一种基于生成对抗网络与注意力机制的可见光和红外图像融合方法

罗 迪[1,2]，王从庆[1,2]，周勇军[2]
（1. 南京航空航天大学 自动化学院，江苏 南京 210016；2. 近地面探测技术重点实验室，江苏 无锡 214000）

摘要：针对低照度可见光图像中目标难以识别的问题，提出了一种新的基于生成对抗网络的可见光和红外图像的融合方法，该方法可直接用于RGB三通道的可见光图像和单通道红外图像的融合。在生成对抗网络中，生成器采用具有编码层和解码层的U-Net结构，判别器采用马尔科夫判别器，并引入注意力机制模块，使得融合图像可以更关注红外图像上的高强度信息。实验结果表明，该方法在维持可见光图像细节纹理信息的同时，引入红外图像的主要目标信息，生成视觉效果良好、目标辨识度高的融合图像，并在信息熵、结构相似性等多项客观指标上表现良好。

关键词：图像融合；可见光/红外图像；低照度图像；生成对抗网络；注意力机制
中图分类号：TN753 **文献标识码：**A **文章编号：**1001-8891(2021)06-0566-09

A Visible and Infrared Image Fusion Method based on Generative Adversarial Networks and Attention Mechanism

LUO Di[1,2]，WANG Congqing[1,2]，ZHOU Yongjun[2]
(1.*College of Automation Engineering of Nanjing University of Aeronautics and Astronautics*, *Nanjing* 210016, *China*;
2. *Science and Technology on Near-Surface Detection Laboratory*, *Wuxi* 214000, *China*)

Abstract： A new fusion method for visible and infrared images based on generative adversarial networks is proposed to solve the problem of recognizing targets in low-light images; the method can be directly applied to the fusion of RGB three-channel visible images and infrared images. In generative adversarial networks, the generator adopts a U-Net structure with encoding and decoding layers. The discriminator adopts a Markovian discriminator, and the attention mechanism is introduced to force the fused image to pay more attention to the high-intensity information on infrared images. The experimental results show that the proposed method not only maintains the detailed texture information of visible images but also introduces the main target information of infrared images to generate fusion images with good visual effects and high target identification, and it performs well in information entropy, structural similarity, and other objective indexes.

Key words： image fusion, visible and infrared image, low-lightimage, generative adversarial networks, attention mechanism

0 引言

可见光图像具有丰富的纹理细节和空间分辨率，符合人类视觉感知方式。但在低照度条件下，图像质量会显著下降，尤其是图像中的有效目标（比如人员目标）将会缺失，变得难以识别。而红外图像是通过目标的热辐射信息成像，可以较为鲜明地区分目标与背景，在全天候条件下都能良好地工作，但图像本身缺乏细节，无法反映场景信息。

图像融合作为一种图像增强技术，可以将由不同传感器在同一场景下采集的不同图像进行组合，生成鲁棒性更强和信息更为丰富的图像，有助于后续处理和决策。图像融合现在被广泛应用于医疗诊断、军事目标检测、生物识别和遥感等领域。因此将可见光图

收稿日期：2020-09-08；**修订日期：**2020-10-12.
作者简介：罗迪（1995-），男，硕士研究生，主要研究方向：深度学习与无人机目标检测。E-mail：1366701808@qq.com。
通信作者：周勇军（1972），男，高级工程师，主要研究方向：近地面目标探测技术。E-mail：478992155@qq.com。
基金项目：近地面探测技术重点实验室基金资助项目（TCGZ2019A006）。

像和红外图像融合，可以在保留细节纹理信息的同时，突出目标信息，有利于人眼感知，提高目标的检测和识别率。

传统图像融合技术发展至今，根据其理论依据的不同，可以分为多尺度变换方法、稀疏表示方法、子空间方法和基于显著性的方法，以及综合以上各类方法的混合模型[1]，其中多尺度变换方法因为其简单与有效性，被广泛用于可见光和红外图像融合中，该方法包括拉普拉斯金字塔变换（Laplace pyramid，LP）[2]、双数复小波变换（the dual-tree complex wavelet transform, DTCWT）[3]、非下采样轮廓波变换（nonsubsampled contourlet transform, NSCT）[4]和曲波变换（Curvelets, CVT）[5]。然而这些融合方法都依赖于特定的图像变换，往往对可见光图像和红外图像采用相同的特征提取与表示，此外，在融合阶段，这些方法都需要手工设计对应的融合规则，且越来越复杂。

随着神经网络和深度学习技术的发展，许多基于神经网络的图像融合方法被提出。这些方法采用卷积神经网络以网络学习的方式提取图像的多维度特征，并结合传统方法的融合规则进行特征重组，或采用解码网络重构融合图像，比如 Hui Li 等人提出的 DenseFuse[6]。Jiayi Ma 等人将生成对抗网络（generative adversarial network, GANs）引入图像融合任务中，该融合模型通过端到端的方式将输入图像直接转换为融合图像，避免了繁复的活动水平测量和融合规则的手工设计[7]。但该模型因为其较为简单的生成器网络结构，在灰度级别的图像融合中虽表现出一定的优势，但在 RGB 图像和红外图像的融合中就缺失了特征提取能力，无法泛化到其他图像数据集和实际应用中。

为了解决这些问题，本文提出了一种基于生成对抗网络的可见光与红外图像融合方法，该方法基于生成器和判别器的极大极小博弈，达到两者的纳什平衡来融合可见光图像和红外图像，实现往低照度图像中引入红外目标信息。在生成对抗网络中的生成器采用 U-Net 网络结构[8]，其所具有的编码层和解码层结构可以更好地提取图像特征，且适用于 RGB 可见光图像和单通道红外图像的融合。同时，在生成对抗网络的基础上，引入了注意力机制模块，使图像的编码解码过程可以更关注于目标信息，对于低照度图像的融合，则可促使网络训练更关注红外图像中高辐射的目标信息。此外，设计了对应的组合损失函数，促使模型的生成图像可以更好地维持可见光图像中的细节和纹理信息。最后，使用公共的红外和可见光图像数据集[9]进行模型训练和实验。

1 原理及方法

1.1 生成对抗网络

生成对抗网络由 Lan Goodfellow 在 2014 年率先提出[10]，引起了深度学习和图像合成领域的极大关注。该网络是一种通过对抗的方式，学习数据分布的生成式模型。其框架包含两个对立的模型：生成器 G 和判别器 D。生成器 G 的作用是尽可能生成符合训练数据分布的样本，判别器 D 的作用是区分样本来源。

为了解决传统 GANs 出现的各种训练问题，GANs 也发展出了各类变体。Radford 等人提出了深度卷积 GANs（DCGANs），该方法将卷积神经网络（CNNs）引入 GANs 中[11]，弥补了用于监督学习的 CNN 与用于无监督学习的 GANs 之间的差距。Mao 等人提出了最小二乘 GANs（LSGANs）[12]，该方法使用最小二乘损失函数取代常规 GANs 使用的交叉熵损失函数。具体形式如下：

$$\min_D V_{\mathrm{LSGAN}}(D)=\frac{1}{2}E_{x\sim p_{\mathrm{data}}(x)}[(D(x)-b)^2]+\frac{1}{2}E_{z\sim p_z(z)}[(D(G(z)-a)^2] \tag{1}$$

$$\min_G V_{\mathrm{LSGAN}}(G)=\frac{1}{2}E_{z\sim p_z(z)}[(DG(z)-c)^2] \tag{2}$$

式中：z 为随机噪声；x 为真实数据；$D(x)$为判别器判断样本是否为真实数据的概率；$E_{x\sim p_{\mathrm{data}}(x)}$、$E_{z\sim p_z(z)}$ 分别为真实数据和生成数据的期望，$p_z(z)$为输入噪声变量先验概率；$p_{\mathrm{data}}(x)$为真实数据概率分布。

参数 a，b，c 的设置有两种方式。一种是满足条件 $b-c=1$ 和 $b-a=2$，常设置为 $a=-1$，$b=1$，$c=0$。另一种设置是满足条件 $b=c$，常设置为 $a=0$，$b=1$，$c=1$。本文采用第二种设置。

1.2 图像翻译

图像翻译通常是指将一张图片通过相应的映射转化成另外一张图片，比如灰度图、梯度图、彩色图和语义标签图之间的转化。传统方法是采用像素和像素之间的映射，并对不同的转化任务设置不同的框架。而 Isola 等人提出了一种使用生成对抗网络解决图像翻译问题的通用解决方法 pix2pix[13]。该方法在语义标签和城市街道图像，黑白图像和彩色图像，线条草图到实物图像等转化任务中都获得了不错的效果。

为了实现可见光图像和红外图像的融合，可以把图像融合也看作一种图像翻译任务，输入图像为在通道上进行拼接的混合域可见光-红外图像对，输出图像

为含有红外信息的可见光图像。此外，用于图像融合的可见光和红外图像数据集大多都是成对采集的图像，所以本文的方法沿用 pix2pix 构建网络的思路，进行网络结构的搭建和损失函数的设计，从而实现高质量的图像融合。

1.3 注意力机制

视觉注意力机制被广泛用在分类网络中。如 Jaderberg 等人提出的空间域转换网络（spatial transformer networks）[14]。Hu Jie 等人提出的 SENet[15]，它通过学习通道之间的相关性，筛选出针对通道的注意力。而 Woo 等人提出的卷积块注意模块（convolutional block attention module，CBAM）[16]，综合了上述两种注意力机制，通过依次使用通道和空间注意力模块，分别推导出注意力图，然后将注意力图与输入特征图相乘，进行自适应特征细化。此外，CBAM 模块作为一个轻量级的通用模块，可以无缝地集成到任何卷积神经网络架构中，因此本文尝试将该模块添加到图像融合框架中。

2 图像融合方案

2.1 数据集准备

目前大多数融合方法所采用的数据集为 TNO 图像融合数据集，并将其作为评价融合方法性能的基准数据集。但该数据集中的可见光图像多为单通道灰度图，且图像所拍摄的目标种类繁杂。以及训练模型需要大量的数据图像。因此，综合多种考虑，本文选择使用多光谱行人检测数据集[9]。该数据集提供了由基于分束器的专用硬件捕捉到的一致的彩色可见光图像和红外图像对，且已经进行物理对齐。数据集大小与其余基于可见光的数据集一样大。此外，该数据集不仅包含白天图像对，也包含夜间图像对。

本文从该数据集中选取 3000 对可见光红外图像作为训练数据集，其中白天和夜晚场景对半，图像对大小为 480×640。在将原始图像输入到网络前，首先将三通道可见光图像和单通道红外图像在通道维度进行拼接，生成四通道图像作为输入图像。对输入图像采用随机翻转、变形、裁剪等预处理后，再将图像缩放至 480×640。这样在每一个周期的训练过程中，网络都会获得不同的图像对输入，根据周期的设置可以成倍扩充训练集的数量。

由于该数据集面向的任务是行人检测，因此图像中（尤其是红外图像）的显著目标信息主要为行人。因此，本文的图像融合任务与以往的融合任务略有不同，不是单纯地将两类图像中的有用信息融合成信息量更全面的图像，而是更关注于在目标并不清晰明确的低照度图像中，引入红外图像中的显著目标信息，即行人目标信息，形成利于人眼感知的融合图像。且该融合过程并没有使用数据集中提供的标签，仅使用可见光和红外图像对。

2.2 生成器的构建

2.2.1 生成器基本结构

本文所采用的基于 U-Net 的卷积神经网络结构如图 1 所示。该网络采用编码层和解码层的对称结构（图中 en 表示编码层，de 表示解码层），并在对应的解码层和编码层之间加上跳跃连接（skip connections），此外，不同于原始 U-Net 网络，该网络取消了所有池化层和传统的上采样操作，取而代之的为卷积步长为 2 的卷积层和反卷积层。

融合过程表述如下：首先将三通道的可见光图像和单通道的红外图像在通道维度进行拼接，形成四通道输入图像，通过第一层卷积层提取特征图后，输入到 CBAM 模块中，经过其中的通道注意力模块和空间注意力模块后，输出同样尺寸大小和通道数的特征图。之后在经过一系列的编码层，不断下采样一直到达瓶颈层，然后进行相对应的反卷积操作，并在每次反卷积操作之前，将前一解码层得到的特征图与编号对应的解码层得到的特征图进行通道维度的拼接，然后输入到下一解码层。这种跳跃连接的操作相当于起到了多尺度融合的目的，可以充分利用编码层网络提取的低层级特征。此外，由于我们在编码层采用了 CBAM 模块，因此最后一层的解码层所拼接的为 CBAM 模块所输出的特征图。最后输出的为与输入图像相同尺寸的三通道彩色融合图像。由于需要处理的图像大小为 480×640，因此没有完全采用文献[13]所提出的框架所采用的网络参数（原参数适用于 256×256 或 512×512 的输入图像）。具体网络参数如表 1 所示。

此外，为了提高模型的收敛速度，并保持每个图像实例之间的独立性，在每层卷积层和反卷积层（除最后一层）的后面使用实例归一化（instance normalization）对数据进行归一化。为了克服训练过程中梯度消失的问题，编码层使用斜率为 0.2 的 LeakyReLU 激活函数，解码层（除最后一层）使用 ReLU 激活函数和来提升网络的非线性程度。解码层中的最后一层则使用 tanh 激活函数。

2.2.2 CBAM 模块的网络结构

CBAM 模块作为一个轻量级的通用模块，我们直接将其添加在第一层卷积层后面。它包含通道和空间注意力模块，其具体网络结构如图 2 所示。

其操作过程如下：第一步首先将由卷积层得到的

240×320×32 的特征图 F 分别经过最大池化层和平均池化层得到两组 1×1×32 的特征图，实现空间维度的压缩，然后将两组特征图分别经过一个共享参数的多层感知器（Multi Layer Perceptron，MLP）进行特征整合，在通道维度上以逐像素求和的方式合并两类输出，并经过 sigmoid 激活函数得到通道注意力特征图 $M_c(F)$，最后将 $M_c(F)$与 F 进行元素点乘，即得到经过通道注意力优化的特征图 F'，其大小为 240×320×32。

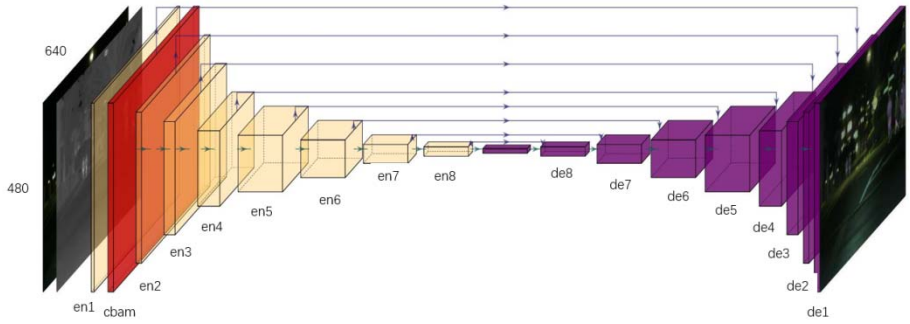


图 1 生成器的网络结构

Fig.1 The network structure of the generator

表 1 生成器网络参数

Table 1 The parameters of generator

Convolution layer	Kernel size/stride	Padding	Input size	Output size
Conv1	4×4/2	(1,1)	480×640×4	240×320×32
CBAM	4×4/2	(1,1)	240×320×32	240×320×32
Conv2	4×4/2	(1,1)	240×320×32	120×160×64
Conv3	4×4/2	(1,1)	120×160×64	60×80×128
Conv4	4×4/2	(1,1)	60×80×128	30×40×256
Conv5	4×4/2	(2,1)	30×40×256	16×20×512
Conv6	4×4/2	(1,1)	16×20×512	8×10×512
Conv7	4×4/2	(1,2)	8×10×512	4×6×512
Conv8	4×4/2	(1,1)	4×6×512	2×3×512
ConvTrans8	4×4/2	(1,1)	2×3×512	4×6×512
ConvTrans7	4×4/2	(1,2)	4×6×1024	8×10×512
ConvTrans6	4×4/2	(1,1)	8×10×1024	16×20×512
ConvTrans5	4×4/2	(2,1)	16×10×1024	30×40×256
ConvTrans4	4×4/2	(1,1)	30×40×512	60×80×128
ConvTrans3	4×4/2	(1,1)	60×80×256	120×160×64
ConvTrans2	4×4/2	(1,1)	120×160×128	240×320×32
ConvTrans1	4×4/2	(1,1)	240×320×64	480×640×3

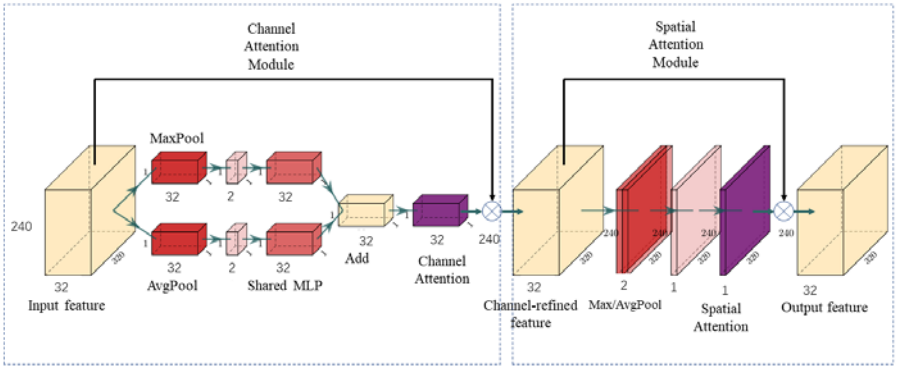


图 2 CBAM 的网络结构

Fig.2 The network structure of the CBAM

第二步继续使用最大池化层和平均池化层分别对 F'进行通道维度的压缩，得到两组 240×320×1 的特征图，将它们进行通道维度的拼接，经过一层卷积核大小为 7×7 的卷积层和 sigmoid 激活函数后，即得到空间注意力特征图 $M_c(F')$，最后将其与 F'进行元素点乘，即得到 CBAM 模块的最终输出特征，其大小为 240×320×32。

通道注意力模块和空间注意力模块的具体计算公式如式(3)和式(4)所示：

$$\begin{aligned} M_c(F)&=\sigma[\mathrm{MLP}(\mathrm{AvgPool}(F))]+\mathrm{MLP}(\mathrm{MaxPool}(F)) \\ &=\sigma[W_1(W_0(F_{\mathrm{avg}}^{c}))]+W_1(W_0(F_{\mathrm{max}}^{c})) \\ F'&=M_c(F)\otimes F \end{aligned} \tag{3}$$

$$\begin{aligned} M_S(F')&=\sigma(f^{7\times7}([\mathrm{AvgPool}(F);\mathrm{MaxPool}(F)])) \\ &=\sigma(f^{7\times7}([F_{\mathrm{avg}}^{s};F_{\mathrm{max}}^{s}])) \\ F''&=M_S(F')\otimes F' \end{aligned} \tag{4}$$

2.3 判别器的构建

本文所采用的判别器结构为文献[13]所提出的中提出的马尔科夫判别器。这种判别器有效地将图像建模为马尔科夫随机场，并假设像素之间的独立性大于补丁（patch）的直径。这种假设也常用于纹理和风格模型。由于我们会通过添加额外的传统损失函数，比如 L1/L2 损失函数来保证图像间的低频正确性（具体见 2.4），所以判别器的目的主要是用来约束图像间高频结构信息，所以在补丁（patch）范围内对结构进行惩罚，被证明是可行的。

判别器的主要流程如下：首先输入由生成器生成的融合图像（或训练集中的可见光图像），对于 480×640×3 的输入，经过 6 层卷积操作后，得到 15×20×1 的特征图矩阵，其中每个像素对应原图像中的 94×94 的补丁，判别器对每个补丁进行判断，最终输出所有结果的均值，用以进行对抗损失的计算。

此外，不同于原文中使用的输入成对图像的条件判断方式，为了让判别器更符合我们的融合任务，仅对判别器输入融合图像，和训练集中的并不与之对应的可见光图像，使得判别器的任务不是判别图像是否相同（若目的为此，对于生成器所输出的融合图像，会减少其所对应的红外信息的引入），而是判别生成的融合图像是否符合可见光图像的特征，以此来保证我们的判别器性能。

判别器的结构如图 3 所示。具体参数如表 2 所示。归一化方法采用 Instance Normalization，激活函数采用斜率为 0.2 的 LeakyReLU。

2.4 损失函数的构建

为了提高 GANs 生成图像的真实性，许多研究尝试将 GANs 的对抗损失和其他传统损失结合起来。添加传统损失后，判别器的目的不变，而生成器的目的不仅是为了生成可以混淆判别器的融合图像，而且还负责约束融合图像和源图像内容上的相似性，保证低频正确性。我们把这部分损失称为内容损失。

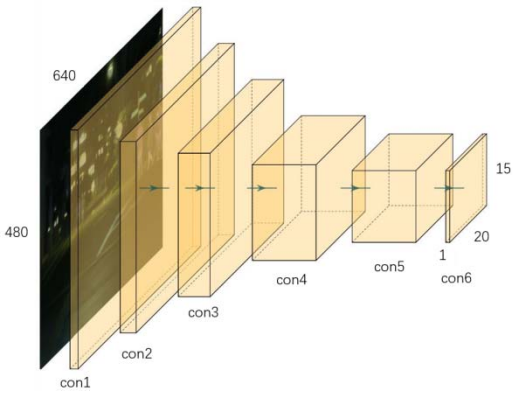


图 3 判别器的网络结构

Fig.3 The network structure of the discriminator

表 2 判别器参数

Table 2 The parameters of discriminator

Convolution layer	Kernel size/stride	Padding	Output size
Conv1	4×4/2	(1,1)	240×320×64
Conv2	4×4/2	(1,1)	120×160×128
Conv3	4×4/2	(1,1)	60×80×256
Conv4	4×4/2	(1,1)	30×40×512
Conv5	4×4/2	(1,1)	15×20×512
Conv6	1×1/1	(0,0)	15×20×1

因此生成器的损失函数由对抗损失函数 $V_{\mathrm{GAN}}(G)$（价值函数）和内容损失函数 L_{con} 两部分组成，如式(5)所示：

$$L_G=V_{\mathrm{GAN}}(G)+L_{\mathrm{con}} \tag{5}$$

对抗损失函数 $V_{\mathrm{GAN}}(G)$采用 2.1 中最小二乘损失函数的第二种形式：

$$\min_G V_{\mathrm{GAN}}(G)=\min V_{\mathrm{LSGAN}}(G)=\frac{1}{2}E[D(G(v,i)-1)^2] \tag{6}$$

其中输入不是噪声 z，而是可见光图像和红外图像的通道拼接图(v,i)；E 表示进行期望计算。。

对于可见光图像和红外图像融合任务，需要分别设计对应的内容损失函数。L1 损失函数常被用于风格迁移任务中，是一个合理的用于约束可见光图像相似性的选择。而红外图像的特征主要为像素强度，可以使用 Frobenius 范数来约束融合图像的灰度图和红外图像的像素强度相似性。

选择内容损失函数如式(7)所示：

$$L_{\text{con}} = E_{\text{wh}}[\eta_1 \left\| G(v,i)_{\text{gray}} - i \right\|_{\text{F}} + \eta_2 \left\| G(v,i) - v \right\|_{\text{L1}}] \quad (7)$$

式中：$G(v,i)$表示融合图像；$G(v,i)_{\text{gray}}$ 表示将融合图像转化为单通道灰度图，因为红外图像为单通道图，要计算两类图像的 Frobenius 范数，需要控制通道数统一；E_{wh} 表示将所得范数在长宽方向求平均；η_1、η_2 为控制两类损失的权重；$\|A\|_{\text{F}}$ 和 $\|A\|_{L1}$ 表示 Frobenius 范数和 L1 范数，计算方式如式(8)和式(9)所示：

$$\|A\|_{\text{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \quad (8)$$

$$\|A\|_{L1} = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}| \quad (9)$$

在融合框架中判别器的目的是用来判断生成的图像是否符合可见光域的图像，通过和生成器的对抗博弈，为融合图像添加更多的高频信息，例如纹理、细节和颜色信息。为了与生成器的对抗损失函数统一，判别器的损失函数如式(10)所示：

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}E[D(v)-1)^2] + \frac{1}{2}E[D(G(v,i))^2] \quad (10)$$

式中：$D[G(v,i)]$表示融合图像的判别结果；$D(v)$表示可见光图像的判别结果。

3 实验与分析

实验所采用的计算机硬件配置为：Inter Core i9-9900k CPU，NVIDIA RTX 2080Ti 11GB GPU。本文所提出的基于生成对抗网络的融合方法，采用 PyTorch 深度学习框架搭建。

3.1 训练参数设置

训练图像和测试图像的大小为 480×640×4，判别器的学习率设为 0.004，生成器的学习率设置为 0.001，这样在训练过程中生成器和判别器可以使用 1:1 的更新间隔。优化器选择自适应矩估计优化器，红外内容损失权重η_1设置为 0.33，可见光内容损失权重η_2设置为 100，批大小（Batch Size）设置为 4，训练周期（epoch）设置为 50。

3.2 融合图像性能对比实验

为了验证本文融合方案的有效性，将其与 7 种先进的图像融合方法进行了对比。其中 3 种为经典的多尺度变换方法，包括拉普拉斯金字塔变换（LP）、双数复小波变换（DTCWT）和非下采样轮廓波变换（NSCT），以及将以上 3 种方法和稀疏表示相结合的混合方法，分别为 LP_SR，DTCWT_SR，NSCT_SR。最后一种为基于神经网络的融合方法 DenseFuse。由于以上方法的设计都只局限于单通道灰度图像的合成，因此，在使用上述方法进行三通道可见光图像和单通道红外图像融合时，进行了部分修改。我们将用于测试的可见光图像分解为 3 个单通道图，分别和对应的单通道红外图像进行融合，最后将 3 个通道融合图进行组合得到三通道彩色融合图像。而本文提出的方法可以直接输入可见光图像和红外图像对，无需进行其他方法所需要的分解组合操作。

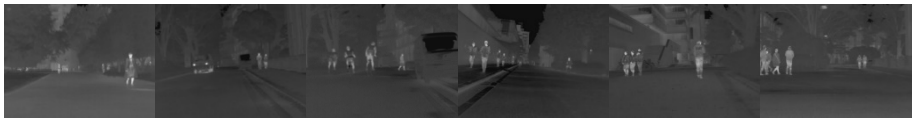
在图 4 中给出了有代表性的 6 组融合图像。观察实验结果可以得到以下结论：首先可以看到同样是基于 GANs 的方法的 FusionGAN，由于其网络结构的限制，在使用本文实验所采用的数据集进行训练时，出现了模式崩塌，造成融合图像大幅度失真，只能在一定程度上看到其有所突出红外目标。而其他方法都实现了将红外目标信息融合进可见光图像中，提高了目标（行人）的辨识度，同时对于强光环境下无法在可见光图像中辨识的车辆，在融合图像中也能进行辨识，如第二张测试图所示。但在细节和纹理上依然存在不同问题。

由于红外图像的灰度值在低照度条件下，普遍高于可见光图像，因此传统方法的融合结果评价指标比较局限于其固定的融合规则，除了保留目标信息的高灰度值，同样会将其余具有高灰度值的无效信息融合进图像中，从而产生肉眼可见的伪影。观察第 4 张测试图所对应的融合图像中的天空部分，除了我们的方法外，其余方法都出现明显分界线。此外还可以观察到，LP_SR、DTCWT_SR、NSCT_SR 这些混合方法，比之 LP、DTCWT、NSCT，虽然使图像目标信息更加突出，但其边缘出现了明显失真和大范围模糊。DenseFuse 的融合效果视觉上比传统方法较好，但在背景部分依然存在少许伪影。本文提出的方法则有效规避以上问题（本文方法称为 CBAM-GAN）。由于本文设计的网络目的是往低照度图像中引入红外图像中的显著目标信息，因此可以有效避免引入无效的红外信息，从而减少伪影的产生。观察本文算法的融合图像可以发现，其背景部分几乎保留了可见光图像中的背景，虽然这会使得其在整体亮度上低于其他融合图像，但目标信息是高亮显示的，且相较于其余方法的高亮部分以灰色输出，图中的高亮部分带有一定的红色，也有助于视觉感知。因为使用生成对抗网络进行训练时，其引入的对抗损失会在原则上意识到灰色输出是不现实的，并鼓励匹配真实的颜色输出。此外，为了验证注意力模块的有效性，我们去掉了生成器中的 CBAM 模块，其余结构保持不变，重新进行

模型训练和实验，实验结果为图 4 中 NORM-GAN 栏所对应的 6 张图像。可以观察到，和 CBAM-GAN 相比，背景部分没有明显差异，但对于红外目标而言，其亮度、色彩鲜艳度、对比度都低于 CBAM-GAN 生成的融合图像，说明了注意力模块在促使融合图像引入显著红外信息时起到积极作用。

最后，为了客观地评价图像融合的性能，我们使用了 6 个融合指标进行了定量比较。6 个融合指标分别为信息熵（EN）、互信息（MI）、特征互信息（FMI）、结构相似性（SSIM）、相关系数（CC）、峰值信噪比（PSNR）。7 种融合方法得到的 20 幅融合图像的 6 个指标的平均值如表 3 所示。

(a) Visible image



(b) Infrared image

(c) LP

(d) LP-SR

(e) NSCT

(f) NSCT-SR

(g) DTCWT

(h) DTCWT-SR

(i) DenseFuse

(j) NORM-GAN(Only for qualitative comparison of the effectiveness of the attention module)

(k) CBAM-GAN(Proposed in this paper)

图 4 融合图像对比

Fig.4 The comparison results of fusion images

表 3 融合图像客观指标值

Table 3 The quantitative comparisons of fusion images

Fusion methods	EN	MI	FMI	SSIM	CC	PSNR
LP	5.918	11.836	0.944	0.681	0.646	68.496
LP-SR	6.393	12.785	0.945	0.823	0.566	67.801
NSCT	5.821	11.643	0.942	0.671	0.652	68.575
NSCT-SR	6.224	12.447	0.940	0.859	0.575	67.472
DTCWT	5.804	11.608	0.942	0.670	0.647	68.570
DTCWT-SR	6.455	12.910	0.945	0.782	0.525	67.338
DenseFuse	6.036	12.071	0.939	0.631	0.684	67.319
CBAM-GAN	5.918	11.836	0.928	0.796	0.649	68.751
Avarage	6.111	12.223	0.941	0.740	0.606	67.967

需要指出的是，这些指标在用于评价可见光和红外图像融合性能的时候，主要用于灰度图像的，且是综合融合图像相对于可见光以及红外图像的相似性。因此，在进行指标计算时，需要将彩色融合图像以及对应的可见光图像转化为灰度图，进行计算。由此过程可见，这些指标仅考虑了灰度层面的比较，而忽视了颜色层面的比较，对于本文实验的评价具有一定的片面性。此外，由于本文算法的融合图像关注于融合红外图像中的显著目标信息，而尽可能减少融合红外图像中的无效信息，因此相对于其他平均融合两类源图像的融合方法，本文算法的融合图像在和红外图像的相关度上是低于其他融合图像的，造成指标值只能处于平均水平。但也表明了我们的融合方法在灰度层面的指标评价也能达到目前先进融合方法的水准。其中 SSIM，CC，PSNR 三项值都在平均值以上，说明本文的融合图像在保留源图像的结构信息方面性能良好，具有更少的失真。

4 结论

本文提出了一种基于生成对抗网络的可见光和红外图像融合方法，该方法不局限于灰度层面的融合，可以直接用于 RGB 三通道的可见光图像和单通道红外图像的融合，是一种端到端的图像融合模型。此外，在网络中引入注意力模块，使得融合图像在维持可见光图像中的背景信息的同时，可以突出目标信息，这有助于对低照度条件下的源图像进行融合，生成整体更干净，伪影更少的融合图像。通过在 6 项指标上和其他融合方法进行实验对比，结果表明了本文提出的方法得到的融合图像具有良好的融合性能和视觉效果。

参考文献：

[1] MA J, MA Y, LI C. Infrared and visible image fusion methods and applications: a survey[J]. *Information Fusion*, 2019, **45**: 153-178.

[2] Burt P J, Adelson E H. The Laplacian pyramid as a compact image code[J]. *Readings in Computer Vision*, 1987, **31**(4): 671-679.

[3] Selesnick I W, Baraniuk R G, Kingsbury N C. The dual-tree complex wavelet transform[J]. *IEEE Signal Processing Magazine*, 2005, **22**(6): 123-151.

[4] A L da Cunha, J Zhou, M N Do. Nonsubsampled contourilet transform: filter design and applications in denoising[C]//*IEEE International Conference on Image Processing* 2005, **749**: (doi: 10.1109/ICIP.2005.1529859).

[5] Hariharan H, Koschan A, Abidi M.The direct use of curvelets in multifocus fusion[C]//16*th IEEE International Conference on Image Processing* (ICIP), 2009: 2185-2188(doi: 10.1109/ICIP.2009.5413840).

[6] LI Hui. Dense fuse: a fusion approach to infrared and visible images[C]//*IEEE Transactions on Image Processing*, 2018, **28**: 2614-2623(doi: 0.1109/TIP.2018.2887342).

[7] MA J, YU W, LIANG P, et al. Fusion GAN: a generative adversarial network for infrared and visible image fusion[J]. *Information Fusion*, 2019, **48**: 11-26.

[8] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C]//*International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015: 234-241.

[9] Hwang S, Park J, Kim N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 1037-1045.

[10] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//*Advances in Neural Information Processing Systems*, 2014: 2672-2680.

[11] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J/OL] [2015-11-07]. arXiv preprint arXiv:1511.06434, 2015: https://arxiv.org/abs/1511.06434v1.

[12] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]//2017 *IEEE International Conference on Computer Vision* (ICCV), 2017: 2813-2821(doi: 10.1109/ICCV.2017.304).

[13] Isola Phillip, ZHU Junyan, ZHOU Tinghui, et al. Image-to-image translation with conditional adversarial networks, 2017: 5967-5976 (doi:10.1109/CVPR.2017.632).

[14] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks [C]//*Advances in Neural Information Processing Systems*, 2015: 2017-2025.

[15] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 7132-7141.

[16] Woo S, Park J, Lee J Y, et al. Cbam: convolutional block attention module[C]//*Proceedings of the European Conference on Computer Vision* (*ECCV*), 2018: 3-19.