基于 Swin Transformer 和混合特征聚合的红外与可见光图像融合方法

李碧草[1,2]，卢佳熙[1]，刘洲峰[1]，李春雷[1]，张 洁[1]
（1. 中原工学院 电子信息学院，河南 郑州 450007；2. 郑州大学 计算机与人工智能学院，河南 郑州 450001）

摘要：红外与可见光图像融合可以生成包含更多信息的图像，比原始图像更符合人类视觉感知也有利于下游任务的进行。传统的基于信号处理的图像融合方法存在泛化能力不强、处理复杂图片融合性能下降等问题。深度学习有很强的特征提取能力，其生成的结果较好，但结果中存在纹理细节信息保存少、图像模糊的问题。针对这一问题，文中提出一种基于多尺度 Swin-transformer 和注意力机制的红外与可见光图像融合网络模型。Swin-transformer 可以在多尺度视角下提取长距离语义信息，注意力机制可以将所提特征中的不重要特征弱化，保留主要信息。此外本文提出了一种新的混合特征聚合模块，针对红外和可见光图像各自的特点分别设计了亮度增强模块和细节保留模块，有效保留更多的纹理细节和红外目标信息。该融合方法包括编码器、特征聚合和解码器三部分。首先，将源图像输入编码器，提取多尺度深度特征；然后，设计特征聚合融合每个尺度的深度特征；最后，采用基于嵌套连接的解码器重构融合后的图像。在公开数据集上的实验结果表明本文提出的方法对比其他先进的方法具有更好的融合性能。其中在客观评价指标中 EI、AG、QP、EN、SD 指标达到最优。从主观感受上，所提红外和可见光图像融合方法能够使结果中保留更多的边缘细节。

关键词：图像融合；红外和可见光图像；Swin-transformer；特征聚合；注意力机制
中图分类号：TP391.41 **文献标志码**：A **文章编号**：1001-8891(2023)07-0721-11

Infrared and Visible Light Image Fusion Method Based on Swin Transformer and Hybrid Feature Aggregation

LI Bicao[1,2]，LU Jiaxi[1]，LIU Zhoufeng[1]，LI Chunlei[1]，ZHANG Jie[1]
(1. *School of Electronic Information*, *Zhongyuan University of Technology*, *Zhengzhou* 450007, *China*;
2. *School of Computer and Artificial Intelligence*, *Zhengzhou University*, *Zhengzhou* 450001, *China*)

Abstract: The fusion of infrared and visible light images can generate images containing more information in line with human visual perception compared with the original images, and is also beneficial for downstream tasks. Traditional image fusion methods based on signal processing have problems such as poor generalization ability and reduced performance of complex image fusion. Deep learning is capable of features extraction and provides good results. However, its results have problems such as reduced preservation of textural details and blurred images. To address these problems, this study proposes a fusion network model of infrared and visible light images based on the multiscale Swin Transformer and an attention mechanism. Swin Transformers can extract long-distance semantic information from a multiscale perspective, and the attention mechanism can weaken the insignificant features in the proposed features to retain the main information. In addition, this study proposes a new hybrid fusion strategy and designs brightness enhancement and detail retention modules according to the respective characteristics of the infrared and visible images to retain more textural details and infrared target information. The fusion method has three parts: the encoder, fusion strategy, and decoder. First, the source image was input into the encoder to extract multiscale depth features. Then, a fusion strategy was

收稿日期：2022-07-30；**修订日期**：2022-09-13.
作者简介：李碧草（1985-），男，博士，副教授，硕士生导师，主要研究方向为医学图像处理、模式识别。E-mail: lbc@zut.edu.cn。
基金项目：国家自然科学基金资助项目（61901537, 62072489）；河南省留学人员科研择优项目资助经费；中国博士后科学基金面上资助（2020M672274）；中国纺织工业联合会科技指导性计划项目（2019059）；中原工学院青年骨干教师培养计划（2019XQG04）；中原工学院学科青年硕导培育计划（SD202207）。

designed to fuse the depth features of each scale. Finally, the fused image was reconstructed using a decoder based on nested connections. The experimental results on public datasets show that the proposed method has a better fusion performance compared with other state-of-the-art methods. Among the objective evaluation indicators, EI, AG, QP, EN, and SD were optimal. From a subjective perspective, the proposed infrared and visible light image fusion method can preserve additional edge details in the results.

Key words: image fusion, infrared and visible light images, Swin-transformer, feature aggregation, attention mechanism.

0 引言

图像融合是一种重要的图像处理技术。旨在通过特定的特征提取和特征融合生成一幅包含源图像互补信息的图像。目前融合算法被广泛应用于自动驾驶、视觉跟踪和医学图像增强等领域。在图像处理领域，红外和可见光图像的融合也是图像融合的研究热点，红外图像中包含热辐射信息，但由于红外成像传感器的特性，采集的红外图像中纹理细节信息不明显。而可见光图像中包含大量细节纹理信息，但是没有热辐射信息，融合后的图像包含二者的互补信息，有利于人类的视觉感知。

现有的融合方法大致可分为两类，传统方法和基于深度学习的方法。常用的传统融合方法包括：基于梯度转移的图像融合[1]（gradient transfer fusion，GTF）；基于显著性检测的图像融合方法[2]（Two-scale Image Fusion，TIF）；基于各向异性扩散和 Karhunen-Loeve 变换[3]的融合方法（Anisotropic Diffusion Fusion，ADF）；基于卷积稀疏表示[4]（Convolutional Sparse Representation，CSR）的图像融合方法；基于高斯滤波和双边滤波混合多尺度分解[5]的图像融合方法等。这些方法虽然都取得了较好的结果，但都需要手工设计繁琐的特征提取和融合规则，且泛化能力不强，当融合图像复杂时融合性能下降。

近年深度学习在图像融合任务中有不错的表现。研究学者们提出了很多相关模型。按网络结构来区分可以分为自编码器和端到端两种。Prabhakar 等提出 DeepFuse[6] 融合方法，采用卷积神经网络来提取两幅 YCbCr 图像中 Y 通道的特征，然后将所提取的特征相加再经过卷积神经网络得到融合后的 Y 通道，Cb、Cr 通道通过加权融合得到，最后将 YCbCr 图像转换成 RGB 图像得到融合结果。Zhang 等提出 IFCNN[7]（Image Fusion based on Convolutional Neural Network）是一种自编码器结构的网络。该方法采用卷积神经网络分别提取两幅源图像的特征，之后通过一定的融合规则将所得到的特征融合，融合后的特征经过卷积神经网络重建出融合图像。

此外，研究者还提出端到端的深度学习融合框架，并取得不错的效果。Xu 等提出 U2Fusion[8]（Unified Unsupervised image Fusion Network）融合算法，通过特征提取和信息测量，自动估计特征对应源图像的重要性，得到了较好的融合效果。Li 等提出 RFN-Fuse[9]（Residual Fusion Network）同样是一种端到端的图像融合方法，先用训练好的编码器提取图像特征，然后输入进融合网络融合特征，再由解码器重建图像。Ma 等提出 FusionGAN[10]（Generative Adversarial Network），一种端到端的方法，将生成对抗网络应用于图像融合，通过构建一个生成器和一个鉴别器使二者相互博弈，迫使生成器生成包含两幅源图像信息的融合图像。Fu 等提出 PerceptionGAN[11]（GAN consistent with perception）通过将可见光图像连接到网络中的不同深度，使融合结果更接近人类的视觉感知，但其结果中红外图像信息较少。此外，基于 GAN 的方法也有其他研究学者提出[12-14]。由于端到端方法存在生成结果模糊、细节保存较少、如果没有很好的约束和大量的训练数据，融合性能并不佳等问题，本文采用自编码器策略。

以上方法忽略了编解码过程中的特征通道注意力信息，并且长距离语义信息没有被充分利用。因此本研究在网络中应用注意力机制和 Swin-Transformer 来缓解这一问题。此外，现有的方法通常只考虑可见光图像的背景信息和红外图像的目标亮度信息，而红外图像的背景亮度信息通常被忽略，导致红外图像中的部分背景信息细节丢失。充分利用红外亮度信息会使背景更加清晰。红外图像的梯度信息也有助于生成更加清晰的图像。因此，一个新的混合特征聚合被提出来融合特征，其中包含红外亮度增强模块和纹理细节增强模块。红外亮度增强模块不仅可以增强红外目标信息，还保留了红外图像中部分背景的亮度。细节保留模块通过梯度算子提取特征图的梯度边缘信息。特征聚合中还加入了注意力机制来融合特征，能够保留更多细节。本文提出一种新的融合方法，主要贡献如下：

1）提出一种注意力巢连接网络，充分利用多尺度

分解和图像重建过程中的注意力信息。

2）在解码器中采用 Swin-transformer 提取图像特征的长距离依赖。增强模型特征提取能力。

3）提出了一种新的混合红外特征增强、纹理细节增强和注意力的特征聚合模块。可以充分保留来自源图像的亮度与细节信息。

4）实验结果表明，所提方法能够更清晰地融合红外和可见光图像，融合结果中的纹理和细节信息更多。与现有的融合方法相比，本文提出的融合框架在公开数据集上的主观视觉评价和客观评价均表现出更好的融合性能。

1 相关工作

随着深度学习被广泛应用于图像融合领域，很多基于深度学习的方法被提出。这些方法大致分为两类，一是端到端的全神经网络，二是深度学习与手工设计融合规则相结合的方法。本章首先介绍几种经典的深度学习图像融合方法。

注意力机制被广泛应用于神经网络中。Hu 等人从通道维度入手提出一种通道注意力机制[15]，该机制可以对特征进行校正，校正后的特征可以保留有价值的特征，剔除没价值的特征。Li 等人提出 CSpA-DN[16]网络将自注意力机制与 DenseNet[17]结合，该方法为端到端的融合方法，大致分为 3 个部分：编码网络、注意力网络和解码网络，编码网络的目的是提取两幅源图像的特征，注意力网络对特征进行校正，解码网络重建图像。该网络采用类似 DenseNet 设计具有密集短连接结构，可以很好地传递特征图，减轻梯度消失，在一定程度上减少了参数量，并且在 PET 和 MRI 融合任务中取得了不错的效果。Li 等提出了一种结合深度学习和手工设计融合规则的方法 DenseFuse[18]。该方法采用两阶段的融合方法，首先训练一个编码和解码网络，源图像经过编码器提取特征，之后将所得特征相加，最后融合后的特征图经过解码网络重建得到融合图像。这些方法都没有充分利用特征图的多尺度信息，并且融合策略相对简单。

其中具有多尺度结构的模型在处理图像任务时有不错的表现。Zhou 等人提出了 Unet＋＋[19]，用于图像分割。Unet＋＋在不同尺度的 Unet 网络上探索并且把这些不同尺度的 Unet 嵌套在一起并使用跳跃连接组合成一个新的巢连接网络。Li 等设计了 NestFuse[20]网络采用巢连接结构，包含一个下采样和上采样过程，能够提取图像的深度特征信息。首先，训练一个提取多尺度信息的编码网络和一个对应的解码网络，在训练过程中没有融合阶段，只有编码解码过程。然后，使用设计的融合策略将编码器提取的每个尺度的特征进行融合。最后，由解码器重建图像并取得了较好的效果。然而，在编解码过程中，该方法并未考虑每个特征图的重要程度。因此，本文提出一种基于注意力的巢连接网络。由于注意力机制能够对特征图进行筛选，将其引入融合模型，充分利用各尺度的通道注意力信息，增强融合性能。

2 融合方法

本章将详细介绍基于注意力机制和巢网络的融合模型，并介绍模型的细节以及特征聚合模块。融合方法的总体框图如图 1。

2.1 网络结构

本文提出的融合方法主融合框架如图 1 所示。其中 EB 为编码器、FA 为特征聚合、DB 为解码器。本节主要介绍编码器、解码器，特征聚合在 2.2 节中详细介绍。

现有的 U 型网络存在相同尺度上卷积层不深导致特征未充分利用的问题，为了缓解这个问题，本文采用巢连接策略，在同一尺度之间增加卷积层，并使用跳跃连接，在不同尺度之间采用上采样连接，来充分利用特征。由于卷积只关注局部的纹理特征没有充分利用长程语义依赖，因此本研究在网络中使用 Swin-transformer 来提取长距离依赖如图 1 所示。Swin-transformer 相比于传统的 Transformer 有更低的计算量和更强的特征提取能力，其结构如图 2 所示。

编码器由 4 个卷积块组成，如图 3(a)所示，其中 Conv 表示卷积层，用来提取图像的浅层特征信息。在编码器中，每个卷积块都包括一个 2×2 的池化层，对特征图进行下采样。图 3(a)中 EB 代表一个卷积块，其结构如图 3(b)。

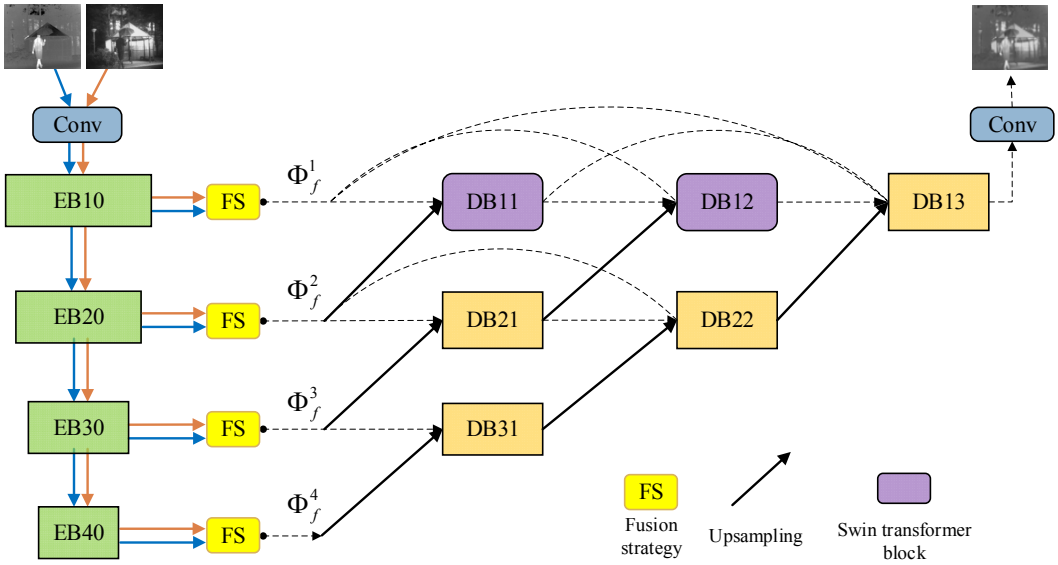
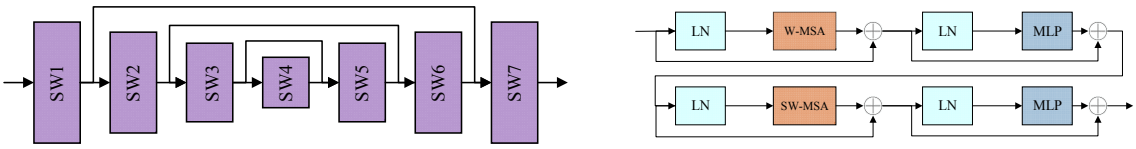


图 1 本文融合方法的网络结构

Fig.1 Network architecture of the fusion method in this paper

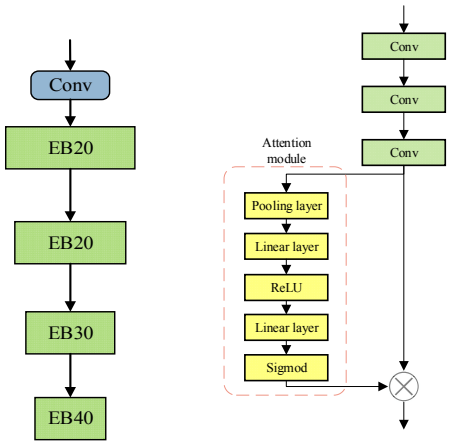


(a) Swin transformer 的结构 (b) Swin transformer 层

(a)The structure of Swin transformer (b) Swin transformer layer

图 2 基于 Swin transformer 的解码块

Fig.2 The decoding block based on Swin transformer



(a)编码器 (b)编码器卷积块结构

(a) Encoder (b) Convolutional block of the encoder

图 3 编码器及编码器中的卷积块结构

Fig.3 Encoder and the structure of convolutional block in encoder

在编码阶段，图像先经过一个输出通道数为 16 的卷积层，再依次经过 EB10，输出通道数为 64，分辨率大小为 224×224。EB20 输出通道数为 112，分辨率为 112×112。EB30 输出通道数为 160，分辨率大小为 56×56，EB40 输出通道数为 208，分辨率大小为 28×28。参数如表 1 所示。

表 1 编码器和解码器网络参数

Table 1 Encoder and decoder network parameters

	Layer	Input channel	Output channel	Resolution
Encoder	Conv	1	16	224×224
	EB10	16	64	224×224
	EB20	64	112	112×112
	EB30	112	160	56×56
	EB40	160	208	28×28
Decoder	DB31	368	160	56×56
	DB21	272	112	112×112
	DB22	384	112	112×112
	DB11	176	64	224×224
	DB12	240	64	224×224
	DB13	304	64	224×224
	Conv	64	1	224×224

编码过程表达式如(1)～(4)所示：

$$\Phi_1 = \mathrm{EB}_1(F_{\mathrm{ATT}}(\mathrm{Conv}(I))) \tag{1}$$

$$\Phi_2 = \mathrm{EB}_2(F_{\mathrm{ATT}}(\Phi_1)) \tag{2}$$

$$\Phi_3 = \mathrm{EB}_3(F_{\mathrm{ATT}}(\Phi_2)) \tag{3}$$

$$\Phi_4 = \mathrm{EB}_4(F_{\mathrm{ATT}}(\Phi_3)) \tag{4}$$

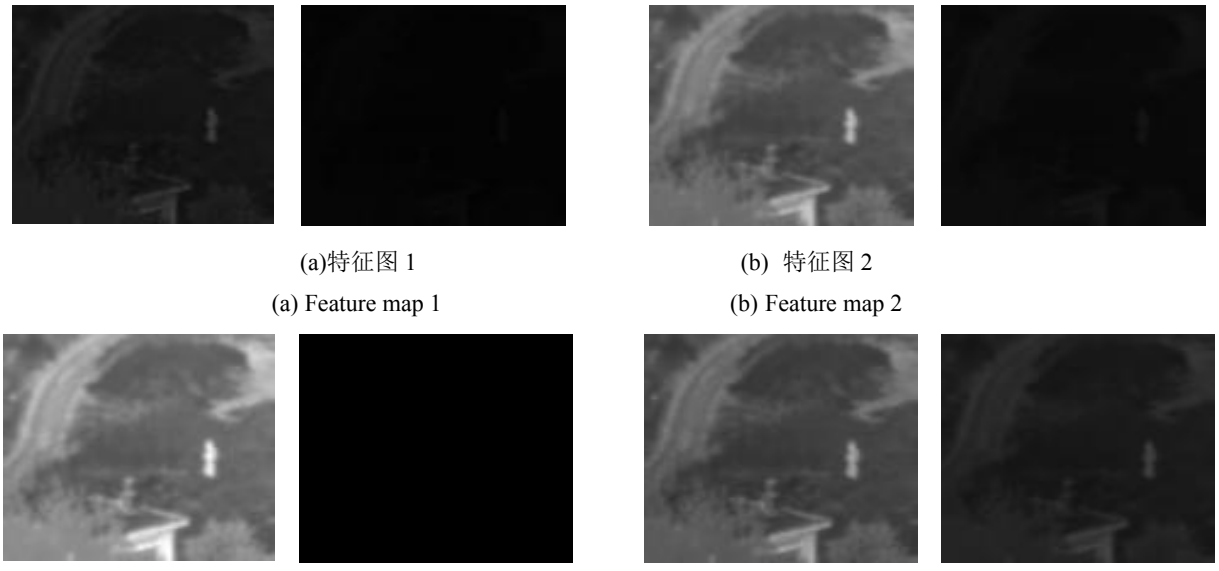
式中：I，Φ分别表示输入图像和多尺度特征；$EB_m(\cdot)$表示多尺度特征提取函数；m 表示多尺度层数 $m\in 1,2,3,4$。Φ_m表示各尺度所得特征图。Conv(·)表示卷积层。

巢连接网络没有筛选特征能力不能突出重要特征，为了提升网络提取特征能力，本文在多尺度网络结构中加入注意力机制，为每个尺度的特征图增加一个权重。本文采用的注意力计算方法如下。对每个特征图取平均池化操作，将得到的结果组成一个特征向量。计算单个 $H\times W$ 特征图对应的公式如(5)所示：

$$F_{\mathrm{ATT}}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u(i,j) \tag{5}$$

式中：i、j 为像素坐标；$u(\cdot,\cdot)$为平均池化操作。对通道数为 C 的特征图按通道进行 $F_{\mathrm{ATT}}(\cdot)$操作，得到 $1\times C$ 维的特征向量。如图 2(b)中所示，使用线性层将所得特征向量的维度压缩，经过激活函数，其目的是增加网络的非线性，拟合通道之间的相关性。经过第一个线形层后维度变为原来的 $1/N$，本文中 $N=16$。之后，再用线性层将特征向量扩展到与原特征图的通道数相同的维度。所得特征向量经过 Sigmoid 函数之后得到与特征图通道数维数一致的权重向量，最后与原特征图相乘。

将图像融合过程中部分特征图可视化，如图 3 所示，输入为 TNO 数据集[21]中的可见光图像。每对图像的左右两幅图片分别为经过注意力机制前后的特征图。可以看出注意力机制能够将模糊的特征弱化，这些特征对重建图像纹理和细节的保留的重要性相对较小。图 4 为解码器中 DB21 卷积块中特征图可视化结果，可视化结果表明注意力机制能够为各通道分配权重，突出重要信息。



(a)特征图 1 (b) 特征图 2

(a) Feature map 1 (b) Feature map 2

(c) 特征图 3 (d) 特征图 4

(c) Feature map 3 (d) Feature map 4

图 4 4 对经过注意力模块前后的特征图

Fig.4 Four pairs of feature maps before and after the attention module

红外和可见光图像分别经过编码器后使用特征聚合 FA 得到融合特征：

$$\Phi_f^m=\mathrm{FA}(\Phi_1^m,\Phi_2^m) \tag{6}$$

式中：FA(·)为特征聚合模块，具体如 2.2 节所示。Φ_1^m 和 Φ_2^m 分别为输入源图像的多尺度特征，m 表示多尺度层数。将 Φ_f^m 输入到解码器中得到最终的融合图像。

解码阶段网络参数与编码阶段相对应。具体参数设置如表 1 所示。解码器由 6 个 DB 卷积块组成，如图 5 所示，用于重建融合图像，解码器的 4 个输入与编码器 4 个卷积块相对应。其中 DB11 和 DB12 由 Swin-transformer 块组成如图 2(a)所示，每个 Swin-transformer 块由 7 层不同尺度的 Swin-transformer 层组成，每个 Swin-transformer 层如图 2(b)所示。

编码阶段和解码阶段的卷积块不完全相同。解码阶段的卷积块由两个卷积层、一个池化层和一个注意力模块组成，注意力模块与图 2(b)中所示的结构相同。如图 5 所示。其中第二个卷积层的核大小为 1×1，用来匹配维度。解码阶段没有用于下采样的池化层，其余卷积层保持不变。特征图上采样后拼接到同尺度特征中。

2.2 特征聚合

大多数特征融合都是基于加权平均算子生成一

个加权图来融合源图像。基于这一理论，权重图的选择成为一个关键问题。而现有的方法忽略了红外图像中的背景亮度信息及红外图像的梯度信息，为此在本研究中设计了红外特征增强模块保留更多红外亮度信息，并且从两幅源图像中分别提取梯度信息，同时混合基于注意力机制[20]的特征聚合，达到保留更多细节的目的。如图 6 所示。在网络训练完成后，测试时将特征聚合加入到网络中，两副原图像经过编码器后得到多尺度特征Φ_1^m和Φ_2^m，通过 l_1-norm 和 Soft-max 算子计算得到的权重映射β_1^m 和β_2^m 权重图由公式(7)表示：

$$\beta_k^m(x,y)=\frac{\left\|\Phi_k^m(x,y)\right\|_1}{\sum_{i=1}^{k}\left\|\Phi_i^m(x,y)\right\|_1} \tag{7}$$

式中：$\|\cdot\|_1$ 表示 L_1 范数；$k\in1,2$。(x,y)表示多尺度深度特征(Φ_1^m 和Φ_2^m)和权重图(β_1^m 和β_2^m)中对应的位置，每个位置表示深度特征中的一个 C 维向量。$\Phi_k^m(x,y)$表示一个 C 维的向量。

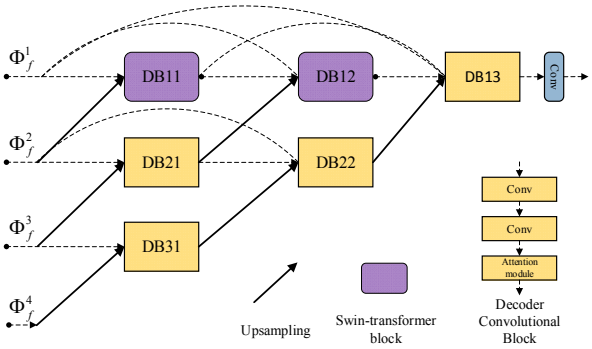


图 5 解码器网络结构

Fig.5 Network structure of decoder

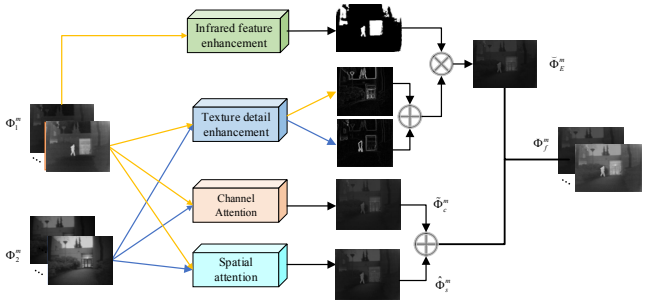


图 6 特征聚合框架

Fig.6 The framework of feature aggregation

$\hat{\Phi}_1^m$ 和 $\hat{\Phi}_2^m$ 表示用β_1^m 和β_2^m 加权的增强深度特征。增强特征 $\hat{\Phi}_k^m$ 通过公式(8)计算：

$$\hat{\Phi}_k^m(x,y)=\beta_k^m(x,y)\times\Phi_k^m(x,y) \tag{8}$$

然后通过这些增强的深度特征计算出融合特征 $\hat{\Phi}_f^m$，公式如公式(9)所示：

$$\hat{\Phi}_f^m(x,y)=\sum_{i=1}^{2}\hat{\Phi}_i^m(x,y) \tag{9}$$

现有方法中特征聚合大都只考虑空间信息。然而，深度特征是三维张量。因此，特征聚合中不仅要考虑空间维度信息，还要考虑通道信息。通道注意力特征计算过程与空间注意力特征计算过程大致相同，如图 6。利用通道注意力模块计算后的结果是一个一维向量，各个值为对应通道的权重。特征聚合输入特征图的权重向量α_1^m 和α_2^m 由公式(10)计算得出。

$$\bar{\alpha}_k^m(n)=P(\Phi_k^m(n)) \tag{10}$$

式中：n 为输入特征中的通道数；$P(\cdot)$为全局池化。全局池化方法是通过每个通道的奇异值求和得到。奇异值往往对应着矩阵中隐含的重要信息，且重要性和奇异值大小正相关。

然后，使用 Soft-max 函数计算得到最终的加权向量α_1^m 和α_2^m 如公式(11)：

$$\alpha_k^m(n)=\frac{\bar{\alpha}_k^m(n)}{\sum_{i=1}^{2}\bar{\alpha}_i^m(n)} \tag{11}$$

最后通道注意力模块的融合特征 $\tilde{\Phi}_f^m$ 由式(12)计算得到：

$$\tilde{\Phi}_f^m(n)=\sum_{i=1}^{2}\alpha_i^m(n)\times\Phi_i^m(n) \tag{12}$$

两幅源图像分别计算空间注意力和通道注意力得到结果和 $\hat{\Phi}_s^m$、$\tilde{\Phi}_c^m$。m 表示多尺度深度特征的层次。

在所提特征聚合中对两幅图像分别进行梯度特征提取得到梯度权重图，如公式所示：

$$\varepsilon_k^m(x,y)=\frac{S(\left\|\Phi_k^m(x,y)\right\|_1)}{\sum_{i=1}^{2}S(\left\|\Phi_i^m(x,y)\right\|_1)} \tag{13}$$

式中：$S(\cdot)$代表 Sobel 函数用于提取特征图的梯度特征。

红外特征增强模块首先将红外特征通过分割的方法分离出来，如公式：

$$\eta_k^m(x,y)=\gamma\times\mathrm{seg}(\Phi_{\mathrm{ir}}^m(x,y)) \tag{14}$$

式中：seg(·)为阈值分割函数，其阈值根据背景和红外目标像素值的最大类间方差获得。γ为平衡权重，在本文中设置为 0.3。

$$\breve{\Phi}_{\mathrm{E}}^{m}(x,y)=\varepsilon_{k}^{m}(x,y)\times\eta_{k}^{m}(x,y) \tag{15}$$

最终的注意力融合特征Φ_{f}^{m}由公式(16)计算得到。

$$\Phi_{\mathrm{f}}^{m}=\frac{1}{3}(\hat{\Phi}_{\mathrm{s}}^{m}+\tilde{\Phi}_{\mathrm{c}}^{m}+\breve{\Phi}_{\mathrm{E}}^{m}) \tag{16}$$

2.3 训练阶段

所提方法采用了两阶段训练策略。首先，训练一个可以提取图片深层特征的自动编码器，和一个可以处理这些特征重建图像的解码器。训练框架如图 7 所示，其中 I 和 O 分别为输入图像和重建图像。训练数据集采用 MS-COCO[22]数据集。

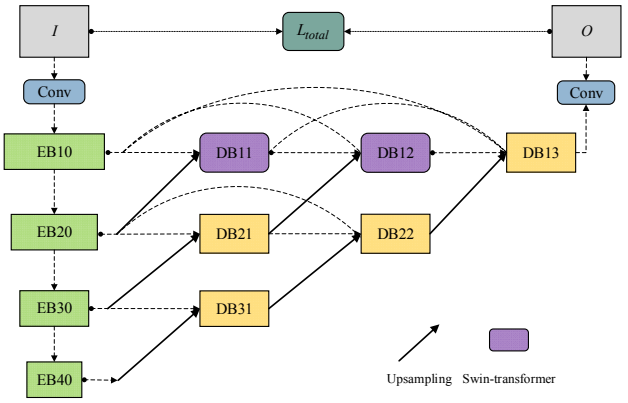


图 7 训练阶段的网络结构

Fig.7 Network structure in the training phase

训练过程没有融合阶段，特征聚合不参与训练。只需训练解码器和编码器。在损失函数的约束下迫使网络能够重建出输入图像。在测试时编码器要分别对两幅源图像进行编码，再经特征聚合后输入到解码器。

在训练阶段，损失函数 L_{total} 定义如下：

$$L_{\mathrm{total}}=L_{\mathrm{pixel}}+\lambda L_{\mathrm{ssim}} \tag{17}$$

式中：L_{pixel} 和 L_{ssim} 分别表示源图像和融合后图像之间的像素损失和结构相似度损失。λ是平衡两个损失的加权因子。在本文中λ取值为 100。

L_{pixel} 由公式(18)得到：

$$L_{\mathrm{pixel}}=\left\|O-I\right\|_{\mathrm{F}}^{2} \tag{18}$$

式中：O 和 I 分别表示输出图像和输入图像。其中$\|\cdot\|_{\mathrm{F}}$为 F 范数。损失函数可以最大程度地使输出图像像素更接近于输入图像。

SSIM 结构相似度损失函数 L_{ssim} 由公式(19)得到。

$$L_{\mathrm{ssim}}=1-\left(\frac{2\mu_{\mathrm{I}}\mu_{\mathrm{O}}+c_{1}}{\mu_{\mathrm{I}}^{2}+\mu_{\mathrm{O}}^{2}+c_{1}}\right)\left(\frac{2\sigma_{\mathrm{IO}}+c_{2}}{\sigma_{\mathrm{I}}^{2}+\sigma_{\mathrm{O}}^{2}+c_{2}}\right) \tag{19}$$

式中：μ_{I}, μ_{O} 和σ_{I}, σ_{O} 分别为输入输出图像的均值和标准差。σ_{IO} 为协方差，c_1，c_2 为常数。L_{ssim} 越小两幅图像的结构越相似。

3 实验结果

本章中，首先介绍本文的实验设置。然后介绍消融研究。在主观评价方面与现有方法进行了比较，并利用多个质量评价指标对融合性能进行了客观评价。图 8 展示了采用的 21 对红外和可见光测试图像的一部分。

(a) 红外图像

(a) Infrared Images

(b) 可见光图像

(b) Visible Images

图 8 TNO 数据集中的 3 对红外和可见光图像

Fig.8 Three pairs of infrared and visible images in TNO dataset

3.1 实验设置

选择 10 种比较典型和先进的融合方法来评价融合性能，包括：GTF[1]，TIF[2]，ADF[3]，FusionGAN[10]，DenseFuse[18]，vggML[23]，RFN-Fuse[9]，DeepFuse[6]，CSF[24]（Classification Saliency-Based Fusion），Dual-branch[25]，这些方法实验结果都由其公开代码得到，其中参数设置与其论文所述相同。网络训练时 epoch 和 batch 大小分别为 2 和 2。实验平台为：E5 2680 v4 CPU, NVIDIA GTX 1080Ti GPU，代码实现使用 PyTorch 框架。

利用以下几个质量指标对本文的融合方法和其他融合方法进行了定量比较。其中包括：边缘强度（Edge Intensity，EI）[26]，视觉保真度（Visual Fidelity，VIF）[27]，平均梯度（Average Gradient，AG）[28]，信息熵（Entropy，EN）[29]，标准差（Standard Deviation，SD），离散余弦特征互信息（Discrete Cosine Feature Mutual Information，FMI_dct）[30]，相位一致（Phase Consistent，QP）[31]。测试采用的是 TNO[21]和 MSRS 数据集[32]，分别取 21 对图像。客观评价结果从其中选取 21 对图像进行测试，取 21 对图像客观结果的平均值进行对比。

3.2 消融研究

如 2.1 节所述，本研究在编解码网络中加入了注意力机制。分别对有注意力机制（Att）和没有注意力机制以及 Swin-transformer（Att＋ST）进行了实验，实验结果如图 9，其中测试图像是从 TNO 数据集中选取的部分图像。左边一列(a)是加上注意力之后的结果，中间一列(b)是加入 Swin-transformer 后的结果，右边一列(c)是所提融合方法的结果。可以看到加上注意力机制之后图像包含更多的纹理信息，背景中的植物细节更加清晰（如图 9 中红框所示）。客观评价方面，两个不同模型的融合结果评价指标如表 2 所示。

(a) Att (b) Att + ST (c) Our

图 9 消融前后融合结果对比

Fig.9 Comparison of fusion results before and after ablation

表 2 消融前后图像评价指标平均值

Table 2 Average values of image evaluation indicators before and after ablation

	SCD	MS-SSIM	MI	VIFF
Att	1.585658489	0.861241115	13.7868369	0.331484695
Att + ST	1.573055161	0.834073744	13.88869037	0.318701695
Ours	1.579132302	0.864855029	13.82841411	0.365041201

可以看出，加入的注意力机制对于客观评价标准的提升非常明显，各个评价标准都有不同程度地提升。客观评价结果表明网络中的注意力机制能够使融合性能得以改善。21 对图片的客观评价指标对比如表 2 所示。可以看到加入注意力后 VIFF、MI、MS-SSIM 三个指标有明显提升。

3.3 结果分析

3.3.1 主观评价

现有融合方法和本文融合方法得到的 TNO 融合结果中选取的一对图像，如图 10 所示。从图中可以看出 FusionGAN 融合结果虽然有一些显著的红外特征但是有些地方比较模糊，例如草丛与路面等部分纹理

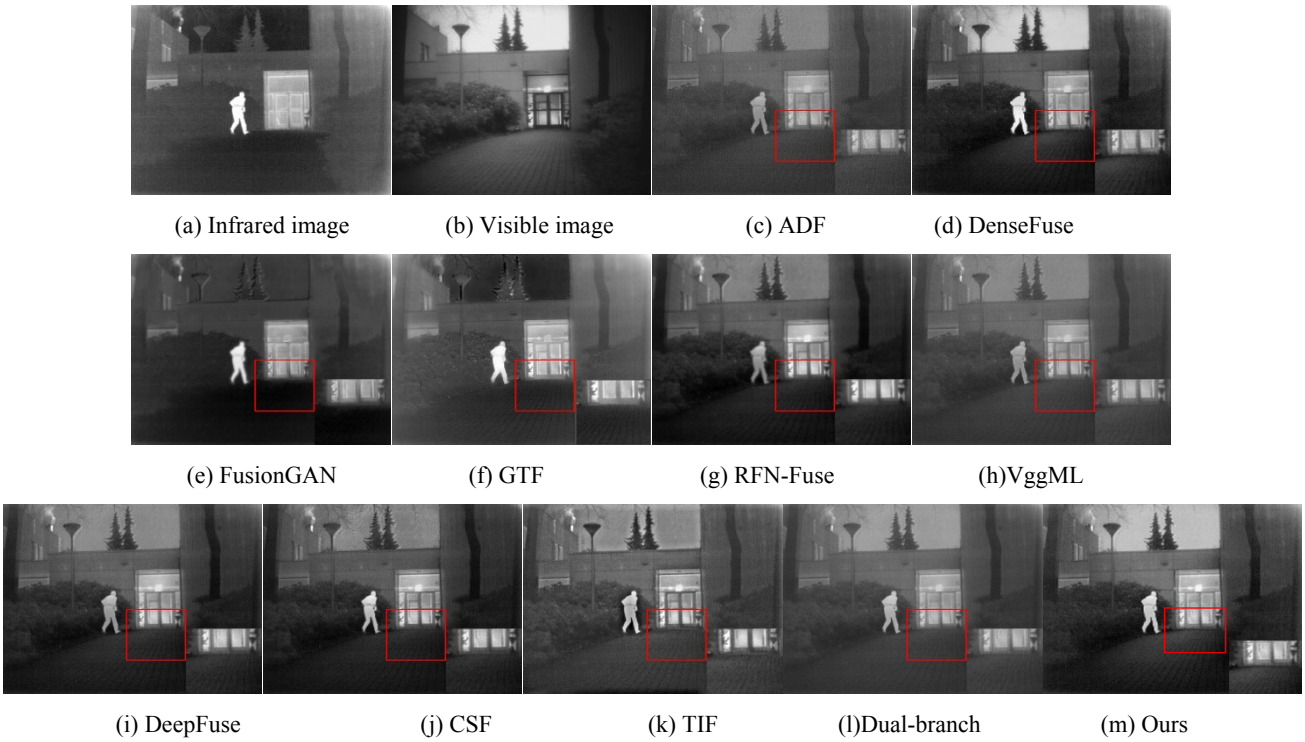
细节不明显。VggML、DenseFuse、Dual-branch 的融合结果中红外信息不突出并且也存在模糊现象。GTF 中丢失部分红外目标信息,例如人物脚部部分。TIF 融合结果较为清晰,但图像中存在噪声和信息融合不均衡现象。

此外,还可以从图 10 红框标记的局部放大区域进行比较。所提方法在主观评价方面比其他融合方法有更好的融合性能,融合结果中的亮度信息也更均衡。RFN-Fuse 融合结果相对较好,但在细节纹理保存方面稍有欠缺。从放大区域可以看出所提方法能较清晰地显示出道路上的条纹,保存更多的纹理细节信息。此外为了体现模型的泛化性能本文还在 MSRS 数据集上做了对比试验如图 11 所示。可以看出相比 FusionGAN、RFN-Fuse 所提方法的红外信息和可见光信息更加平衡,融合结果中可以保留更多细节。

3.3.2 客观评价

本文采用了客观评价指标进行对比,实验结果如表 3 所示。采用的评价指标有 7 种同 3.1 节所示指标。其中每个评价标准最好的结果用红色字体表示。

从表 3 可以看出本文方法有 5 个指标是最优的,用红色字体标出。视觉保真度高说明融合结果具有更高的视觉保真度。平均梯度、边缘强度越高表明图像质量越高,也就更清晰。表 4 展示了 MSRS 数据集上的客观评价结果可以看到所提方法的 5 个指标达到最好结果与在 TNO 数据集得出结果一致,说明所提方法的泛化性能较好。



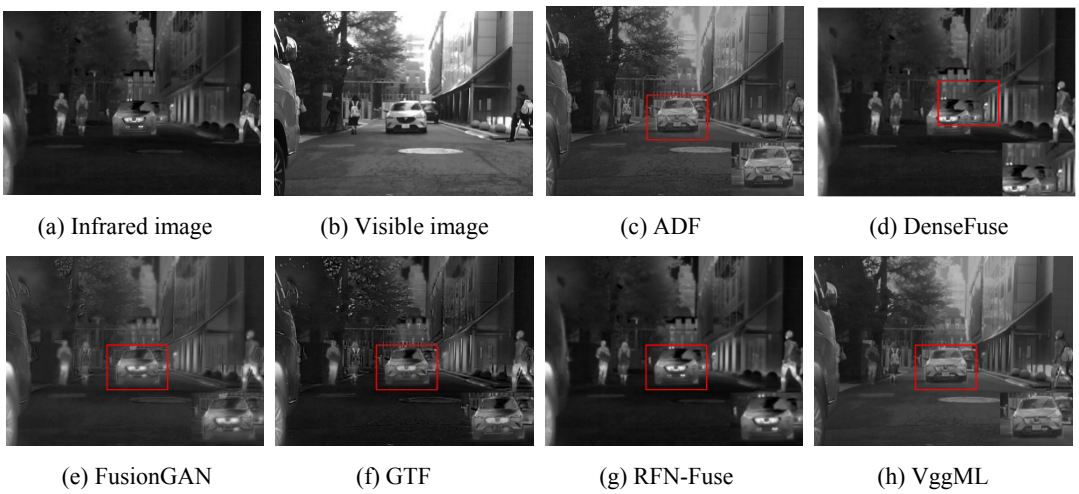
(a) Infrared image (b) Visible image (c) ADF (d) DenseFuse

(e) FusionGAN (f) GTF (g) RFN-Fuse (h)VggML

(i) DeepFuse (j) CSF (k) TIF (l)Dual-branch (m) Ours

图 10 红外和可见光图像的融合结果

Fig.10 Fusion results of infrared and visible images



(a) Infrared image (b) Visible image (c) ADF (d) DenseFuse

(e) FusionGAN (f) GTF (g) RFN-Fuse (h) VggML

(i) DeepFuse (j) CSF (k) TIF (l) Dual-branch (m) Ours

图 11 MSRS 数据集红外和可见光图像融合结果

Fig.11 Fusion results of infrared and visible light images from MSRS dataset

表 3 TNO 数据集 21 对图像评价指标平均值

Table 3 Average value of 21 pairs of image evaluation indicators in TNO dataset

	EI	FMI_dct	QP	VIF	AG	EN	SD
GTF	32.52770	0.10836	0.02177	0.45364	3.35874	6.63534	31.57911
TIF	39.23519	0.19743	0.11410	0.74760	3.89565	6.52602	28.24174
ADF	35.26416	0.28190	0.16059	0.31281	3.67947	6.27304	23.42029
VggML	24.00504	0.40463	0.28970	0.29509	2.42635	6.18260	22.70687
FusionGAN	22.14833	0.36334	0.09887	0.45354	2.20517	6.36285	26.06731
DenseFuse	23.30637	0.40727	0.28615	0.28695	2.35330	6.17403	22.54629
RFN-Fuse	29.14734	0.10639	0.01774	0.34545	2.73375	6.84134	35.27043
DeepFuse	34.73729	0.41501	0.28615	0.28695	2.35330	6.17403	33.65323
Dual-branch	25.07866	0.30116	0.29138	0.35070	2.47084	6.33231	27.02308
CSF	36.81830	0.25636	0.24811	0.71146	3.60953	6.79053	35.71607
Ours	50.76634	0.254905	0.303399	0.684504	5.38937	6.91420	38.77089

表 4 MSRS 数据集 21 对图像评价指标平均值

Table 4 Average value of 21 pairs of image evaluation indicators in MSRS dataset

	EI	FMI_dct	QP	VIF	AG	EN	SD
GTF	28.45466	0.19621	0.15700	0.44730	2.71035	5.73625	24.19185
TIF	43.39727	0.22136	0.33786	1.04271	4.09034	6.58252	35.54339
ADF	32.29431	0.21340	0.29474	0.45374	3.08234	6.29048	28.62276
VggML	26.05613	0.38575	0.40246	0.45717	2.46865	6.24643	28.33981
FusionGAN	16.97583	0.31703	0.13058	0.33249	1.59356	5.60325	19.71231
DenseFuse	30.93252	0.09862	0.02089	0.13650	3.16776	5.65645	24.04045
RFN-Fuse	16.06580	0.26362	0.35816	0.53009	1.47516	5.60288	25.07045
Deep-fuse	28.63384	0.39021	0.39733	0.59795	2.70763	6.42196	32.44943
Dual-branch	26.34184	0.28525	0.36961	0.50415	2.47727	6.21497	31.06896
CSF	28.93600	0.24274	0.34685	0.58995	2.71384	6.25018	32.16605
Ours	55.88537	0.35160	0.47274	0.74274	5.66437	6.73437	41.75073

4 结语

本文提出一种基于 Swin-transformer 和混合特征聚合的融合网络并提出了一种新的混合特征聚合。将 Swin-transformer 与注意力机制引入到多尺度网络中，充分利用长距离语义信息与通道注意力信息，解决基于卷积神经网络方法中细节丢失的问题。所提特征聚合将注意力与特征增强模块混合，能够保留更多背景细节信息。所提方法首先利用一个解码器来提取特征图的多尺度信息。再将各个尺度的特征用所提特征聚合进行融合，分别输入到解码器的对应接口进行解码。由于在编解码过程中使用了注意力机制，突出对结果有重要影响的通道，使得融合结果保留了更多细节和纹理特征。利用提出的网络结构，可以在重构过

程中保留更多的显著特征，提高图像融合的性能。

参考文献：

[1] 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.

[2] Bavirisetti D P, D Huli R. Two-scale image fusion of visible and infrared images using saliency detection [J]. *Infrared Physics & Technology*, 2016, **76**: 52-64.

[3] Bavirisetti D P, Dhuli R. Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform [J]. *IEEE Sensors Journal*, 2015, **16**(1): 203-9.

[4] LIU Y, CHEN X, WARD R K, et al. Image fusion with convolutional sparse representation [J]. *IEEE Signal Processing Letters*, 2016, **23**(12): 1882-6.

[5] ZHOU Z, WANG B, LI S, et al. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters [J]. *Information Fusion*, 2016, **30**: 15-26.

[6] Prabhakar K R, Srikar V S, Babu R V. DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs[C/OL]//*Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*), 2017, https://arxiv.org/abs/1712.07384.

[7] 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.

[8] 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 - 18.

[9] 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.

[10] 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.

[11] FU Y, WU X J, DURRANI T. Image fusion based on generative adversarial network consistent with perception [J]. *Information Fusion*, 2021, **72**: 110-25.

[12] SONG A, DUAN H, PEI H, et al. Triple-discriminator generative adversarial network for infrared and visible image fusion [J]. *Neurocomputing*, 2022, **483**: 183-94.

[13] XUE W, HUAN XIN C, SHENG YI S, et al. MSFSA-GAN: multi-scale fusion self attention generative adversarial network for single image deraining [J]. *IEEE Access*, 2022, **10**: 34442-8.

[14] ZHANG H, YUAN J, TIAN X, et al. GAN-FM: infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators [J]. *IEEE Transactions on Computational Imaging*, 2021, **7**: 1134-47.

[15] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(8): 2011-23.

[16] LI B, LIU Z, GAO S, et al. CSpA-DN: channel and spatial attention dense network for fusing PET and MRI images[C]//*Proceedings of the* 25th *International Conference on Pattern Recognition*, 2021, DOI: 10.1109/ICPR48806.2021.9412543.

[17] HUANG G, LIU Z, MAATEN L V D, et al. Densely connected convolutional networks[C/OL]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, https://arxiv.org/abs/1608.06993.

[18] LI H, WU X. DenseFuse: a fusion approach to infrared and visible images[J]. *IEEE Transactions on Image Processing*, 2019, **28**(5): 2614-23.

[19] ZHOU Z, Rahman Siddiquee M M, Tajbakhsh N, et al. UNet++: A Nested U-Net architecture for medical image segmentation[J/OL]. *Computer Vision and Pattern Recognition*, 2018, https://arxiv.org/abs/1807.10165.

[20] 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): 9645-56.

[21] TOET A. TNO Image Fusion Dataset[EB/OL]. 2014, https://doi.org/10.6084/m9.figshare.1008029.v2.

[22] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[J/OL]. *European Conference on Computer Vision*, 2014, https://arxiv.org/abs/1405.0312.

[23] LI H, WU X, KITTLER J. Infrared and visible image fusion using a deep learning framework[C]// *Proceedings of the 24th International Conference on Pattern Recognition* (ICPR), 2018: 2705-2710, DOI: 10.1109/ICPR.2018.8546006.

[24] XU H, ZHANG H, MA J. Classification saliency-based rule for visible and infrared image fusion [J]. *IEEE Transactions on Computational Imaging*, 2021, **7**: 824-36.

[25] FU Y, WU X J. A dual-branch network for infrared and visible image fusion [J/OL]. *International Conference on Pattern Recognition* (ICPR), 2021, https://arxiv.org/abs/2101.09643.

[26] Xydeas C S, Petrović V. Objective image fusion performance measure [J]. *Electronics Letters*, 2000, **36**(4): 308-309.

[27] HAN Y, CAI Y, CAO Y, et al. A new image fusion performance metric based on visual information fidelity [J]. *Information Fusion*, 2013, 14(2): 127-135.

[28] CUI G, FENG H, XU Z, et al. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition [J]. *Optics Communications*, 2015, 341: 199-209.

[29] AARDT V, JAN. Assessment of image fusion procedures using entropy, image quality, and multispectral classification [J]. *Journal of Applied Remote Sensing*, 2008, **2**(1): 1-28.

[30] Haghighat M, Razian M A. Fast-FMI: Non-reference image fusion metric[C]//*Proceedings of the IEEE 8th International Conference on Application of Information and Communication Technologies* (AICT), 2014: 1-3, DOI: 10.1109/ICAICT.2014.7036000.

[31] ZHAO J, LAGANIERE R, LIU Z. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement[J]. *International Journal of Innovative Computing Information & Control Ijicic*, 2006, **3**(6): 1433-1447.

[32] TANG L, YUAN J, ZHANG H, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware[J]. *Information Fusion*, 2022, **83-84**: 79-92.