对所提出的方法进行定量和定性评估, 包括 DENSE [10], DIDFuse [24], FGAN [13], GAN-FM [22], MFEIF [12], RFN [11] 和 U2F [20], 配备 FlowNet+STN 和 VoxelMorph 这两种典型的图像对齐算法用于跨模态图像配准. 为了公平比较, 我们使用相同的测试样本 (来自 RoadScene 和 TNO 数据集的 211 和 24 张图像) 来评估上述方法.

1) http://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.

2) https://github.com/hanna-xu/RoadScene.

表 2 本文方法与八种最先进的融合方法在两个常见数据集上的定量比较 (使用 FlowNet+STN 和 VoxelMorph 算法作为基本配准模型).

	DENSE	DIDFuse	FGAN	GAN-FM	MFEIF	PMGI	RFN	U2F	Ours
	①: **Aligned Method: (FlowNet+STN) / VoxelMorph** ②: **Dataset: RoadScene**								
CC↑	0.589/**0.597**	0.573/0.582	0.502/0.582	0.565/0.575	0.588/0.596	0.530/0.538	0.587/0.593	0.558/0.569	**0.621**
VIF↑	0.488/0.521	0.445/0.469	0.493/0.513	0.564/0.597	0.619/**0.659**	0.408/0.447	0.563/0.590	0.430/0.464	**0.895**
SSIM↑	0.311/0.355	0.301/0.335	0.255/0.281	0.334/0.368	0.342/**0.380**	0.266/0.314	0.319/0.347	0.297/0.343	**0.507**
	①: **Aligned Method: (FlowNet+STN) / VoxelMorph** ②: **Dataset: TNO**								
CC↑	0.390/0.463	0.380/0.452	0.315/0.400	0.350/0.419	0.381/0.451	0.363/0.447	0.392/**0.465**	0.378/0.456	**0.481**
VIF↑	0.675/0.742	0.545/0.597	0.588/0.654	0.556/0.627	0.701/**0.780**	0.518/0.606	0.671/0.728	0.535/0.601	**1.016**
SSIM	0.318/**0.382**	0.266/0.328	0.236/0.294	0.295/0.372	0.297/0.379	0.277/0.363	0.300/0.349	0.299/0.372	**0.473**

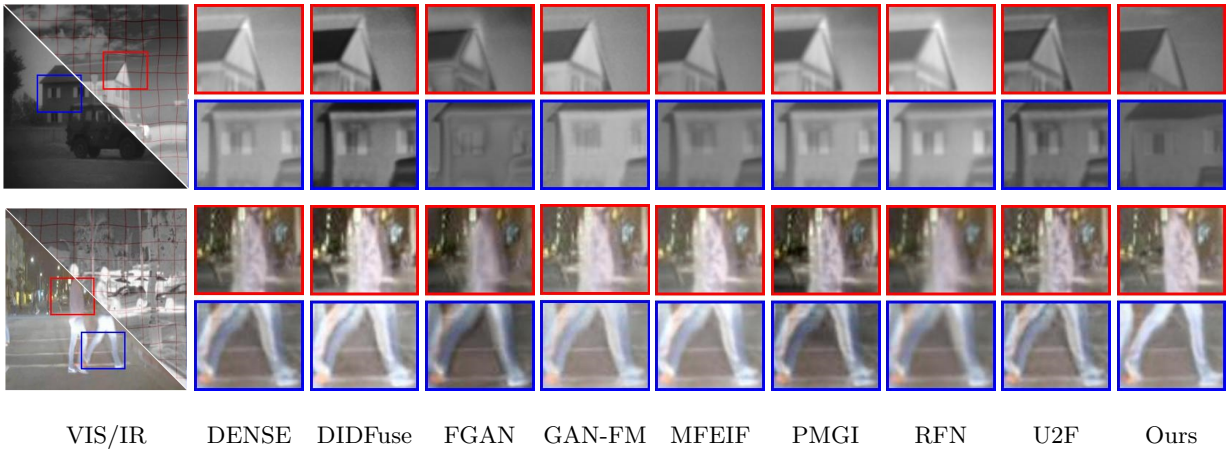


图 2 不同 IVIF 方法在 TNO 和 RoadScene 数据集上的定性比较 (使用 FlowNet+STN 算法作为基本配准模型).

定量评估. 我们在 Table 2中报告了 TNO 和 RoadScene 数据集的定量融合结果. 可以看出, 我们的方法在数值上大大优于现有的 IVIF 方法, 并且在三个指标中均排名第一, 包括互相关 (CC), 视觉信息保真度 (VIF) [5] 和结构相似性指数 (SSIM) [18]. 特别是与最近的一项研究 GAN-FM [22] 相比, 我们的方法在 TNO 和 RoadScene 数据集上的 VIF 获得了 **0.33** 和 **0.46** 的改进, 这证明了所提出方法的优越性.

定性评估. 由于篇幅限制, Figure 2 仅给出了使用 FlowNet+STN 作为基础配准模型的定性比较. 通过观察局部放大区域, 我们可以发现现有工作的方法未能达到理想的对齐和鬼影消除. 相比之下, 所提出的方法显示出良好的对齐和融合能力, 同时保留了清晰的结构.

模型效率分析. 为了验证模型效率, 我们在 Table 3中报告了参数量 (M) 和运行时间 (s) 的统计数据. 我们在单个 2080ti GPU 上对大小为 64×64 的配对图像进行测量. 需要注意的是, 我们的结果是通过联合测量配准和融合网络获得的, 而其他方法只有一个融合网络. 结果表明, 我们的方法至少在运行时间上具有竞争力.

表 3 本文提出的联合配准融合模型与最先进的融合方法的效率分析.

	DENSE	FGAN	DIDFuse	U2F	PMGI	RFN	MFEIF	GAN-FM	Ours
Param.	0.925	**0.074**	0.261	0.659	**0.042**	10.94	0.158	10.21	0.80
Time	0.124	0.251	0.055	0.123	0.182	0.239	**0.045**	0.334	**0.024**

3.4 消融实验

CPSTN 的有效性. 我们将 CPSTN 插入 FlowNet 和 VoxelMorph 作为其改进版本, 并采用它们进行跨模态 IR-VIS 图像配准. 如表 1 所示，与 CPSTN 合作的定量结果比原始版本有很大的提高。因此，图 3 提供的视觉比较表明了 CPSTN 的有效性。我们观察到配备 CPSTN 的 FlowNet 模型生成的注册结果消除了明显的变形。此外, 我们从图像融合的角度验证了 CPSTN 的有效性. 如表 4 中的量化结果, 使用 CPSTN 在 TNO 和 RoadScene 数据集上分别获得了 VIF 上 **0.14** 和 **0.22** 的改进。图 4(b) 和 (c) 中相应的定性比较表明, 对于未对准的红外和可见图像, CPSTN 有助于产生可忽略重影的良好融合结果. 上述结果从配准和融合的角度全面揭示了 CPSTN 的有效性.

表 4 CPSTN ($\mathcal{T}$) 和 MRRN ($\mathcal{R}$) 在 TNO 和 RoadScene 数据集上的消融研究.

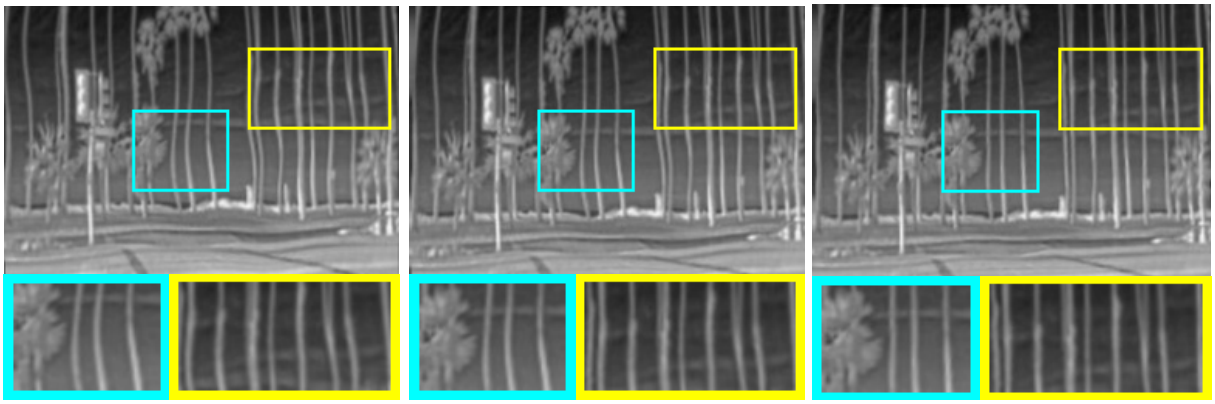
Datasets	CPSTN	MRRN	DIFN	Metrics		
				CC↑	SSIM↑	VIF↑
TNO	✗	✗	✓	0.438	0.369	0.732
	✗	✓	✓	$0.476_{\uparrow 0.038}$	$0.418_{\uparrow 0.049}$	$0.876_{\uparrow 0.144}$
	✓	✓	✓	$\mathbf{0.481}_{\uparrow 0.044}$	$\mathbf{0.473}_{\uparrow 0.103}$	$\mathbf{1.016}_{\uparrow 0.374}$
Road	✗	✗	✓	0.563	0.367	0.543
	✗	✓	✓	$0.601_{\uparrow 0.038}$	$0.422_{\uparrow 0.055}$	$0.673_{\uparrow 0.130}$
	✓	✓	✓	$\mathbf{0.621}_{\uparrow 0.085}$	$\mathbf{0.507}_{\uparrow 0.140}$	$\mathbf{0.895}_{\uparrow 0.352}$

MRRN 的有效性. 假设在没有 CPSTN 的情况下, 我们研究了 MRRN 对未对齐的 IVIF 任务的影响. 需要注意的是, MRRN 在相同设置下将成对的红外图像和可见图像作为输入执行跨模态图像配准. 正如 Table 4报告, 我们对比了前两个实验在 TNO 和 RoadScene 数据集上的结果, 与直接融合未对齐的红外和可见图像相比, 实验结果表明使用 MRRN 实现跨模态配准也一定程度上提高了 IVIF 的性能.

CPSTN 中损失函数的有效性.

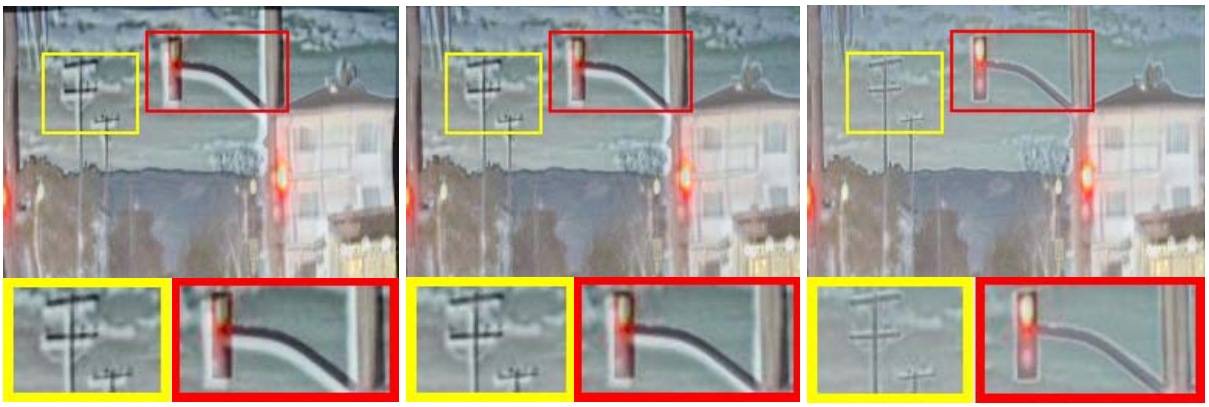
我们分析了 CPSTN 中使用的 $\mathcal{L}_{pst}$ 和 $\mathcal{L}_{cross}$ 的作用. 如图 5, 不使用 $\mathcal{L}_{pst}$ 和 $\mathcal{L}_{cross}$ 的约束, 生成的伪红外图像会受到影响, 与参考图像 (图 5(a)) 相比, 严重的结构退化 (图 5(c)). 使用 $\mathcal{L}_{cross}$, 模型保留了一般的结构信息, 而微妙的结构维护得不够好, 明显引入了“棋盘伪影”(图 5(d). 相比之下, 我们的模型生成的伪红外图像 (图 5(e)) 具有更清晰的几何结构, 这符合红外图像“重结构轻纹理”的特性.

交互融合模块的有效性. 本文通过拼接操作替换了 DIFN 中的 IFM, 以研究其对 IVIF 的有效性. 如图 6 所示, 不使用 IFM 的模型倾向于生成低对比度的平滑纹理, 而使用 IFM 的结果具有更丰富, 更清晰的纹理, 具有正高对比度. 结果表明自适应选择待融合特征对于提高融合图像质量更为有效.



(a) FlowNet (b) FolwNet+CPSTN (c) Ours

图 3 在 RoadScene 数据集上从配准的角度验证 CPSTN 的有效性.



(a) w/o $\mathcal{T}$,$\mathcal{R}$ (b) w/o $\mathcal{R}$ (c) Ours

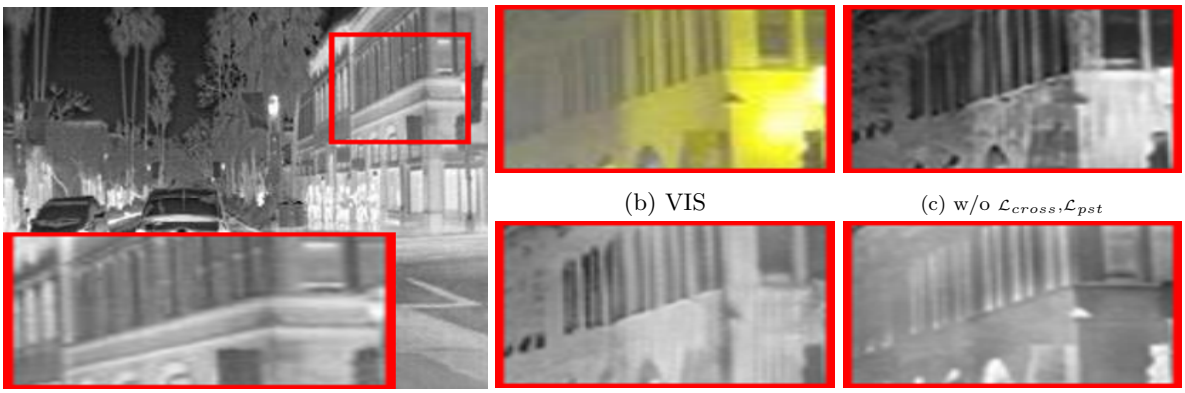
图 4 CPSTN ($\mathcal{T}$) 和 MRRN ($\mathcal{R}$) 在 RoadScene 数据集上的有效性分析.

3.5 额外分析

为了验证所提出的 CPSTN 的泛化性, 我们将其插入两个典型的图像对齐算法 (即 FlowNet [6] 和 VoxelMorph [2]) 来实现单模态扭曲红外图像和伪红外图像之间的图像配准, 并进一步对配准的红外图像和可见图像在八种比较方法上进行融合. 如图 7, 每个方法的第一个块是没有使用 CPSTN 的模型的结果, 第二个块是使用 CPSTN 的模型的结果. 可以观察到：i) 这些配备 CPSTN 的对齐模型可以有效地促进现有 IVIF 方法对未对齐图像的融合性能, 这表明 CPSTN 对未对齐 IVIF 具有良好的泛化能力. ii) 尽管以 i) 为条件, 所提出的方法在未对齐的跨模态图像上仍然优于这些改进的 IVIF 方法, 这解释了我们整体框架的优越性依赖于每个组件的合作.

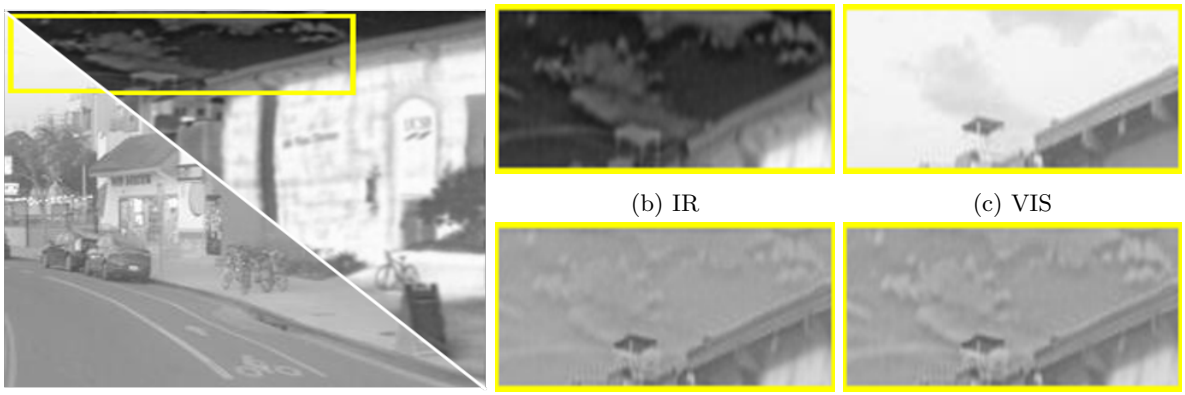
4 Conclusion

在本文中, 我们提出了一种高度鲁棒的无监督未对准红外和可见图像融合框架, 用于减轻融合图像的重影。我们利用生成配准范式将跨模态图像对齐简化为单模态配准. 开发了一个特征交互融合



(a) Reference (b) VIS (c) w/o $\mathcal{L}_{cross}$,$\mathcal{L}_{pst}$ (d) w/o $\mathcal{L}_{pst}$ (e) Ours

图 5 在 RoadScene 数据集上验证 CPSTN 中损失函数的有效性.



(a) VIS/IR (b) IR (c) VIS (d) w/o IFM (e) w/ IFM

图 6 在 RoadScene 数据集上分析 IFM 的有效性.

模块, 以自适应地从红外和可见图像中选择有意义的特征进行融, 避免特征平滑并强调忠实的纹理. 广泛的实验结果证明了我们的方法在未对齐的跨模态图像融合方面的卓越能力. 重要的是, 我们的生成配准范式可以很好地扩展到现有的 IVIF 方法, 以提高它们在未对齐的跨模态图像上的融合性能.

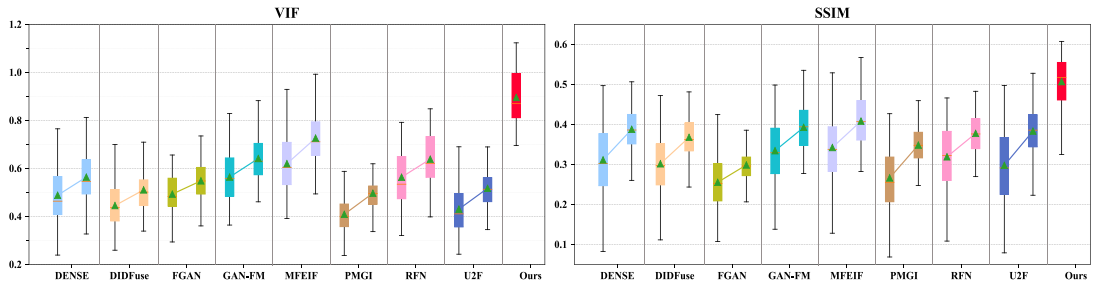


图 7 CPSTN 的泛化性分析.

参考文献

1 M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *CVPR*, pages 13407–13416, 2020.

2 G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca. An unsupervised learning model for deformable medical image registration. In *CVPR*, pages 9252–9260, 2018.

3 X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, and D. Shen. Deep learning based inter-modality image registration supervised by intra-modality similarity. In *MLMI*, pages 55–63, 2018.

4 S. Ghosh, R. G. Gavaskar, and K. N. Chaudhury. Saliency guided image detail enhancement. In *NCC*, pages 1–6, 2019.

5 Y. Han, Y. Cai, Y. Cao, and X. Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013.

6 E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 1647–1655, 2017.

7 P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017.

8 M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *NeurIPS*, pages 2017–2025, 2015.

9 W. Lai, J. Huang, N. Ahuja, and M. Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2599–2613, 2019.

10 H. Li and X. Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

11 H. Li, X. Wu, and J. Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.

12 J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

13 J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

14 D. Mahapatra, Z. Ge, S. Sedai, and R. Chakravorty. Joint registration and segmentation of xray images using generative adversarial networks. In Y. Shi, H. Suk, and M. Liu, editors, *MICCAI*, volume 11046, pages 73–80, 2018.

15 M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, pages 4501–4510, 2017.

16 K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun, editors, *ICLR*, 2015.

17 G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3622–3631, 2019.

18 Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

19 Z. Wang, Z. Wang, Y. Zheng, Y. Chuang, and S. Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019.

20 H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022.

21 Q. Yang, N. Li, Z. Zhao, X. Fan, E. I. Chang, and Y. Xu. MRI image-to-image translation for cross-modality image registration and segmentation. *CoRR*, abs/1801.06940, 2018.

22 H. Zhang, J. Yuan, X. Tian, and J. Ma. GAN-FM: infrared and visible image fusion using GAN with full-scale skip connection and dual markovian discriminators. *IEEE Transactions on Computational Imaging*, 7:1134–1147, 2021.

23 Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2021.

24 Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, and P. Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976, 2020.

25 J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017.