Website：ycyk.brit.com.cn

基于CNN-Transformer的自编码器红外和可见光图像融合方法

李 霖，沈永健，张鹏宇，原 昊，王 超
（北京遥测技术研究所 北京 100076）

摘要：基于自编码器的图像融合模型因无需手动设计融合规则而受到更多关注。然而，该融合网络编码器采用的卷积神经网络仅对局部感受野敏感，无法提取图像全局特征，且缺乏从红外图像和可见光图像中提取独特特征的能力。本文构建了一种基于自动编码器的新型图像融合网络，该网络由编码器模块、融合模块和解码器模块组成。在编码器模块中，结合使用CNN和Transformer模块以同时捕捉原图像的局部和全局特征。此外，为提取原图像特定信息，分别为原红外和可见光图像设计对比度增强和梯度增强模块。编码器模块获得的特征图像经融合模块串联后输入解码器模块，从而获得融合图像。在三个数据集上的实验结果表明，本文提出的融合网络能较好地保留了红外图像和可见光图像的清晰目标和细节信息，在主观和客观评价方面均优于其他先进方法。同时，本文提出的网络所获得的融合图像在目标检测中获得了最高的平均精度，证明图像融合有利于下游任务。

关键词：图像融合；卷积神经网络；Transformer；红外图像；可见光图像
中图分类号：TP75；TN957.52 **文献标志码：**A **文章编号：**2095-1000(2024)05-0109-11
DOI：10.12347/j.ycyk.20240326002
引用格式：李霖，沈永健，张鹏宇，等. 基于CNN-Transformer的自编码器红外和可见光图像融合方法[J]. 遥测遥控，2024，45(5)：109–119.

Infrared and Visible Image Fusion Based on Autoencoder Composed of CNN-transformer

LI Lin，SHEN Yongjian，ZHANG Pengyu，YUAN Hao，WANG Chao
（Beijing Research Institute of Telemetry, Beijing 100076, China）

Abstract: Image fusion model based on autoencoder network gets more attention because it does not need to design fusion rules manually. However, most autoencoder-based fusion networks use two-stream CNNs with the same structure as the encoder, which are unable to extract global features due to the local receptive field of convolutional operations and lack the ability to extract unique features from infrared and visible images. A novel autoencoder-based image fusion network which consist of encoder module, fusion module and decoder module is constructed in this paper. In the encoder module, the CNN and Transformer are combined to capture the local and global feature of the source images simultaneously. In addition, novel contrast and gradient enhancement feature extraction blocks are designed respectively for infrared and visible images to maintain the information specific to each source images. The feature images obtained by encoder module are concatenated by the fusion module and input to the decoder module to obtain the fused image. Experimental results on three datasets show that the proposed network can better preserve both the clear target and detailed information of infrared and visible images respectively, and outperforms some state-of-the-art methods in both subjective and objective evaluation. Meanwhile, the fused image obtained by the proposed network can acquire the highest mean average precision in the target detection which proves that image fusion is beneficial for downstream tasks.

Keywords: Image fusion; Convolutional neural network; Transformer; Infrared image; Visible image
Citation: LI Lin, SHEN Yongjian, ZHANG Pengyu, et al. Infrared and Visible Image Fusion Based on Autoencoder Composed of CNN-transformer[J]. Journal of Telemetry, Tracking and Command, 2024, 45(5): 109–119.

收稿日期：2024-03-26；修回日期：2024-07-12

0　引言

图像融合技术可将来自不同场景的图像合并成具有多个原图像特征的融合图像。结合利用多模态数据可进一步增强特征表征，是提高图像分割和目标检测等任务性能的有效手段之一。具有互补性的图像有很多，如红外图像和可见光图像、红外图像和合成孔径雷达图像、可见光图像和合成孔径雷达图像、医学中的CT和MRI图像等。其中，红外图像有利于提高目标探测和识别能力，可以避免烟雾、光线、雨水等环境的影响，同时也存在一些不足，如像素分辨率低、对比度差、背景纹理不足等[1]。可见光图像分辨率高，能反映出丰富纹理、细节等场景信息，但是它容易受到天气、烟雾、遮挡等环境因素的影响[1]，导致目标不能被突出。因此，融合技术成为结合二者互补特征的必要选择，从而获得目标明亮、背景细致的融合图像。实验结果表明：与单个模态相比，该融合图像的人脸识别性能有了显著提高，如图1所示。目前，红外与可见光图像融合技术被广泛应用于图像增强、农业自动化、遥感探测、目标识别、检测和跟踪等领域[2]。

（a）红外图像　（b）可见光图像　（c）融合图像
（a）Infrared image　（b）Visible image　（c）Fused image

图1　红外与可见光图像融合示意图

Fig. 1　Diagram of infrared and visible image fusion

过去有许多不同的图像融合方法，包括多尺度变换、稀疏表示、神经网络、子空间、显著性、混合模型以及其他融合方法。尽管传统的融合方法已经取得了一些进展，但这些方法依赖于人工设计活动水平测量和融合规则策略，从而使得融合过程越来越复杂，且经过这些方法提取的特征缺乏多样性，往往导致融合后的图像对比度低、纹理模糊、目标出现伪影等[2]。

基于深度学习的图像融合模型通过自适应训练，借助网络的学习能力更新模型参数，从而形成端到端的融合模型。与传统方法相比，深度学习融合方法避免了活动水平测量和融合规则设计，大大降低了人为因素对融合结果的影响。此外，深度学习方法借助深度学习网络强大的特征提取能力，在融合图像中充分保留了原图像丰富内容信息，提高了融合图像的质量。

目前，基于深度神经网络的图像融合方法可分为基于CNN(卷积神经网络)的方法、基于Auto-encoder(自动编码器)的方法、基于GAN(生成对抗网络)的方法和基于Transformer的方法。基于自动编码器的图像融合网络无需人工设计融合规则，已成为当今广泛研究的一种方法。然而，现有的自动编码器大多使用卷积运算，由于局部感受野的特性，无法完全提取全局特征。此外，目前融合模型的特征提取子网络没有对不同的原图像进行区分，因此融合图像中的互补信息体现不足。为解决上述问题，本文提出了一种基于CNN-Transformer的红外与可见光图像的融合网络。本文方法的特点如下：

① 建立了一种由CNN和Transformer组合的新型编码器，同时提取红外图像和可见光图像的局部和全局信息；

② 对比度增强模块和梯度残差模块分别提取红外图像和可见光图像的独特信息，以在融合图像中保持原图像的互补信息；

③ 在三个数据集(TNO、OTCBVS 和 RoadScene)上进行的广泛实验表明，本文提出的方法能够获得既包含清晰目标又包含丰富纹理的融合图像，且优于近几年的一些先进方法，包括 U2Fusion、DDcGAN、DenseFuse、RFN-Nest、STDFusion 和 SwinFusion。

1　本文方法

本文提出的融合模型包括特征提取、特征融

合和特征重组。特征提取过程分为两个通道，每个通道结合使用CNN特征增强模块和Transformer模块提取原图像特征。

特征融合模块对提取的特征进行堆叠，并输入由四个卷积层组成的解码器，最终还原生成包含红外和可见光图像特征的融合图像。本文提出的融合模型的整体框架如图2所示。

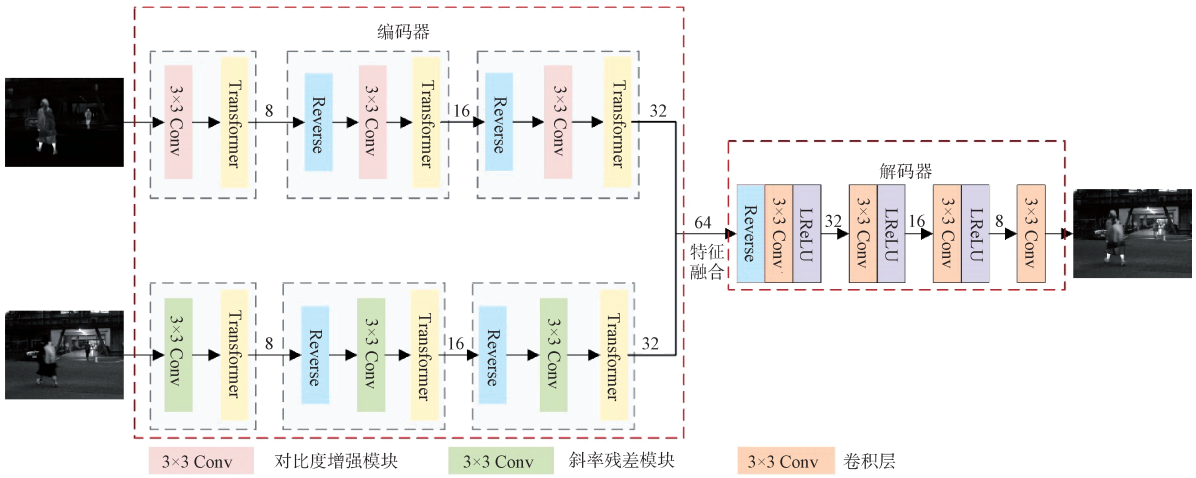

图2　红外与可见光图像融合网络模型

Fig. 2　Network model of infrared and visible image fusion

1.1　编码器

① 对比度增强模块

对比度增强模块设计目的是增强红外图像的对比度，如图3所示，使用步长分别为2、4和8的池化层对特征图像进行最大池化(最大池化操作可以保留红外图像中较大的像素信息，并过滤掉不重要的信息)，并进行卷积操作，将这些特征图像线性插值到输入特征图像大小，并与输入特征图像相加，使用3×3卷积核避免叠加像素中出现伪影，最终得到对比度增强的特征图像。

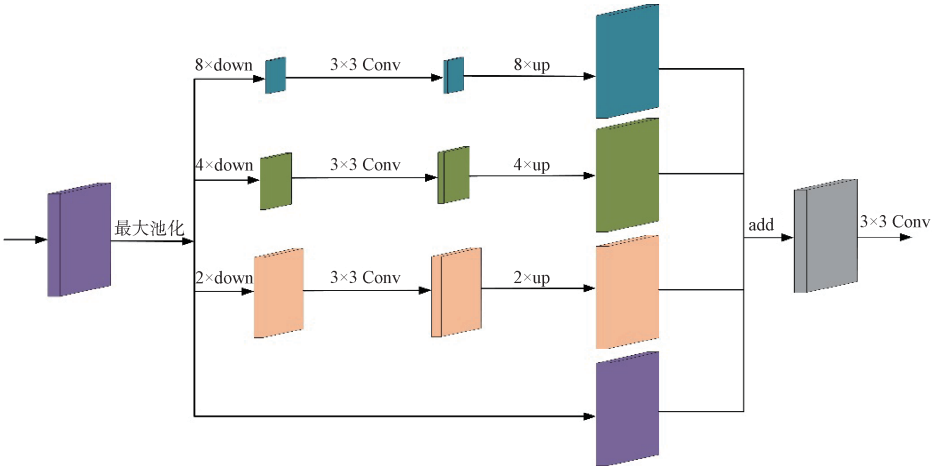

图3　对比度增强模块

Fig. 3　Contrast enhanced block

② 斜率残差模块

梯度残差模块旨在增强可见图像的细节，该块采用残差连接模式，其结构如图4所示。Sobel滤波器可以提取可见图像或特征的梯度信息，主通道特征提取器采用大小为3×3的卷积核，激活函数为 Leaky ReLU。使用大小为1×1的卷积核作为辅助通道特征提取器，激活函数为Leaky ReLU。

③ Transformer模块

本文设计的Transformer模块与ViT模型相似，但在输入和注意力计算方法上有所不同。首先，

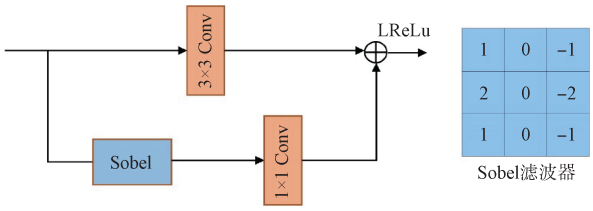

图 4 斜率残差模块

Fig. 4 Gradient residual block

二维源图像被拉伸为一维矩阵，并将批次数和每批图像数等信息一起输入Transformer模型。此外，受SwinFusion的启发，特征图像被划分为多个小窗口，并针对这些窗口计算全局注意力(基于窗口的多头自注意力机制W-MSA)，以克服传统模型的计算复杂性。滑动窗口的大小设置为8，与CNN相比，它具有更大的感知域。图5展示了Transformer模块，以下是多头注意力的计算过程。

对于特征图像$X\in\mathbb{R}^{M^2\times C}$，将三个可学习的权重矩阵$W^Q\in\mathbb{R}^{C\times C},W^K\in\mathbb{R}^{C\times C}$和$W^V\in\mathbb{R}^{C\times C}$映射至查询矩阵$Q$、键矩阵$K$和值矩阵$V$，其表达式如式(1)：

$$\{Q,K,V\}=\{XW^Q,XW^K,XW^V\} \tag{1}$$

注意力机制定义如式(2)：

$$\text{Attention}(Q,K,V)=\text{Softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V \tag{2}$$

其中，d_k是一个常量，便于在Softmax操作后存在梯度。将特征图拉伸为固定长度的向量后，进行计算，如式(3)、(4)所示：

$$\hat{f}^l=W\text{-}MSA(LN(f^{l-1}))+f^{l-1} \tag{3}$$

$$f^l=MLP(LN(\hat{f}^l))+\hat{f}^l \tag{4}$$

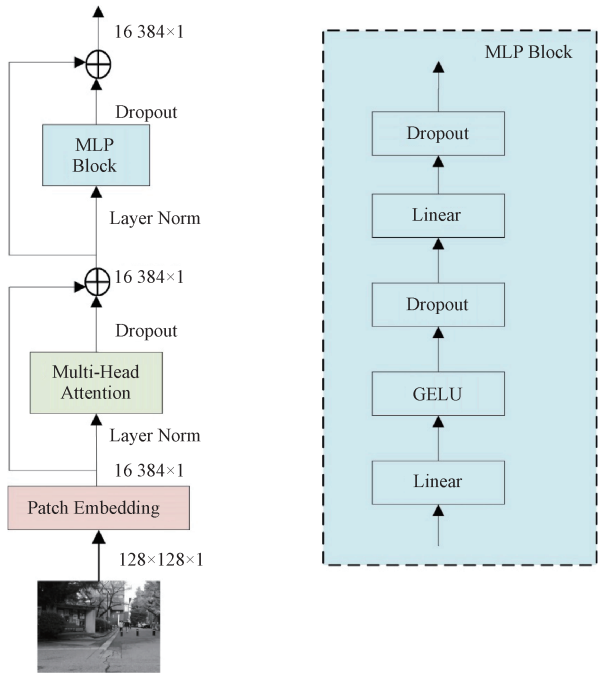

图 5 Transformer模块

Fig. 5 Transformer Block

其中，$\hat{f}^l$和f^l分别表示$W\text{-}MSA$和MLP模块的输出特征图像。

1.2 解码器

原图像经过编码器特征提取后生成32张特征图，采用堆叠的方式对两个通道的特征图进行融合，解码器对堆叠融合后的特征图序列进行重组得到融合图像。如图6所示，解码层通过四次卷积操作把64张的特征图整合成一张结果图，每一层的输出通道数分别为32、16、8、1。

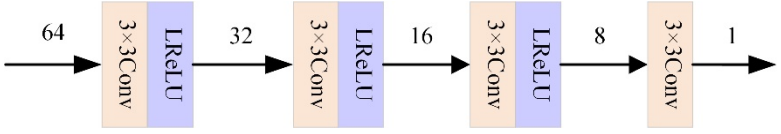

图 6 解码模块

Fig. 6 Decoder block

1.3 损失函数

在本文损失函数设置中，使融合图像的灰度与原图像灰度值较大者相似，同时使融合图像的梯度与原图像梯度较大者相似。损失函数定义如式(5)：

$$L=\alpha L_{\text{int}}+\beta L_{\text{texture}} \tag{5}$$

其中α和β是超参数，用于平衡像素损失函数L_{int}和梯度损失函数L_{texture}，L_{int}和L_{texture}的定义分别如式(6)、(7)。

$$L_{\text{int}}=\frac{1}{HW}\|I_{\text{f}}-\max(I_{\text{ir}},I_{\text{vis}})\|_1 \tag{6}$$

$$L_{\text{texture}}=\frac{1}{HW}\left\||\nabla I_{\text{f}}|-\max(|\nabla I_{\text{ir}}|,|\nabla I_{\text{vis}}|)\right\|_1 \tag{7}$$

其中，H和W分别表示特征图像的宽度和高度；I_{ir}、I_{vis}和I_{f}分别表示红外图像、可见光图像和融合图像；∇I_{ir}、∇I_{vis}和∇I_{f}分别表示红外图像、可见光图像和融合图像的梯度。

2　实验结果

2.1　训练数据

训练数据集选择公开且配准度较高的MFNet数据集[3]，其中包括1 083对不同场景的图像，场景目标包括人、车等。数据集中的图像大小为640×480，位深度为24。为了适合网络训练，这些图像被处理成灰度图像并将图像裁剪为128×128大小。图7展示了一些MFNet数据集样本，前四列是光线不足场景下的图像对，后四列是光线充足场景下的图像对。

图7　红外与可见光图像示例[3]

Fig. 7　Examples of infrared and visible images[3]

2.2　模型参数

在本文提出的融合模型中，学习率设置为2×10^{5}，衰减率为0.99，权值更新规则为Adam，批量图像大小设置为8，α、β分别设置为1和9。实验操作系统为Windows 10，硬件平台为AMD Ryzen Threadripper PRO 3945WX，主频为4.0 GHz，GPU为RTX 3080，显存为10 G，软件平台为Python 3.7，模型通过Pytorch进行搭建，从三个公开数据集(TNO[4]、OTCBVS[5]和RoadScene[6])中随机抽取样本作为测试集。

2.3　算法对比

为了验证所提模型相对于其他先进方法的优势，选择两种传统方法和七种深度学习方法进行主客观比较，包括两种传统方法：MST_SR[7]、GTF[8]；一种基于CNN的方法：U2Fusion[9]；两种基于GAN的方法：DDcGAN[10]、SDDGAN[11]；三种基于自动编码器的方法：DenseFuse[12]、RFN-Nest[13]和STDFusion[14]；一种基于Transformer的方法：SwinFusion[15]。客观评价指标包括互信息(MI)[16]、标准偏差(SD)[17]、平均梯度(AG)[18]、峰值信噪比(PSNR)[19]、融合的视觉信息保真度(VIFF)[20]和基于梯度的融合性能($Q^{AB/F}$)[21]。经验证表明：本文融合后的图像不仅具有高质量的视觉效果，而且在客观指标上的评价也优于其他方法。此外，实验部分对比了原图像、本文融合图像以及其他方法生成的融合图像在目标检测任务中的表现，验证了本文融合方法的优势以及在下游任务中的应用价值。

2.3.1　融合图像主观对比

图8分别展示了上述算法在TNO数据集上的一对红外和可见光图像融合结果。从红外目标角度，图(c)、(e)、(f)和(g)的目标亮度较低，与背景灰度值相似；另外，融合图像的目标局部信息被放大，如融合图像左下角绿色框所示，图(c)、(e)、(f)和(g)的目标信息弱，图(d)目标信息几乎丢失，图(h)目标边缘出现虚影，图(i)、(j)和(k)目标边缘不够完整。对比上述融合图像目标，本文算法目标明亮，结构清晰。融合图像右下角红色框对图像背景局部区域进行了放大，相对其他算法，图(c)、(e)、(f)、(g)和(i)背景层次性不够突出；图(h)、(j)、(k)和本文方法背景纹理清晰。整体而言，本文方法在目标亮度信息及图像纹理信息上优于其他算法。

如图9所示，第一行展示的是从OTCBVS视频截取的单帧红外和可见光图像。图(c)、(d)、(e)、(f)和(g)目标亮度较低，图(d)和(g)几乎没有体现红外目标亮度信息；即便图(h)、(i)和(j)的目标明亮，但是图(h)目标边缘受光晕影响，存在一定失真；图(i)目标信息不完整，目标模糊；图(j)目标仅保留红外图像信息，纹理丢失。图(k)和本文方法目标明亮，信息丰富。在红色框位置，图(e)、(g)、(i)树枝纹理模糊。相对原可见光图像，图(d)、(f)、(g)、(h)和(i)图像灰度值较低，视觉效果不够理想。本文方法的融合结果在保留目标特征亮度、纹理信息的同时，背景清晰，细节明显。因此，本文方法的融合效果优于其他对比方法。

图10第一行展示了一对马路场景下的红外和

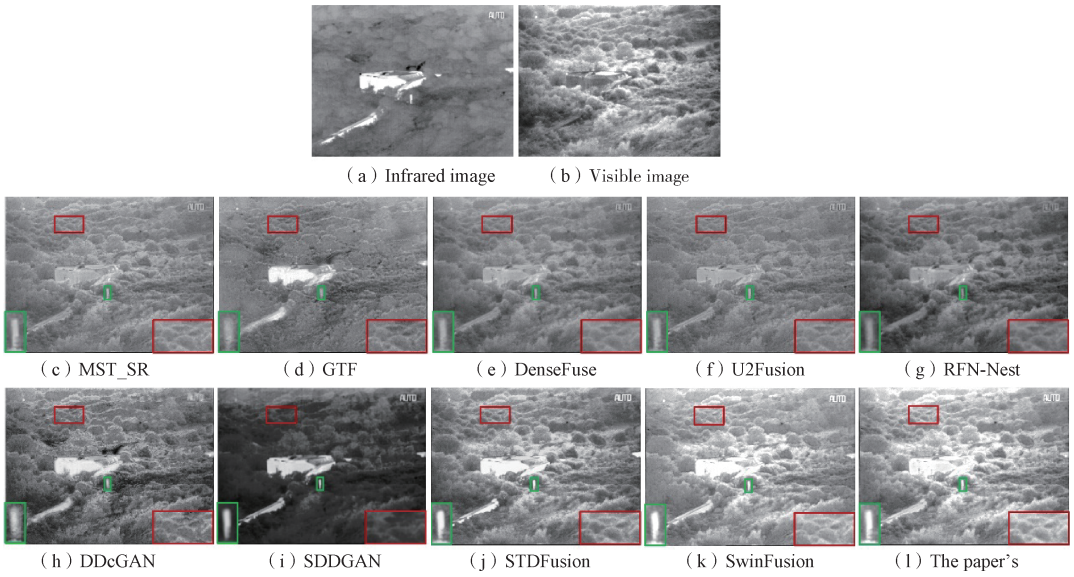

图 8　TNO 数据集的融合结果

Fig. 8　Fusion results for the TNO dataset

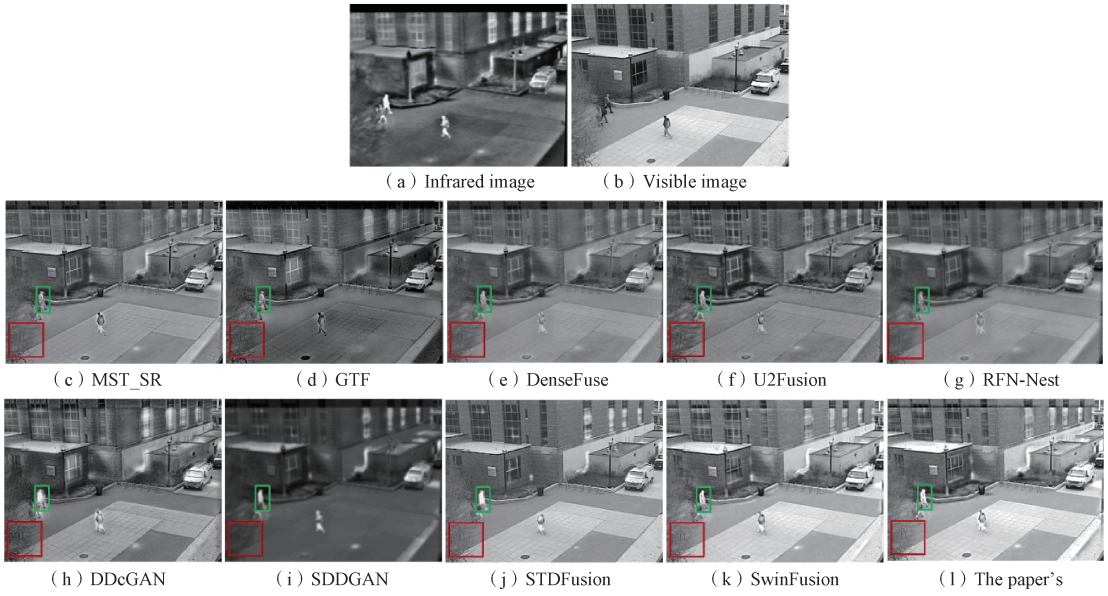

图 9　OBTCVS 数据集的融合结果

Fig. 9　Fusion results for the OBTCVS dataset

可见光图像，该图像来自 RoadScene 数据集。绿色、红色和蓝色框分别标定了融合图像的目标、路灯和树木背景。图(c)、(e)、(f)、(g)和(h)目标灰度值偏低，图(g)、(h)、(i)目标模糊，图(d)、(h)、(j)和(k)路灯信息不完整。图(j)树木背景信息丢失，图(d)、(g)、(i)和(k)树木背景模糊，图(h)背景出现错误，与原图像结构不一致。综上三个方面，本文方法的融合图像目标明亮、路灯信息完整且树木背景纹理清晰。

2.3.2　融合图像客观评价

基于融合主观评价，可对图像质量做出一定判断。为了进一步验证本文方法的优势，本节对各算法在上述融合图像上的客观评价指标进行对比，做出综合性的评价。指标值越大，说明图像在该指标上的性能更好。

表 1 记录了各算法在上述 TNO 数据集上融合图像的客观评价指标。在 SD、VIFF 以及 $Q^{AB/F}$ 三个指标上，本文方法表现最优，表明其融合图像目标明亮、局部信息丰富，视觉效果良好。另外，

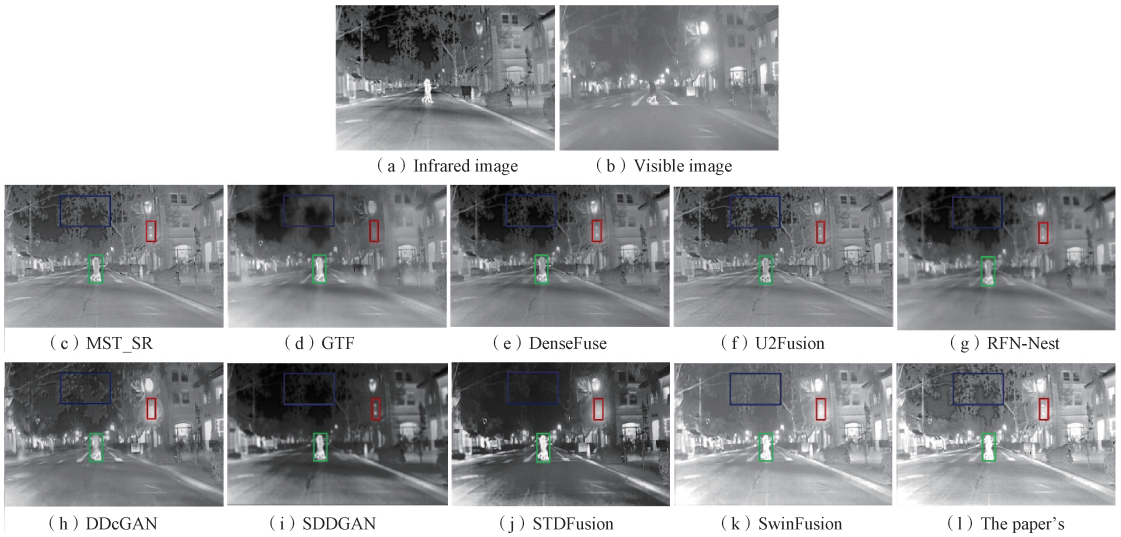

图 10　RoadScene 数据集的融合结果

Fig. 10　Fusion results for the RoadScene dataset

本文算法的其他指标在所有算法中处于前列，比如：MI 数值排在第二位置，仅比 STDFusion 算法小了 0.001；AG、PSNR 均排在第三位置，因此，本文算法融合图像的信息量、梯度等指标均已得到验证。

表 1　TNO 数据集融合结果的客观指标

Table 1　Objective indicator of fusion results for TNO datasets

	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
MST_SR	13.605 6	9.229 6	4.820 0	17.296 3	0.240 0	0.568 0
GTF	13.559 9	8.467 7	5.265 6	20.612 7	0.154 6	0.487 2
DenseFuse	13.633 2	9.350 4	3.543 4	17.711 3	0.263 9	0.348 9
U2Fusion	13.170 9	8.943 7	4.275 0	17.576 4	0.235 4	0.373 5
RFN-Nest	14.050 1	9.634 3	3.190 4	16.786 3	0.284 2	0.331 3
DDcGAN	14.656 0	9.451 5	**6.869 5**	15.072 6	0.333 3	0.428 0
SDDGAN	13.329 9	7.900 4	2.065 0	10.941 4	0.137 1	0.119 6
STDFusion	**14.664 2**	9.632 1	6.154 8	12.796 8	0.303 1	**0.633 0**
SwinFusion	14.310 4	9.629 4	5.470 6	**21.393 6**	0.328 8	0.497 1
本文方法	14.659 7	**10.065 1**	5.685 9	19.129 0	**0.384 1**	0.615 5

表 2 记录了各算法在上述 OTCBVS 视频数据集上融合图像的客观评价指标。从客观指标来看，本文方法除 MI、VIFF 值排在第二位置，其他指标都能达到最优。与主观评价一致，本文方法融合图像的信息量、纹理、对比度以及视觉效果均优于最新算法。

表 3 记录了各算法在上述 RoadScene 数据集上融合图像的客观评价指标。本文方法在 AG 和 VIFF 指标上表现最优，MI 和 $Q^{AB/F}$ 指标值均排在第二，因此，本文方法融合图像质量高，客观评价良好。

表 2　OTCBVS 数据集融合结果的客观指标

Table 2　Objective indicator of fusion results for OTCBVS datasets

	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
MST_SR	**15.514 7**	10.696 0	7.772 7	21.285 5	0.253 0	0.517 4
GTF	13.955 5	9.760 0	7.512 3	21.087 0	0.163 2	0.426 8
DenseFuse	14.247 3	10.708 6	5.481 8	21.363 9	0.286 2	0.358 5
U2Fusion	14.310 6	10.811 1	7.132 6	21.463 3	0.292 2	0.400 7
RFN-Nest	14.389 8	10.664 9	4.397 5	21.161 8	0.265 7	0.252 5
DDcGAN	14.770 3	9.728 4	8.249 6	19.037 8	0.230 2	0.403 7
SDDGAN	12.977 4	8.663 1	3.053 4	21.261 4	0.165 3	0.130 9
STDFusion	14.748 8	10.187 6	7.925 0	20.285 5	0.166 4	0.532 1
SwinFusion	15.063 6	10.763 8	8.120 0	21.780 0	0.232 4	0.477 0
本文方法	15.173 3	**10.959 2**	**8.832 7**	**21.952 2**	0.279 9	**0.555 6**

表 3　RoadScene 数据集融合结果的客观指标

Table 3　Objective indicator of fusion results for RoadScene datasets

	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
MST_SR	14.340 4	10.622 4	5.347 9	**18.447 6**	0.501 7	**0.639 7**
GTF	**15.116 3**	10.580 3	3.186 9	18.242 6	0.315 1	0.337 7
DenseFuse	14.636 5	**10.824 7**	4.506 6	18.242 6	0.561 9	0.540 4
U2Fusion	14.320 2	10.750 3	5.409 9	17.927 9	0.504 7	0.569 0
RFN-Nest	14.752 7	10.946 0	2.912 0	18.347 8	0.460 1	0.280 9
DDcGAN	14.492 4	9.459 5	4.557 2	15.426 2	0.333 2	0.324 4
SDDGAN	14.776 1	9.642 1	3.322 1	15.579 9	0.437 2	0.224 1
STDFusion	14.646 8	8.709 4	5.388 3	10.259 8	0.494 2	0.378 7
SwinFusion	14.445 0	10.664 8	4.604 8	17.096 8	0.454 1	0.388 7
本文方法	14.790 7	10.644 5	**6.388 4**	15.176 9	**0.677 2**	0.605 4

除上述单幅图像指标对比之外，本章从测试集随机挑选20对红外和可见光图像，以对比各算法得到的融合图像在不同指标上的表现，如图11所示。可以注意到，本文方法在SD、AG、VIFF以及$Q^{AB/F}$指标上整体优于其他方法，并且有13幅融合图像的MI指标值排在对比算法前列。除个别融合图像外，本文方法融合图像的PSNR值与最高值差距较小，甚至与第4、5、6幅图像相比，本文方法PSNR值最高。最后，对这些图像指标取均值，结果如表4所示。显然，本文方法的SD、VIFF和$Q^{AB/F}$值均高于其他方法，并且MI和AG值排在第二位。综上，本文方法的融合图像客观评价整体优于其他先进算法。

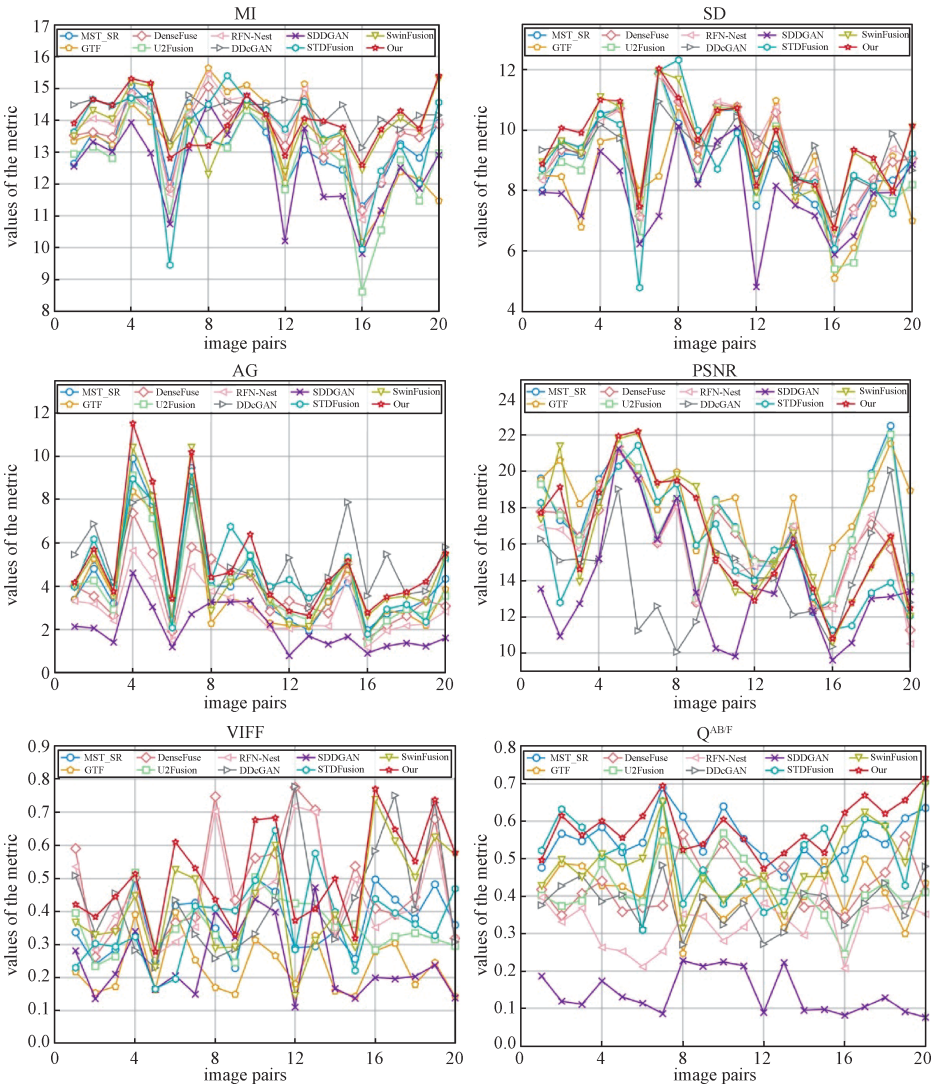

图11　融合图像客观评价指标图

Fig. 11　Map of objective evaluation metrics for fused images

2.3.3　目标检测对比

为证明本文方法得到的融合图像在应用领域的价值，以目标检测任务为例，对比红外图像、可见光图像和其他融合方法得到的融合图像的目标检测结果。从MFNet数据集中选择50张用于目标检测的图像对，这些图像反映的是一些城市场景，笔者手动标注了一些关键的目标，比如：人、车等等。把红外图像、可见光图像和其他算法的融合图像分别输入到Faster-RCNN[21]检测器，得到对应的检测图。利用所有目标的平均精度(mAP)衡量检测性能，检测的评估结果如表5所示。显然，两幅原图像的检测精度是最低的，因此图像融合

表 4　融合图像客观评价指标均值

Table 4　Mean values of objective evaluation indexes for fused images

实验方法	MI	SD	AG	PSNR	VIFF	$Q^{AB/F}$
MST_SR	13.182 4	8.911 7	4.235 4	17.062 7	0.366 8	0.553 9
GTF	13.352 8	8.615 9	3.836 1	**18.040 1**	0.231 2	0.410 8
DenseFuse	13.590 1	9.260 3	3.754 4	16.205 1	0.483 5	0.431 5
U2Fusion	12.766 6	8.747 9	4.085 2	17.069 7	0.355 4	0.424 7
RFN-Nest	13.787 7	9.343 4	2.940 3	15.868 4	0.448 6	0.325 2
DDcGAN	**14.303 2**	9.233 7	**5.189 7**	14.168 3	0.431 1	0.378 6
SDDGAN	12.614 2	7.856 9	2.058 0	13.842 4	0.246 0	0.139 2
STDFusion	13.503 3	9.004 9	4.772 1	15.624 6	0.368 1	0.502 5
SwinFusion	13.809 2	9.428 8	4.618 3	16.338 4	0.428 2	0.489 1
Our	14.001 8	**9.523 7**	5.069 8	16.256 2	**0.510 0**	**0.585 6**

表 5　目标检测精度

Table 5　Target detection accuracy

	mAP
红外图像	0.494
可见光图像	0.561
MST_SR	0.773
GTF	0.725
DenseFuse	0.699
U2Fusion	0.788
RFN-Nest	0.667
DDcGAN	0.654
SDDGAN	0.606
STDFusion	0.712
SwinFusion	0.730
本文方法	**0.800**

的必要性已被验证。其次，在所有方法中，本文方法的精度最高，因此本文的融合方法相较其他对比方法更利于应用。

此外，图 12 提供了两个可视化的例子来说明本文的融合方法在帮助物体目标检测方面的优势。在第一个场景中，红外和可见光原图像各检测到两个人，且红外图像检测到车的类型错误，而本文方法的融合图像检测到三个人，且检测到车的类别正确。对于其他方法，RFN-Nest、DDcGAN、SDDGAN、STDFusion 和 SwinFusion 方法的融合图像仅检测到一个人，MST_SR、GTF 和 DenseFuse 方法的融合图像仅检测出两个人。在第二个场景中，红外图像和可见光图像各检测到一个不同位置的人，而本文方法融合结果可以同时检测到两个人。在对比方法中，除 MST_SR 和 SDDGAN 方

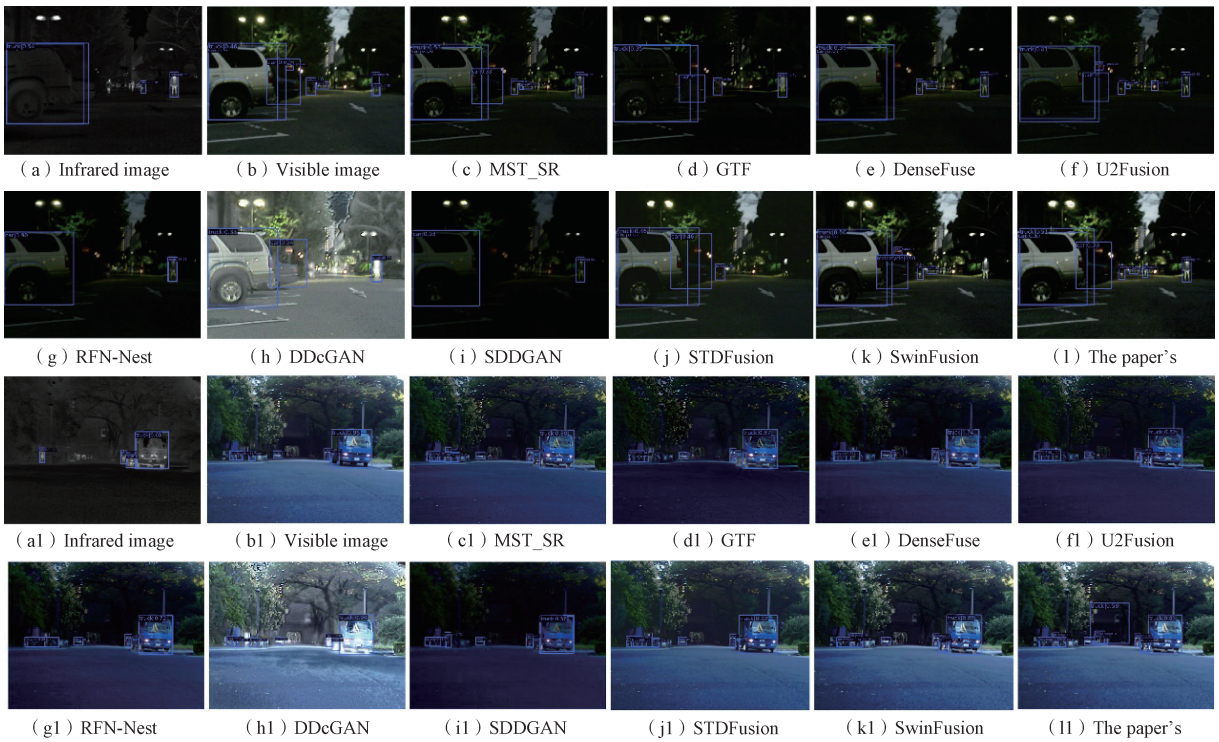

图 12　融合图像目标检测结果

Fig. 12　Fusion image target detection results

法外，其他方法均不能检测到红外图像中的目标。综上所述，本文方法的融合结果在目标检测领域的应用价值已被验证。

3 结束语

本文提出了一种基于CNN-Transformer架构的新型自动编码器，用于红外图像和可见光图像的融合。一方面，针对自动编码器网络中CNN对全局信息不敏感的问题，结合Transformer模块以构建既能保留局部信息又能保留全局信息的编码器，提高融合图像的质量。另一方面，针对目前编码器使用孪生结构而无法从红外图像和可见光图像中充分提取独特特征问题，分别为红外和可见光图像设计对比度增强模块和梯度残差模块，使特征提取过程更具针对性。与其他先进的方法相比，本文提出的融合方法可以获得主观和客观评价都较优的融合图像，同时有利于包括物体检测和识别等其他下游任务。

参考文献

[1] 沈英, 黄春红, 黄峰, 等. 红外与可见光图像融合技术的研究进展[J]. 红外与激光工程, 2021, 50(9): 20200467.
SHEN Ying, HUANG Chunhong, HUANG Feng, et al. Research progress of infrared and visible image fusion technology [J]. Infrared and Laser Engineering, 2021, 50(9): 20200467.

[2] 李霖, 王红梅, 李辰凯. 红外与可见光图像深度学习融合方法综述[J]. 红外与激光工程, 2022, 51(12): 20220125.
LI Lin, WANG Hongmei, LI Chenkai. A review of deep learning fusion methods for infrared and visible images [J]. Infrared and Laser Engineering, 2022, 51(12): 20220125.

[3] HA Q, WATANABE K, KARASAWA T, et al. MFNet: Towards realtime semantic segmentation for autonomous vehicles with multi-spectral scenes[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2017: 5108-5115.

[4] TOET A. TNO Image fusion dataset[DS/OL]. [2024-03-28].https://figshare.com/articles/dataset/TNO_Image-_Fusion_Dataset/1008029, 2014.

[5] DAVIS J W, SHARMA V. OTCBVS benchmark dataset collection[DS/OL]. [2024-03-28]. http://www. cse. ohio-state. edu/otcbvs-bench, 2007.

[6] 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(1): 502-518.

[7] LIU Y, LIU S, WANG Z. A general framework for image fusion based on multi-scale transform and sparse representation[J]. Information Fusion, 2015, 24: 147-164.

[8] MA J, CHEN C, LI C. Infrared and visible image fusion via gradient transfer and total variation minimization[J]. Information Fusion, 2016, 31: 100-109.

[9] MA J J, XU H, JIANG J J, et al. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995.

[10] ZHOU H B, WU W, ZHANG Y D, et al. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network[J]. IEEE Transactions on Multimedia, 2021, 25: 635-648.

[11] 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.

[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: 72-86.

[13] MA J, TANG L, XU M, et al. STDFusionNet: An Infrared and visible image fusion network based on salient target detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.

[14] MA J Y, TANG L F, FAN F, et al. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.

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

[16] ESKICIOGLU A M, FISHER P S. Image quality measures and their performance[J]. IEEE Transactions on Communications, 1995, 43(12): 2959-2965.

[17] GUO W, XIONG N, CHAO H C, et al. Design and

analysis of self-adapted task scheduling strategies in wireless sensor networks[J]. Sensors, 2011, 11(7): 6533-6554.

[18] 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.

[19] 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.

[20] XYDEAS C S, PETROVIC V. Objective image fusion performance measure[J]. Electronics Letters, 2000, 36 (4): 308-309.

[21] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//29th Annual Conference on Neural Information Processing Systems, NIPS 2015, Canad. 2015: 91-99.

[作者简介]

李 霖 1995年生，硕士，助理工程师。

沈永健 1985年生，博士，研究员。

张鹏宇 1996年生，硕士，工程师。

原 昊 1998年生，硕士，工程师。

王 超 1994年生，硕士，助理工程师。

(本文编辑：潘三英)

(英文编辑：赵尹默)