基于多尺度及多头注意力的红外与可见光图像融合

李秋恒[1,2]，邓 豪[1,2]，刘桂华[1,2]，庞忠祥[3]，唐 雪[1,2]，赵俊琴[4]，卢梦圆[1]
（1. 西南科技大学 信息工程学院，四川 绵阳 621010；2. 特殊环境机器人技术四川省重点实验室，四川 绵阳 621010；
3. 中国电信股份有限公司成都分公司，四川 成都 610066；
4. 中国空气动力研究与发展中心 空天技术研究所，四川 绵阳 621006）

摘要：针对红外与可见光图像融合容易出现细节丢失，且现有的融合策略难以平衡视觉细节特征和红外目标特征等问题，提出一种基于多尺度特征融合与高效多头自注意力相结合的红外与可见光图像融合方法。首先，为提高目标与场景的描述能力，采用了多尺度编码网络提取源图像不同尺度的特征；其次，提出了基于 Transformer 的多头转置注意力结合残差密集块的融合策略以平衡融合细节与整体结构；最后，将多尺度特征融合图输入基于巢式连接的解码网络，重建具有显著红外目标和丰富细节信息的融合图像。基于 TNO 与 M^3FD 公开数据集与 7 种经典融合方法进行实验，结果表明，本文方法在视觉效果与量化评价指标上表现更佳，生成的融合图像在目标检测任务上取得更好的效果。

关键词：图像融合；红外与可见光图像；多尺度特征；多头自注意力；Transformer
中图分类号：TP391 **文献标识码**：A **文章编号**：1001-8891(2024)07-0765-10

Infrared and Visible Images Fusion Method Based on Multi-Scale Features and Multi-head Attention

LI Qiuheng[1,2]，DENG Hao[1,2]，LIU Guihua[1,2]，PANG Zhongxiang[1,2]，TANG Xue[1,2]，
ZHAO Junqin[4]，LU Mengyuan[1]
(1. *School of Information Engineering, Southwest University of Science and Technology, Mianyang* 621010, *China*;
2. *Sichuan Key Laboratory of Special Environmental Robotics, Southwest University of Science and Technology, Mianyang* 621010, *China*; 3. *China Telecom Corporation, Chengdu Branch, Chengdu* 610066, *China*; 4. *Institute of Aerospace Technology, China Aerodynamics Research and Development Center, Mianyang* 621006, *China*)

Abstract: To address the challenges of detail loss and the imbalance between visual detail features and infrared (IR) target features in fused infrared and visible images, this study proposes a fusion method combining multiscale feature fusion and efficient multi-head self-attention (EMSA). The method includes several key steps. 1) Multiscale coding network: It utilizes a multiscale coding network to extract multilevel features, enhancing the descriptive capability of the scene. 2) Fusion strategy: It combines transformer-based EMSA with dense residual blocks to address the imbalance between local details and overall structure in the fusion process. 3) Nested-connection based decoding network: It takes the multilevel fusion map and feeds it into a nested-connection based decoding network to reconstruct the fused result, emphasizing prominent IR targets and rich scene details. Extensive experiments on the TNO and M^3FD public datasets demonstrate the efficacy of the proposed method. It achieves superior results in both quantitative metrics and visual comparisons. Specifically, the proposed method excels in targeted detection tasks, demonstrating state-of-the-art performance. This approach not only enhances the fusion quality by effectively preserving detailed information and balancing visual and IR features but also establishes a benchmark in the field of infrared and visible image fusion.

Keywords: image fusion, visible and infrared images, multi-scale features, multi-head self-attention, transformer

收稿日期：2023-08-24；**修订日期**：2023-09-20.
作者简介：李秋恒（2002-），女，硕士研究生，研究方向为图像处理、深度学习，E-mail：1050920982@qq.com。
通信作者：刘桂华（1972-），女，教授，研究方向为计算机视觉、图像处理和传感器融合技术，E-mail：liughua_Swit@163.com。
基金项目：装备预先研究共用技术项目（50927010302）。

0 引言

视觉传感器通过捕捉环境中的光信号来获取具有丰富视觉信息的图像，但不同类型的传感器在感知能力上存在差异，其中红外传感器通过捕捉物体的热辐射，提供了对于热量分布和热效应的非接触式检测和分析手段。可见光传感器具有反映场景细节和纹理信息的优势，但容易受到极端环境影响。因此融合技术成为必要选择，它可以结合二者的互补性优势，从而得到一幅目标明亮、背景丰富的融合图像。目前红外与可见光融合技术在电力巡检、医学与军事等多个领域具有广泛的应用前景[1]。

传统的图像融合算法发展成熟，基于多尺度变换的图像融合方法主要包括拉普拉斯金字塔变换（Laplacian Pyramid，LP）、小波变换（Wavelet Transform，WT）、多尺度几何分解 3 种方法。金字塔变换中的冗余和无方向性分解可能会导致对图像的描述不准确，而 WT 可以很好地解决该问题并具有多方向性，因此受到了广泛关注和研究。例如，Kumar 等人[2]（2013）提出了基于离散余弦谐波小波变换（Disrcret Cosine Harmonic Wavelet Transform, DCHWT）的方法，该方法虽然可以增强稀疏性表达，但采样过程中的数据冗余会导致融合图像信息丢失、轮廓模糊等问题。之后，Kumar 等人[3]（2015）提出交叉双边滤波器（Cross Bilateral Filter，CBF），其在量化评价指标方面得到了较好的表现，但在融合图像中会出现伪影及细节信息丢失问题，而且计算时间相对较长。Li 等人[4]（2016）通过多个低层特征来设计活动度量提出了一种有效的图像融合方案，融合结果的成像质量和客观评价效果显著。因此，不能仅仅依赖于单一特征，而需要设计更全面的特征提取与细节描述的方法，更完整地描述图像属性。

深度学习具有很强的特征提取和数据表示能力，在图像融合领域得到了飞速发展。其方法大致分为 3 类：卷积神经网络方法、生成对抗网络方法以及自编码解码网络方法。基于 CNN 的端到端图像融合框架因其网络容量有限和训练优化的限制可能导致图像重建中的失真和细节损失，影响融合结果的质量和清晰度。Ma 等人[5]（2017）提出了基于生成对抗网络的 FusionGAN，该网络虽然设计了内容损失和对抗损失来约束网络，但难以平衡不同特征的贡献。基于自编码器（Auto Encoder，AE）[6]的图像融合算法可在大型数据集上进行预训练，从而获得良好的特征提取能力，因此其在图像融合领域得到广泛研究。Li 等人[7]（2018）将密集块（Dense block）[8]融入编码器，提出了一种新的融合框架 DenseFuse，该方法通过手工设计融合策略，且其只关注单一尺度的特征融合，因此融合图像在细节信息与整体结构方面的表现不佳。Li 等人[9]（2019）提出了一种基于残差架构的残差融合网络（An end-to-end residual fusion network for infrared and visible images，RNF_Nest），该网络在自编码器中引入多尺度结构进行特征提取。融合图像在可见光细节信息上获得较好的表现，但红外图像的信息丢失，难以突出红外显著目标。Vibashan V. S.等人[10]（2021）提出了一种基于 Transformer 的图像融合方法（Image Fusion Transformer，IFT），其融合图像保留了丰富的细节信息，但是难以突出红外显著目标，人眼视觉感知效果不足。

因此，本文提出了一种新的图像融合模型。首先，构建了一个多尺度编解码网络，编码器采用多次下采样，实现图像多尺度的特征提取。解码器通过多尺度密集网络连接，对融合特征进行最大程度的重建，防止细节信息丢失。其次，通过引入多头注意力与密集卷积块，设计了一个有效的双分支融合策略，对局部细节信息以及全局依赖进行特征加强。最后，通过实验表明，本文方法比其他有代表性的对比方法在视觉效果与量化评价指标上均有所提高。

1 本文算法

本文将红外与可见光图像作为源图像分别输入到双编码结构中，从源图像中提取多尺度的深度特征。融合层采用了基于 Transformer 的多头转置注意力与残差密集块相结合的双分支结构，将每个尺度上提取到的多模态浅层与深层特征进行融合。最后使用基于巢式连接[11]的解码网络对融合特征进行更全面的学习，解码得到具有突出红外目标和丰富细节信息的融合图像。

1.1 多尺度特征融合框架

图像融合方法中通常直接使用训练好的 VGG（Visual Geometry Group）或 ResNet 等深度卷积网络进行特征提取[12]。这些网络使用多层卷积提取出高级语义特征。但仅使用最后一层的深度特征进行图像融合可能会导致信息丢失，图像融合效果不佳。因此，本文基于特征金字塔结构和巢式连接，构建的多尺度特征融合结构框图如图 1 所示，该结构主要包括编码网络、融合层和解码网络 3 个部分。

首先，将红外与可见光源图像输入到编码网络中，其中 1×1 卷积实现特征维度的转换，每个编码卷积模块（Encoding Convlusion Block，ECB）使用一个 3×3 与 1×1 的卷积进行特征提取，并使用最大池

化的方式对源图像进行 3 次下采样，逐步缩小图像分辨率的同时扩充通道数，从而提取多尺度的深度特征。随后，将提取到的多个尺度的红外与可见光特征图输入双分支融合层，得到增强后的多尺度特征融合图。最后，利用解码网络将不同尺度的融合特征图进行相应倍数的上采样，然后与相同尺度的融合特征图进行连接，使用可促进多层次特征交互和信息流动的巢式连接网络来重建红外图像热辐射目标和可见光图像细节纹理。图 1 中 Conv1 表示 1×1 卷积，ECB1～ECB4 表示 4 个使用最大池化的下采样层组成的编码网络，TFS（Transformer Fusion Layers）代表本文提出的双分支融合策略，DCB31～DCB11 表示由上采样层组成的解码网络。其中编码网络和解码网络的设置如表 1 所示。其中 Ch_i 与 Ch_o 分别代表输入输出通道数。

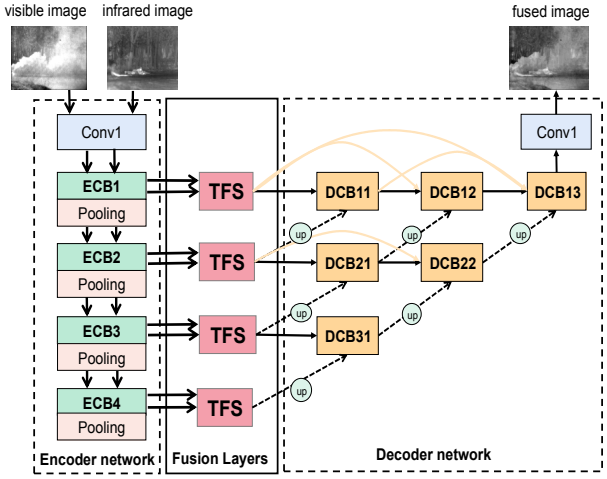


图 1 多尺度特征融合结构框图

Fig.1 Block diagram of multi-scale feature fusion structure

表 1 编码网络(E)和解码网络(D)的设置

Table 1 Settings of encoding network (E) and decoding network (D)

	Layer	Size	Stride	Ch_i	Ch_o
E	ECB1	-	-	16	64
	ECB2	-	-	64	112
	ECB3	-	-	112	160
	ECB4	-	-	160	208
D	DCB31	-	-	368	160
	DCB22	-	-	384	112
	DCB21	-	-	272	112
	DCB13	-	-	304	64
	DCB12	-	-	240	64
	DCB11	-	-	176	64
ECB	Conv	3	1	N_{in}	16
	Conv	1	1	16	N_{out}
DCB	Conv	3	1	N_{in}	16
	Conv	1	1	16	N_{out}

1.2 局部-全局双分支融合策略

传统的融合网络难以在关注局部特征的同时平衡全局建模的重要性。因此，本文提出了一种基于 Transformer 的多头转置注意力结合密集卷积块的双分支特征融合层。融合层的结构框图如图 2 所示，其中“c”表示拼接，“＋”表示元素相加。首先，将编码后的红外与可见光特征图分别输入到融合层中，其中全局分支提出了高效的视觉 Transformer 对长距离依赖关系进行建模，以学习全局语境特征。局部分支提出了残差密集块来捕获空间信息，加强局部特征的学习。通过双分支策略的特征学习之后将不同模态的特征图进行拼接，并使用卷积与激活函数进一步加工和提取拼接后的特征，增强特征表达能力。最后，将增强后的特征进行相加，得到包含增强的局部和全局上下文信息的融合特征图。这种融合方式不仅可以提高融合结果的一致性、语义理解能力，也能够适应不同图像的特征分布差异，提升图像融合的质量和视觉效果。

1.2.1 长距离依赖捕获

基于 Zamir 等人[13]（2021）提出的多头转置注意（multi-dconv head transposed attention, MDTA）模块。设计了一个高效的多头自注意力机制（Efficient Multi-Head Self-Attention，EMSA），其网络结构如图 3 所示。与传统 Transformer 中的多头自注意力模块相比，EMSA 使用了深度卷积压缩内存，且该模块在通道维度上进行操作，因此可以显著减小计算量。其具体过程如图 3 所示，其中 R 表示 reshape，T 表示转置，“＋”“×”分别表示元素相加与相乘。首先，将输入的 Token 尺寸为 $\boldsymbol{X}\in\boldsymbol{R}^{C\times H\times W}$ 的特征图通过深度卷积和层归一化的预处理，为多头注意力提供更丰富的输入特征和更稳定的训练环境。其次，通过线性变换得到 query($\boldsymbol{Q}$)、key($\boldsymbol{K}$)和 value($\boldsymbol{V}$)，并使用 reshape 操作后得到 $\boldsymbol{K}$、$\boldsymbol{Q}$、$\boldsymbol{V}\in\boldsymbol{R}^{M\times C\times HW}$，其中 M 表示注意力头个数，本文中 4 个不同尺度的 EMSA 中自注意头个数依次设置为 1，2，4，8。之后，通过将矩阵 $\boldsymbol{K}$ 转置后与矩阵 $\boldsymbol{Q}$ 进行矩阵相乘，可以生成一个维度为 $\boldsymbol{R}^{C\times C}$ 的转置特征图 $\boldsymbol{A}$。最后，将 $\boldsymbol{A}$ 经过 softmax 激活函数后与 $\boldsymbol{V}$ 相乘，通过 reshape 和线性层后与原始输入特征图 X 进行残差连接，得到 EMSA 的输出 X'。

采用 1×1 卷积替代 Transformer 块中的全连接层可以有效防止空间结构被破坏，且减少计算量的同时可以保持较好的性能。因此，本文基于该思想采用了一种高效通道注意力（Efficient Channel Attention，ECA）[14]模块，如图 4 所示。该模块结合 EMSA 组成视觉 Transformer。EMSA 模块提供了全局的上下文信

息，使得模型能够更好地理解特征中的依赖关系。而 ECA 模块通过自适应地调整特征通道之间的关系，帮助模型更好地理解特征之间的重要性和相互作用。在本文方法中，ECA 模块直接在平均池化之后使用 1×1 卷积层取代了传统的全连接层，这样可以避免维度的缩减，并通过一维卷积来实现跨通道间的信息交互。卷积核的大小可以通过一个函数自适应地调整，这种方式只需要很少的参数就能有效地捕捉跨通道的交互关系。且该方法可进一步强化由 EMSA 获得的全局特征。

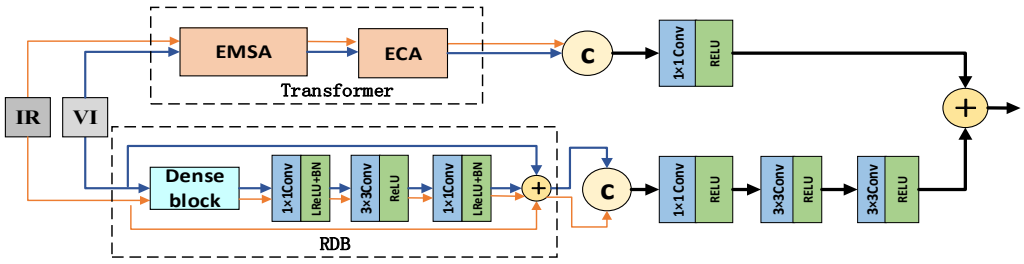


图 2 TFS 融合层结构框图

Fig.2 Block diagram of TFS fusion layer structure

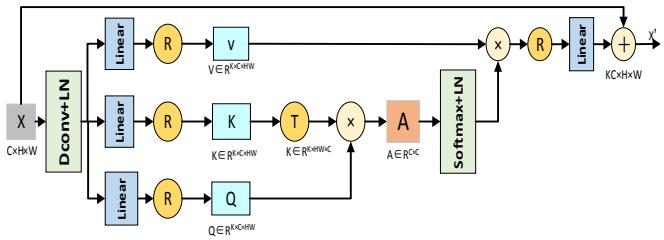


图 3 EMSA 结构框图

Fig.3 Block diagram of EMSA structure

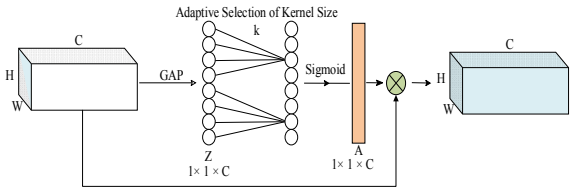


图 4 ECA 结构框图

Fig.4 Block diagram of ECA structure

1.2.2 局部细节纹理保留

针对现有融合方法容易出现细节丢失等问题，本文基于 DenseNet[8]密集卷积块设计的残差密集块（RDB）如图 2 所示。RDB 模块中的 Dense block 结构如图 5 所示。首先，密集卷积块内的卷积层可以直接访问前面所有层的输出，这种密集连接的方式可以促进信息在网络中的流动，有助于信息的传递和重用，提高模型的表示能力；其次，为充分提取源图像细节信息，通过引入残差学习提高特征学习能力；最后，密集连接块之后使用两个 1×1 的卷积实现渐进式的通道缩减，可以减少信息损失和特征混淆的风险。

根据图 5 中 Dense block 的结构可知第 q 层输出为：$X_q=F_q(\mathrm{cat}(X_0, X_1, X_2, \cdots, X_{q-1}))$，其中，$F_q$ 使用了一个 3×3 卷积、LeakRelu 激活函数与 Batch Norm 实现非线性变换。$\mathrm{cat}(X_0, X_1, X_2, \cdots, X_{q\text{-}1})$表示将之前所有层的输出特征图进行拼接。

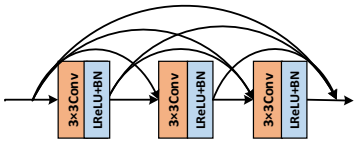


图 5 Dense block 结构框图

Fig.5 Block diagram of dense block structure

1.3 损失函数

本文采用了 3 种损失函数来训练特征融合网络，分别为特征相似性损失函数 L_{feat}、最大梯度损失函数 L_{grad} 以及结构相似度损失函数 L_{ssim}。总损失函数 L_{loss} 表达为：

$$L_{\mathrm{loss}}=L_{\mathrm{feat}}+\lambda_1 L_{\mathrm{grad}}+\lambda_2 L_{\mathrm{ssim}} \tag{1}$$

式中：λ_1，λ_2 为超参数，用来控制损失之间的比例大小。

L_{ssim} 计算融合图像和源图像之间的结构相似性，其表达式为：

$$L_{\mathrm{ssim}}=(1-f_{\mathrm{ssim}}(I_{\mathrm{if}}, I_{\mathrm{iv}}))+(1-f_{\mathrm{ssim}}(I_{\mathrm{if}}, I_{\mathrm{ir}})) \tag{2}$$

式中：I_{if} 为融合图像；I_{ir} 为红外图像；I_{iv} 为可见光图像。

L_{feat} 通过限制融合后的深度特征以保留显著结构，其表达式为：

$$L_{\mathrm{feat}}=\sum_{m=1}^{M}\omega_1(m)\,\|\Phi_{\mathrm{f}}^m-(\omega_{\mathrm{vi}}\Phi_{\mathrm{vi}}^m+\omega_{\mathrm{ir}}\Phi_{\mathrm{ir}}^m)\|_{\mathrm{F}}^2 \tag{3}$$

式中：Φ_{f}^m 表示融合特征图；Φ_{ir}^m，Φ_{vi}^m 分别表示红外与可见光的特征图。M 为多尺度深度特征的个数。

L_{grad} 函数可以计算重建图像和输入图像之间的梯度损失，其表达式为：

$$L_{\text{grad}}=\frac{1}{HW}\left\|\nabla I_{\text{f}}\right|-\max(\left|\nabla I_{\text{ir}}\right|,\left|\nabla I_{\text{vi}}\right|)\Big\|_1 \tag{4}$$

式中：∇代表 Sobel 边缘算子。

2 实验与分析

2.1 实验设置

本算法硬件平台为 CPU（Intel Xeon E5-2620）和 GPU（NVIDIA TITAN XP*2 12G），操作系统为 Ubuntu18.04，使用 Pytorch1.12.0 框架构建模型，CUDA 版本为 11.3，所有实验均在相同实验环境中进行训练、验证和测试。使用 Microsoft COCO[15]数据集作为训练集用于训练编解码网络，从中选择 80000 张图片用于训练，输入图像尺寸为 256×256。针对融合网络，选择了 KAIST dataset[16]数据集中的 20000 对图像进行训练，初始学习率为 1×10^{-4}，batch_size＝4，epoch＝2。

为测试所提方法的融合效果，本文选择 TNO[17]数据集中的 40 对图像和 M^{3}FD[18]公开数据集中的 20 对图像进行融合实验，并使用 M^{3}FD[18]数据集的融合结果图进行目标检测任务来进一步验证本文所提融合方法的有效性。本文选择了多种有代表性的融合方法进行对比，这些方法分别是 CBF[3]、DCHWT[2]、Densefuse[7]、RFN-Nest[8]、IFT[10]、FusionGAN[3]和 U2fusion[19]。7 种对比方法都是公开可用的，训练的数据集与本文方法相同，分别从视觉效果、量化对比、检测效果以及消融实验等 4 个方面对融合结果进行分析。

2.2 评价指标

熵（EN）用于衡量融合图像包含的信息量。EN 越大，表明融合图像所包含的信息量越多。其定义为：

$$\text{EN}=-\sum_{L=0}^{L-1}p_l\log_2 p_l \tag{5}$$

式中：L 表示图像的灰度级数；p_l 表示融合图像中相应灰度的归一化直方图。

标准差（SD）反映融合图像的单个像素值与平均值的差异性。SD 越高代表融合结果具有更好的对比度。其定义为：

$$\text{SD}=\sqrt{\sum_{i=1}^{H}\sum_{j=1}^{W}(F(i,j)-\mu)^2} \tag{6}$$

式中：$F(i,j)$表示融合图像 F 在(i,j)处的像素值；μ表示融合图像的均值。

互信息（MI）用于度量两幅图像之间的相似程度。当融合图像保留了更多源图像的信息量时，互信息值越大。其定义为：

$$\begin{aligned}\text{MI}=0.5\times&(\sum_{i,f}p_{\text{I,F}}(i,f)\log\frac{p_{\text{I,F}}(i,f)}{p_{\text{I}}(i)p_{\text{F}}(f)}+\\&\sum_{v,f}p_{\text{V,F}}(v,f)\log\frac{p_{\text{V,F}}(v,f)}{p_{\text{V}}(v)p_{\text{F}}(f)})\end{aligned} \tag{7}$$

式中：$p_{\text{V}}(v),p_{\text{I}}(i)$和 $p_{\text{F}}(f)$分别代表可见光图像、红外图像和融合图像的边缘直方图；$p_{\text{I,F}}(i,f)$和 $p_{\text{V,F}}(v,f)$分别表示红外图像、可见光图像与融合图像的联合直方图。

差异相关性总和（sum of correlation differences, SCD）通过计算源图像及其对融合图像的影响来表征图像质量。SCD 越高，意味着融合图像包含源图像中的信息越丰富。其定义为：

$$D_1=F-S_1,\ D_2=F-S_2 \tag{8}$$

$$\text{SCD}=r(D_1,S_1)+r(D_2,S_2) \tag{9}$$

式中：D_1、D_2 分别表示融合图像 F 与输入源图像 S_1、S_2 的差分图像。$r(\cdot)$函数计算 S_1 和 D_1、S_2 和 D_2 之间的相关性，其表达式为：

$$r(D_K,S_K)=\frac{\sum_i\sum_j(D_K(i,j)-\overline{D}_K)(S_K(i,j)-\overline{S}_K)}{\sqrt{(\sum_i\sum_j(D_K(i,j)-\overline{D}_K)^2)(\sum_i\sum_j(S_K(i,j)-\overline{S}_K)^2)}} \tag{10}$$

式中：$K=1,2$，$\overline{D}_K$ 与 $\overline{S}_K$ 表示 D_K 与 S_K 像素值的平均值。

多尺度结构相似性度量（multi-scale structural similarity index measure, MS-SSIM）能更好地与人眼视觉系统的视觉感知相一致，并且在一定的尺度下，评价效果优于 SSIM。其定义为：

$$\begin{aligned}\text{MS-SSIM}(x,f)=&[l_M(x,f)]^{\alpha_{M'}}\times\\&\prod_{j=1}^{M'}[c_j(x,f)]^{\beta_j}\times\left[s_j(x,f)\right]^{\gamma_j}\end{aligned} \tag{11}$$

式中：$l_M(x,f)$表示在第 M'个尺度上的亮度相似度，$c_j(x,f)$和 $s_j(x,f)$分别表示在第 j 个尺度上的对比度和结构相似度。α、β、γ用于平衡上述 3 个分量的参数。设置 $\alpha_{M'}=\beta_j=\gamma_j$，$\sum_{j=1}^{M'}\gamma_j=1$。

VIF（Visual Information Fidelity）是一种用于评估融合图像信息保真度的指标。它通过对融合图像和源图像进行分块，并比较图像块之间的视觉信息，来衡量融合图像的整体质量。VIF 值越大，表示融合图像与原始图像之间的信息保持得越好。

2.3 视觉效果

对比实验结果如图 6 所示，其中前 4 列来自 TNO 数据集，后 4 列来自 M^{3}FD 数据集。(a)、(b)为用于测试的红外与可见光图像对。首先，本文方法成功地展现红外显著信息与可见光纹理信息之间的互补效果。

如第 1、4 列的图像中，本文方法能够清晰地显示可见光图像中的建筑物与灌木丛等物体的细节信息，同时有效融合了红外热辐射目标。而对比方法中的一些方法如 FusionGAN，虽然能有效突出红外目标，但背景模糊，整体表现更偏向于红外源图像。CBF 算法的融合效果不佳，存在大量噪声与伪影。DCHWT 和 DenseFuse、RFN_Nest 等算法同样存在轮廓模糊，细节不清晰等问题。此外，第 2、3 列的融合图像结果显示，本文方法在保持整体对比度方面也具有一定优势。对比方法如 DensFuse、RFN_Nest、U2Fusion、IFT 等算法融合结果对比度低，人眼视觉难以锁定目标。而本文方法不仅能够突出红外显著目标，实现保留图像的整体对比度的同时能够更好地保留细节信息。同理，由后 4 列融合图像中的人物和车辆等目标可以发现，本文方法在 M^3FD 数据集上同样可以有效实现红外显著信息的表达。如图中红外目标突出且轮廓清晰。同时，从融合结果中的建筑物、树叶和车辆等的融合效果可以证明本文方法在保留细节纹理方面同样具有优势。

综上所述，根据图 6 中在 TNO 数据集和 M^3FD 数据集上的视觉结果分析，可以得出结论：本文方法的融合图像在视觉效果上与对比算法相比表现最佳，能够有效地实现红外显著目标与可见光细节纹理上的互补融合，有助于人眼视觉感知与在高级视觉任务上的表现，且该算法避免了融合图像中红外目标不显著、边缘和背景模糊等缺陷。

(a) Infrared image

(b) Visible image

(c) CBF

(d) DCHWT

(e) DenseFuse

(f) U2Fusion

(g) FusionGAN

(h) IFT

(i)RFN_Nest

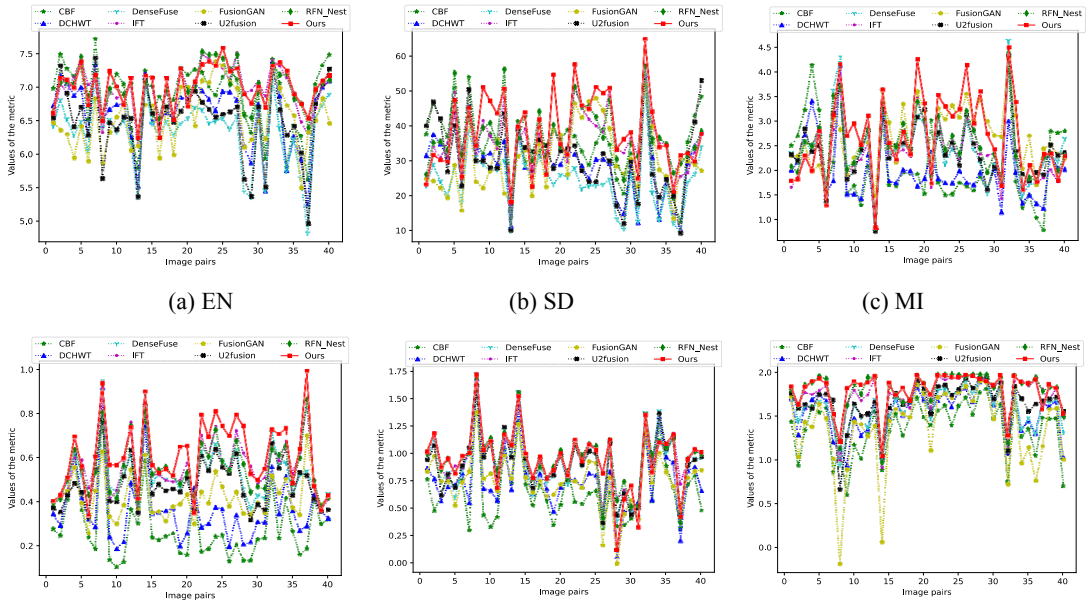
(j)Ours

图 6 对比实验的融合结果

Fig.6 Fusion results of comparison experiments

2.4 量化对比

在 TNO 与 M^{3}FD 数据集中的 2 组图像上的指标对比结果如图 7 和图 8 所示。表 2 列出了这两组图像在 6 个评价指标上的均值，其中 average 代表所有方法的指标平均值。对于 TNO 数据集，本文方法在 EN、SD、MI、SCD 和 VIF 指标上取得了最优结果。通过分析各个指标可以得出以下结论：首先，本文获得最佳的 EN、MI 和 SCD 值表明融合图像能够很好地保留红外图像和可见光图像中的信息，这也是本文引入多尺度特征融合和视觉 Transformer 的意义所在。此外，本文方法在 SD、VIF 指标上也获得最高值，表明融合图像有较高的对比度与视觉保真度。MS_SSIM 考虑了不同尺度下的结构信息，分析表 2 可得，RFN_Nest、IFT 以及本文算法都获得了不错的效果，表明了基于多尺度的编解码网络对于图像中的细节和纹理具有更好的感知能力。对于 M^{3}FD 数据集，本文方法在 MI、SCD、VIF、MS_SSIM 等指标上仍然取得了最佳结果。EN 相较于 average 提高了 0.243，SD 相较于 average 提高了 4.765。总体量化评价结果与在 TNO 数据集上的表现大致相同，本文方法在 6 个量化指标上均大于所有方法的指标平均值。综上，根据表 2 中的量化比较可以进一步说明，本文方法在 TNO 与 M^{3}FD 公开数据集上的实验评估中取得了具有竞争性的效果，这充分证明了本文方法的有效性。

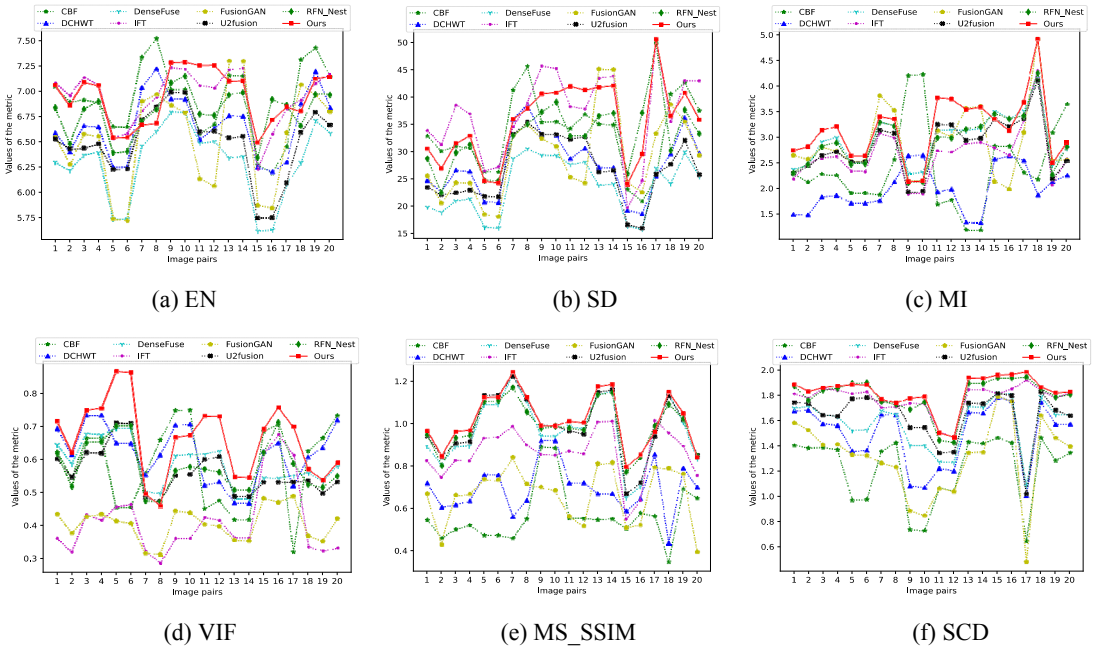


(a) EN (b) SD (c) MI

(d) VIF (e) MS_SSIM (f) SCD

图 7 不同融合方法在 TNO 数据集中 40 对红外与可见光图像的指标比较

Fig.7 Comparison of metrics between 40 pairs of infrared and visible images in TNO dataset with different fusion methods



(a) EN (b) SD (c) MI

(d) VIF (e) MS_SSIM (f) SCD

图 8 不同融合方法在 M³FD 数据集中 20 对红外与可见光图像的指标比较

Fig.8 Comparison of metrics between 20 pairs of infrared and visible images in M3FD dataset with different fusion methods

表 2 不同融合方法在 TNO 与 M³FD 数据集上各指标均值

Table 2 Mean values of indicators on TNO and M3FD datasets with different fusion methods

Dataset	Methods	EN	SD	MI	SCD	MS-SSIM	VIF
TNO	CBF	6.890	34.010	2.115	1.326	0.665	0.285
	DCHWT	6.626	29.402	1.993	1.542	0.759	0.369
	FusionGAN	6.548	30.699	2.593	1.382	0.755	0.425
	DenseFuse	6.347	24.707	2.423	1.595	0.918	0.529
	U2Fusion	6.511	31.186	2.411	1.654	0.923	0.490
	RFN_Nest	6.997	37.42	2.484	1.799	0.967	0.555
	IFT	6.981	36.301	2.357	1.745	0.962	0.566
	Ours	7.015	38.559	2.683	1.805	0.957	0.614
	Average	6.739	32.785	2.382	1.606	0.863	0.479
M³FD	CBF	6.920	33.339	2.427	1.217	0.566	0.583
	DCHWT	6.668	27.781	2.003	1.493	0.699	0.621
	FusionGAN	6.551	29.344	2.909	1.307	0.668	0.406
	DenseFuse	6.307	23.722	2.969	1.582	0.975	0.587
	U2Fusion	6.496	26.606	2.834	1.643	0.989	0.562
	RFN_Nest	6.795	32.995	2.919	1.799	0.993	0.581
	IFT	6.950	36.943	2.677	1.776	0.866	0.413
	Ours	6.947	35.550	3.181	1.829	1.020	0.665
	Average	6.704	30.785	2.739	1.565	0.847	0.552

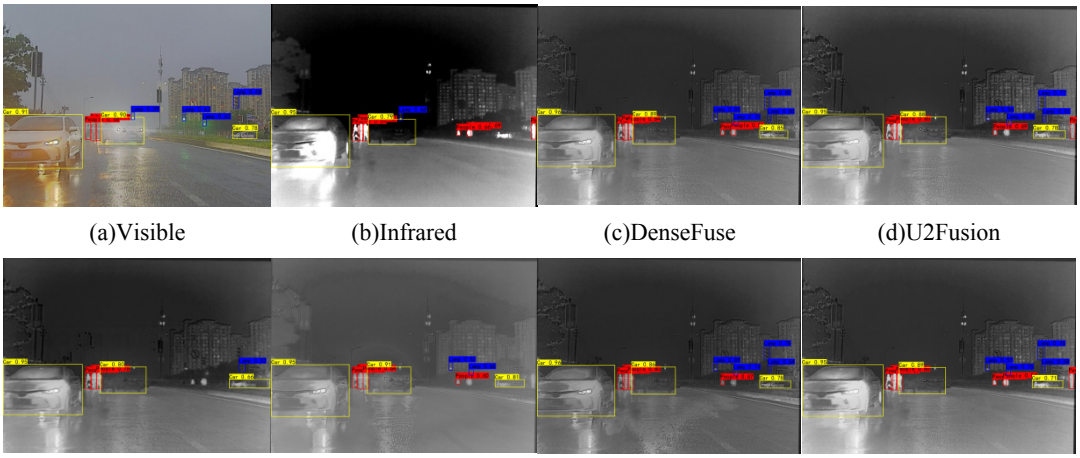
2.5 检测效果

为进一步验证本文融合方法的有效性，选择 YOLO-v7[20]检测算法对上述基于深度学习算法的融合图像进行目标检测。实验采用 M³FD[18]公开数据集进行训练与检测，其图像分辨率为 1024×768。选择 420 对红外与可见光图像融合图像进行目标检测，使用平均精准率（Average Precision，AP）、平均精度均值 mAP（mean Average Precision）作为检测结果的评价指标。其中 AP 度量是由精准率与召回率（Precision-Recall，P-R）刻画曲线的面积，用于衡量目标检测任

务中模型的精确度和召回率之间的平衡。mAP 是多个类别的 AP 的平均值。本文选择了一张有代表性的检测效果图进行展示，从图 9 可知，在本文融合图像上可准确地识别出在雨雾等恶劣环境下行人、车辆，以及路灯等目标物体。不同方法的 AP 和 mAP 结果如表 3 所示。结果显示，融合图像相比红外与可见光图像在提高目标检测性能方面具有潜在的优势。相比 5 种经典融合算法，本文融合图像在目标检测任务上获得了最高的 mAP，与对比算法中检测任务上效果最好的 DenseFuse 相比提高了 0.56。综上，本文所提图像融合方法在目标检测任务上取得了更好的效果，表明本文方法可实现有效的图像融合。

2.6 消融实验

为验证本文所提模块的有效性，对局部信息保留分支的密集卷积块模块与捕获长距离依赖分支的视觉 Transformer 模块进行消融实验，结果如表 4 所示，表中无视觉 Transformer 表示去除长距离依赖捕获分支的融合策略，无 RDB 表示去除局部细节分支的融合策略。消融实验结果表明，本文提出的双分支融合策略可达到最佳效果，除去任何一个分支量化指标都会降低，从而进一步证实了本文所提出的融合策略的有效性。



(a)Visible (b)Infrared (c)DenseFuse (d)U2Fusion

(e)IFT (f)FusionGan (g)RFN_Nest (h)Ours

图 9 目标检测对比结果

Fig.9 Comparison results of target detection

表 3 融合效果目标检测实验结果评价

Table 3 Evaluation of experimental results of fusion effect target detection

Models	AP						mAP
	Bus	People	Car	Truck	Motorcycle	Lamp	
Visible	0.839	0.639	0.877	0.794	0517	0.681	72.45%
Infrared	0.826	0.763	0.835	0.740	0.452	0.351	66.17%
DenseFuse	0.827	0.776	0.898	0.830	0.636	0.569	75.62%
FusionGan	0.831	0.687	0.883	0.763	0.550	0.425	69.02%
RFN_Nest	0.834	0.683	0.895	0.814	0.592	0.650	74.50%
IFT	0.844	0.765	0.891	0.824	0.589	0.580	74.94%
U2Fusion	0.836	0.754	0.900	0.818	0.612	0.587	75.16%
Ours	0.837	0.739	0.889	0.831	0.665	0.607	76.18%

表 4 消融实验结果评价

Table 4 Evaluation of ablation experiment results

Dataset	Methods	EN	SD	MI	SCD	MS_SSIM	VIF
TNO	Exclude Transformer	6.948	38.159	2.675	1.787	0.948	0.606
	Exclude RDB	6.941	38.036	2.705	1.780	0.944	0.608
	Ours	7.015	38.559	2.683	1.805	0.957	0.614
M^3FD	Exclude Transformer	6.745	33.325	3.108	1.783	1.007	0.643
	Exclude RDB	6.74	33.365	3.16	1.774	1.005	0.635
	Ours	6.947	35.550	3.181	1.829	1.020	0.665

3 结语

针对单一的融合策略难以平衡局部细节与整体结构等问题，本文提出一种基于多尺度特征与多头转置注意力模型相结合的红外与可见光图像融合方法。一方面，该方法采用了多尺度编解码网络，用来提取多尺度特征并重建具有丰富信息的融合图像。另一方面，为捕获全局信息设计了视觉 Transformer 模块，用于获取长距离依赖关系，并结合残差密集块得到更加全面的融合特征。选择了 7 种经典的融合算法在公开 TNO 和 M^3FD 数据集上进行图像融合与融合图像目标检测的对比实验。结果显示，生成的融合图像可突出红外显著目标的同时保留可见光纹理信息，并在 6 个量化指标上均取得了较好的效果。此外，本文方法的融合图像在目标检测任务上的 mAP 相比对比算法中效果最好的 DenseFuse 提高了 0.56。综上，本文方法可有效地融合红外与可见光图像。

参考文献：

[1] 王天元, 罗晓清, 张战成. 自注意力引导的红外与可见光图像融合算法[J]. **红外技术**, 2023, **45**(2): 171-177.
WANG T Y, LUO X Q, ZHANG Z C. Self-attention guided fusion algorithm for infrared and visible images[J]. *Infrared Technology*, 2023, **45**(2): 171-177.

[2] KUMAR B K S. Multifocus multispectral image fusion based on pixel significance using disrcret cosin harmonic wavelet transform[J]. *Signal Image&.Video Processing*, 2013, **7**(6): 1125-1143.

[3] KUMAR B K S. Image fusion based on pixel significance using cross-bilateral filter[J]. *Signal Image & Video Processing*, 2015, **9**(5): 1193-1204.

[4] LI H, QIU H, YU Z, et al. Infrared and visible image fusion scheme based on NSCT and low-level visual features[J]. *Infrared Physics & Technology*, 2016, **76**: 174-184.

[5] HOU J L, ZHANG D Z, WEI W, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. *Information Fusion*, 2019, **48**: 11-26.

[6] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural net-works[J]. *Science*, 2006, **313**(5786): 504-507.

[7] LI H, WU X J. DenseFuse: A fusiona pproach to infrared and visible images[J]. *IEEE Transactions on Image Processing*, 2018, **28**(5): 2614-2623.

[8] HUANG G, LIU Z, LAURENSVD M, et al. Densely connected convolutional networks[C]// *IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2261-2269.

[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] Vibashan V S, Valanarasu J, Oza P, et al, et al. Image fusion transformer [J/OL]. arXiv preprint arXiv: 2107.09011. 2021. https://ieeexplore.ieee.org/document/9897280.

[11] 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.

[12] 黄玲琳, 李强, 路锦正, 等. 基于多尺度和注意力模型的红外与可见光图像融合[J]. **红外技术**, 2023, **45**(2): 143-149.
HUANG L L, LI Q, LU J Z, et al. Infrared and visible image fusion based on multi-scale and attention modeling[J]. *Infrared Technology*, 2023, **45**(2): 143-149.

[13] Zamir S W, Arora A, Khan S, et al. Restormer: efficient transformer for high-resolution image restoration[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2022: 5718-5729.

[14] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2020: 11531-11539.

[15] LIN T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]//*Computer Vision-ECCV*, 2014: 740-755.

[16] WANG S H, Park J, Kim N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//*IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2015: 1037-1045.

[17] TOET A. The TNO multi band image data collection[J]. *Data in Brief*, 2017, **15**: 249-251.

[18] LIU J, FAN X, HUANG Z B, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2022: 5792-5801.

[19] XU H, MA J Y, JIANG J J ,et al. U2Fusion: a unified unsupervised image fusion network[J]. *IEEE Trans. Pattern Anal. Mach. Intell.,* 2022, **44**(1): 502-518.

[20] WANG C Y , Bochkovskiy A , LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2023: 7464-7475, DOI: 10.1109/CVPR52729.2023.00721.