文章编号:1001-5078(2024)03-0457-09　　·图像与信号处理·

基于 Transformer 和空间注意力的红外与可见光图像融合

耿　俊,吴子豪,李文海,李晓瑜
(新疆大学软件学院,新疆 乌鲁木齐 830091)

摘　要:目前,已经有很多研究人员将卷积神经网络应用到红外与可见光图像融合任务中,并取得了较好的融合效果。其中有很多方法是基于自编码器架构的网络模型,这类方法通过自监督方式进行训练,在测试阶段需要采用手工设计的融合策略对特征进行融合。但现有的基于自编码器网络的方法很少能够充分地利用浅层特征和深层特征,而且卷积神经网络受到感受野的限制,建立长距离依赖较为困难,因而丢失了全局信息。而 Transformer 借助于自注意力机制,可以建立长距离依赖,有效获取全局上下文信息。在融合策略方面,大多数方法设计的较为粗糙,没有专门考虑不同模态图像的特性。因此,在编码器中结合了 CNN 和 Transformer,使编码器能够提取更加全面的特征。并将注意力模型应用到融合策略中,更精细化地优化特征。实验结果表明,该融合算法相较于其他图像融合算法在主观和客观评价上均取得了优秀的结果。

关键词:图像融合;深度学习;Transformer;红外图像;可见光图像
中图分类号:TP391.12;TN29　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2024.03.018

Infrared and visible image fusion based on transformer and spatial attention model

GENG Jun, WU Zi-hao, LI Wen-hai, LI Xiao-yu
(College of Software, Xinjiang University, Urumqi 830091, China)

Abstract: Currently, the applications of convolutional neural networks to the task of fusing infrared and visible images have achieved better fusion results. Many of these methods are based on network models with self-encoder architectures, which are trained in a self-supervised methods and require the use of hand-designed fusion strategies to fuse features in the testing phase. However, existing methods based on self-encoder networks rarely make full use of both shallow and deep features, and convolutional neural networks are limited by the receptive field, making it more difficult to establish long-range dependencies and thus losing global information. In contrast, Transformer, with the help of self-attention mechanism, can establish long-range dependencies and effectively obtain global contextual information. In terms of fusion strategies, most of the methods are designed in a crude way and do not specifically consider the characteristics of different modal images. Therefore, CNN and Transformer are combined in the encoder to enable the encoder to extract more comprehensive features. And the attention model is applied to the fusion strategy to optimize the features in a more refined way. The experimental results show that the fusion algorithm achieves excellent results in both subjective and objective evaluations compared to other image fusion algorithms.

Keywords: image fusion; deep learning; Transformer; infrared image; visible image

1　引　言

可见光图像富含清晰的纹理和细节信息,但在光照弱或伪装条件下,信息丢失严重。红外图像可以凸显复杂背景下的热目标,例如人和车辆等,但噪

基金项目:新疆维吾尔自治区自然科学基金项目(No. 2021D01C077)资助。
作者简介:耿　俊(1977 -),男,硕士,高级工程师,研究方向为人工智能、计算机网络。E-mail:gengjun@xju.edu.cn
通讯作者:吴子豪(1997 -),男,硕士,研究生,研究方向为红外与可见光图像融合。E-mail:761545864@qq.com
收稿日期:2023-03-20;**修订日期:**2023-04-15

声较大,细节模糊,视觉效果较差。它们分别是使用不同种类的传感器得到的,而不同种类传感器对同一个场景往往会有截然不同的场景描述,仅依靠单种传感器难以对场景进行全面表征。因此,红外与可见光图像融合的目标就是通过有效地提取和融合不同模态的图像中互补的特征信息,生成单幅信息量更丰富、场景表达更完整的融合图像,提升用户的视觉感知体验[1]。目前,红外与可见光图像融合已广泛应用于军事侦察和目标检测等领域。

根据所采用的方法不同,融合方法分为传统图像融合方法和基于深度学习的图像融合方法。在传统方法中,主要包括多尺度变换、稀疏表示、低秩表示、基于显著性的方法和混合方法。这些方法通常都设计一个固定的表示模型来提取特征,然后采用手工设计的融合策略对特征进行融合,最后通过逆变换重构得到融合图像。但这些方法未充分考虑不同模态图像之间的差异,仅对它们采用相同的特征提取方法,导致特征提取不充分,融合效果也不稳定。并且由于设计的方法比较复杂,导致计算成本较高,还可能会引入大量噪声。不同于传统的融合方法,基于深度学习的方法可通过一系列可学习的卷积核自动提取不同模态的特征,并通过强大的非线性表示能力来建立输入和输出的复杂关系,缓解了以上所提的传统方法的缺陷。比如,Li 等于 2018 年和 2020 年相继提出的 DenseFuse[2] 和 NestFuse[3],两者都采用自编码器架构。其中 DenseFuse 使用了带有密集连接块的编码器,增强了编码器的特征提取能力。NestFuse 通过提取多尺度特征和具备嵌套连接结构的解码器,提高了编码器的特征提取能力和解码器的重建能力。2019 年,Ma 等提出了 FusionGan[4],采用生成对抗策略进行图像融合,这种方法是一种端到端的方式,免去了融合规则的设计,通过判别器和生成器不断地对抗训练,最终让生成器生成以假乱真的图像,即融合图像与两幅源图像都非常相似。

相比于已经存在的融合方法,这些基于 CNN 和 GAN 的融合框架取得了较好的融合效果。但是在这些方法中,也有一些被忽视的地方,主要分为三点。首先,大多数方法没有很好地利用多尺度特征,并且没有很好地利用浅层特征和深层特征。其次,这些方法都没有关注到全局依赖,因为 CNN 的感受野较小,主要提取的是局部特征,所以会丢失部分全局上下文信息,这也是 CNN 固有的缺陷。最后,大多数方法的融合策略较为粗糙,未充分考虑红外图像和可见光图像不同模态的特性,这也导致融合效果受到约束。

为此,提出了基于 Transformer 和空间注意力的融合框架(Fusion Architecture Based on Transformer and Spatial Attention Model, TAFuse)。该框架包含三个关键部分,分别是编码器、融合策略和解码器。一方面,引入了多尺度和跳跃连接,更充分的提取和利用浅层特征和深层特征。其中编码器模块把 CNN 和 Transformer 进行结合,两者取长补短,发挥各自优势,更全面地提取了局部信息和全局信息。另一方面,通过空间注意力融合策略,对编码器提取到的特征进行精细化地优化调整,得到合理的权重图,使其更符合红外与可见光图像融合的目的。

2 相关技术

2.1 自编码器网络

在图像融合领域,自编码器网络引起了极大的关注,近年来涌现了各种基于自编码器的融合方法。编码图像是指获得稀疏系数,然后用适当的融合规则进行融合,最后通过解码器重建得到融合图像。其一般步骤如下:

$$X_{\text{features}} = Encode(X_{\text{input}}) \tag{1}$$

$$Y_{\text{features}} = Encode(Y_{\text{input}}) \tag{2}$$

$$Z_{\text{features}} = Fuse(X_{\text{features}}, Y_{\text{features}}) \tag{3}$$

$$Z_{\text{out}} = Decode(Z_{\text{features}}) \tag{4}$$

式(1)和式(2)代表对不同的源图像提取特征。式(3)代表通过融合策略得到融合特征。式(4)代表解码操作,对融合特征解码得到融合图像。

2.2 Transformer

Transformer[5] 首先由 Vaswani 提出,这是一个纯自注意力机制模型,起初在自然语言处理领域广泛应用。目前席卷了计算机视觉领域,挑战了 CNN 的地位,比如 Dosovitskiy 等提出用于图像分类任务的 Vision Transformer[6],极大地促进了在视觉任务中采用 Transformer 模型。这主要得益于其自注意力机制,它可以建立长距离依赖,弥补 CNN 较难获取全局感受野的缺陷。

3 TAFuse 模型介绍

为了清晰展现 TAFuse 网络模型,先介绍该模型的训练框架,再介绍模型的测试框架。同时,为了清晰地展现具体流程细节,分别用虚线箭头和向右的

实线箭头表示上下采样和跳跃连接，其他实线箭头表示普通连接。

图 1 所示，训练框架包含编码器和解码器两个部分，融合层不参与训练，因此被移除。I 和 O 分别代表源图像和重建图像，在它们之间通过损失函数约束网络，以使重建图像和源图像更相似。编码器网络包含一个 Conv_in 模块，四个 ERB（Encoder Residual Block）模块，一个 Transformer 模块和四个由 Transformer 模块学习到的全局空间关系图 Map。Conv 模块代表一层普通卷积层，紧跟着一层激活函数。ERB 模块代表残差卷积块，包含两个 Conv 模块和一个残差连接。通过三次下采样提取多尺度特征。Transformer 模块代表视觉 Transformer，考虑到计算资源的限制，把经过三次下采样得到的特征作为 Transformer 模块的输入，通过学习空间关系得到全局空间关系图 Map，然后将 Map 上采样与不同 ERB 模块提取到的特征相乘得到优化后的特征。在解码器网络中，DCB（Decoder Convolution Block）代表两个 Conv 卷积模块。通过三次上采样使得特征图尺寸和编码器中的特征图尺寸保持一致，并通过跳跃连接将编码器优化后的特征与解码器中对应的特征进行拼接。最后通过一层 Conv_out 卷积模块得到重建图像。网络中，所有卷积核大小为 3 × 3，激活函数都采用 ReLU 激活函数。

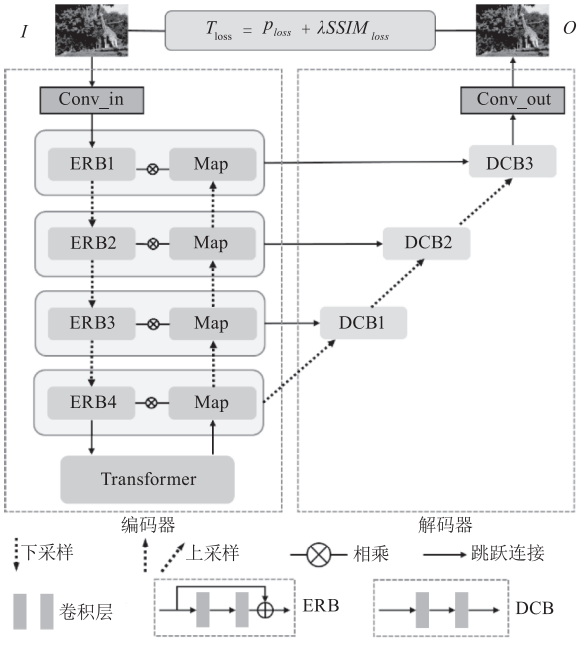


图 1　训练框架

Fig. 1 Training framework

如图 2 所示，在测试框架中加入了融合层 FS（Fusion Strategy）。红外图像 I_i 和可见光图像 I_v 共用一个共享权重的孪生编码网络。首先，将红外图像和可见光图像输入到编码器中，通过 ERB 残差卷积块和三次下采样提取多尺度特征，并且通过 Transformer 模块学习空间关系，对特征进行优化。然后把各个尺度的特征分别输入到融合层，采用特定的融合规则分别对各个尺度的特征进行融合。最后，将融合后的特征输入解码器网络，并且通过跳跃连接把相同尺度的融合特征拼接到解码器，通过解码器重建得到融合图像 O 。

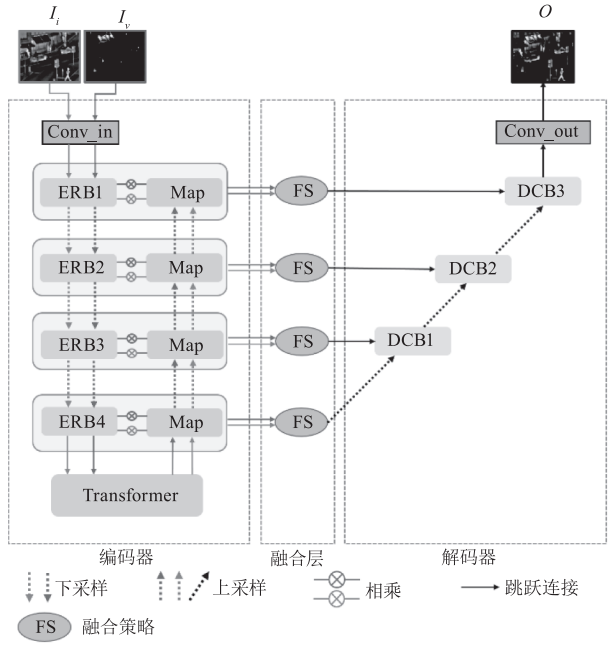


图 2　测试框架

Fig. 2Testing framework

3.1　Transformer 模块

Transformer 模块的结构如图 3 所示，其中，p 表示对特征图分块后小图像块的个数，w 和 h 分别表示特征图在宽度上和高度上分割的图像块数，E 表示维度，Split 表示分割特征图的操作，Flatten 表示把小图像块映射为一维向量，Reshape 表示维度转换。该模块先把输入特征图进行分割，并且将其映射成一组向量。然后，把这组向量输入 Transformer 模型学习全局空间关系。最后，使用全连接层把这组向量还原成原来的维度，并通过 Reshape 操作把这组向量转换成和输入特征图大小一致的全局空间关系图。

3.2　损失函数

红外与可见光图像没有参考的融合图像，所以损失函数是监督网络的关键。总损失函数定义如下：

$$T_{\text{loss}} = P_{\text{loss}} + \lambda SSIM_{\text{loss}} \tag{5}$$

其中 P_{loss} 和 $SSIM_{\text{loss}}$ 代表输入图像 I_{in} 和输出图像 I_{out} 之间的像素损失和结构相似性损失。P_{loss} 计算了输入图像和输出图像之间的距离,在像素级别约束重建图像,使其和输入图像更相似。$SSIM_{loss}$ 表示结构相似性度量。其值越大,输出图像和输入图像在结构上就有更大的相似性。λ 为权重系数。

P_{loss} 的计算公式如下所示:

$$P_{\text{loss}} = \left\| I_{\text{out}} - I_{\text{in}} \right\|_F^2 \tag{6}$$

$SSIM_{\text{loss}}$ 的计算公式如下所示:

$$SSIM_{\text{loss}} = 1 - SSIM(I_{\text{out}}, I_{\text{in}}) \tag{7}$$

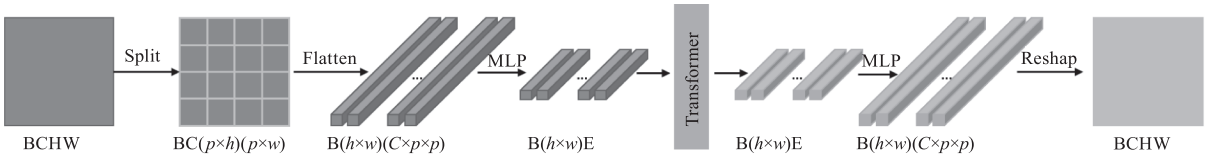


图 3 Transformer 模块的结构

Fig. 3 Structure of the Transformer module

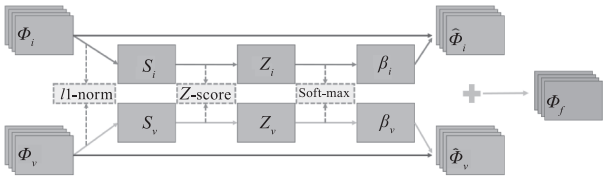


图 4 空间注意力融合策略

Fig. 4 Spatial attention fusion strategy

3.3 融合策略

融合策略是图像融合中的关键环节,它对最终得到的融合图像会产生很大的影响。因此,需要精心地设计合理的融合规则,使其更有针对性地融合重要特征。注意力机制可以赋予重要特征更大的权重,赋予不相关特征更小的权重。通过这种自适应方式进一步对特征进行优化。TAFuse 在融合策略中引入了空间注意力机制。融合策略的具体细节如图 4 所示。其中,$\boldsymbol{\Phi}_i \in R^{M\times N\times C}$ 和 $\boldsymbol{\Phi}_v \in R^{M\times N\times C}$ 分别代表红外与可见光特征图,M、N 和 C 代表高、宽和通道数。

首先,使用 l_1 范数沿着通道方向分别计算出红外和可见光各自的活动水平图 S_i 和 S_v,由如下式计算得到:

$$S_i(x,y) = \left\| \boldsymbol{\Phi}_i^{1:C}(x,y) \right\|_1 \tag{8}$$

$$S_v(x,y) = \left\| \boldsymbol{\Phi}_v^{1:C}(x,y) \right\|_1 \tag{9}$$

其中,$S_i(x,y)$ 和 $S_v(x,y)$ 分别代表相应像素位置 (x,y) 在 C 维通道上的活动水平。$\|\cdot\|_1$ 代表 l_1 范数。

其次,对红外和可见光的活动水平图进行 Z-score 标准化。由如下式计算得到:

$$Z_i(x,y) = \frac{S_i(x,y) - \mu_i}{\sigma_i} \tag{10}$$

$$Z_v(x,y) = \frac{S_v(x,y) - \mu_v}{\sigma_v} \tag{11}$$

其中,μ 代表活动水平图的均值,σ 代表活动水平图的标准差。

接着,通过 softmax 函数分别计算红外和可见光特征各自的权重图 β_i 和 β_v,其计算方式如下所示:

$$\beta_i(x,y) = \frac{\exp(Z_i(x,y))}{\exp(Z_i(x,y)) + \exp(Z_v(x,y))} \tag{12}$$

$$\beta_v(x,y) = \frac{\exp(Z_v(x,y))}{\exp(Z_i(x,y)) + \exp(Z_v(x,y))} \tag{13}$$

其中,$\beta_i(x,y)$ 和 $\beta_v(x,y)$ 分别代表 C 维向量的权重图。

然后,将权重图与提取到的特征图相乘得到优化的红外特征 $\hat{\boldsymbol{\Phi}}_i$ 和可见光特征 $\hat{\boldsymbol{\Phi}}_v$,由如下式得到:

$$\hat{\boldsymbol{\Phi}}_i(x,y) = \beta_i(x,y) \cdot \boldsymbol{\Phi}_i(x,y) \tag{14}$$

$$\hat{\boldsymbol{\Phi}}_v(x,y) = \beta_v(x,y) \cdot \boldsymbol{\Phi}_v(x,y) \tag{15}$$

最后,融合的特征图 $\boldsymbol{\Phi}_f$ 由如下公式计算得到:

$$\boldsymbol{\Phi}_f(x,y) = \hat{\boldsymbol{\Phi}}_i(x,y) + \hat{\boldsymbol{\Phi}}_v(x,y) \tag{16}$$

4 实验设置

在训练阶段,选用 MS-COCO 数据集进行训练,选用其中 80000 张不同场景的可见光图像,把它们转换为灰度图,并裁剪成 256×256 尺寸。批大小和训练次数分别设置为 4 和 2,学习率设为 1×10^{-4}。超参数 λ 的值设为 1000。本文硬件环境:显卡为 NVIDIA GeForce RTX 3090,CPU 为 Intel(R)Xeon(R)E5-2680 v4,主频为 2.40 GHz。

在测试阶段,从 TNO 数据集中选取了 21 对红外与可见光图像进行测试。同时,为了验证 TAFuse 的泛化能力,还从 RoadScene 数据集中选取了 50 对

红外与可见光图像进行测试。

TAFuse 使用了 Transformer 模块，而输入 Transformer 模块的图像必须可以被均匀分块，否则无法正确处理。Transformer 中小图像块尺寸设置为 4 × 4。由于各个测试图像的尺寸大小不同，不能被均匀分块，因此需要提前对测试图像的边缘填充。对于不同尺寸的测试图像，其宽和高对 128 取余得到的值即为填充区域的大小。TAFuse 采用像素值 128 对边缘进行填充，像素值范围为 0 ~255。测试图像的填充结果如图 5 所示。选取一张尺寸为 632 × 496 的测试图像进行填充，其中 a 和 b 分别表示需要填充的宽度和高度，H 和 W 分别表示填充后的尺寸。

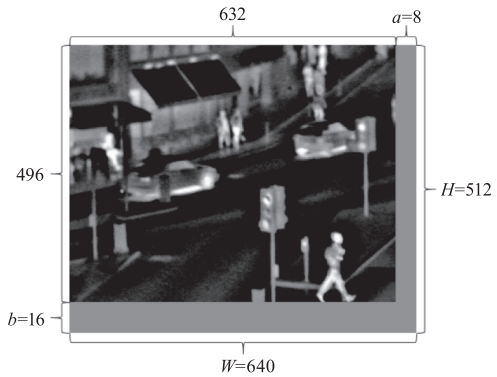


图 5 测试图像填充结果

Fig. 5 Result of filling test image

对于融合效果的主观评价，主要从对比度，亮度，纹理和噪声等方面评价。对于融合效果的客观评价，选择十个重要客观指标比较。分别是边缘强度[7]（EI），空间频率[8]（SF），熵（EN）[9]，边缘信息保持度[7]（$Q^{AB/F}$），小波特征互信息[10]（FMIw），结构相似性[11]（SSIM），互信息[12]（MI），标准差（SD），视觉信息保真度[13]（VIF）和非线性相关信息熵[14]（NCIE）。

5 消融实验

探究损失函数中 λ 取值、Transformer 模块、对活动水平图 Z-score 标准化和不同融合策略对融合效果的影响。从测试集中选取有代表性的红外和可见光图像展示融合效果，并给出相应客观指标。融合图像中使用实线方框标注显著目标，使用虚线方框标注纹理细节，并在左下角放大。客观指标中最优值用实线加粗字体标注，次优值用斜体标注。

1）损失函数中 λ 取值的影响：不同 λ 值对应的损失函数折线图如图 6 所示。其中（a）、（b）和（c）分别代表总损失、Pixel 损失和 SSIM 损失。λ 分别取值 1、10、100、1000，使用不同颜色表示，迭代次数取为 1000。由图可知，不同 λ 的值都快速收敛。

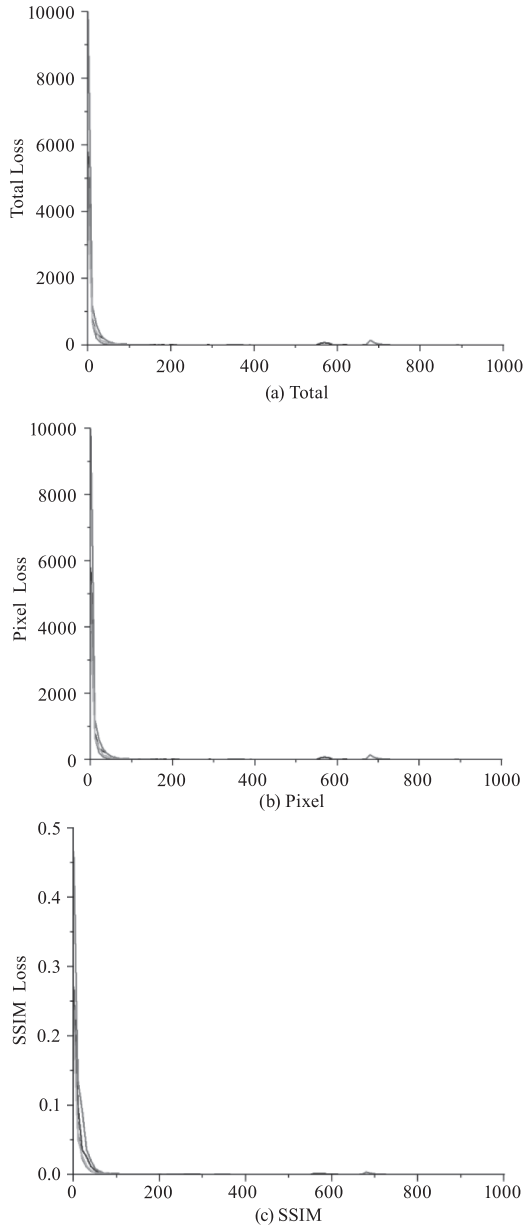


图 6 总损失，pixel 损失和 SSIM 损失的折线图

The Line charts of Total loss，Pixel loss and SSIM loss

表 1 展示了不同 λ 值对应的客观指标。由表可知，λ 值为 1000 时，在大多数客观指标上取得了最优值和次优值。因此本文 λ 取值为 1000。

2）Transformer 模块的影响：融合结果如图 7 所示，用“无 T”代表不使用 Transformer 模块的实验。由图可见，“无 T”的融合图像略微泛白，而使用了 Transformer 模块的融合图像没有出现这种情况，并且对比度和清晰度还得到提升，更利于视觉观察。

表 2 展示了是否使用 Transformer 模块对客观

指标的影响。加入 Transformer 后,大多数指标均得到了不同程度的提升,这表明加入 Transformer 模块后,融合图像的质量优于没有使用 Transformer 模块的融合图像。

综合来看。将 Transformer 和 CNN 结合,可以提取更全面的特征,进而改善融合图像质量。

表 1 不同 λ 取值的客观指标

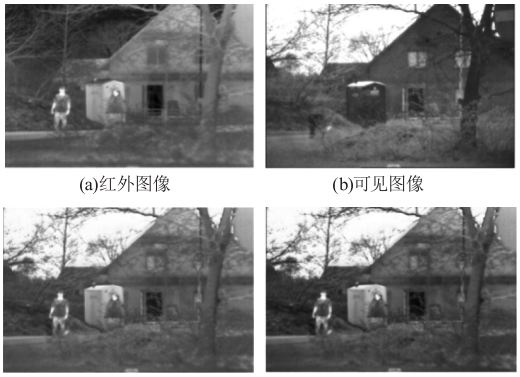
Tab. 1 Objective metrics for different values of λ

	EI	SF	EN	$Q^{AB/F}$	FMIw	SSIM	MI	SD	VIF	NCIE
λ =1	**41.6072**	***11.2217***	7.0595	0.4892	**0.4449**	0.6902	14.1189	46.7246	1.0531	0.8109
λ =10	40.8940	10.9258	***7.0733***	0.4937	0.4430	**0.6936**	***14.1466***	***47.0132***	1.0510	0.8113
λ =100	41.2737	11.1678	**7.0781**	0.4947	0.4431	***0.6926***	**14.1562**	47.0117	***1.0543***	***0.8117***
λ =1000	***41.5596***	**11.2394**	7.0646	**0.4948**	***0.4436***	0.6907	14.1293	**47.0165**	**1.0550**	**0.8119**

表 2 Transformer 模块的客观指标

Tab. 2 Objective metrics of the Transformer module

	EI	SF	EN	$Q^{AB/F}$	FMIw	SSIM	MI	SD	VIF	NCIE
无 T	***37.5832***	***9.8409***	***7.0519***	***0.4801***	***0.4334***	**0.7065**	***14.1004***	***45.6348***	***1.0161***	***0.8096***
TAFuse	**41.5596**	**11.2394**	**7.0646**	**0.4948**	**0.4436**	***0.6907***	**14.1293**	**47.0165**	**1.0550**	**0.8119**



(a)红外图像 (b)可见图像

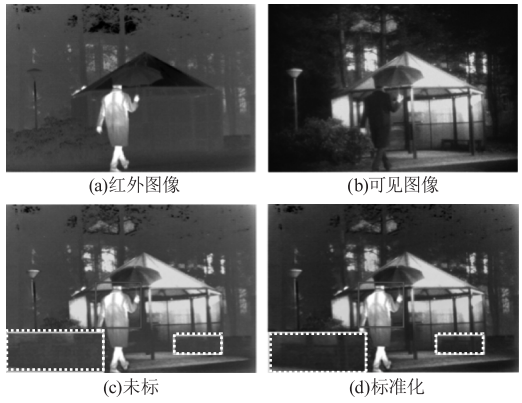
(c)无T (d)TAFuse

图 7 Transformer 模块的可视化结果

Fig. 7 Visualization results of the Transformer module

3)是否对活动水平图进行 Z-score 标准化的影响:为了公平地比较对活动水平图进行标准化的影响,此处没有选择 TAFuse 进行比较,而是选择编码器中去除了 Transformer 模块的方法。这是因为 TAFuse 没有对数据集进行归一化处理,并且得到的特征还会和 Transformer 模块学习到的全局空间关系图相乘,这会让活动水平图数值过大,如果直接使用 softmax 函数计算会导致计算溢出问题。而去掉了 Transformer 模块后,可以直接根据活动水平图计算 softmax 函数,进而得到权重图。

融合结果如图 8 所示,用“未标”和“标准化”分别代表未标准化的融合结果和标准化的融合结果,从图中左下角放大的虚线方框可以看到,未标准化的融合图像中板凳已经失真,只能看到模糊黑色区域,而标准化后的融合图像的板凳相对清晰。值得注意的是,两种方式都很好地保留了显著目标。从图像整体来看,未标准化的融合图像背景亮度较高,但对比度和视觉效果不如标准化后的融合图像。所以标准化后的融合图像可以更充分地保留纹理细节,同时也保留了显著的红外目标,这也更符合红外与可见光图像融合的目的。



(a)红外图像 (b)可见图像

(c)未标 (d)标准化

图 8 标准化操作的可视化结果

Fig. 8 Visualization of the results of standardized operations

表 3 展示了是否使用标准化对客观指标的影响。从客观指标看,未标准化和标准化的融合图像都取得了优秀的效果,但是它们各有侧重,很难一比高下。未标准化的融合图像侧重于图像的亮度,而标准化的融合图像侧重于图像的对比度和视觉效果。

综合来看,标准化后的融合图像更符合红外与可见光图像融合的目的,这也表明在 TAFuse 中对活动水平图使用 Z-score 标准化是合理且有效的。

表 3 标准化操作的客观指标

Tab. 3 Objective indicators for standardized operations

	EI	SF	EN	$Q^{AB/F}$	FMI_w	SSIM	MI	SD	VIF	NCIE
未标	**39.4349**	**10.8357**	***6.7028***	**0.5369**	**0.4802**	***0.7061***	***13.4056***	***43.9682***	***0.9458***	**0.8240**
标准化	***37.5832***	***9.8409***	**7.0519**	***0.4801***	***0.4334***	**0.7065**	**14.1004**	**45.6348**	**1.0161**	***0.8096***

4)采用不同的融合策略的影响:分别采用相加策略和空间注意力策略来探究不同的融合策略的影响。融合结果如图 9 所示。从图中左下角放大的黄色方框可以看到,相加策略的融合图像中,松树周围的树枝不清晰,而且含有很多噪声,图像的对比度也较低,实线方框中的热目标也不显著。而利用空间注意力机制的融合策略得到的融合图像中松树很好地保留了可见光的纹理信息,没有受到噪声干扰,图像对比度也较高,并且保留的热目标较为显著,视觉效果也更好。

表 4 展示了使用不同融合策略对客观指标的影响。从客观指标看,使用空间注意力机制的融合策略在大多数指标上取得了最优值,与相加融合策略相比有很大的提升,尤其是标准差和视觉保真度。

综合来看,采用不同的融合策略对融合图像会产生巨大的影响。因此,精心地设计融合策略是很有必要的,而且也是图像融合中的关键步骤。

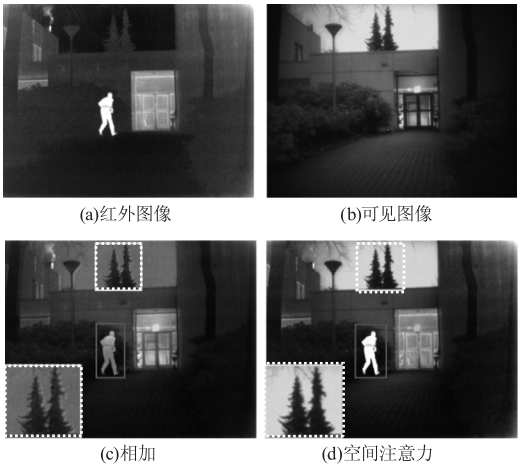


图 9 融合策略的可视化结果

Fig. 9 Visualization results of the fusion strategy

表 4 融合策略的客观指标

Tab. 4 Objective indicators of fusion strategy

	EI	SF	EN	$Q^{AB/F}$	FMI_w	SSIM	MI	SD	VIF	NCIE
相加	***34.3928***	***8.8621***	***6.7491***	***0.4396***	***0.4261***	**0.7422**	***13.4983***	***33.1385***	***0.6440***	***0.8006***
空间注意力	**41.5596**	**11.2394**	**7.0646**	**0.4948**	**0.4436**	***0.6907***	**14.1293**	**47.0165**	**1.0550**	**0.8119**

6 对比实验

将 TAFuse 与九种主流和先进的方法进行对比。分别是双树复小波变换[15](DCHWT),基于梯度传递和总变差最小化[16](GTF),DenseFuse[2],FusionGan[4],IFCNN[17],NestFuse[3],PMGI[18],U2Fusion[19]和 SwinFuse[20]。所有方法都基于公开代码测试,在 Matlab R2020a 上计算各个客观指标值。

在 TNO 数据集和 RoadScene 数据集中分别选取一个例子,分别是“街道”和“行人”图像,并在图 10 和图 11 中进行展示。图中第一行前两张分别是红外图像和可见光图像。为了更清晰地展示融合图像的效果,分别使用实线方框和虚线方框标记显著信息和纹理信息。

从图 10 中可以看到,DCHWT 方法的融合图像细节信息较为模糊,而且还引入了大量的噪声。GTF 和 FusionGan 方法的融合图像中红外目标的边缘锐化,较为显著,但是纹理信息丢失。DenseFuse

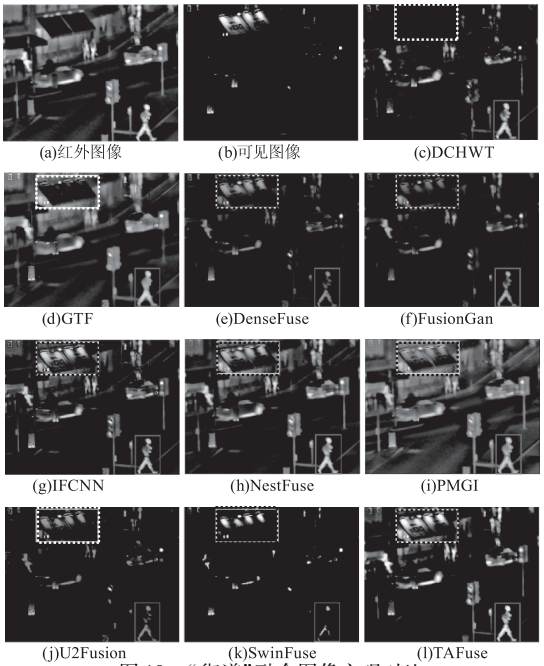


图 10 “街道”融合图像主观对比

Fig. 10 Subjective comparison of "street" fusion images

方法的融合图像没有很好地突出红外目标,并且纹理信息不够清晰,图像整体亮度偏低,对比度也较低。U2Fusion 和 SwinFuse 方法的融合图像红外目标不显著,而且纹理信息丢失严重,整体图像偏暗,视觉效果较差。IFCNN、NestFuse、PMGI 和 TAFuse 方法的融合效果相对较好,可以同时保留显著的目标信息和纹理细节信息,但是 IFCNN 的融合图像噪声较多,不够平滑,NestFuse 和 PMGI 的融合图像中广告牌亮度较低,丢失了部分纹理信息。相比于其他方法,TAFuse 在保留显著红外热辐射信息的同时,可见光图像的纹理细节也得到了很好的保留,并且图像对比度高,具有很好的视觉感知体验。

在图 11 中,TAFuse 的融合图像纹理特征更为清晰,很好地突出了行人的热辐射信息,而且看起来更自然,更符合人类视觉感知。

随后选取十个客观评价指标对主流和先进的方法在 TNO 和 RoadScene 两个数据集上进行比较。TAFuse 在 TNO 数据集上的结果如表 5 所示。其在 6 个指标(EN,FMIw,MI,SD,VIF 和 NCIE)上取得最优值,EN 和 NCIE 表示融合图像包含了较丰富的信息,FMIw 和 MI 表示从源图像中保留了更多特征信息,SD 表示融合图像对比度高,VIF 表示融合图像很符合人类视觉感受。在 $Q^{AB/F}$ 上取得次优值,并在 EI 和 SF 上取得较优值,这意味着 TAFuse 可以保留更多边缘和纹理细节。但 TAFuse 的 SSIM 指标并不理想,分析认为 TAFuse 会对测试图像填充补齐,会引入新的区域从而对整体图像造成影响,而 SSIM 是计算融合图像与两幅源图像之间的亮度、对比度和结构三个方面的相似性,因此其值会偏低。

如表 6 所示,TAFuse 在 RoadScene 数据集上取得了 4 个最优值和 5 个次优质,同样获得了优秀的结果,这也意味着 TAFuse 的泛化能力较强。

图 11 “行人”融合图像主观对比

Fig. 11 Subjective comparison of "pedestrian " fusion images

表 5 TNO 数据集的客观指标对比

Tab. 5 Comparison of objective metrics for the TNO dataset

	EI	SF	EN	$Q^{AB/F}$	FMI_w	SSIM	MI	SD	VIF	NCIE
DCHWT	36.0139	9.7467	6.5678	0.4664	0.4015	**0.7540**	13.1355	30.5206	0.5056	0.8050
GTF	32.5279	9.2044	6.6353	0.4265	0.4362	0.7274	13.2707	31.5792	0.4136	0.8061
DenseFuse	34.5570	8.9140	6.6773	0.4410	0.4276	0.7315	13.3546	33.1777	0.6461	0.8050
FusionGan	22.1482	5.7909	6.3629	0.2189	0.3708	0.6718	12.7257	26.0676	0.4536	0.8052
IFCNN	42.1677	***11.4913***	6.5954	**0.5041**	0.4017	***0.7462***	13.1909	31.4000	0.5902	0.8054
NestFuse	36.3888	9.7450	6.9203	0.4892	***0.4373***	0.7308	13.8406	40.1708	0.7957	***0.8087***
PMGI	36.5525	8.7099	***6.9339***	0.4104	0.3984	0.7311	***13.8679***	34.8728	0.7928	0.8051
U2Fusion	**48.3617**	11.3013	6.7571	0.4249	0.3620	0.7053	13.5142	31.7084	0.8543	0.8040
SwinFuse	***44.3590***	**12.6399**	6.8820	0.4452	0.4273	0.6811	13.7640	***46.9457***	***1.0044***	0.8056
TAFuse	41.5596	11.2394	**7.0646**	***0.4948***	**0.4436**	0.6907	**14.1293**	**47.0165**	**1.0550**	**0.8119**

表 6 RoadScene 数据集的客观指标对比

Tab. 6 Comparison of objective metrics for the RoadScene dataset

	EI	SF	EN	$Q^{AB/F}$	FMI_w	SSIM	MI	SD	VIF	NCIE
DCHWT	49.4608	11.8975	7.1710	0.4588	0.3671	***0.7265***	14.3420	39.7779	0.5241	0.8067
GTF	37.2992	10.1343	**7.6346**	0.3816	0.3923	0.7198	**15.2693**	***59.7582***	0.4247	***0.8112***
DenseFuse	34.0233	8.5547	6.6757	0.3817	0.4192	0.7152	13.3514	30.7038	0.6694	0.8079
FusionGan	35.4048	8.6400	7.1753	0.2737	0.3410	0.6515	14.3507	42.3040	0.4256	0.8077
IFCNN	57.6653	15.0677	6.9730	**0.5150**	0.4032	**0.7301**	13.9460	35.8183	0.6249	0.8076
NestFuse	54.7351	14.6151	7.3848	0.4911	**0.4344**	0.7031	14.7695	51.7192	0.9625	0.8105
PMGI	47.2067	10.9368	7.3493	0.4248	0.3774	0.6899	14.6986	49.3262	0.6461	0.8100
U2Fusion	**66.2529**	15.8242	7.1969	0.4805	0.3717	0.7038	14.3938	42.9368	0.8317	0.8075
SwinFuse	61.0275	***16.4591***	7.3113	0.4493	0.4201	0.6993	14.6226	53.7563	***0.9928***	0.8085
Ours	***61.2184***	**17.0160**	***7.4541***	***0.5050***	***0.4333***	0.6911	***14.9081***	**60.6918**	**1.1685**	**0.8137**

7 结 论

本文提出了一种基于 Transformer 和空间注意力的红外与可见光图像融合框架。可以在大规模的自然图像数据集上进行训练。编码器网络结合了 CNN 和 Transformer 各自的优点,CNN 模块用于提取多尺度局部特征,Transformer 模块用于学习图像的空间关系,获取全局特征。进而使编码器能够更全面地提取特征。此外,引入了多尺度特征和跳跃连接,使解码器能够充分地利用浅层特征和深层特征,进而提升了图像重建能力。大量的对比实验表明,TAFuse 的融合结果可以同时保留源图像中显著目标区域和丰富的纹理细节信息,并且在与其他方法的定性比较中,TAFuse 的结果对比度高,也更符合人类视觉感知。同时在多个数据集上都取得了较好的效果,验证了 TAFuse 具有较好的泛化能力。

参考文献:

[1] Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: A survey[J]. Information Fusion, 2019, 45: 153 – 178.

[2] Li H, Wu X J. DenseFuse: a fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614 – 2623.

[3] Li H, Wu X J, Durrani T. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645 – 9656.

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

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[J]. arXiv Preprint arXiv: 2010. 11929, 2020.

[7] Xydeas C S, Petrovic V. Objective image fusion performance measure[J]. Electronics Letters, 2000, 36(4): 308 – 309.

[8] Eskicioglu A M, Fisher P S. Image quality measures and their performance[J]. IEEE Transactions on Communications, 1995, 43(12): 2959 – 2965.

[9] Roberts J W, Van Aardt J A, 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.

[10] Haghighat M, Razian M A. Fast-FMI: non-reference image fusion metric[C]//2014 IEEE 8th International Conference on Application of Information and Communication Technologies(AICT). IEEE, 2014: 1 – 3.

[11] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600 – 612.

[12] Qu G, Zhang D, Yan P. Information measure for performance of image fusion[J]. Electronics Letters, 2002, 38(7): 1.

[13] Sheikh H R, Bovik A C. Image information and visual quality[J]. IEEE Transactions on Image Processing, 2006, 15(2): 430 – 444.

[14] Wang Q, Shen Y, Jin J. Performance evaluation of image fusion techniques[J]. Image Fusion: Algorithms and Applications, 2008, 19: 469 – 492.

[15] Shreyamsha Kumar B K. Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform[J]. Signal, Image and Video Processing, 2013, 7: 1125 – 1143.

[16] Ma J, Chen C, Li C, et al. Infrared and visible image fusion via gradient transfer and total variation minimization[J]. Information Fusion, 2016, 31: 100 – 109.

[17] Zhang Y, Liu Y, Sun P, et al. IFCNN: a general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99 – 118.

[18] Zhang H, Xu H, Xiao Y, et al. Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 12797 – 12804.

[19] Xu H, Ma J, Jiang J, et al. U2Fusion: a unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 502 – 518.

[20] Wang Z, Chen Y, Shao W, 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.