文章编号　1004-924X(2023)24-3651-11

非成对训练样本条件下的红外图像生成

蔡　伟，姜　波*，蒋昕昊，杨志勇
(火箭军工程大学 兵器发射理论与技术国家重点学科实验室，陕西 西安 710025)

摘要：针对实测红外图像数据集构建难度大、测试制作成本高的问题，本文提出了一种非成对训练样本条件下的生成对抗网络VTIGAN，实现不同场景下可见光到红外(Visible-to-Infrared，VTI)的高质量图像转换。VTIGAN在transformer模块基础引入了一个新的生成器学习图像内容映射关系，通过重组目标风格特性实现图像风格转换，同时使用PathGAN作为判别器强化模型的图像细节信息生成能力，最后联合对抗损失、多层对比损失、风格相似性损失、同一性损失四种损失函数对模型训练过程进行约束。将VTIGAN与其他主流算法在可见光-红外数据集上进行了广泛的对比实验，同时使用峰值信噪比、结构相似度和Frechét inception distance 3个评价指标进行定量评估和主观定性评价。实验结果表明，VTIGAN相比于次优的UGATIT算法在PSNR，SSIM和FID三个指标上分别提升了3.1%，2.8%和11.3%，有效实现了非成对训练样本条件下可见光到红外的图像转换，且对于复杂场景的抗干扰能力更强，生成的红外图像清晰度高、细节特征完整、真实感强。

关　键　词：图像处理；红外图像仿真；图像风格迁移；生成对抗网络；Transformer
中图分类号：TP391　　**文献标识码**：A　　**doi**：10.37188/OPE.20233124.3651

Infrared image generation with unpaired training samples

CAI Wei，JIANG Bo*，JIANG Xinhao，YANG Zhiyong

(*Armament Launch Theory and Technology Key Discipline Laboratory of PRC*，
Rocket Force University of Engineering，*Xi'an* 710025，*China*)
* *Corresponding author*，*E-mail*：*jiang*20202033@163.*com*

Abstract：To address the difficulties in constructing an infrared image dataset from measurements and the high cost of testing and production，this study proposes a generation countermeasure VTIGAN for unpaired training samples to achieve high-quality image conversion from visible-to-infrared in different scenarios. VTIGAN introduces a new generator inspired by the transformer module to learn the mapping relationship of image content，whereby image style conversion is realized by reorganizing the characteristics of the target style. PathGAN is used as a discriminator to strengthen the image detail information generation ability of the model. Finally，four loss functions，namely，resistance，multi-layer contrast，style similarity，and identity losses are combined to constrain the model training process. VTIGAN was compared with other mainstream algorithms in a wide range of experiments on visible infrared datasets. Three evaluation indicators，namely，peak signal-to-noise ratio (PSNR)，structural similarity (SSIM)，and Fréchet inception distance (FID)，were used for quantitative and subjective qualitative evaluation. The experimental results show that VTIGAN improves PSNR，SSIM，and FID by 3.1%，2.8%，and 11.3%，respective-

收稿日期：2022-06-28；**修订日期**：2022-07-26.
基金项目：国防科技173基础加强计划技术领域基金资助项目(No. 2021-JCJQ-JJ-0871)

ly, compared with the suboptimal UGATIT algorithm, which effectively realizes image conversion from visible light to infrared under the condition of unpaired training samples, demonstrating stronger anti-interference ability for complex scenes. The generated infrared images have high definition, complete details, and strong realism.

Key words: image processing; infrared image simulation; image style transfer; generative adversarial network; transformer

1 引　言

红外成像技术利用红外探测器获取目标与环境的热辐射差异,通过光电转换技术形成可观测的红外图像,具有抗干扰能力强、探测距离远、可全天候工作等优势,因此在军事侦察、公共安防、医疗成像等方面发挥了重要作用。但在红外成像技术的发展过程中,经常需要丰富多样的红外图像数据作为验证测试的样本。同时基于深度学习的算法也必须使用大量的红外数据对模型进行训练,从而确保模型的训练效果。然而,在许多场景中,由于测试环境、搭载设备等客观因素的限制,采用实拍红外图像的获取方式是十分困难的,构建大规模的红外数据集成本极高。因此利用红外图像仿真技术生成高质量的红外数据,不仅能够有效地降低获取红外数据的成本,还可以生成很多自然环境以及外场试验难以获得的红外数据,这些数据将为红外相关设备在军事、民用领域的应用提供有力支撑。

Vega/VegaPrime, SE-Workbench, MuSES, DIRSIG等传统的红外仿真技术[1-3]利用人工构建的三维模型计算每个单位的热辐射,模拟真实场景下的红外数据。但这些方法的计算框架较为复杂,需要大气温度、相对湿度、风速等诸多额外的环境信息。而利用图像风格迁移的方法将种类多样的可见光图像直接转换为红外图像相比于传统的红外仿真方法更加快捷方便。近几年,随着深度学习技术的迅速崛起,生成对抗网络在图像生成和转换、信息安全、视觉计算等领域展现出显著的应用价值[4-6]。基于生成对抗网络的图像风格迁移模型取得了巨大成功。Pix2pix[7]首次使用U-Net作为生成器,PathGAN(Generative Adversarial Network)作为判别器,实现了语义图到真实场景图、卫星图到地图、日景图到夜景图的转换,但Pix2pix对训练样本要求严格,必须使用成对图像作为输入。然而,针对图像跨模态转换问题,很难采集和建立充足的成对训练数据,因此Pix2pix在实际应用中具有一定局限性。而CycleGAN,DiscoGAN,DualGAN等[8-10]方法利用一种循环一致性损失函数对非成对图像进行训练,实现了四季场景转换、照片与艺术画互转。但基于循环一致性损失函数的方法通常要求图像之间的双向映射,这种限制严格的双向映射对于模态差异较大的图像转换效果并不理想,尤其面对复杂场景的红外图像转换时模型的性能受到极大影响。如图1所示,CycleGAN在进行可见光图像到红外图像转换时,由于双向映射的严格要求,反而造成了图中红色方框内的生成结果失真(彩图见期刊电子版)。因此,如何在非成对样本训练的条件下高质量的实现红外图像的转换,对于红外数据集的构建具有重要的研究价值。

基于以上分析,本文提出了一种使用非成对样本进行训练的可见光图像到红外图像转换的生成对抗网络(Visible to Infrared Generative Adversarial Network, VTIGAN),有效实现了红外图像仿真生成。VTIGAN相比于循环一致性损失的方法而言,并不要求模态间的双向映射,而是引入多层对比损失和风格相似性损失约束网络训练,重点关注可见光图像到红外图像的单向映射,因此不仅降低了网络复杂度,红外图像的生成质量也更高。本文的贡献总结如下:

(1)针对可见光图像到红外图像转换时模态差异较大,现有算法使用非成对数据训练效果不佳的问题,本文提出了一种新颖的红外图像仿真生成算法VTIGAN,实现非成对训练样本下可见光图像到红外图像的转换;

(2)在生成对抗网络的基础上引入多层对比损失、风格相似性损失和同一性损失约束网络模

型的训练,在图像转换过程中最大程度保留输入的可见光图像语义内容不变,同时生成逼真的红外风格特征;

(3)在公开的可见光-红外数据集上进行了大量实验,证明了 VTIGAN 生成的红外图像在定量评价和视觉效果上的优越性,相比于其他算法更加适合非成对训练样本条件下可见光图像到红外图像的转换。

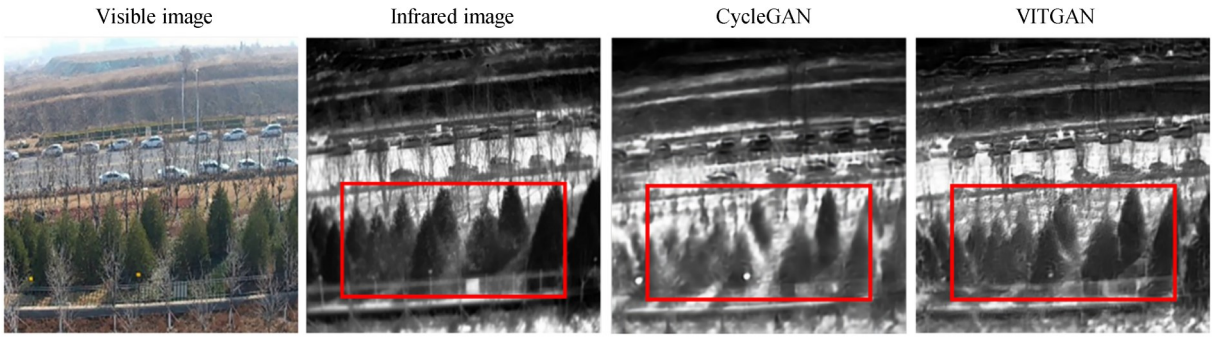

图 1　两种红外仿真算法生成图像效果对比

Fig. 1　Comparison of image effects generated by two infrared simulation algorithms

2　算法原理

假设可见光域为 $\chi\subset R^{H\times W\times C}$、红外域为 $\gamma\subset R^{H\times W\times C}$,给定非成对可见光图像数据集和红外图像数据集分别为 $X=\{x\in\chi\}$, $Y=\{y\in\gamma\}$。本文的目的是通过 VTIGAN 学习一个映射 G: $X\rightarrow Y$,从而实现可见光图像到红外图像的转换。VTIGAN 包含一对生成器和判别器 $\{G,D_Y\}$,生成器 G 完成 $X\rightarrow Y$ 的映射,判别器 D_Y 负责确保转换后的图像属于红外域。生成器 G 由编码器、转换器、解码器组成,分别表示为 G_{enc}, G_{con}, G_{dec},生成红外图像 $\tilde{y}$ 的过程如式(1)所示:

$$\tilde{y}=G(x)=G_{dec}\left(G_{con}\left(G_{enc}(x)\right)\right). \tag{1}$$

在图像转换过程中,VTIGAN 从编码器中提取输入图像的多层特征信息,并使用一个双层 MLP 网络[11](H_X 和 H_Y)将提取到的特征信息投影到共享的嵌入空间,从而计算输入和输出图像间的多层对比损失。此外,VTIGAN 还引入了两个轻量的特征映射网络 $\{H_f,H_r\}$,提取仿真红外图像和真实红外图像间的公共信息,并以此构成风格相似性损失。特征映射网络 $\{H_f,H_r\}$ 依次由卷积层、ReLU 激活函数、平均池化层、双层点卷积级联而成。

图 2 显示了 VTIGAN 的整体网络架构。

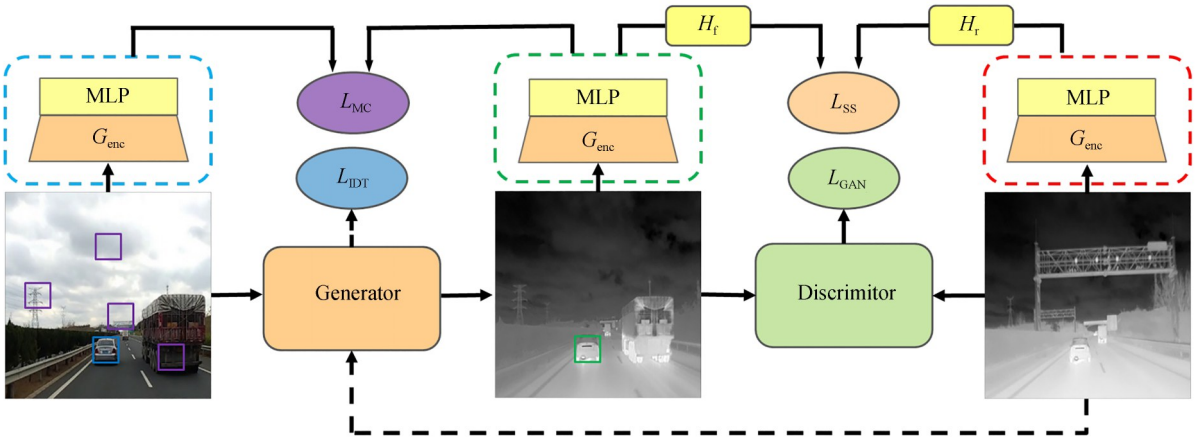

图 2　VTIGAN 的基本框架

Fig. 2　Basic framework of VTIGAN

VTIGAN 主要结合了对抗损失、多层对比损失、风格相似性损失以及同一性损失实现网络模型的训练。下面将从网络结构和损失函数两个方面详细介绍本方法的具体原理。

2.1　生成器

VTIGAN 利用卷积神经网络和 transformer[12]的组合，构建了一种新颖的生成器，主要由编码器、转换器、解码器三部分组成。具体网络结构如图 3 所示。

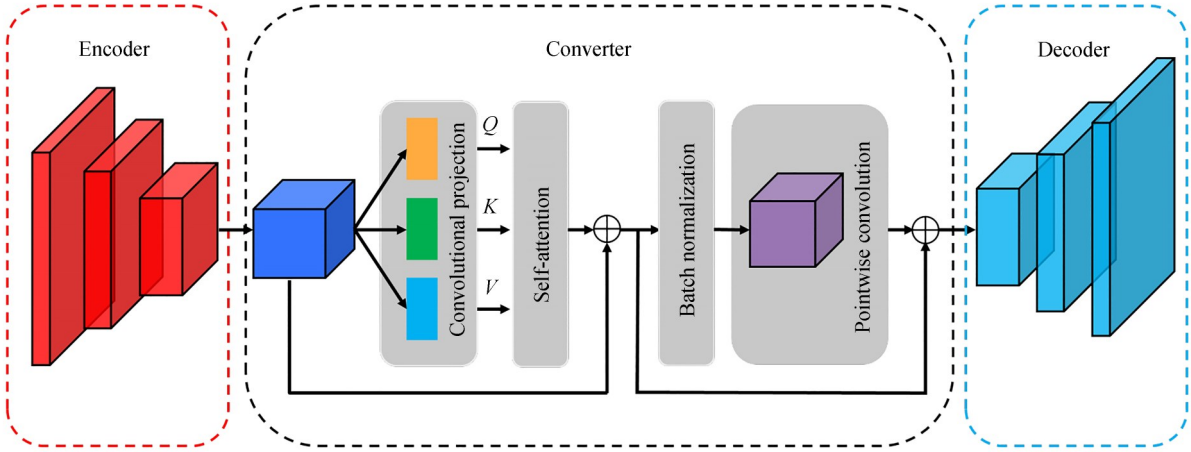

图 3　生成器网络结构

Fig. 3　Network structure of generator

2.1.1　编码器

编码器主要利用卷积神经网络对图像进行特征提取，便于之后的转换器进行相应的转换。第 1 个卷积层采用 64 个尺度为 7×7 的卷积核提取特征，并加入 3 个像素的填充保留特征图的空间维度。接下来的两个卷积层分别由 128 和 256 个卷积核组成，所有卷积核尺度为 3×3，步幅设置为 2，填充设置为 1。每个卷积层后面还包括一个批量归一化层和 Relu 激活函数。设 $T\in R^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ 表示编码器输出的张量，则输入尺寸为 256×256×3 的可见光图像，输出 T 的尺寸为 64×64×256。

2.1.2　转换器

转换器利用 transformer 架构在潜在空间对深度特征进行解纠缠并重新组合，实现可见光到红外的风格转换。为方便后续的矩阵运算，首先使用卷积投影模块(conv_projection)将 T 映射为三组向量，这些向量称为查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$。投影模块主要由深度可分离卷积[13]构成，其中特别的是生成 $\boldsymbol{Q}$ 的投影模块使用步幅为 1 的深度卷积，其他投影模块的深度卷积步幅为 2。设 W_Q，W_K，W_V 分别为投影模块的训练参数，则三组向量的学习过程如式(2)所示：

$$\begin{aligned}\boldsymbol{Q}&=\mathrm{reshape}\left[\mathrm{Depthwise}\left(T,W_Q\right)\right]\\ \boldsymbol{K}&=\mathrm{reshape}\left[\mathrm{Depthwise}\left(T,W_K\right)\right],\\ \boldsymbol{V}&=\mathrm{reshape}\left[\mathrm{Depthwise}\left(T,W_V\right)\right]\end{aligned}\tag{2}$$

其中：Depthwise(·)表示深度可分离卷积，reshape(·)表示矩阵重组。向量 $\boldsymbol{Q}\in R^{s_q\times t_q}$，$\boldsymbol{K}\in R^{s_k\times t_k}$，$\boldsymbol{V}\in R^{s_v\times t_v}$，$s_q$=4 096，$s_k$=$s_v$=1 024，$t_q$=$t_k$=$t_v$=256。

卷积投影模块之后设置了一个自注意力模块(Self-Attention)。在此模块中，查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 通过矩阵运算对全局特征进行解纠缠，并生成自注意力矩阵更新值向量 $\boldsymbol{V}$ 中的特征信息。此过程如式(3)所示：

$$\boldsymbol{\Theta}=\mathrm{reshape}\left[\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{t_q}}\right)\cdot\boldsymbol{V}\right]+T,\tag{3}$$

其中：$\boldsymbol{\Theta}\in R^{64\times64\times256}$，$\boldsymbol{K}^{\mathrm{T}}$ 为键 $\boldsymbol{K}$ 的转置，softmax(·)表示归一化函数，reshape(·)表示矩阵重组。

与常规的 transformer 架构不同，为进一步降低计算量，本文使用双层点卷积代替 MLP 作为 transformer 中的最后处理模块。转换器的最终

输出如式(4)所示:

$$\boldsymbol{\Theta}^{*}=\Gamma_{\mathrm{BN}}\left[\Gamma_{\mathrm{conv512}}\left(\Gamma_{\mathrm{GELU}}\left(\Gamma_{\mathrm{conv256}}(\boldsymbol{\Theta})\right)\right)\right]+\boldsymbol{\Theta},\quad(4)$$

其中:$\Gamma_{\mathrm{conv256}}$和$\Gamma_{\mathrm{conv512}}$分别表示卷积核数量为256和512的卷积层,$\Gamma_{\mathrm{GELU}}$表示高斯误差线性单元激活函数,$\Gamma_{\mathrm{BN}}$为批量归一化层。

2.1.3　解码器

解码器的构成与编码器恰好相反,采用3层逐级连接的反卷积,将转换器输出的特征图逐级还原为红外图像,完成整个生成过程。编码器、转换器、解码器的结构参数详见表1。

表1　生成器内部参数

Tab. 1　Generators internal parameters

Module	Encoder			Convert		Decoder		
	Convolution kernel	Network Connection	Number of convolution kernels	Module	Number	Convolution kernel	Network Connection	Number of convolution kernels
Instruction	7×7	Conv ,BN ,ReLU	64	Transformer	1	3×3	ConvTranspose ,BN ,ReLU	128
	3×3	Conv ,BN ,ReLU	128			3×3	ConvTranspose ,BN ,ReLU	64
	3×3	Conv ,BN ,ReLU	256			7×7	Conv ,BN ,ReLU	3

2.2　判别器

在一般的生成对抗网络中,判别器通常由多个卷积层构成,最终输出一个介于0~1之间的标量作为判定样本真伪的概率。这种判别方式针对整张图像进行加权得出结果,无法关注到图像的局部特征。由于可见光图像与红外图像模式差异大,此种判别方法精度要求过低无法实现高质量的图像生成。因此,本文使用70×70的PathGAN[6]作为VTIGAN的判别器,具体结构如图4所示,其输入为256 pixel×256 pixel的图像,输出结果为30×30的矩阵,每个矩阵元素代表输入图像中一个70×70感受野的图像补丁为真的概率,最终以全部矩阵元素的均值判定整张图像的真假。这种补丁级判别器针对每一个图像补丁精准判别,重点关注了图像局部特征的提取和表征,更加适合可见光到红外这两种模式差异较大的图像转换。

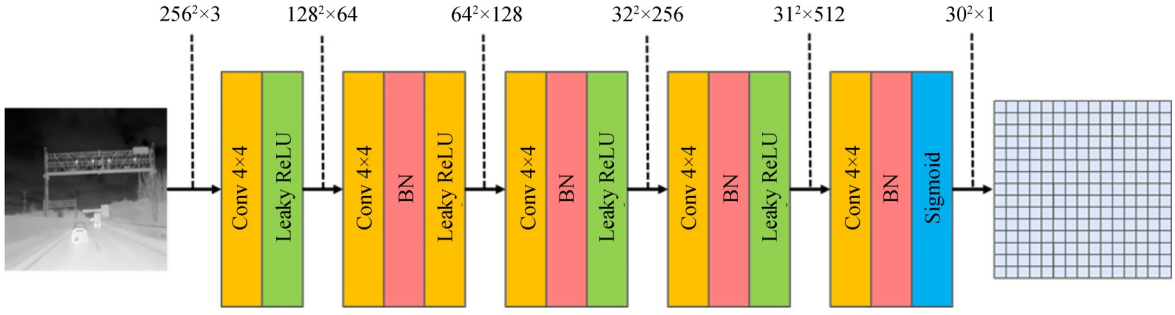

图4　判别器网络结构

Fig. 4　Network structure of discriminator

2.3　损失函数

2.3.1　对抗损失

在传统生成对抗网络中通常使用交叉熵损失函数,这会导致生成的图像虽然被判定为真实图像,但实际仍远离判别器的决策边界。由于此时交叉熵损失函数已经拟合,生成器不会继续进行优化,最终限制了生成的图像质量。通过理论分析和实际测试发现最小二乘损失函数可以对判定为真实样本的图像进行惩罚,使距离决策边界较远的样本不断靠近决策边界,从而进一步提

高图像转换的效果。因此，VTIGAN选择最小二乘损失函数构建对抗损失，其表达形式如式(5)～式(7)所示：

$$L_{D_Y}=\frac{1}{2}E_{y}\left\{\left[D_{Y}(y)-1\right]^{2}\right\}+\frac{1}{2}E_{\hat{y}}\left\{\left[D_{Y}(\hat{y})\right]^{2}\right\},\tag{5}$$

$$L_{G}=\frac{1}{2}E_{\hat{y}}\left\{\left[D_{Y}(\hat{y})-1\right]^{2}\right\},\tag{6}$$

$$L_{\mathrm{GAN}}=L_{G}+L_{D_Y},\tag{7}$$

其中：y为真实红外图像，$\hat{y}$为生成红外图像，$E_{\xi}(\cdot)$表示计算期望，$D_{Y}(\cdot)$表示判别器判定图像为真的概率。

2.3.2　多层对比损失

图像转换过程中，要求输入图像和输出图像在对应空间位置上应具有相似的语义信息。因此VTIGAN引入多层对比损失最大化输入和输出图像对应位置间的互信息。通过将可见光和红外图像对应位置互信息的最大化，可以监督生成网络提取两个对应位置的共性部分(图像转换时应该被保留的内容)，同时忽略其他位置上的特征对图像转换的影响，具体实现如下。

多层对比损失的目的是关联查询样本和所对应的正样本，并排除数据中的其他负样本。如图2所示(彩图见期刊电子版)，设定查询样本为生成的红外图像中随机选取的绿色图像框，正样本则为可见光图像中相同位置所对应的蓝色图像框，而可见光图像中除正样本位置外随机位置选取的紫色图像框均为负样本。首先将查询样本、正样本和N个负样本映射为P维向量，分别表示为$\boldsymbol{q}\in\boldsymbol{R}^{P}$，$\boldsymbol{\nu}^{-}\in\boldsymbol{R}^{N\times P}$，$\boldsymbol{\nu}^{+}\in\boldsymbol{R}^{P}$，接着对$P$维向量进行L2正则化，此时可构建一个$(N+1)$类的分类问题，并计算排除负样本选出正样本的概率。在数学上可使用交叉熵损失[13]表示为：

$$\Gamma\left(\boldsymbol{q},\boldsymbol{\nu}^{-},\boldsymbol{\nu}^{+}\right)=-\log\left[\frac{\exp\left(\boldsymbol{q}\cdot\boldsymbol{\nu}^{+}/\delta\right)}{\exp\left(\boldsymbol{q}\cdot\boldsymbol{\nu}^{+}/\delta\right)+\sum_{n=1}^{N}\exp\left(\boldsymbol{q}\cdot\boldsymbol{\nu}^{-}/\delta\right)}\right],\tag{8}$$

其中，δ为查询样本和其他样本间的距离缩放因子。

VTIGAN使用G_{enc}和双层MLP的组合分别对可见光和红外图像进行特征提取，同时为了更加精确的进行特征表示，可见光和红外图像的特征提取网络之间并不共享权重。以可见光图像为例，选取$G_{\mathrm{enc}}(\boldsymbol{x})$中的L层传递到MLP中，将其投影为特征集合，如式(9)所示：

$$\left\{\boldsymbol{w}_{l}\right\}_{L}=\left\{H_{X}^{l}\left[G_{\mathrm{enc}}^{l}(\boldsymbol{x})\right]\right\}_{L},\tag{9}$$

其中，$G_{\mathrm{enc}}^{l}(\boldsymbol{x})$表示第$l$层的输出。对选定层中的空间位置信息记作$t\in\{1,\cdots,T_l\}$，此时正样本特征可表示为$\boldsymbol{w}_{l}^{t}\in\boldsymbol{R}^{C_l}$，负样本特征为$\boldsymbol{w}_{l}^{T\setminus t}\in\boldsymbol{R}^{(T_l-1)C_l}$，其中$T_l$为每一层的空间位置数、$C_l$为每一层的通道数。同理，对生成的红外图像进行特征提取可得到一个新的特征集合，如式(10)所示：

$$\left\{\hat{\boldsymbol{w}}_{l}\right\}_{L}=\left\{H_{Y}^{l}\left[G_{\mathrm{enc}}^{l}\left(G(\boldsymbol{x})\right)\right]\right\}_{L}.\tag{10}$$

由上述公式可得到最终输入与输出图像间的多层对比损失，如式(11)所示：

$$L_{\mathrm{MC}}=E_{x}\left[\sum_{l=1}^{L}\sum_{t=1}^{T_l}\Gamma\left(\hat{\boldsymbol{w}}_{l},\boldsymbol{w}_{l}^{t},\boldsymbol{w}_{l}^{T\setminus t}\right)\right].\tag{11}$$

2.3.3　风格相似性损失

在红外图像数据集$Y=\{\boldsymbol{y}\in\gamma\}$中，不同红外图像虽然语义内容各不相同，但具有相似的红外风格特征。为了提高生成红外图像的真实性，VTIGAN使用风格相似性损失充分挖掘生成图像和真实样本之间的联系，最大程度上实现红外风格特征的转换。如图1所示，在网络中首先使用编码器G_{enc}和双层MLP将生成的红外图像和真实红外图像转换为两组特征集合，接着利用两个轻量的特征映射网络$\{H_{\mathrm{f}},H_{\mathrm{r}}\}$将特征集合映射为64维的风格特征向量，此时利用风格相似性损失度量二者之间距离，其形式化表达如式(12)所示：

$$L_{\mathrm{SS}}=\left\|H_{\mathrm{r}}\left(H_{Y}\left(G_{\mathrm{enc}}(\boldsymbol{y})\right)\right)-H_{\mathrm{f}}\left(H_{Y}\left(G_{\mathrm{enc}}\left(G_{\mathrm{enc}}(\boldsymbol{x})\right)\right)\right)\right\|_{1}.\tag{12}$$

通过在深层信息上约束生成图像和真实图像间风格特征的相似，可以促使生成红外图像更加真实，而不是简单的色彩叠加。

2.3.4　同一性损失

生成器以可见光图像作为输入，可以实现可见光图像到红外图像的映射，然而若将红外图像作为输入，理想的生成器将不进行任何更改而直

接输出原始图像。这种以红外图像直接作为输入，得到的输出图像与输入图像之间的L1损失被定义为同一性损失。其表达式如式(13)所示：

$$L_{\mathrm{IDT}}=E_{y}\left[\left\|G(\boldsymbol{y})-\boldsymbol{y}\right\|_{1}\right]. \tag{13}$$

同一性损失的加入可以在训练过程中纠正生成器的偏差，激励生成器生成更加真实的红外图像的特征。

2.3.5　总损失函数

上述各种损失函数都会对红外仿真图像生成效果产生影响，将各项损失进行加权，最终构成VTIGAN的整体损失函数，如式(14)所示：

$$L_{\mathrm{total}}=\lambda_{\mathrm{GAN}}\left(L_{D_{Y}}+L_{G}\right)+\lambda_{\mathrm{MC}}L_{\mathrm{MC}}+\lambda_{\mathrm{SS}}L_{\mathrm{SS}}+\lambda_{\mathrm{IDT}}L_{\mathrm{IDT}}, \tag{14}$$

其中，λ是调整各项损失函数在训练中占比的权值。

3　实验结果与分析

3.1　准备工作

本文算法的有效性检验在Pytorch深度学习开发框架下进行，硬件平台配置如表2所示。

表2　实验平台配置

Tab. 2　Experimental platform configuration

Names	Related configurations
GPU	NVIDIA Quadro GV100
CPU	Inter Xeon Silver 4210/128 G
GPU Memory size	32 G
Operating systems	Win10
Computing platform	CUDA11.0

3.1.1　数据集制备

本文使用的数据集为艾睿光电红外开源平台提供的可见光-红外数据集。该数据集内容丰富，包含了车辆、建筑、植物等多种物体的可见光和红外图像，图像尺寸均为256 pixel×256 pixel。为充分验证算法的有效性，本实验将该数据集划分为交通场景和建筑场景两大类，其中交通场景主要包含车辆和公路的可见光与红外图像，建筑场景主要包含水泥建筑、树木、草坪等物体。两类场景中均含有5 000组图像作为实验样本，其中训练集包含可见光和红外图像各4 500张，且训练过程中图像均采用随机输入的方式，确保在非成对训练样本条件下进行训练。测试集则包含500组配对的可见光和红外图像，其中500张可见光图像作为测试样本，而与之对应的500张红外图像作为参考样本对生成红外图像进行定量评价。

3.1.2　评价指标选取

为比较不同算法的图像迁移效果，本文使用全参考评价方法定量分析生成红外图像质量，主要选取以下三个指标：峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)、结构相似度(Structure Similarity Index Measure，SSIM)和Frechét inception distance(FID)。

PSNR[15]作为使用最广泛的全参考评价指标之一，在图像压缩、超分辨率重建等领域发挥了重要作用。PSNR基于对应像素点间的均方误差衡量生成图像与参考图像之间的差异，其值越大代表生成图像失真越小，即生成图像的质量越高。PSNR的形式化表达如式(15)所示：

$$PSNR=10\cdot\lg\left(\frac{MAX_{\mathrm{I}}^{2}}{\sqrt{MSE}}\right)=20\cdot\lg\left(\frac{MAX_{\mathrm{I}}}{\sqrt{MSE}}\right), \tag{15}$$

其中：MSE表示参考图像和生成图像之间的均方误差，MAX_{I}为图像中的最大像素值。

SSIM[16]用于计算生成图像与参考图像之间的相似性，取值范围在0～1之间，值越大图像相似性越高。SSIM用均值作为亮度的估计，标准差作为对比度的估计，协方差作为结构相似程度的度量，更加符合人类视觉提取图像结构信息的方式。假设参考图像和生成图像分别为$\boldsymbol{x}$和$\boldsymbol{y}$，其结构相似性定义如下：

$$\begin{aligned} l(\boldsymbol{x},\boldsymbol{y})&=\frac{2\mu_{x}\mu_{y}+C_{1}}{\mu_{x}^{2}+\mu_{y}^{2}+C_{1}},\\ c(\boldsymbol{x},\boldsymbol{y})&=\frac{2\sigma_{x}\sigma_{y}+C_{2}}{\sigma_{x}^{2}+\sigma_{y}^{2}+C_{2}},\\ s(\boldsymbol{x},\boldsymbol{y})&=\frac{\sigma_{xy}+C_{3}}{\sigma_{x}\sigma_{y}+C_{3}}, \end{aligned} \tag{16}$$

$$SSIM(\boldsymbol{x},\boldsymbol{y})=[l(\boldsymbol{x},\boldsymbol{y})]^{\alpha}[c(\boldsymbol{x},\boldsymbol{y})]^{\beta}[s(\boldsymbol{x},\boldsymbol{y})]^{\gamma}, \tag{17}$$

其中：$l(\boldsymbol{x},\boldsymbol{y})$，$c(\boldsymbol{x},\boldsymbol{y})$，$s(\boldsymbol{x},\boldsymbol{y})$分别$\boldsymbol{x}$和$\boldsymbol{y}$的亮度比较、对比度比较以及结构比较，$\mu_{x}$和$\mu_{y}$分别为$\boldsymbol{x}$和$\boldsymbol{y}$的均值，$\sigma_{x}$和$\sigma_{y}$分别为$\boldsymbol{x}$和$\boldsymbol{y}$的标准差，$\sigma_{xy}$表

示x和y的协方差，C_1，C_2，C_3，α，β，γ均为常数。

FID[17]通过Inception模型度量生成图像与参考图像在深度特征空间中的距离，具有较好的鲁棒性，其分数越低，说明生成图像的质量越高且多样性丰富。FID的形式化表达如式(18)所示：

$$FID = /\!/ \mu_r - \mu_g /\!/^2 + Tr(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{\frac{1}{2}}), \tag{18}$$

其中：μ_r，μ_g分别为参考图像和生成图像的特征的均值，$\boldsymbol{\Sigma}_r$，$\boldsymbol{\Sigma}_g$分别为参考图像和生成图像的特征的协方差矩阵，Tr表示求矩阵对角线元素和的运算。

3.1.3　实验参数设置

使用动量为0.5的自适应矩估计优化器(Adam)[18]对训练过程中神经网络损失进行优化，初始网络参数随机选自均值为0，方差为0.02的高斯分布。经调试后，实验具体参数设置如表3所示。

表3　实验参数设置

Tab. 3　Setting of experimental parameters

Parameter	Epoch	Batch size	Learning rate	λ_{GAN}	λ_{MC}	λ_{SS}	λ_{IDT}
Value	300	1	0.000 2	1	2	10	1

3.2　实验结果分析

为了验证VTIGAN算法在可见光到红外图像迁移任务上的优良性能，本文将其与其他图像迁移算法进行了对比实验分析。主要对比网络包含CycleGAN，DSMAP(Domain-Specific Mappings)[19]，UGATIT(Unsupervised Generative Attentional networks)[20]，GLANet(Global and Local Alignment Networks)[21]，CUT(Contrastive Learning for Unpaired Image-to-Image Translation)[22]。为了充分对比不同算法之间的性能差异，本文从客观定量分析和主观视觉评价两方面对图像风格迁移结果进行了评价。

3.2.1　客观定量分析

使用PSNR，SSIM和FID三个指标对各模型的测试结果进行定量分析，对比结果如表4所示。

表4　图像评价指标对比

Tab. 4　Comparison of image evaluation indexes

Method	Pub'Year	Architectural scene			Traffic scene		
		PSNR/dB	SSIM	FID	PSNR/dB	SSIM	FID
GLANet	2021	12.720	0.631	75.39	13.823	0.680	68.27
CUT	ECCV'2020	12.895	0.634	71.25	13.995	0.675	62.21
CycleGAN	ICCV'2017	13.138	0.646	66.52	14.779	0.766	50.11
DSMAP	ECCV'2020	14.351	0.738	50.38	15.364	0.780	42.88
UGATIT	ICLR'2019	15.210	0.765	46.92	16.641	0.805	38.57
VTIGAN	**Our**	**15.850**	**0.796**	**39.36**	**16.750**	**0.818**	**36.48**

从不同方法的实验结果来看VTIGAN在三项指标的评价结果上达到了最优，相比于CycleGAN在PSNR，SSIM和FID三个指标上分别提升了16.8%，14.3%和35.0%，相比于排名第二的方法UGATIT在PSNR、SSIM和FID三个指标上分别提升了3.1%，2.8%和11.3%。从两组不同场景实验的评价结果对比来看，由于交通场景相比于建筑场景复杂程度略低，因此在交通场景上的图像生成结果均优于建筑场景。但场景复杂度对不同方法的影响程度并不相同，从表4中数据分析可以看出，VTIGAN从交通场景到建筑场景三项图像评价指标下降幅度分别为5.3%，2.7%和7.9%，指标下降幅度最小。而CycleGAN的表现受场景复杂度的影响最大，在PSNR，SSIM和FID三个指标上分别下降了11.1%，15.7%和32.7%。从定量评价结果来看本文提出的VTIGAN针对可见光图像到红外图像的迁移取得了最优的表现，且该方法具有

较好的鲁棒性,对于复杂场景下的抗干扰能力更强。

3.2.2　主观视觉评价

为了更加直观比较各模型在可见光图像到红外图像迁移任务上的性能差异,本节选取了水泥建筑、树木、草坪、工地、交通车辆五类主要实例的红外图像生成结果直接进行比对,各方法具体实验结果示例如图5所示。

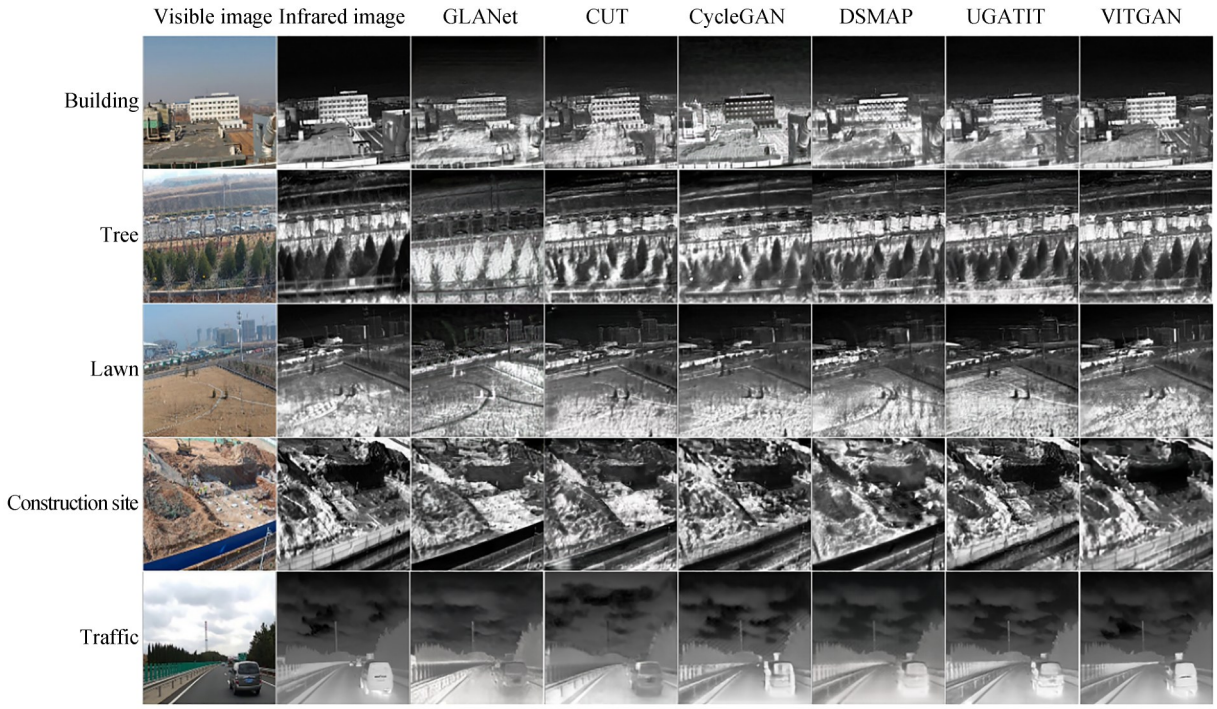

图5　实验结果示例

Fig. 5　Example of experimental results

图5中第一列为测试时输入的可见光图像,第二列为真实的红外参考样本,后面依次排列的是不同方法的测试结果。从五类实例的生成效果来看,GLANet,CUT,CycleGAN三种方法对于5类实例的红外特征生成均有明显的错误,例如水泥建筑的结果对比中CycleGAN未能正确生成楼房的红外特征,GLANet和CUT对于车辆的红外信息均生成错误。而DSMAP,UGATIT,VTIGAN三种方法对于红外特征均能正确的生成,但从图像的逼真程度和清晰度来看VTIGAN优于DSMAP和UGATIT。这些方法对于红外图像仿真均有一定的应用价值,但也存在一些共性的不足,从图5中可以看出交通车辆的红外图像中远处行驶的两辆车,由于目标较小,六种方法均未正确生成其红外特征,甚至直接造成了信息的缺失。总之,从主观评价上来看,本文提出的VTIGAN相比于其他方法生成图像的红外特征精确且清晰度和逼真度更优。

3.3　消融实验

本文联合对抗损失、多层对比损失、风格相似性损失、同一性损失四种损失函数对网络模型进行训练。为验证各个损失函数的有效性,本节采取消融实验的方法测试多层对比损失、风格相似性损失、同一性损失对算法性能的影响。实验结果如表5所示,在本实验中依次去除多层对比损失、风格相似性损失、同一性损失,并保留其他损失函数不变,从实验结果可知去除任意一项损失函数均导致算法性能的下降。其中,多层对比损失和风格相似损失作为图像迁移过程中的主要约束,去除后导致算法性能大幅下降,进一步验证了本文算法的有效性。

表5　消融实验结果

Tab. 5　Results of ablation experiments

Loss function				PSNR/dB	SSIM	FID
Adversarial loss	Multi-layer contrast loss	Style similarity loss	Identity loss			
√	√	√		15.182	0.721	45.84
√	√		√	10.146	0.513	96.47
√		√	√	9.851	0.494	98.18
√	√	√	√	**15.850**	**0.796**	**39.36**

4　结　论

可见光图像迁移生成红外图像具有重要的理论研究意义和实际应用价值。本文针对非成对数据条件下可见光图像迁移生成红外图像的任务要求，提出了一种新颖的红外图像仿真算法VTIGAN。VTIGAN以特征提取能力强大的transformer架构为基础构建了一种新的生成器，使用PathGAN作为判别器对生成图像进行更精细化的鉴别，并联合对抗损失、多层对比损失、风格相似性损失、同一性损失四种损失函数对模型的训练加以约束，确保红外图像高质量的生成。在可见光-红外数据集上进行实验，与目前主流的图像迁移算法相比VTIGAN在客观定量分析和主观视觉评价两方面均取得明显优势，对于复杂场景下的抗干扰能力优于基于循环一致性损失函数的方法，相比于CycleGAN在PSNR，SSIM和FID三个指标上分别提升了20.6%，23.2%和40.8%，能够生成清晰度更高、红外特征更准确的红外仿真图像。在下一步的研究中将关注以下两方面：(1)现有算法对于图像中的小目标转换效果不佳，下一步研究可加强模型对小目标红外特征转换的能力，以进一步提高红外仿真图像的使用价值；(2)可见光-红外数据集的种类和数量仍然较少，下一步可从大型数据集制备的角度来提高模型性能。

参考文献：

[1] LATGER J, CATHALA T, DOUCHIN N, *et al*. Simulation of active and passive infrared images using the SE-WORKBENCH[C]. *Defense and Security Symposium. Proc SPIE* 6543, *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XVIII, Orlando, Florida, USA*. 2007, 6543: 11-25.

[2] JOHNSON K, CURRAN A, LESS D, *et al*. MuSES: a new heat and signature management design tool for virtual prototyping[C]. *Proceedings of the 9th Annual Ground Target Modelling & Validation Conference*. 1998.

[3] SCHOTT J R, BROWN S D, RAQUEÑO R V, *et al*. An advanced synthetic image generation model and its application to multi/hyperspectral algorithm development[J]. *Canadian Journal of Remote Sensing*, 1999, 25(2): 99-111.

[4] 杨艳春，高晓宇，党建武，等. 基于WEMD和生成对抗网络重建的红外与可见光图像融合[J]. 光学精密工程，2022，30(3)：320-330.
YANG Y C, GAO X Y, DANG J W, *et al*. Infrared and visible image fusion based on WEMD and generative adversarial network reconstruction[J]. *Opt. Precision Eng.*, 2022, 30(3): 320-330. (in Chinese)

[5] 杨植凯，卜乐平，王腾，等. 基于循环一致性对抗网络的室内火焰图像场景迁移[J]. 光学精密工程，2020，28(3)：745-758.
YANG Z K, BU L P, WANG T, *et al*. Scenemigration of indoor flame image based on Cycle-Consistent adversarial networks[J]. *Opt. Precision Eng.*, 2020, 28(3): 745-758. (in Chinese)

[6] 徐哲，耿杰，蒋雯，等. 联合训练生成对抗网络的半监督分类方法[J]. 光学精密工程，2021，29(5)：1127-1135.
XU Z, GENG J, JIANG W, *et al*. Co-training generative adversarial networks for semi-supervised classification method[J]. *Opt. Precision Eng.*, 2021, 29(5): 1127-1135. (in Chinese)

[7] ISOLA P, ZHU J Y, ZHOU T H, *et al*. Image-to-image translation with conditional adversarial networks[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu, HI, USA*. IEEE, 2017: 5967-5976.

[8] ZHU J Y, PARK T, ISOLA P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 2242-2251.

[9] KIM T, CHA M, KIM H, *et al*. Learning to discover cross-domain relations with generative adversarial networks[C]. *Proceedings of the* 34*th International Conference on Machine Learning - Volume* 70. *August* 6-11, 2017, *Sydney*, *NSW*, *Australia*. *New York*: ACM, 2017: 1857-1865.

[10] YI Z L, ZHANG H, TAN P, *et al*. DualGAN: unsupervised dual learning for image-to-image translation[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 2868-2876.

[11] CHEN T, KORNBLITH S, NOROUZI M, *et al*. A simple framework for contrastive learning of visual representations[C]. *Proceedings of the* 37*th International Conference on Machine Learning*. *New York*: ACM, 2020: 1597-1607.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. *An image is worth* 16*x*16 *words*: *transformers for image recognition at scale*[EB/OL]. 2020: *arXiv*: 2010.11929. https://arxiv.org/abs/2010.11929.pdf

[13] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 1800-1807.

[14] GUTMANN M, HYVÄRINEN A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models[C]. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. *JMLR Workshop and Conference Proceedings*, 2010: 297-304.

[15] PANG Y, LIN J, QIN T, *et al*. Image-to-Image Translation: Methods and Applications[EB/OL]. 2021: *arXiv*: 2101.08629. https://arxiv.org/abs/2101.08629.pdf

[16] WANG Z, 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.

[17] CHE Z P, PURUSHOTHAM S, CHO K, *et al*. Recurrent neural networks for multivariate time series with missing values[J]. *Scientific Reports*, 2018, 8: 6085.

[18] HEUSEL M, RAMSAUER H, UNTERTHINER T, *et al*. *GANs Trained by A two Time-Scale Update Rule Converge to a Local Nash Equilibrium* [EB/OL]. 2017: *arXiv*: 1706.08500. https://arxiv.org/abs/1706.08500.pdf

[19] CHANG H Y, WANG Z X, CHUANG Y Y. *Domain-Specific Mappings for Generative Adversarial Style Transfer*[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 573-589.

[20] KIM J, KIM M, KANG H, *et al*. *U-GAT-IT*: *Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation*[EB/OL]. 2019: *arXiv*: 1907.10830. https://arxiv.org/abs/1907.10830.pdf

[21] YANG G, TANG H, SHI H, *et al*. *Global and Local Alignment Networks for Unpaired Image-to-Image Translation* [EB/OL]. 2021: *arXiv*: 2111.10346. https://arxiv.org/abs/2111.10346.pdf

[22] PARK T, EFROS A A, ZHANG R, *et al*. *Contrastive Learning for Unpaired Image-to-Image Translation*[EB/OL]. 2020: *arXiv*: 2007.15651. https://arxiv.org/abs/2007.15651.pdf

作者简介:

蔡　伟(1974－),男,陕西西安人,博士,教授,2004 年获西安交通大学博士,主要研究方向为神经网络、计算机视觉、光学材料和光电技术。E-mail: XHT807@outlook.com

通讯作者:

姜　波(1999－),男,安徽滁州人,硕士研究生,主要研究方向为神经网络、计算机视觉。E-mail: jiangxinhao2020@outlook.com