红外与可见光图像多层感知机交互融合方法

孙 婧[1]，王志社[1]，杨 帆[1]，余朝发[2]
（1. 太原科技大学 应用科学学院，山西 太原 030024； 2. 陆军工程大学 军械士官学校，湖北 武汉 430075）

摘要：现有的 Transformer 融合方法利用自注意力机制建立图像上下文的全局依赖关系，从而产生优越的融合性能。然而由于与注意力机制相关的模型高复杂度，导致训练效率较低，限制了图像融合的实际应用。为此，本文提出了红外与可见光图像多层感知机交互融合方法。首先，构建轻量化多层感知机网络架构，利用全连接层建立全局依赖关系，在获得较高的计算效率时，具有较强的特征表征能力。其次，设计了级联空间通道交互模型，实现不同空间位置和独立通道之间的特征交互，从而聚焦源图像各自的内在特征，增强模态间特征的互补性。与其他 7 种典型的融合方法相比，TNO、MSRS 数据集以及目标检测任务的实验结果表明，本文方法在主观视觉描述和客观指标评价都优于其他融合方法。本方法利用多层感知机建立图像的长距离依赖关系，构建了级联空间通道交互模型，从空间和通道维度提取图像全局特征，比其他典型融合方法具有更优越的融合性能和更高的计算效率。

关键词：图像融合；多层感知机；特征交互；红外图像；可见光图像
中图分类号：TP394.1；TH691.9 **文献标识号：**A **文章编号：**1001-8891(2025)05-0619-09

Multi-layer Perceptron Interactive Fusion Method for Infrared and Visible Images

SUN Jing[1]，WANG Zhishe[1]，YANG Fan[1]，YU Zhaofa[2]
(1. *School of Applied Science, Taiyuan University of Science and Technology, Taiyuan* 030024, *China*;
2. *Ordnance NCO Academy, Army Engineering University of PLA, Wuhan* 430075, *China*)

Abstract: Existing Transformer-based fusion methods employ a self-attention mechanism to model the global dependency of the image context, which can generate superior fusion performance. However, due to the high complexity of the models related to attention mechanisms, the training efficiency is low, which limits the practical application of image fusion. Therefore, a multilayer perceptron interactive fusion method for Infrared and visible images, called MLPFuse, is proposed. First, a lightweight multilayer perceptron network architecture is constructed that uses a fully connected layer to establish global dependencies. This framework can achieve high computational efficiency while retaining strong feature representation capabilities. Second, a cascaded token- and channel-wise interaction model is designed to realize feature interaction between different tokens and independent channels to focus on the inherent features of the source images and enhance the feature complementarity of different modalities. Compared to seven typical fusion methods, the experimental results on the TNO and MSRS datasets and object detection tasks show that the proposed MLPFuse outperforms other methods in terms of subjective visual descriptions and objective metric evaluations. This method utilizes a multilayer perceptron to model the long-distance dependency of images and constructs a cascaded token-wise and channel-wise interaction model to extract the global features of images from spatial and channel dimensions. Compared with other typical fusion methods, our MLPFuse achieves remarkable fusion performance and competitive computational efficiency.

Key words: image fusion, multi-layer perceptron, feature interaction, infrared image, visible image

收稿日期：2023-12-12；**修订日期：**2024-01-19.
作者简介：孙婧（2000-），女，河南开封人，硕士，研究方向为图像融合、深度学习。E-mail: sunjing@stu.tyust.edu.cn。
通信作者：王志社（1982-），男，安徽定远人，博士，教授，主要从事机器视觉、深度学习、图像融合研究。E-mail：wangzs@tyust.edu.cn。
基金项目：山西省应用基础研究计划项目（202203021221144），山西省专利转化计划项目（202405012）。

0 引言

图像融合技术能够综合两种不同传感器的成像优势，弥补单一传感器的不足，从而获得更好的场景表达和目标特性描述。红外传感器通过捕获热辐射成像，能在低照度、恶劣天气条件下工作，目标显著性强，但存在纹理细节少，对比度低的问题。可见光传感器捕获光反射信息，成像分辨率高，纹理特征清晰，但易受光照环境的影响。红外和可见光图像融合技术能将源图像不同波段，不同频率的目标信息以适当的策略进行互补融合，获得具有丰富纹理细节和突出典型目标的融合图像，可广泛应用于目标检测[1]、行人再识别[2]和语义分割[3]等领域。

传统融合方法通过空间域和变换域对源图像进行特征提取，采用特定的融合规则进行特征合并。典型的传统融合方法主要包括基于多尺度变换的方法[4]和基于稀疏表示的方法[5]。这些融合方法往往依靠人工测量活动水平或人为设计融合规则来实现特征整合，无法根据源图像特征自适应变化，特征提取能力有限，难以适应复杂的成像场景。

由于神经网络具备强大的学习能力，能有效克服传统融合方法的不足。深度学习融合方法可大致分为自编码器（Auto Encoder, AE）融合方法[6-7]、卷积神经网络（Convolutional Neural Network, CNN）方法[8-12]、生成对抗网络（Generative Adeversarial Network, GAN）方法[13-17]和 Transformer 图像融合方法[18-20]。Wang 等人[6]设计了统一的多尺度密集编码解码器，采用全局注意力模型作为融合策略。Xu 等人[7]设计了一种分类显著性融合规则，更好地保留了源图像各自特征。尽管上述方法取得了较好的融合性能，但仍需人为设计相应的融合策略。Xu 等人[11]提出无监督端到端网络，自适应保持融合图像与源图像的相似性，利用信息保存度控制权重分配。Li 等人[12]构建了一种两阶段学习训练网络，分别训练特征提取与融合网络，取代了人为设计的手动融合策略。然而这些方法仅仅采用卷积操作提取局部特征，无法有效建模全局特征，导致图像上下文信息丢失，限制了图像的融合性能。Ma 等人[13]通过生成器和鉴别器建立对抗博弈，生成的融合图像易偏向于红外图像。Ma 等人[14]随后构建了双鉴别器网络，将图像融合转化为多分类问题，虽然能平衡融合结果，但仍存在融合图像边缘信息模糊，目标纹理边缘信息丢失等问题。Wang 等人[16]设计了迭代特征注意力模型，采用双路注意力模块来传递和补偿三重路径的特征信息。Wang 等人[17]通过跨尺度迭代方式逐步优化源图像的活动水平。虽然这些方法取得了优越的融合性能，但忽略了跨模态的特征交互，影响了融合性能的进一步提高。

近年来，Vision Transformer[18]通过自注意力机制可有效提取全局上下文信息，克服了卷积神经网络的局限性，因此被广泛应用于图像融合领域。Wang 等人[19]设计了一种基于 L1 范数的序列矩阵特征融合策略，利用 Transformer 特征编码模块构建长距离依赖关系，具有较强的特征表征能力。Tang 等人[20]，遵循图像级框架，通过动态 CNN-Transformer 模块提取图像局部和全局特征。基于 Transformer 的图像融合方法，可以对图像特征进行全局建模，但自注意力中矩阵 $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ 的计算量与图像尺寸成平方关系，导致模型复杂度高。

针对上述问题，本文提出了红外与可见光图像多层感知机交互融合方法（Multi-Layer perceptron interactive fusion method for infrared and visible images，MLPFuse）。首先，构建了轻量化多层感知机网络架构，利用卷积操作将低维图像映射为高维特征，提取图像浅层特征，再使用多层感知机建模全局依赖关系，获得的全局特征更聚焦于红外的典型目标和可见光的场景细节。同时，由于多层感知机忽略了注意力机制相关计算，模型具有更高的计算效率。此外，设计了级联空间通道交互模型，允许不同空间和独立通道之间以交互方式进行特征传递，增强了融合图像信息的互补性，获得质量更高的融合图像。

1 融合方法

1.1 网络结构

多层感知机交互融合方法的原理如图 1 所示，网络框架由编码模块、融合层和解码模块 3 部分构成。将红外和可见光图像 $I_{\mathrm{ir}} \in R^{H\times W\times C}$ 和 $I_{\mathrm{vis}} \in R^{H\times W\times C}$ 分别输入到编码器中。其中，H、W、C 分别为输入图像的高、宽和通道数。首先，利用卷积将低维图像映射到高维特征空间，提取浅层特征信息 Φ_{ir} 和 Φ_{vis}，如公式(1)所示：

$$\{\Phi_{\mathrm{ir}},\Phi_{\mathrm{vis}}\}=\{H_{\mathrm{SE}}(I_{\mathrm{ir}}),H_{\mathrm{SE}}(I_{\mathrm{vis}})\} \tag{1}$$

式中：H_{SE} 为浅层提取操作，由两个卷积核大小为 3×3 卷积层组成。

然后将浅层特征输入级联空间通道交互模型，提取图像全局特征 $\Phi_{\mathrm{ir}}^{\mathrm{c}}$ 和 $\Phi_{\mathrm{vis}}^{\mathrm{c}}$。最后，将提取到的全局特征经过通道合并实现特征融合，并使用卷积层解码得到融合特征图 I_{f}，如公式(2)所示：

$$I_{\mathrm{f}}=H_{\mathrm{Conv}}(\mathrm{Concate}[\Phi_{\mathrm{ir}}^{\mathrm{c}},\Phi_{\mathrm{vis}}^{\mathrm{c}}]) \tag{2}$$

式中：Concate 为通道合并操作；H_{Conv} 是解码模块，为两个卷积核 3×3、步长为 1 的卷积。

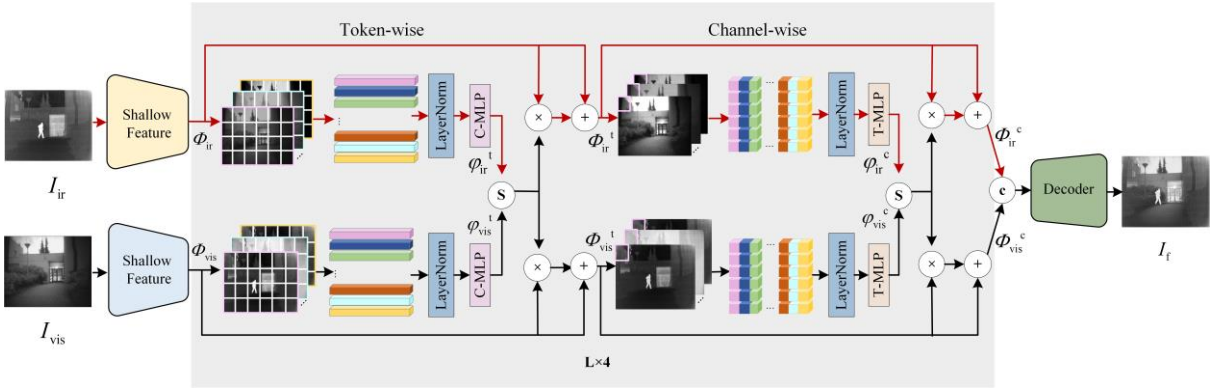


图 1 多层感知机融合方法框架

Fig.1 Network framework for multi-layer perceptron fusion method

1.2 级联空间通道交互模型

级联空间通道交互模型由 token-wise 和 channel-wise MLPs 组成，分别从空间和通道维度构建全局依赖关系，并通过 SoftMax 函数进行红外和可见光全局特征交互。如图 2 所示，每个 MLP 都由两个全连接层和激活函数 GELU 构成。因此，MLP 包含输入输出层和一个隐藏层，层与层之间都是全连接的，设定输入输出层的神经元个数相同，隐藏层神经元为输入层神经元个数的 2 倍。通过 MLP 全连接层提取图像全局特征，抛弃了自注意力相关计算，模型复杂度低，提高了计算效率。

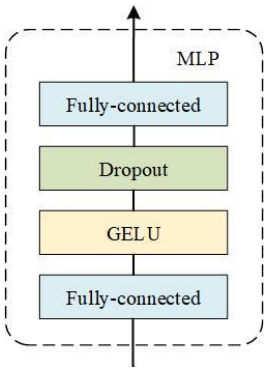


图 2 MLP 模块示意图

Fig.2 Schematic diagram of multi-layer perceptron module

对于输入局部特征 Φ_{ir}, Φ_{vis}，首先将其划分为 $P\times P$ 大小的图像块，并投影到二维矩阵 $\boldsymbol{X}\in R^{T\times C}$。其中，$T=HW/P^2$ 表示图像块数，C 表示通道数。然后向量组进入 token-wise MLP 中，在每个 token 上映射为：$R^{T\times C}\rightarrow R^{T}$，通过全连接层建模全局依赖关系，获得空间维度的特征序列 $\varphi_{\mathrm{ir}}^{\mathrm{t}}$ 和 $\varphi_{\mathrm{vis}}^{\mathrm{t}}$，如公式(3)和(4)所示：

$$\varphi_{\mathrm{ir}}^{\mathrm{t}}=\text{T-MLP}(\text{LN}(\Phi_{\mathrm{ir}})) \tag{3}$$

$$\varphi_{\mathrm{vis}}^{\mathrm{t}}=\text{T-MLP}(\text{LN}(\Phi_{\mathrm{vis}})) \tag{4}$$

式中：T-MLP 表示 token-wise MLP 操作；LN 表示层归一化 LayerNorm。

随后，利用 SoftMax 函数计算出的空间维度红外与可见光图像的各自权重，如公式(5)所示：

$$[\beta_{\mathrm{ir}}^{\mathrm{t}},\ \beta_{\mathrm{vis}}^{\mathrm{t}}]=\text{SoftMax}(\varphi_{\mathrm{ir}}^{\mathrm{t}},\ \varphi_{\mathrm{vis}}^{\mathrm{t}}) \tag{5}$$

将生成的特征权重与输入的红外和可见光浅层特征 Φ_{ir} 和 Φ_{vis} 分别进行相乘，再通过短连接得到空间维度的红外与可见光图像全局特征，如公式(6)和(7)所示：

$$\Phi_{\mathrm{ir}}^{\mathrm{t}}=\Phi_{\mathrm{ir}}\times(\beta_{\mathrm{ir}}^{\mathrm{t}}+1) \tag{6}$$

$$\Phi_{\mathrm{vis}}^{\mathrm{t}}=\Phi_{\mathrm{vis}}\times(\beta_{\mathrm{vis}}^{\mathrm{t}}+1) \tag{7}$$

接着，将空间维度提取的图像特征输入到 channel-wise MLP 中，在每个通道维度上映射为：$R^{T\times C}\rightarrow R^{\mathrm{C}}$，获得通道维度的特征序列 $\varphi_{\mathrm{ir}}^{\mathrm{c}}$ 和 $\varphi_{\mathrm{vis}}^{\mathrm{c}}$，如公式(8)和(9)所示：

$$\varphi_{\mathrm{ir}}^{\mathrm{c}}=\text{C-MLP}(\text{LN}(\Phi_{\mathrm{ir}}^{\mathrm{t}})) \tag{8}$$

$$\varphi_{\mathrm{vis}}^{\mathrm{c}}=\text{C-MLP}(\text{LN}(\Phi_{\mathrm{vis}}^{\mathrm{t}})) \tag{9}$$

式中：C-MLP 表示 channel-wise MLP 操作。

类似地，再通过 SoftMax 函数计算出通道维度红外与可见光图像的各自权重，如公式(10)所示：

$$[\beta_{\mathrm{ir}}^{\mathrm{c}},\ \beta_{\mathrm{vis}}^{\mathrm{c}}]=\text{SoftMax}\ (\varphi_{\mathrm{ir}}^{\mathrm{c}},\ \varphi_{\mathrm{vis}}^{\mathrm{c}}) \tag{10}$$

将生成的特征权重与输入的空间维度特征分别进行相乘和相加，得到通道维度的红外与可见光图像全局特征 $\Phi_{\mathrm{ir}}^{\mathrm{c}}$ 和 $\Phi_{\mathrm{vis}}^{\mathrm{c}}$，如公式(11)和公式(12)所示：

$$\Phi_{\mathrm{ir}}^{\mathrm{c}}=\Phi_{\mathrm{ir}}^{\mathrm{t}}\times(\beta_{\mathrm{ir}}^{\mathrm{c}}+1) \tag{11}$$

$$\Phi_{\mathrm{vis}}^{\mathrm{c}}=\Phi_{\mathrm{vis}}^{\mathrm{t}}\times(\beta_{\mathrm{vis}}^{\mathrm{c}}+1) \tag{12}$$

最后，将生成的红外和可见光图像全局特征 $\Phi_{\mathrm{ir}}^{\mathrm{c}}$ 和 $\Phi_{\mathrm{vis}}^{\mathrm{c}}$ 经过 L 次全局建模后，提取到图像特征通道合并进行融合，随后输入到解码器中，由卷积层解码得到融合图像。

1.3 损失函数

为了获得更好的融合性能，网络采用 3 种损失函数来约束融合图像与源图像之间的差异性，分别是结构相似度损失 L_{ssim}、纹理损失 L_{grad} 和亮度损失 $L_{\mathrm{intensity}}$，总损失函数如公式(13)所示：

$$L_{\text{total}}=L_{\text{ssim}}+\lambda_1 L_{\text{grad}}+\lambda_2 L_{\text{intensity}} \tag{13}$$

结构相似度通过比较两幅图像的亮度、对比度和结构等相似性，评估生成图像与真实图像的相似程度。L_{ssim} 函数用于计算源图像和融合图像的结构相似性，公式如(14)所示：

$$L_{\text{ssim}}=\omega_1\cdot(1-\text{ssim}(I_{\text{f}}, I_{\text{ir}}))+\omega_2\cdot(1-\text{ssim}(I_{\text{f}}, I_{\text{vis}})) \tag{14}$$

式中：ssim(·)表示结构相似度操作，是衡量两幅图像相似性的指标。ω_1 和 ω_2 为超参数，且设置为 $\omega_1=\omega_2=0.5$。

设计纹理损失函数更好地保留源图像的纹理细节和边缘信息。公式如(15)所示：

$$L_{\text{grad}}=\frac{1}{HW}\left\|\nabla I_{\text{f}}\right|-\max\left(\left|\nabla I_{\text{ir}}\right|,\left|\nabla I_{\text{vis}}\right|\right)\Big\|_1 \tag{15}$$

式中：∇表示 Sobel 梯度算子；|·|表示绝对值算子；$\|\cdot\|_1$ 表示 L1 范数，max(·)表示最大值函数。

最后，亮度损失函数具体如公式(16)所示：

$$L_{\text{intensity}}=\frac{1}{HW}\|I_{\text{f}}-\text{mean}(I_{\text{ir}}, I_{\text{vis}})\|_1 \tag{16}$$

式中：mean(·)表示元素平均操作。

2 实验验证

2.1 实验参数设定

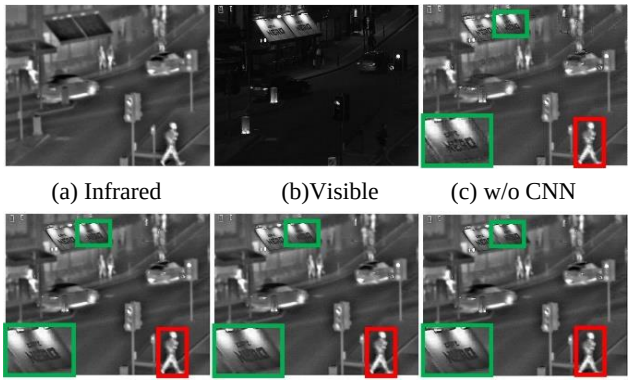
在训练阶段，采用 TNO 数据集进行训练。为了扩大数据集，采用滑动步长为 12，将训练图像裁剪为分辨率大小为 128×128 图像块，同时将灰度值范围转化为[0,1]，得到 18813 组红外和可见光图像。窗口 P 大小设置为 8，损失函数的权重参数设置为 $\lambda_1=30$，$\lambda_2=5$。采用 Adam 优化器更新模型参数，初始学习率设置为 1×10^{-5}，batch size 和 epoch 分别设置为 4 和 8。所有实验都在 NVDIA GeForce GTX 3090 GPU 和 Inter i9-10850 K CPU 上进行。

在测试阶段，从 TNO[21]和 MSRS[22]数据集中分别选取 25 和 361 组红外和可见光图像作为测试集。选择 7 种具有代表性的方法进行比较，分别是基于 AE 的融合方法 CSF[7]，基于 CNN 的融合方法 U2Fusion[11]和 RFN-Nest[12]，基于 GAN 的融合方法 FusionGan[13]和 GanMcC[14]，基于 Transformer 的融合方法 SwinFuse[19]和 YDTR[20]。选择 8 个定量指标进行性能评估，分别是平均梯度（average gradient, AG）、相位一致性（phase congruency, PC）[23]、视觉信息保真度（visual information fidelity, VIF）[24]、结构相似度度量（structural similarity index measure, SSIM）[25]、标准差（standard deviation, SD）[26]、互信息（mutual information, MI）[27]、基于梯度的相似度度量（gradient-based similarity measurement, Q^{abf}）[28]和基于边缘的相似度度量（edge-based similarity measurement, Q^{e}）[29]。

2.2 消融实验

为了验证网络模型各个组件的有效性，采用 3 个模型进行对比，在原有模型的基础上，分别去除 CNN 模块（记作 w/o CNN），去除 token-wise MLP 模块（记作 w/o Token）和去除 channel-wise MLP 模块（记作 w/o Channel）。利用 TNO 数据集上的 25 组红外和可见光图像进行定性定量实验，定性对比结果如图 3 所示。w/o CNN 由于缺乏多维特征空间，融合结果部分局部细节信息丢失，边缘模糊。而 w/o Token 和 w/o Channel 融合图像结果差异不明显，这是因为只保留一个 MLP 模块，仍可以从通道或空间维度进行全局依赖关系建模。相比之下，MLPFuse 融合结果有更高的对比度和清晰的背景信息，能更好地保留红外显著目标和可见光纹理细节。

各种模型的定量对比结果如表 1 所示，最优值和次优值分别以黑体加粗和下划线标注。从表中看出，当去除任一组件，不同模型的融合结果都有所下降。MLPFuse 方法在指标 PC、VIF、SD、MI 和 Q^{abf} 均排名第一，AG 低于 w/o CNN，SSIM 和 Q_{e} 仅次于 w/o Channel。与其他模型相比，MLPFuse 方法具有更好的融合性能，说明模型框架设计的有效性和合理性。



(a) Infrared (b)Visible (c) w/o CNN

(d) w/o Token (e) w/o Channel (f) MLPFuse

图 3 TNO 数据集上的定性对比结果

Fig.3 The qualitative comparison results on the TNO dataset

表 1 TNO 数据集 4 种模型的定量对比结果

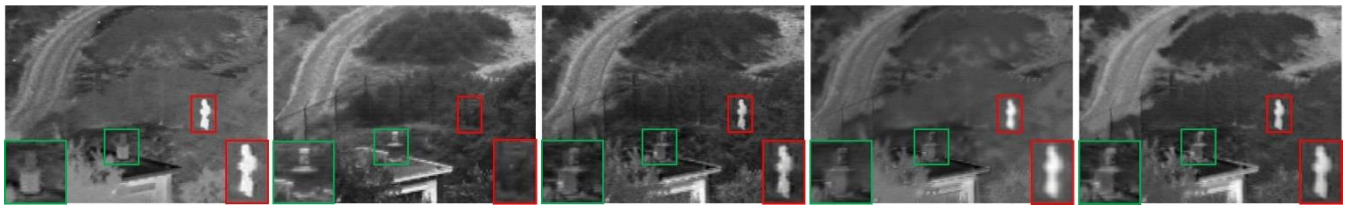
Table 1 The quantitative comparison results of four fusion models on the TNO dataset

Metrics	w/o CNN	w/o Channel	w/o Token	MLPFuse
AG	**6.1248**	5.1800	5.1256	<u>5.2920</u>
PC	0.1592	<u>0.3504</u>	0.3238	**0.3552**
VIF	0.3298	<u>0.4492</u>	0.4432	**0.4527**
SSIM	0.6664	**0.7222**	0.7173	<u>0.7185</u>
SD	36.4795	<u>37.2235</u>	37.0068	**37.5581**
MI	2.2632	<u>3.6139</u>	3.3471	**3.8572**
Q^{abf}	0.5234	<u>0.5379</u>	0.5120	**0.5411**
Q^{e}	0.2425	**0.4934**	0.4850	<u>0.4923</u>

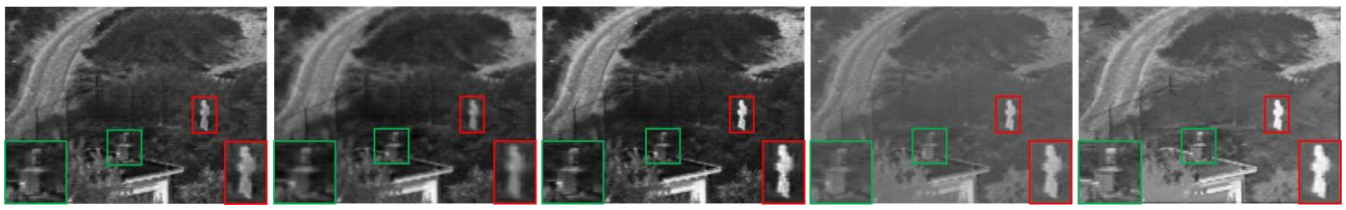
2.3 TNO 数据集实验对比

为了验证方法的优越性，选取 TNO 数据集中“Nato_camp”和“Street”这两组具有代表性的场景进行定性评价对比，其对比结果如图 4 和图 5 所示。分别用红色和绿色框标注红外显著目标和可见光纹理细节，并进行放大以便于观察。CSF 方法设计了基于显著性的融合规则，但融合图像仍产生了有限的亮度和较差的对比度。FusionGan 和 GanMcC 融合图像保留了显著的热目标，但存在边缘和背景模糊，且保留纹理细节信息较差。U2Fusion 和 RFN-Nest 较好地保留了如“烟囱”、“广告牌”和“树枝”等纹理细节信息，但丢失了红外典型目标的亮度信息。SwinFuse 由于采用 L1 正则化融合策略，融合结果背景亮度较低，图像对比度较差。YDTR 融合图像保留了可见光的场景信息，但红外目标不明显，且边缘轮廓不清晰。MLPFuse 方法取得了更好的视觉效果，有较高的亮度和清晰的边缘信息，能同时保留红外图像显著目标和可见光图像丰富的纹理细节。

TNO 数据集的定量评价结果如图 6 所示。从表中可以看出，MLPFuse 方法的指标 PC、VIF、MI、Q^{abf} 和 Q_e 取得了最优值，指标 SSIM 和 SD 取得次优值，分别次于 YDTR 和 SwinFuse。而指标 AG 取得第三，低于 SwinFuse 和 U2Fusion。PC 和 MI 指标取得最优值，表明方法保留了更多源图像的特征信息。Q^{abf} 和 Q^e 指标取得最优值，表明了该方法生成的融合图像更好地融合了边缘信息。VIF 指标取得最优值，表明融合图像有较好的视觉保真度，更符合人类视觉感知系统。SD 和 SSIM 指标取得了次优值，表明具有更高的对比度和更好的场景结构保留能力。实验结果表明，相比其他 7 种先进的融合方法，MLPFuse 方法具有更优的融合性能。



(a) Infrared (b) Visible (c) CSF (d) Fusion GAN (e) GANMcC

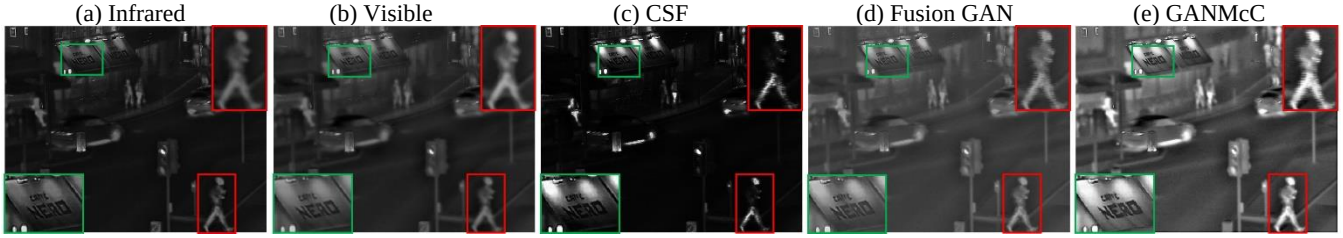


(f) U2Fusion (g) RFN-Nest (h) SwinFuse (i) YDTR (j) MLPFuse

图 4 TNO 数据集“Nato_camp”不同方法的定性对比结果

Fig.4 The qualitative comparison results of different methods for Nato_camp

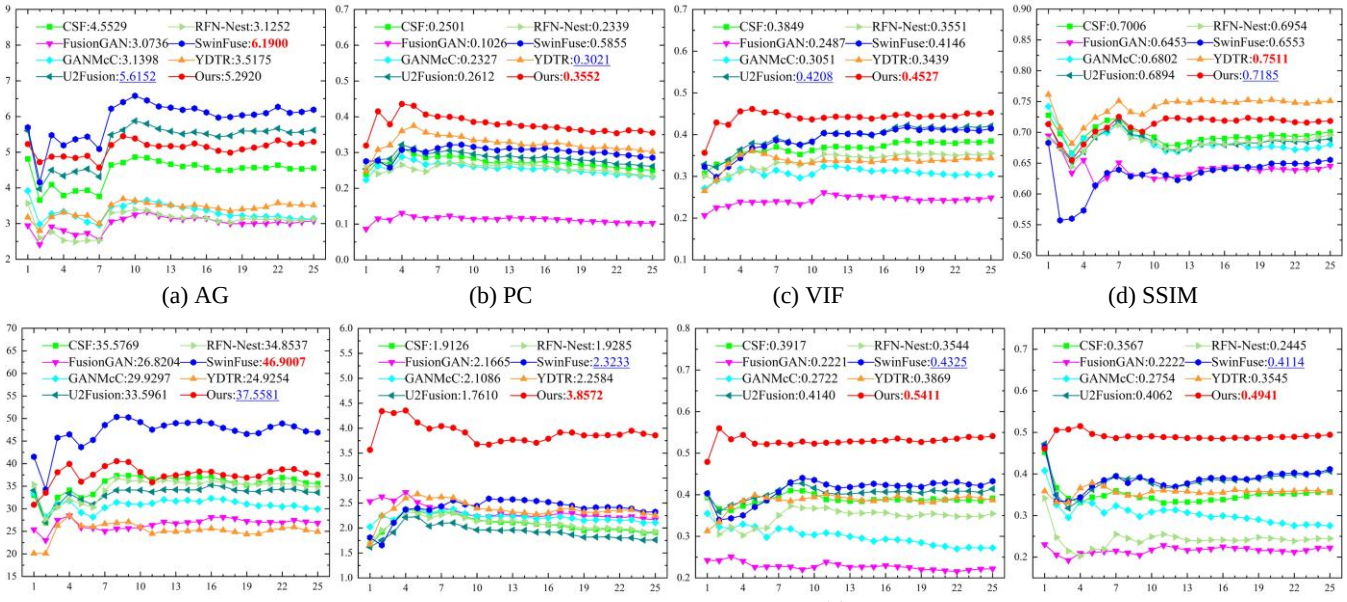
(a) Infrared (b) Visible (c) CSF (d) Fusion GAN (e) GANMcC



(f) U2Fusion (g) RFN-Nest (h) SwinFuse (i) YDTR (j) MLPFuse

图 5 TNO 数据集“Street”不同方法的定性对比结果

Fig.5 The qualitative comparison results of different methods for Street



(a) AG (b) PC (c) VIF (d) SSIM

(e) SD (f) MI (g) Q^{abf} (h) Q^{e}

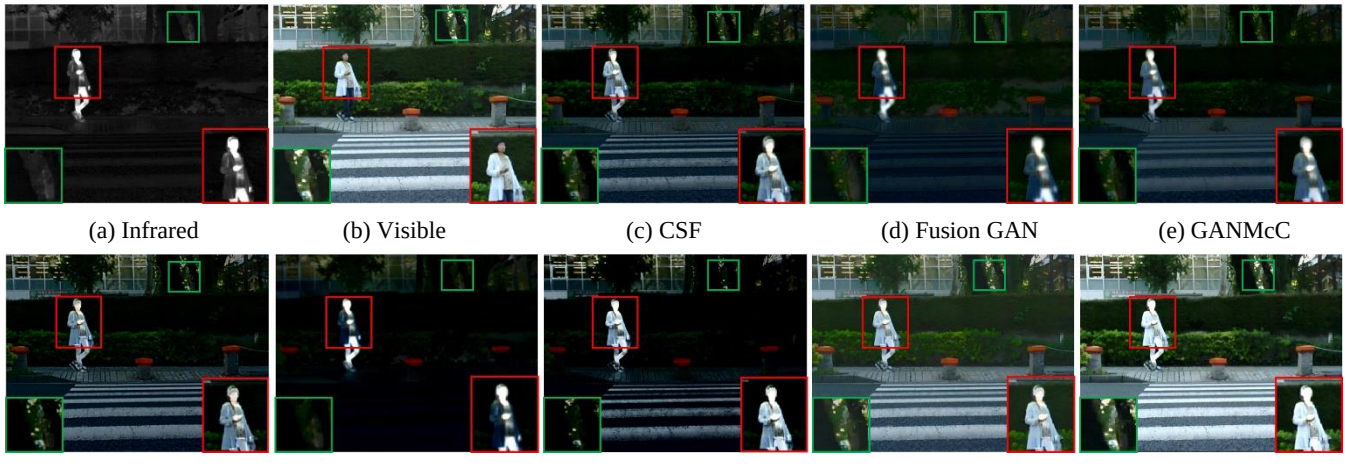
图 6 TNO 数据集上不同方法的定量对比结果

Fig.6 The quantitative comparison results of different methods on the TNO dataset

2.4 MSRS 数据集实验对比

为了进一步验证方法的有效性，对 MSRS 数据集进行实验验证。其中，MSRS 数据集中可见光图像为 RGB 三通道图像。首先要实现可见光图像颜色通道转换，将图像转化为 Y、C_b 和 C_r 通道，然后，将 Y 通道分量作为可见光与红外图像输入到网络中，得到的融合结果与 C_b 和 C_r 通道合并，进行颜色反变换，获得最终的 RGB 融合图像。

从 MSRS 数据集中选取“00123D”和“00591D”进行定性评价，对比结果如图 7 和图 8 所示。CSF 融合图像背景亮度较低，对比度较差。对于红外目标，GanMcC 和 FusionGan 较好地保留了人物显著目标，但融合图像中细节信息较少，出现了边缘模糊。U2Fusion 和 RFN-Nest 更倾向于可见光图像，能够保留“树枝”和“建筑”的细节信息，但红外显著目标不明显。基于 Transformer 融合方法，YDTR 和 SwinFuse 有良好的融合结果，但背景信息丢失，图像亮度低。相比之下，MLPFuse 方法同时保留了红外图像和可见光图像的各自特征，获得的彩色融合图像既有清晰的纹理特征又有显著的红外目标，呈现更好的视觉效果。表 2 给出了各种方法在 MSRS 数据集的定量对比结果，最优值和次优值分别以黑体加粗和下划线标注，MLPFuse 方法在指标 AG、PC、VIF、SSIM、SD、MI、Q^{abf} 和 Q^{e} 上取得最优值。指标平均值越大，表明图像的融合性能越好。总体上，主客观评价结果表明，本文方法取得了良好的结果。

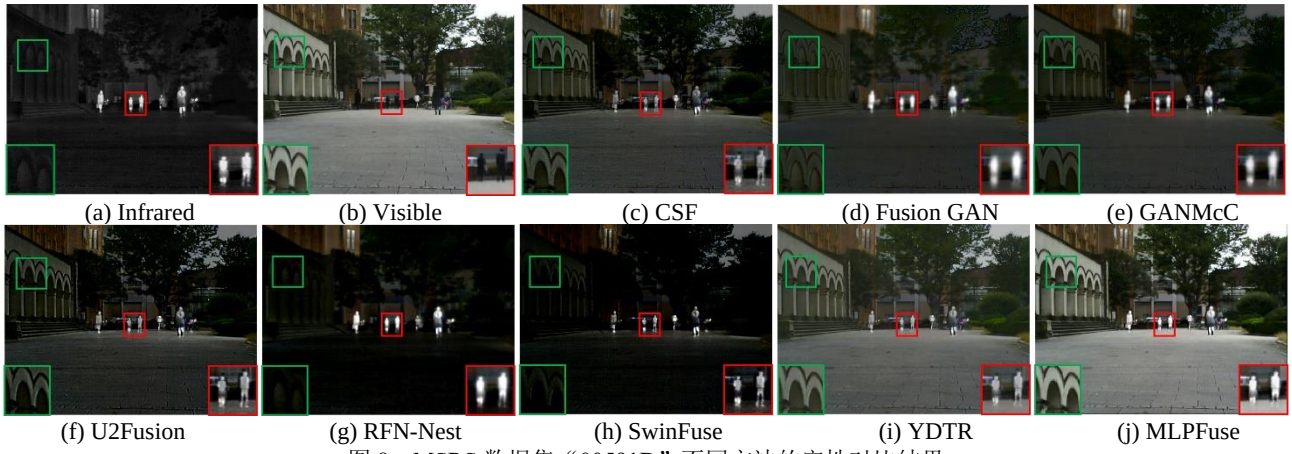


(a) Infrared (b) Visible (c) CSF (d) Fusion GAN (e) GANMcC

(f) U2Fusion (g) RFN-Nest (h) SwinFuse (i) YDTR (j) MLPFuse

图 7 MSRS 数据集“00123D”不同方法的定性对比结果

Fig.7 The qualitative comparison results of different methods for 00123D



(a) Infrared (b) Visible (c) CSF (d) Fusion GAN (e) GANMcC

(f) U2Fusion (g) RFN-Nest (h) SwinFuse (i) YDTR (j) MLPFuse

图 8 MSRS 数据集“00591D”不同方法的定性对比结果

Fig.8 The qualitative comparison results of different methods for 00591D

表 2 MSRS 数据集上不同方法的定量对比结果

Table 2 The quantitative comparison results of different methods on the MSRS dataset

Metrics	CSF	FusionGAN	GanMcC	U2Fusion	RFN-Nest	SwinFuse	YDTR	MLPFuse
AG	2.7039	1.6765	2.3093	3.2542	1.4409	2.3004	2.5967	**4.2991**
PC	0.3465	0.1347	0.3218	0.3369	0.3278	0.2555	0.3800	**0.4184**
VIF	0.3458	0.2269	0.3328	0.3462	0.3818	0.2025	0.2992	**0.4697**
SSIM	0.6872	0.6126	0.6863	0.6910	0.6711	0.3197	0.5969	**0.7124**
SD	26.6847	17.0763	26.3381	25.5250	19.8085	29.7195	25.3717	**42.5426**
MI	2.4007	1.8926	2.5656	2.0158	3.3227	1.7803	2.7674	**3.8856**
Q^{abf}	0.3799	0.1405	0.3044	0.4191	0.2457	0.1790	0.3489	**0.6065**
Q_e	0.2843	0.1446	0.2921	0.3191	0.2364	0.1348	0.2694	**0.5092**

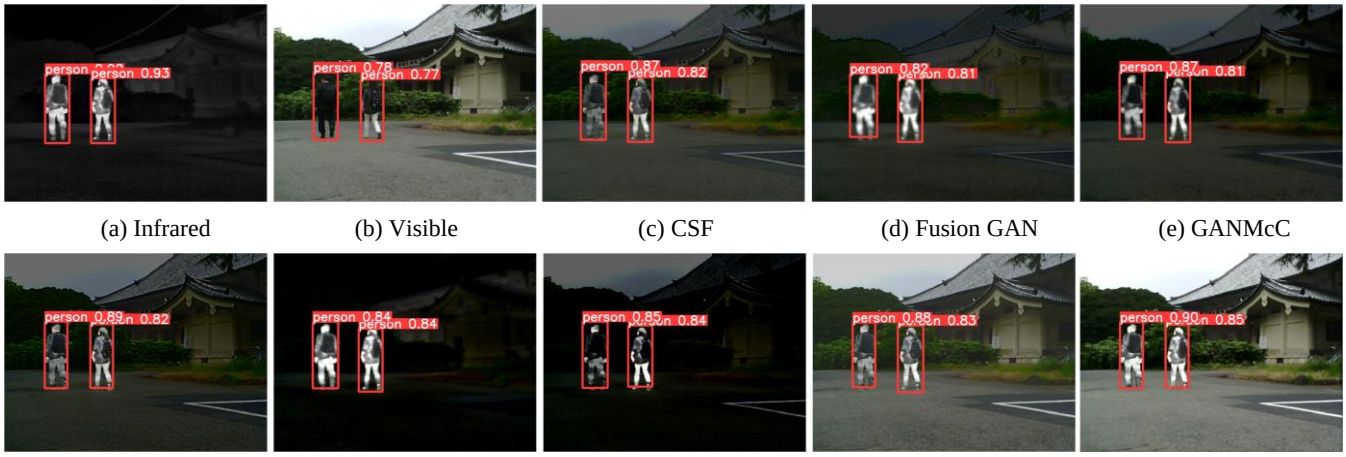
2.5 目标检测实验对比

采用 YOLOv5 检测器对融合图像的目标检测性能进行评估，选择 MSRS 数据集 80 组红外和可见光图像作为训练和测试集，其中，标注的目标类别是行人和车辆。将红外图像、可见光图像和融合图像分别输入到 YOLOv5 检测器中，使用平均精度均值（the mean average precision，mAP）评估检测性能，其中 mAP@0.5 表示 IoU（intersection over nuion，IoU）阈值为 0.5 时的 mAP 值，mAP@[0.5:0.95]表示不同 IoU 阈值下所有 mAP 的平均值（从 0.5 到 0.95，以 0.05 为步长）。MSRS 数据集上源图像和不同融合结果的目标检测定量结果如表 3 所示，最优值和次优值分别以黑体加粗和下划线标注，红外图像在不同的 IoU 阈值下对行人的检测性能较好，表示红外图像可以为检测器提供显著目标，而可见光图像中包含汽车的信息。不同的融合方法将红外图像和可见光图像的互补信息进行融合，为图像检测提供了更全面的场景表达。与其他典型融合方法的检测结果相比，MLPFuse 方法在 mAP@0.5 和 mAP@[0.5:0.95]中都有较高的值，表明方法有更优越的检测性能。图 9 给出了目标检测视觉对比结果，从图中可以看出，MLPFuse 的融合图像有更好的检测结果。从主观评价看，本文融合方法目标检测结果有更优越的性能。

表 3 MSRS 数据集上源图像和不同融合结果的目标检测定量对比结果

Table 3 The quantitative comparison results of object detection in infrared, visible and fused images on the MSRS dataset

Method	mAP@0.5			mAP@ [0.5:0.95]		
	Person	Car	All	Person	Car	All
Infrared	**0.983**	0.946	0.965	0.631	0.666	0.649
Visible	0.908	0.979	0.944	0.492	**0.687**	0.590
CSF	0.977	0.939	0.958	0.623	0.655	0.639
FusionGAN	0.974	0.955	0.965	0.615	0.626	0.620
GanMcC	0.974	0.940	0.957	0.628	0.665	0.646
U2Fusion	0.976	0.949	0.963	0.628	0.635	0.631
RFN-Nest	0.979	0.912	0.945	0.652	0.606	0.629
SwinFuse	0.948	0.828	0.888	0.590	0.479	0.534
YDTR	0.976	0.947	0.962	0.641	0.660	0.641
MLPFuse	0.968	**0.985**	**0.977**	**0.711**	0.653	**0.682**



(a) Infrared (b) Visible (c) CSF (d) Fusion GAN (e) GANMcC

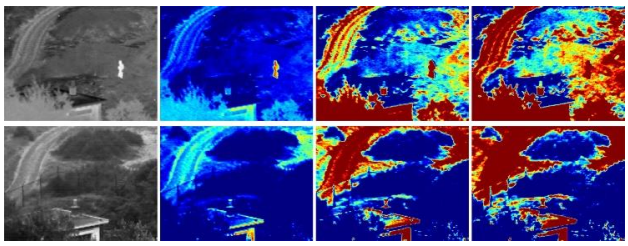
(f) U2Fusion (g) RFN-Nest (h) SwinFuse (i)YDTR (j) MLPFuse

图 9 MSRS 数据集上红外、可见光和融合图像的目标检测定性对比结果

Fig.9 The qualitative comparison results of object detection in infrared, visible and fused images on the MSRS dataset

2.6 特征可视化

网络编码层依赖卷积操作将低维图像映射到高维特征，并构建了级联空间和通道 MLP 交互模型，从空间和通道维度进行特征提取和交互。红外与可见光特征可视化结果如图 10 所示。从图中可以看出，第 2 列是卷积操作的特征图，卷积操作更倾向于保留源图像的部分边缘信息和背景等浅层特征；第 3 列和第 4 列分别是空间维度和通道维度的特征图，通过 MLP 分别在空间和通道维度交互特征信息，保留了更丰富的纹理特征和亮度等信息。



(a) Source images (b) CNN (c) Token_MLP (d) Channel_MLP

图 10 不同组件的特征可视化

Fig.10 Feature visualization of different components

2.7 计算效率

此外，图像融合任务中计算效率也是重要的评价标准，不同融合方法的计算效率如表 4 所示，最优值和次优值分别以黑体加粗和下划线标注，所有方法都在 GPU 上进行测试。通过对比，本文方法的运行速度明显高于其他 7 种融合方法，主要原因是构建了轻量化多层感知机模型，全连接层权重共享，减少了参数量，模型简单。此外，仅利用多层感知机提取全局特征，忽略注意力机制相关计算，简化了运算过程，提高了计算效率。因此，本方法具有更好的融合性能和更高的计算效率。

表 4 不同融合方法计算效率对比结果

Table 4 The comparison results of computation efficiency for different fusion methods

Method	TNO	MSRS
CSF	4.129	11.976
FusionGAN	0.513	1.779
GanMcC	0.785	0.404
U2Fusion	1.515	<u>0.148</u>
RFN-Nest	0.235	0.218
SwinFuse	0.223	0.302
YDTR	<u>0.201</u>	0.360
MLPFuse	**0.149**	**0.121**

3 结论

本文提出红外与可见光图像多层感知机交互融合方法。与 CNN 和 Transformer 图像融合方法不同，设计了一个轻量化多层感知机模型，模型简单且参数量少，通过全连接层提取图像全局上下文信息，具有更强的特征表征能力，同时，大大提高了计算效率。此外，构建了级联空间通道交互模型，从不同空间和独立通道之间进行特征交互。通过 TNO 和 MSRS 数据集的实验对比，与其它 7 种典型融合方法相比，MLPFuse 方法获得了更优越的融合性能，且具有较强泛化能力和更高的计算效率。

参考文献：

[1] 宁大海，郑晟. 可见光和红外图像决策级融合目标检测算法[J]. **红外技术**, 2023, **45**(3): 282-291.

NING D H, ZHENG S. An object detection algorithm based on decision-level fusion of visible and infrared images[J]. *Infrared Technology*, 2023,

45(3): 282-291.

[2] FENG Z, LAI J, XIE X. Learning modality-specific representations for visible-infrared person re-identification[J]. *IEEE Transactions on Image Processing*, 2020(29): 579-590.

[3] 周华兵, 侯积磊, 吴伟, 等. 基于语义分割的红外和可见光图像融合[J]. **计算机研究与发展**, 2021, **58**(2): 436-443.
ZHOU H B, HOU J L, WU W, et al. Infrared and visible image fusion based on semantic segmentation[J]. *Journal of Computer Research and Development*, 2021, **58**(2): 436-443.

[4] WANG Z S, XU J W, JIANG X L, et al. Infrared and visible image fusion via hybrid decomposition of NSCT and morphological sequential toggle operator[J]. *Optik*, 2020, **201**: 1-11.

[5] LI H, WU X J, Kittler J. MDLatLRR：A novel decomposition method for infrared and visible image fusion[J]. *IEEE Transactions on Image Processing*, 2020, **29**: 4733-4746.

[6] WANG Z S, WANG J Y, WU Y Y, et al. UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(6): 3360-3374.

[7] XU H, ZHANG H, MA J Y. Classification saliency-based rule for visible and infrared image fusion[J]. *IEEE Transactions on Computational Imaging*, 2021(7): 824-836.

[8] 杨艳春, 李永萍, 党建武, 等.基于快速交替引导滤波和 CNN 的红外与可见光图像融合[J]. **光学精密工程**, 2023, **31**(10):1548-1562.
YANG Y C, LI Y P, DANG J W, et al. Infrared and visible image fusion based on fast alternating guided filtering and CNN[J]. *Optics and Precision Engineering*, 2023, **31**(10): 1548-1562.

[9] WANG Z S, WU Y Y, WANG J Y, et al. Res2Fusion: Infrared and visible image fusion based on dense Res2net and double non-local attention models[J]. *IEEE Transactions on Instrumentation and Measurement*, 2022, **71**: 1-12.

[10] WANG Z S, YANG F, WANG J Y, et al. A dual-path residual attention fusion network for infrared and visible images[J]. *Optik*, 2023, **33**(7): 3159-3172.

[11] XU H, MA J Y, JIANG J J, et al. U2Fusion: A unified unsupervised image fusion network[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **4**(11): 502-518.

[12] 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): 1566-2535.

[13] MA J Y, YU W, LIANG P W, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. *Information Fusion*, 2019(48): 11-26.

[14] MA J Y, ZHANG H, SHAO Z F, et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021(70): 1-14.

[15] 陈欣. 基于双注意力机制的红外与可见光图像融合方法[J]. **红外技术**, 2023, **45**(6): 639-648.
CHEN X. Infrared and visible image fusion using double attention generative adversarial networks[J]. *Infrared Technology*, 2023, **45**(6): 639-648.

[16] WANG Z S, SHAO W Y, CHEN Y L, et al. Infrared and visible image fusion via interactive compensatory attention adversarial learning[J]. *IEEE Transactions on Multimedia*, 2023, 25: 7800-7813.

[17] WANG Z S, SHAO W Y, CHEN Y L, et al. A cross-scale iterative attentional adversarial fusion network for infrared and visible images[J]. *Transactions on Circuits and Systems for Video Technology*, 2023, **33**(8): 3677-3688.

[18] Dosovitskiy A, Beyer L, A Kolesnikov, et al. An image is worth 16×16 words: Transformers for image recognition at Scale[J]. ArXiv, abs/2010.11929.

[19] WANG Z S, CHEN Y L, SHAO W Y, et al. SwinFuse: A residual swin transformer fusion network for infrared and visible images[J]. *IEEE Transactions on Instrumentation and Measurement*, 2022, **71**: 1-12.

[20] TANG W, HE F Z, LIU Y. YDTR: Infrared and visible image fusion via Y-shape dynamic transformer[J]. *IEEE Transactions on Multimedia*, 2023, **25**: 5413-5428.

[21] TOET A (2014). TNO Image Fusion Dataset. Data[DB/OL]. [2023-12-01]. https://figshare.com/articles/TNO Image Fusion Dataset/1008029.

[22] TANG L F. MSRS Dataset. Data [DB/OL]. [2023-12-01]. https://github.com/Linfeng-Tang/MSRS. 2022.

[23] ZHENG L, FORSYTH D S, Laganière R. A feature-based metric for the quantitative evaluation of pixel-level image fusion[J]. *Computer Vision and Image Understanding*, 2008, **109**(1): 56-68.

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

[25] ZHOU W, 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.

[26] RAO Y J. In-fibre bragg grating sensors[J]. *Measurement Science and Technology*, 1997(8): 355-375.

[27] QU G H, ZHANG D L, YAN P F. Information measure for performance of image fusion[J]. *Electronics Letters*, 2002, **38**(7): 313-315.

[28] PIELLA G, HEIJMANS H. A new quality metric for image fusion[C]//*International Conference on Image Processing*, 2023: 111-173.

[29] XYDEAS C, PETROVIC V. Objective image fusion performance measure[J]. *Electron. Lett.*, 2000, **36**: 308-309.