计算机科学
COMPUTER SCIENCE

基于双重语义对比学习的无监督红外图像生成方法

程梓萌, 杨馨悦, 艾浩军, 王中元

引用本文

程梓萌, 杨馨悦, 艾浩军, 王中元. 基于双重语义对比学习的无监督红外图像生成方法[J]. 计算机科学, 2026, 53(4): 260-268.

CHENG Zimeng, YANG Xinyue, AI Haojun, WANG Zhongyuan. Unsupervised Infrared Image Generation Method Based on Dual Semantic Contrastive Learning [J]. Computer Science, 2026, 53(4): 260-268.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250700172

基于双重语义对比学习的无监督红外图像生成方法

程梓萌[1,2]　杨馨悦[1,2]　艾浩军[1,2]　王中元[3]
1 武汉大学国家网络安全学院　武汉 430072
2 武汉大学空天信息安全与可信计算教育部重点实验室　武汉 430072
3 武汉大学计算机学院　武汉 430079
(zmcheng@whu.edu.cn)

摘　要　红外图像在计算机视觉领域应用广泛。受制于采集条件,高质量红外图像数据集规模较小。把可见光图像转换为红外图像,是扩充红外数据集的有效手段。现有生成方法多依赖有监督学习,需要大量配对数据。为此,提出基于双重语义对比学习的无监督红外图像生成方法DSCGAN。该方法采用双向转换架构,通过语义对比学习增强图像内容保持能力和红外特征学习能力。损失函数增加几何一致性损失,协助保留可见光图像的原始结构与细节。同时,构建多尺度PatchGAN判别器,增强判别能力,提升生成图片的真实感。在AVIID-1,AVIID-2和Day-DroneVehicle数据集上的实验表明,DSCGAN在多项指标上优于对比方法,生成的红外图像热辐射分布更合理,视觉质量更优。在AVIID-1数据集中,DSCGAN的SSIM值提升至0.8144,FID分数降低至0.1456。在Day-DroneVehicle数据集中,DSCGAN的PSNR值提升至18.14,LPIPS值降低至0.2949。所提方法为无监督红外图像生成提供了新思路,可进一步应用于红外目标检测和场景分割等下游任务。

关键词:图像到图像转换;语义对比学习;红外图像生成;多尺度判别器;几何一致性约束
中图分类号　TP391.41

Unsupervised Infrared Image Generation Method Based on Dual Semantic Contrastive Learning

CHENG Zimeng[1,2], YANG Xinyue[1,2], AI Haojun[1,2] and WANG Zhongyuan[3]
1 School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
2 Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, Wuhan University, Wuhan 430072, China
3 School of Computer Science, Wuhan University, Wuhan 430079, China

Abstract　Infrared images are widely used in computer vision, but high-quality infrared image datasets are limited in scale due to restricted acquisition conditions. To address this problem, converting visible datasets to infrared datasets has become an effective way. Existing generation methods generally rely on supervised learning, which requires a large amount of paired data that is extremely difficult to obtain in practical applications. This paper proposes an unsupervised infrared image generation method named DSCGAN. This method adopts a bidirectional transformation architecture and introduces semantic contrast learning to enhance the ability to preserve image content and learn discriminative infrared features. The geometric consistency loss is introduced to preserve the original structure and details of visible images effectively. Meanwhile, a multi-scale PatchGAN discriminator is constructed to improve discriminative capability and enhance the realism of generated images. Experimental results on the AVIID-1, AVIID-2, and Day-DroneVehicle datasets show that DSCGAN outperforms the comparison methods in several metrics, and the generated infrared images exhibit a more reasonable thermal radiation distribution and better visual quality. In the AVIID-1 dataset, the SSIM value increases to 0.8144, and the FID score decreases to 0.1456. In the Day-DroneVehicle dataset, the PSNR value improves to 18.14, while the LPIPS value drops to 0.2949. This study provides a new idea for unsupervised infrared image generation, with potential applications in infrared target detection, infrared scene segmentation, and other downstream tasks.

Keywords　Image-to-image translation, Semantic contrastive learning, Infrared image generation, Multi-scale discriminator, Geometric consistency constraint

到稿日期:2025-07-25　返修日期:2025-09-22
基金项目:湖北省国际科技合作项目(2025EHA043)
This work was supported by the Hubei Province International Science and Technology Collaboration Program(2025EHA043).
通信作者:艾浩军(aihj@whu.edu.cn)

1 引言

红外图像，特别是热红外图像，凭借强大的抗干扰能力和全天候成像特性，在目标检测与跟踪[1]、光伏面板故障检测[2]及行人重识别[3]等领域具有广泛应用。然而，红外相机的高成本与复杂成像条件不仅导致高质量红外数据集采集困难，其独有的热辐射特性更使得数据标注面临严峻挑战。将可见光图像转换为红外图像，不仅能有效补充现有红外数据资源，还能直接继承可见光数据的标签，降低数据标注的成本。

当前红外图像生成方法主要分为两类。第一类是基于物理建模的仿真方法[4-5]，这类方法流程复杂，生成的图像容易出现纹理失真的问题，难以满足大规模数据仿真的需求。第二类是基于图像到图像转换的方法[6-9]，其通过学习跨域特征映射，在保留可见光图像内容的同时，模拟红外热特征的分布，从而实现更自然的红外图像合成。生成对抗网络（Generative Adversarial Network，GAN）[10]利用生成器与判别器的对抗机制，在可见光到红外的图像转换任务中取得显著进展。早期研究主要采用基于 Pix2Pix[11]的有监督学习框架，通过引入 L1 重建损失实现了较高质量的风格转换，但现有的公开数据难以满足有监督学习对数据对齐的严格要求。此外，当前主流的图像转换方法[12-16]更侧重风格转换的多样性，而对图像语义特征的关联性考虑不足，导致转换前后的语义一致性难以保证。近年来，一些无监督方法[8-9]被用于可见光到红外的图像转换任务，但这类方法普遍存在生成器与判别器复杂度失衡的结构缺陷，容易导致判别器过拟合，从而限制模型的生成质量和泛化性能。

针对上述问题，提出基于双重语义对比学习的无监督红外图像生成方法 DSCGAN（Dual-Semantic-Contrast Generative Adversarial Network）。该方法设置双向转换结构，结合红外图像中不同物体具有差异化热辐射分布的特点，在对比学习中引入语义一致性约束，有效增强了内容保持能力和红外特征学习能力。同时，添加几何一致性损失，进一步保留了可见光图像的原始结构与细节。此外，构建多尺度 PatchGAN 判别器，通过并行处理不同尺度的图像，显著提升了模型的判别能力。实验结果表明，DSCGAN 在红外图像生成质量方面取得了显著提升。

本文的主要贡献如下：

1）提出一种用于可见光到红外图像生成的双向转换架构，通过强制约束转换前后图像的语义分布一致，解决了现有方法因忽略语义关联导致的热辐射特征失配问题，显著提升了生成图像对语义特征的保持能力。

2）设计几何一致性损失与其他损失函数协同约束的机制，通过增强生成器对空间变换的鲁棒性，进一步提升了模型的语义保持能力和双向转换架构的稳定性。同时，构建多尺度 PatchGAN 鉴别器，通过综合多个尺度输出的判别信息来增强判别能力，提升生成图片的真实感。

3）在 AVIID-1（白天）、AVIID-2（夜晚）以及 Day-Drone Vehicle 3 个数据集上的实验表明，DSCGAN 在红外图像生成质量方面综合优于其他方法。

2 相关工作

2.1 无监督图像到图像的转换方法

图像到图像的转换是计算机视觉领域的重要研究方向。早期的有监督方法需要成对的训练数据，这种严格的数据对齐要求限制了模型的泛化能力。为解决无配对数据的转换问题，研究者们提出了多种无监督方法。Zhu 等[12]提出的 CycleGAN，通过设置双向生成器和循环一致性约束，实现了源域和目标域之间的可逆转换。UNIT[13]采用共享潜在空间假设，隐式结合循环一致性约束，为跨域转换提供了新思路。Fu 等[14]提出 的 GcGAN，从图像属性出发，引入几何一致性损失来维持语义结构的稳定性。近年来，基于对比学习的图像转换方法取得了重要突破，CUT[15]将互信息最大化思想引入该领域，提升了转换前后的跨域语义对齐能力。为了进一步利用对比学习，Han 等[16]提出 DCLGAN，将对比学习扩展为双向约束，实现了更优的双重无监督图像转换效果。此外，基于扩散模型的图像转换方法，如 BBDM[17]和 DMT[18]，虽然在复杂转换任务中取得了优异的生成质量与细节表现，但存在计算效率较低及条件控制精度不稳定等问题，在实际应用中仍面临挑战。

2.2 可见光到红外图像的转换方法

可见光到红外图像的转换作为多模态视觉领域的关键任务，在航空遥感和智能交通等场景中具有重要的应用价值。目前的主流方法可分为两类：基于 Pix2Pix 框架的有监督方法和基于 CycleGAN 的无监督方法。在有监督方法方面，ThermalGAN[6]通过多模态热成像生成和两步温度预测，在行人重识别任务中实现了多峰值热图像的合成。Ma 等[7]采用 ConvNeXt-Unet 生成器、梯度向量损失与多尺度判别器相结合的方法，进一步突破了纹理细节提取的瓶颈，提升了生成图像的质量。这些有监督方法始终面临泛化能力不足和对训练数据要求过高的局限性。在无监督方法方面，Wang 等[8]通过融合 Swin Transformer 与特征分层映射，增强了热特征的真实性；Han 等[9]提出的 DR-AVIT，通过解耦表示与几何一致性损失，实现了航空图像的多样性生成。但现有无监督方法普遍存在生成器与判别器能力失衡的问题：一方面，生成器为提升性能，不断引入残差块和 Transformer 等复杂结构进行优化；另一方面，判别器仍沿用普通 PatchGAN 架构，导致判别能力难以匹配生成器的发展。这种不平衡容易引发过拟合，并影响生成图像的真实感。因此，未来研究需要着重解决生成器与判别器的协同优化问题，以进一步提升可见光到红外图像转换的质量。

2.3 对比学习

对比学习通过构建正负样本对来学习判别性特征表示，在图像转换领域展现出显著优势。在对比学习中，正负样本的构造策略会直接影响模型性能。SimCLR[19]利用随机数据增强生成正样本对，借助对比损失提升视觉特征学习能力。

CUT[15]运用噪声对比估计(Noise Contrastive Estimation, NCE),最大化输入与输出块之间的映射关系,有效改善了循环一致性约束导致的模糊问题。DCLGAN[16]通过设置双向基于图像块的对比学习机制,显著提升了特征学习能力。QS-Attn[20]设计查询选择注意力模块,通过选择重要锚点进行补丁级对比学习,进一步优化了特征对齐。然而,现有方法过度关注正负样本的简单分类,未能充分考虑全局语义信息的整合,导致在处理复杂场景图像转换任务时性能受限。

3 DSCGAN

本文提出一种基于双重语义对比学习的无监督图像转换框架 DSCGAN,其整体架构如图 1 所示。该框架主要包含 4 个核心组件(两个生成器(G_V 和 G_I)、两个多尺度判别器(D_V 和 D_I))以及两个用于获取特征嵌入的多层感知机(MLP_V 和 MLP_I)。生成器由前半部分的编码器 enc 和后半部分的解码器 dec 构成,负责将图像从一个域映射至另一个域,其中 enc 可用于生成可见光域和红外域的嵌入表示;多尺度判别器通过并行分析不同尺度的图像特征,在各域中实现对生成图像局部细节与全局结构的联合判别;多层感知机模块把提取到的嵌入特征投影到共享空间,为语义对比学习提供跨域对齐的特征表示。该双向架构参考了 CycleGAN 的设计思路,但摒弃了循环一致性约束。它通过两条对称的转换路径实现:路径 G_V :$VIS \rightarrow IR$ 将可见光图像转换为红外图像,路径G_I:$IR \rightarrow VIS$ 执行相反的转换。该机制能够更全面地捕捉域间的特征对应关系,同时提供更丰富的监督信号,进而有效促进模型收敛。框架的每个方向主要包含 3 个关键模块:基于图像块的语义对比学习、多尺度 PatchGAN 判别器,以及包含几何一致性损失在内的若干损失函数。

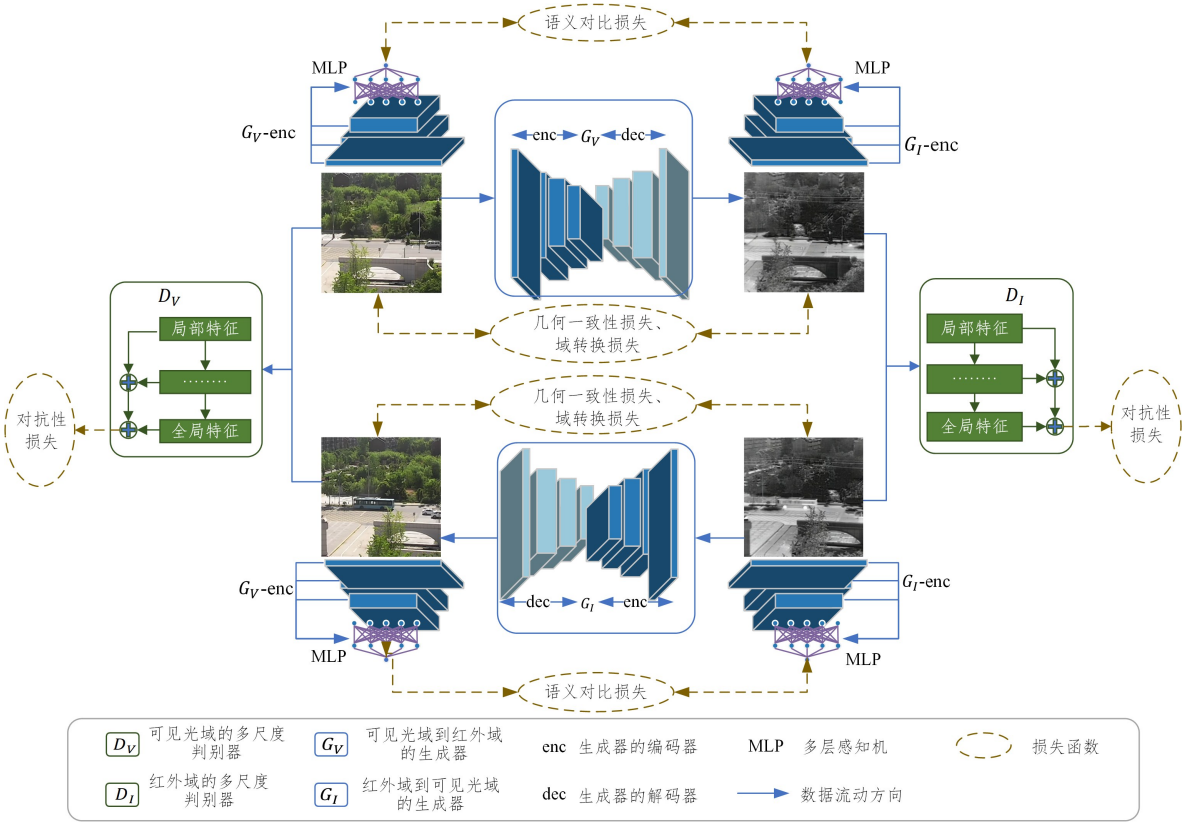

图 1 DSCGAN 的总体架构

Fig. 1 Overall architecture of DSCGAN

3.1 语义对比学习

可见光到红外图像的转换不仅是纹理特征的改变,其核心还在于精准捕捉并呈现物体深层的热值特征。不同材质(如金属与非金属)、不同物理性质(如高温发热源与常温物体)以及它们与环境的不同热交互,都会形成独特的热辐射分布,这些热辐射分布构成了图像的关键语义信息。因此,进行图像转换时,保证转换前后图像的语义关系一致性极为重要。

受 Jung 等[21]研究的启发,将图像块的异质语义关系约束引入 DSCGAN 的双向转换架构中。如图 2 所示,语义对比学习框架中包含多种不同来源的图像块,查询图像块(亮色标注)源自卡车头,而其他图像块代表与卡车头无关的元素。需要注意的是,同属卡车的图像块中蕴含的语义信息也具有多样性,例如卡车头和卡车货仓。在众多与卡车相关的元素中,卡车与绿化带的语义关系最远。

在图 2 中,$\boldsymbol{z}_k$ 与 $\boldsymbol{w}_k$ 分别为可见光图像块 $v_k \in V$ 和红外图像块 $i_k \in I$ 对应的嵌入特征,其中 V 表示可见光图像集合,I 表示红外图像集合,$k(k \in L)$表示图像块在各自图像中的位置索引。具体而言,首先从生成器的编码器 enc 中选择 4 层特征,对每层特征图均匀采样 256 个空间位置 s,将其输入双层 MLP 网络进行非线性投影,最终构建共享嵌入空间。最终,嵌入特征 $\boldsymbol{z}_k$ 和 $\boldsymbol{w}_k$ 的表达式如下:

$$\boldsymbol{z}_k = MLP_V(enc_V(G_I(i))) \tag{1}$$

$$\boldsymbol{w}_k = MLP_I(enc_I(G_V(v))) \tag{2}$$

当前主流的对比学习方法（如 CUT[15] 和 DCLGAN[16]）通常遵循示例分类的方式，即为对应位置的图像块分配标签 1（正样本），其他位置的图像块分配标签 0（负样本）。这种分类方法过于严苛，直接将所有来自不同位置的图像块简单视为负样本，忽视了不同语义区域之间固有的关联性差异。本文将图像块的语义分布关系以相似度分布的形式进行捕捉，将传统的硬标签转化为包含全局信息的软标签，使每个相似度评分都包含全局信息，比 DCLGAN 的离散二元判断更符合实际语义关系。

对于给定可见光图像块 v_k，其与负块 v_i（i 为另一个位置索引）的特征相似度关系由 softmax 函数定义为：

$$Z_k(i) = \frac{\exp(\boldsymbol{z}_k^{\mathrm{T}}\boldsymbol{z}_i)}{\sum_{j\in L}\exp(\boldsymbol{z}_k^{\mathrm{T}}\boldsymbol{z}_j)} \tag{3}$$

同样，红外图像中的逐块语义关系定义为：

$$W_k(i) = \frac{\exp(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_i)}{\sum_{j\in L}\exp(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_j)} \tag{4}$$

其中，$Z_k(i)$ 与 $W_k(i)$ 采用相同的位置索引集 L 来定义。为了在图像转换前后保持图像块之间多样的语义分布关系，采用 Jensen-Shannon 散度（JSD）来度量两个概率分布 $Z_k(i)$ 和 $W_k(i)$ 之间的差异，通过最小化该散度来强制图像转换前后语义关系的一致性。本文对所有 K 个采样向量的 JSD 求和，得到语义一致性损失 L_{SRC}，其表达式为：

$$L_{\mathrm{SRC}} = \sum_{k\in M} JSD(Z_k \mid W_k) \tag{5}$$

在图像转换任务中，不仅要确保转换前后图像的语义关系保持一致，生成后的图像块还应与用于生成它的输入图像块相似，而非与随机图像块相似。为此，在全局语义约束框架下，采用 NCE 损失优化局部特征对应关系，以最大化输入-输出互信息。全局语义约束框架减轻了简单负样本对梯度更新的主导作用，从而缓解了 NCE 损失中存在的负-正样本耦合效应。以图 2 为例，在可见光到红外图像的转换场景中，将红外图像的卡车头图像块设定为查询对象，可见光图像中与之对应的卡车头图像块为其正样本，其余的图像块作为 n 个负样本。查询、正样本和负样本图像块各自对应的嵌入特征向量分别记为 $\boldsymbol{z}_k$，$\boldsymbol{w}_k{}^+$ 和 $\boldsymbol{w}_n{}^-$。其中，$\boldsymbol{w}_n{}^-$ 表示为 N 个样本中的第 n 个负值。在使用 L2 归一化对向量进行归一化处理后，该损失可简化为一个 $N+1$ 分类问题。通过计算选择正例而非负例的可能性，并将其以交叉熵损失的形式呈现，可见光到红外方向的 NCE 损失的定义如下：

$$L_{\mathrm{NCE\_V}} = -\sum_{k=1}^{K}\log\frac{\exp(sim(\boldsymbol{z}_k,\boldsymbol{w}_k^+)/\tau)}{\exp(sim(\boldsymbol{z}_k,\boldsymbol{w}_k^+)/\tau)+\sum_{n-1}^{N}\exp(sim(\boldsymbol{z}_k,\boldsymbol{w}_n^-)/\tau)} \tag{6}$$

其中，$sim(\boldsymbol{z},\boldsymbol{w}) = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{w}/|\boldsymbol{z}||\boldsymbol{w}|$ 表示向量 $\boldsymbol{z}$ 和 $\boldsymbol{w}$ 之间的余弦相似度；τ 是温度参数，用于调整样本间相似性的度量尺度。同样，在红外到可见光方向的 NCE 损失定义为：

$$L_{\mathrm{NCE\_I}} = -\sum_{k=1}^{K}\log\frac{\exp(sim(\boldsymbol{w}_k,\boldsymbol{z}_k^+)/\tau)}{\exp(sim(\boldsymbol{w}_k,\boldsymbol{z}_k^+)/\tau)+\sum_{n-1}^{N}\exp(sim(\boldsymbol{w}_k,\boldsymbol{z}_n^-)/\tau)} \tag{7}$$

因此，总的语义对比损失定义为：

$$L_{\mathrm{semantic}}(\tau) = \lambda_{\mathrm{SRC}}L_{\mathrm{SRC}} + \lambda_{\mathrm{NCE}}(L_{\mathrm{NCE\_V}}(\tau)+L_{\mathrm{NCE\_I}}(\tau)) \tag{8}$$

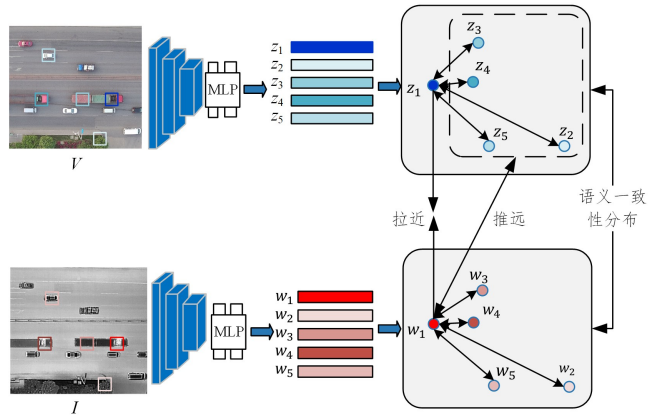

图 2 语义对比学习（电子版为彩图）

Fig. 2 Semantic contrastive learning

3.2 多尺度判别器

DSCGAN 在可见光域和红外域分别构建了多尺度 PatchGAN 判别器。与通用的 PatchGAN 判别器仅依赖单一判别网络不同，多尺度判别器通过设计图像金字塔结构，分别在 1×，1/2× 和 1/4× 尺度上对图像进行处理，并分别由 D1，D2 和 D3 这 3 个判别网络进行判别。这种设计，使网络能够同时捕获从细节到整体结构的多层次特征信息。

如图 3 所示，DSCGAN 的多尺度判别器采用相同的三层卷积网络结构，具体内部结构如图中虚线框所示。各尺度图像经判别器处理后，通过计算输出矩阵的平均值实现对细节特征到整体结构特征的多层次信息的综合。相较于增加单一判别器的深度，这种多尺度策略仅需线性增加计算量就能获得多样化的感受野。这种策略不仅保持了单个判别器的结构复杂度，还增强了判别器对真实和虚假图像的辨别能力。

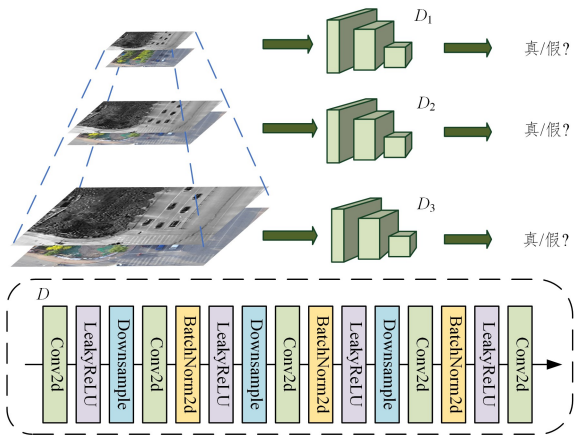

图 3 多尺度判别器的网络架构

Fig. 3 Network architecture for multi-scale discriminator

3.3 损失函数

DSCGAN 的总目标函数由对抗性损失 L_{adv}、语义对比损失 $L_{semantic}$（见 3.1 节）、几何一致性损失 L_{geo} 和领域转换损失 L_{dom} 组成。

3.3.1 对抗性损失

对抗性损失用于衡量生成器生成图像与真实图像的差异，以及判别器判断的准确性。对于生成器 G_V：$VIS \rightarrow IR$ 和判别器 D_I，对抗性损失的计算式为：

$$L_{adv\_V}=E_{i\sim I}[\log D_I(i)]+E_{v\sim V}[\log(1-D_I(G_V(v))] \quad (9)$$

其中，生成器试图最小化此损失以生成逼真红外图像，判别器则试图最大化它来精准识别虚假红外图像。对于生成器 G_I：$IR \rightarrow VIS$ 和判别器 D_V，引入类似的对抗性损失：

$$L_{adv\_I}=E_{v\sim V}[\log D_V(v)]+E_{i\sim I}[\log(1-D_V(G_I(i))] \quad (10)$$

因此，总对抗性损失的表达式如下：

$$L_{adv}=\frac{1}{2}L_{adv\_V}+\frac{1}{2}L_{adv\_I} \quad (11)$$

3.3.2 几何一致性损失

在无监督图像转换中，现有约束通常忽略了图像的几何不变性，即简单的几何变换不会改变图像语义结构。因此，转换图像应保持与输入图像一致的几何关系。通过约束生成器对旋转等变换具备等变性，可避免车辆轮廓、建筑边缘等刚性结构在模态转换中发生形变。对于给定的图像 v、一个几何变换函数 $f(\cdot)$ 和生成器 G_V，几何一致性可以定义为 $f(G_V(v))\approx G_V(f(v))$ 和 $G_V(v)\approx f^{-1}(G_V(f(v)))$，其中 $f^{-1}(\cdot)$ 是 $f(\cdot)$ 的反函数。该约束简单且严格，不仅可以减轻 GAN 所面临的模式崩溃问题，还可以进一步缓解域适应过程中的语义失真。为了强制执行这一约束，将两个方向的几何一致性损失分别表示为如下形式：

$$L_{geo\_V}(v,G_V,f)=\| G_V(v)-f^{-1}(G_V(f(v)))\|_1 \quad (12)$$

$$L_{geo\_I}(i,G_I,f)=\| G_I(i)-f^{-1}(G_I(f(i)))\|_1 \quad (13)$$

在本研究中，使用 90°顺时针旋转作为预定义的几何变换函数。总几何一致性损失为：

$$L_{geo}=\frac{1}{2}L_{geo\_V}(v,G_V,f)+\frac{1}{2}L_{geo\_I}(i,G_I,f) \quad (14)$$

3.3.3 领域转换损失

DSCGAN 采用双向转换架构，为防止生成器产生不必要的变化，引入领域转换损失。具体而言，生成器 G_V 将可见图像转换为红外图像。若将红外图像输入 G_V，理想的输出应仍为红外图像，且 $G_V(i)$ 需与原始红外图像 $i\in I$ 高度相似。同理，$G_I(v)$ 的输出应与可见图像 $v\in V$ 相似。该损失旨在增强对输入图像固有属性的保持能力，进而在双向转换过程中维持稳定性。领域转换损失的设计如下：

$$L_{dom}=\frac{1}{2}E_{v\sim V}[\| G_I(v)-v\|_1]+\frac{1}{2}E_{i\sim I}[\| G_V(i)-i\|_1] \quad (15)$$

3.3.4 总目标损失

结合 3.1 节中式(8)所定义的语义关系一致性损失，DSCGAN 的总损失 L_{total} 计算如下：

$$L_{total}=\lambda_1\cdot L_{adv}+\lambda_2\cdot L_{semantic}+\lambda_3\cdot L_{geo}+\lambda_4\cdot L_{dom} \quad (16)$$

其中，λ_1、λ_2、λ_3 和 λ_4 分别为对抗性损失、语义对比损失、几何一致性损失和领域转换损失的权重系数，分别设为 $\lambda_1=1$，$\lambda_2=2$，$\lambda_3=1$ 和 $\lambda_4=1$，语义对比损失的权重系数取 $\lambda_{SRC}=0.1$，$\lambda_{NCE}=0.5$，并设置温度参数 $\tau=0.07$。DSCGAN 的总体算法流程如算法 1 所示。

算法 1 DSCGAN 算法流程

输入：(V, I, epochs, λ_1, λ_2, λ_3, λ_4)

输出：(G_V, G_I) # 训练好的生成器

```
G_V ← Generator() # VIS→IR 生成器
G_I ← Generator() # IR→VIS 生成器
D_V, D_I ← MultiScaleDiscriminator() # 多尺度判别器
MLP_V, MLP_I ← MLP() # 特征嵌入网络
for epoch ← 1 to epochs do: //训练循环
    fake_i ← G_V(v), v∈V //可见光到红外域
    fake_v ← G_I(i), i∈I //红外到可见光域
    L_adv ← Eq.(11) # 对抗性损失
    L_semantic ← Eq.(8) # 语义对比损失
    L_geo ← Eq.(14) # 几何一致性损失
    L_dom ← Eq.(15) # 领域转换损失
    L_total ← Eq.(16) # 总损失
    Update(G_V, G_I, D_V, D_I via Adam) //参数更新
end for
```

4 实验

4.1 数据集

本研究采用 AVIID[22] 基准数据集的两个子集（AVIID-1，AVIID-2）和大型数据集 Day-DroneVehicle[9]，来验证所提出方法的有效性及泛化能力。

1）AVIID 数据集：该数据集是第一个为无人机航拍可见光到红外图像翻译任务而设计的基准数据集。为全面评估方法的泛化性能，选取 AVIID-1 和 AVIID-2 进行实验，其分别涵盖了白天和夜晚两种典型场景。

AVIID-1 包含 993 对白天拍摄的可见光-红外图像对，分辨率为 434×434，主要为道路场景，目标车辆类型包括汽车、公交车和货车等。

AVIID-2 包含 1090 对夜间采集的可见光-红外图像对，分辨率为 434×434。该数据集在低光照条件下采集，所采集的图像存在较多噪声，目标物体呈现模糊特征，为可见光到红外图像转换任务带来了挑战。

2）Day-DroneVehicle 数据集：该数据集来自 DRAVIT[9]，由公开数据集收集而来，包含无人机在白天 10 个不同高度和角度拍摄的 7660 对可见光-红外图像对，分辨率为 640×512。数据集覆盖了城市道路、居民区、工业区和停车场等多种复杂场景，目标车辆类别包括汽车、货车、卡车、公交车和面包车 5 类，且采集时采用了多种高度和角度，充分体现了场景的多样性，更有助于测试所提出方法的泛化能力。值得注意的是，该数据集对每一类目标车辆均提供了精确的标签

注释，这使其不仅可用于可见光到红外图像转换任务的验证，未来还可用于下游任务的性能评估。

4.2 评估指标

为了全面评估实验结果，使用保真度指标和感知指标来评估实验结果。

结构相似性指数(SSIM)：该指标用于评估生成图像与真实图像在结构上的相似性，值域为[0,1]，值越高表明两图的结构一致性越高。

峰值信噪比(PSNR)：该指标基于均方误差计算失真程度，值越高表明图像失真越低。

Fréchet Inception 距离(FID)：该指标通过 Inception-v3 模型提取图像特征，计算真实图像与生成图像的特征分布差异，值越低表明两者特征分布越接近。

核 Inception 距离(KID)：该指标基于平方最大均值差异(MMD)衡量特征分布差异，在小数据集场景中更具鲁棒性，值越低表明性能越优。

学习感知图像块相似度(LPIPS)：该指标通过深度卷积神经网络提取图像特征并计算加权 L2 距离。LPIPS 已被广泛验证与人类视觉感知高度一致，值越低表明感知质量越好。

4.3 实验结果

DSCGAN 的实验采用以下设置：将数据集按 4∶1 划分为训练集和测试集，并采用非配对数据模式进行训练。训练过程中，将所有输入图像的大小调整为 256×256，批量大小设置为 1，初始学习率为 1×10^{-4}，使用 Adam 优化器进行参数更新。训练周期共 400 轮，其中前 200 轮保持恒定学习率 0.0001，后 200 轮逐步衰减学习率以加速模型收敛。所有实验均在配备了 INVIDA 3090 的服务器上实现。

4.3.1 定量结果

为评估方法的有效性，在上述数据集上复现 5 种经典图像转换方法(pix2pix[11]，CycleGAN[12]，GcGAN[14]，CUT[15] 和 DCLGAN[16])，一种基于扩散模型的图像转换方法(BBDM[17])，以及两种最新的开源可见光到红外图像的转换方法(IR-GAN[7] 和 DR-AVIT[9])。所有对比实验均基于开源代码实现，严格遵循原始论文的参数配置。不同方法在数据集上的转换评估结果如表 1 和表 2 所列，其中粗体为最优结果，下划线为次优结果。

表 1 不同方法在 AVIID-1 和 AVIID-2 上的定量结果

Table 1 Quantitative results of different methods on AVIID-1 and AVIID-2

Method	AVIID-1					AVIID-2				
	SSIM↑	*PSNR*↑	*FID*↓	*KID*↓	*LPIPS*↓	*SSIM*↑	*PSNR*↑	*FID*↓	*KID*↓	*LPIPS*↓
pix2pix[11]	0.7299	25.50	71.75	0.0733	0.2572	0.6915	23.32	69.05	0.0666	0.2763
CycleGAN[12]	0.5183	21.57	63.57	0.0534	0.2968	0.5505	20.19	64.58	0.0552	0.3022
GcGAN[14]	0.4975	21.39	51.46	0.0373	0.2556	0.5798	20.66	53.11	0.0408	0.2718
CUT[15]	0.7372	24.73	32.68	**0.0111**	0.1972	0.6504	19.63	48.42	0.0307	0.2519
DCLGAN[16]	0.7481	25.11	33.54	0.0121	0.1889	0.7456	24.86	43.49	0.0226	0.2195
DR-AVIT[9]	0.5272	21.12	76.47	0.0657	0.3406	0.5909	20.47	73.52	0.0575	0.2964
IR-GAN[7]	0.8123	25.86	35.76	0.0178	0.1628	0.7594	24.28	44.50	0.0338	0.2217
BBDM[17]	0.6912	24.90	46.23	0.0329	0.2476	0.6617	23.91	65.09	0.0471	0.2842
DSCGAN	**0.8144**	**27.22**	**28.47**	**0.0111**	**0.1456**	**0.7683**	**25.64**	**40.64**	**0.0212**	**0.1954**

表 2 不同方法在 Day-DroneVehicle 上的定量结果

Table 2 Quantitative results of different methods on Day-DroneVehicle

Method	*SSIM*↑	*PSNR*↑	*FID*↓	*KID*↓	*LPIPS*↓
pix2pix[3]	0.5088	16.25	67.30	0.0406	0.3425
CycleGAN[4]	0.3964	13.31	128.86	0.0835	0.5622
DCLGAN[5]	0.4752	16.18	88.33	0.0420	0.3821
DR-AVIT[2]	0.2303	10.91	**45.08**	**0.0142**	0.3007
IR-GAN[6]	0.4540	15.75	116.17	0.0696	0.3520
BBDM[17]	0.4273	17.02	110.05	0.0651	0.4204
DSCGAN	**0.5371**	**18.14**	65.64	0.0289	**0.2949**

可以看出，DSCGAN 在白天和黑夜两种经典场景中都取得了领先的结果。在黑夜场景下的高质量转换结果表明，DSCGAN 在抗噪声方面具有一定的积极作用。从方法特性来看，基于对比学习的无监督图像转换方法，如 CUT，DCLGAN 以及 DSCGAN，都在感知指标上表现优异。这有力地证明了对比学习机制在学习域特征映射方面具有巨大作用。由 Day-DroneVehicle 数据集上的实验结果可知，相较于单一场景，各方法在复杂场景中的转换效果均有不同程度的下降。其中，DR-AVIT 因具备多样化的转换风格，在 FID 和 KID 指标上表现突出，但该方法生成的图像在细节和结构保留方面存在不足，难以满足真实红外图像的模拟需求。值得一提的是，BBDM 虽然在 PSNR 指标上表现良好，但其 SSIM 值低于当前最先进的基于 GAN 的图像转换方法，这类方法在生成图像的感知质量方面也存在一定的局限性。相比之下，DSCGAN 展现出综合最优的结果。

4.3.2 定性结果

图 4 和图 5 分别展示了不同方法在 AVIID-1 和 AVIID-2 数据集上的图像转换结果。从图中可以观察到，DSCGAN 通过引入几何一致性损失，有效保持了目标物体在转换过程中的几何形状的稳定性，有效避免了无监督图像转换中常见的纹理失真问题。其生成结果与采用 L1 损失的各类有监督方法相当，能够清晰地保留地面路砖的边缘。与其他无监督方法相比，DSCGAN 生成的热辐射分布更为合理，能够准确模拟汽车车身以及前后车轮的热辐射特征。在低光环境下，DSCGAN 不仅能够精确模拟汽车的热值分布，还能清晰地展现出背景中楼房和树林的红外图像层次，提升了图像的视觉层次感和真实感。相较于其他方法，DSCGAN 在生成红外图像的真实性和内容保持能力方面展现出显著优势。

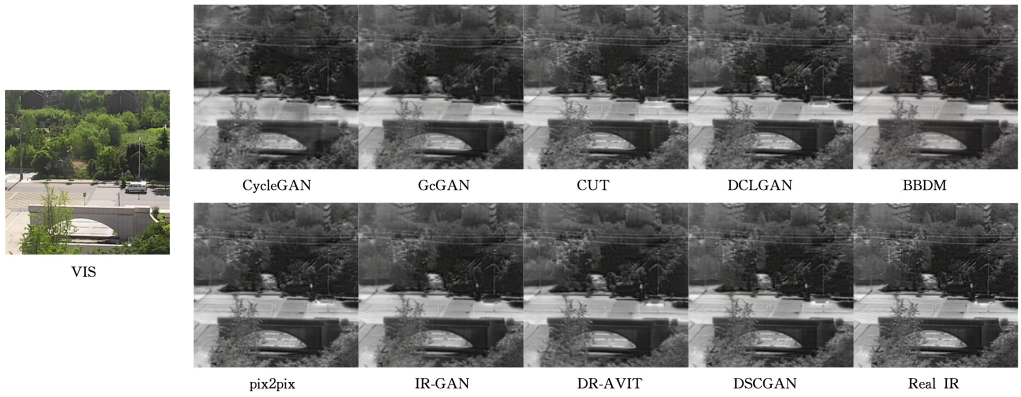

图 4 不同方法在 AVIID-1 数据集上的定性结果

Fig. 4 Qualitative results of different methods on the AVIID-1 dataset

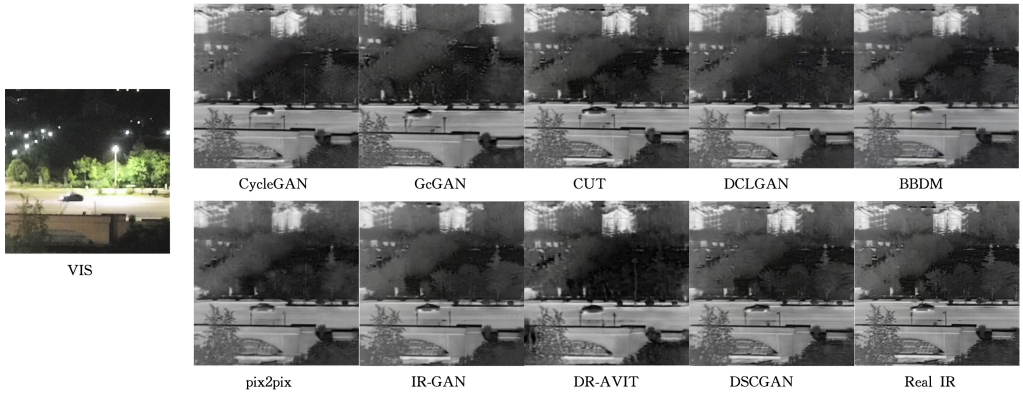

图 5 不同方法在 AVIID-2 数据集上的定性结果

Fig. 5 Qualitative results of different methods on the AVIID-2 dataset

图 6 展示了不同方法在 Day-DroneVehicle 数据集上的图像转换结果。实验结果表明，在复杂场景下，各方法均存在不同程度的模糊、失真或几何变形等问题。基于扩散模型的图像转换方法虽然在图像生成质量方面表现良好，能够有效避免模糊现象，但在内容保持能力方面存在明显不足（如第二组样本中卡车轮廓的几何形变问题），这种现象可能源于扩散模型的条件约束机制不如基于 GAN 的方法严格。在无监督方法中，基于双向转换架构的方法展现出较好的稳定性，未出现模式崩溃现象，这进一步验证了双向架构的鲁棒性。DSCGAN 表现出总体最佳性能。以第四组样本为例，在可见光图像存在橘红色光源干扰的情况下，仅 DSCGAN 准确翻译了路灯的形状和路面的斜向箭头，同时保持了车辆轮廓的完整性。

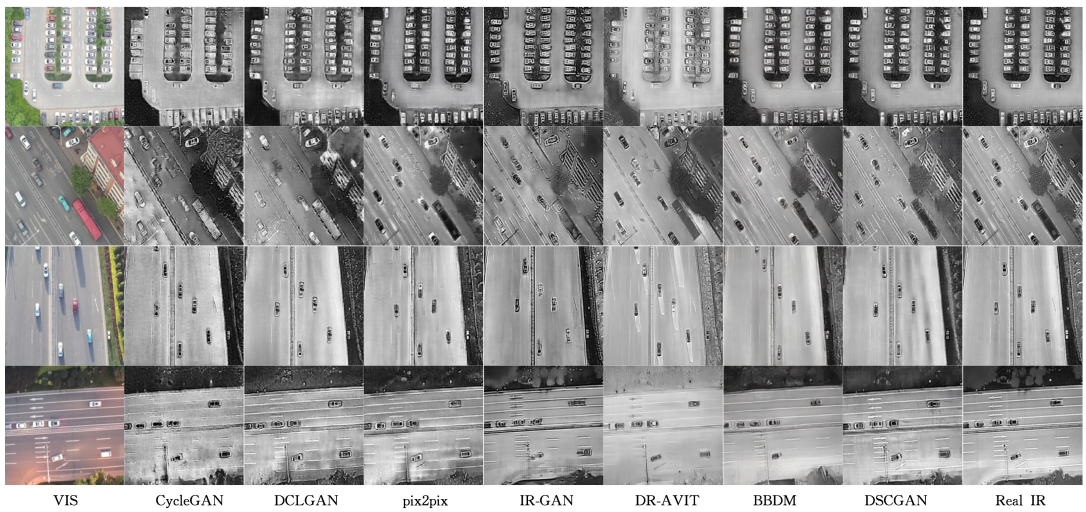

图 6 不同方法在 Day-DroneVehicle 数据集上的定性结果

Fig. 6 Qualitative results of different methods on the Day-DroneVehicle dataset

4.4 消融实验

为验证 DSCGAN 各模块的有效性，在 AVIID-1 和 AVIID-2 数据集上进行渐进式组合消融实验。实验以 resnet_9blocks 为基线生成网络，以 PatchGAN 为判别器，采

用对抗性损失和领域转换损失作为损失函数。在此设定下，基线方法相当于 CycleGAN。表 3 所列数据显示，语义对比学习（SCL）模块展现出最强的独立改进能力，单独添加后各项评估指标均获得显著提升，这验证了该模块在可见光到红外转换任务中的优越性能。虽然单独添加几何一致性损失（GCL）时指标提升幅度相对有限，但当组合使用语义对比学习和几何一致性损失（Model-4）时，模型在夜间场景下取得了 0.7410 的 SSIM 值，显著优于单独使用任意其中某模块的效果，表明语义对齐与几何约束存在互补性：语义对比学习确保热特征符合物理规律，几何一致性损失维持刚性结构稳定，二者共同解决了无监督转换中的语义失真问题。

表 3 消融实验结果

Table 3 Ablation experiment results

Method	SCL	GCL	MSD	AVIID-1					AVIID-2				
				SSIM	*PSNR*	*FID*	*KID*	*LPIPS*	*SSIM*	*PSNR*	*FID*	*KID*	*LPIPS*
baseline				0.5183	21.57	63.57	0.0534	0.2968	0.5505	20.19	64.58	0.0552	0.3022
Model-1	√			0.7213	25.21	35.95	0.0195	0.1886	0.7222	23.26	47.92	0.0265	0.2284
Model-2		√		0.5575	23.39	42.64	0.0267	0.2157	0.5828	21.76	53.79	0.0394	0.2735
Model-3			√	0.6780	24.24	40.75	0.0217	0.2074	0.6858	23.01	50.95	0.2736	0.2318
Model-4	√	√		0.7278	24.81	30.25	**0.0097**	0.1764	0.7410	23.78	42.57	0.0229	0.2153
DSCGAN	√	√	√	**0.8144**	**27.22**	**28.47**	0.0111	**0.1456**	**0.7683**	**25.64**	**40.64**	**0.0212**	**0.1954**

多尺度判别器（MSD）的加入使两个数据集的 FID 分数都大幅下降，表明其多尺度判别机制能显著提升生成图像的真实性。完整模型 DSCGAN 整合 3 个模块后，在 AVIID-1 数据集上的 SSIM 提升至 0.8144，LPIPS 首次降至 0.15 以下，展现了 DSCGAN 在结构纹理保持和视觉感知方面的突出表现。综合所有模块，DSCGAN 取得了最佳结果，从而证明了 3 个模块具有互补性，能促进模型性能的提升。

本文在 AVIID-1 数据集上研究了语义对比损失权重 λ_2 和几何一致性损失权重 λ_3 对红外图像生成质量的影响。如表 4 所列，当 λ_2 从 1.0 增至 2.0 时，各指标均有明显提升，表明增强语义对比学习能有效改善生成质量。而当 λ_3 增至 2.0 时，模型性能下降，说明过强的几何约束会对生成效果产生负面影响。当 $\lambda_2=2.0$ 且 $\lambda_3=1.0$ 时，模型性能最优。

表 4 超参数分析

Table 4 Hyperparameter analysis

λ_2	λ_3	*SSIM*↑	*FID*↓	*LPIPS*↓
1.0	1.0	0.7826	36.15	0.1624
1.5	0.5	0.8039	32.23	0.1531
2.0	1.0	**0.8144**	**28.47**	**0.1456**
2.0	2.0	0.8002	34.8	0.1518

结束语 本文提出基于双重语义对比学习的无监督红外图像生成方法 DSCGAN，用于可见光到红外图像的转换任务。该方法通过构建双向转换框架，引入图像块级语义对比学习机制，构建多尺度 PatchGAN 判别器，并结合几何一致性损失，显著提升了红外图像的生成质量。在 AVIID-1，AVIID-2 和 Day-DroneVehicle 数据集上的实验结果表明，DSCGAN 在昼夜场景下优于多种经典及最新方法，且在保真度指标和感知指标上表现突出，LPIPS 值最低达 0.1456，证明生成的图像质量与人类视觉感知质量极为接近。消融实验验证了各模块对模型性能的积极贡献，表明各关键模块具有有效性和互补性。未来可结合红外图像的下游任务等方向进一步提升模型的泛化能力和实际应用价值。

参 考 文 献

[1] ZHAO M J，LI W，LI L，et al. Single-frame infrared small-target detection：a survey [J]. IEEE Geoscience and Remote Sensing Magazine，2022，10(2)：87-119.

[2] ZHAO X F，ZHAO Y J，HU S C，et al. Progress in active infrared imaging for defect detection in the renewable and electronic industries [J]. Sensors，2023，23(21)：8780.

[3] TANG W，HE F Z，LIU Y，et al. DATFuse：infrared and visible image fusion via dual attention transformer [J]. IEEE Transactions on Circuits and Systems for Video Technology，2023，33(7)：3159-3172.

[4] HOU Y，VOLK R，SOIBELMAN L. A novel building temperature simulation approach driven by expanding semantic segmentation training datasets with synthetic aerial thermal images [J]. Energies，2021，14(2)：353.

[5] POGLIO T，MATHIEU-MARNI S，RANCHIN T，et al. OSIrIS：a physically based simulation tool to improve training in thermal infrared remote sensing over urban areas at high spatial resolution [J]. Remote Sensing of Environment，2006，104(2)：238-246.

[6] KNIAZ V V，KNYAZ V A，HLADUVKA J，et al. ThermalGAN：multimodal color-to-thermal image translation for person re-identification in multispectral dataset [C]// Proceedings of the European Conference on Computer Vision（ECCV） Workshops. Munich：Germany，2018：606-624.

[7] MA D C，XIAN Y，LI B，et al. Visible-to-infrared image translation based on an improved CGAN [J]. The Visual Computer，2024，40(2)：1289-1298.

[8] WANG H N，LI N，ZHAO H J，et al. MappingFormer：learning cross-modal feature mapping for visible-to-infrared image translation [C]// Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne：Australia，2024：10745-10754.

[9] HAN Z H，ZHANG S，SU Y R，et al. DR-AVIT：toward diverse and realistic aerial visible-to-infrared image translation [J]. IEEE Transactions on Geoscience and Remote Sensing，2024，62(5)：1-13.

[10] GOODFELLOW I J，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Nevada：USA，2014：2672-2680.

[11] ISOLA P，ZHU J Y，ZHOU T H，et al. Image-to-image transla-

tion with conditional adversarial networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: USA, 2017: 1125-1134.

[12] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: Italy, 2017: 2223-2232.

[13] LIU M Y, BREUEL T, KAUTZ J. Unsupervised image-to-image translation networks [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. California: USA, 2017: 700-708.

[14] FU H, GONG M M, WANG C H, et al. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. California: USA, 2019: 2427-2436.

[15] PARK T, EFROS A A, ZHANG R, et al. Contrastive learning for unpaired image-to-image translation [C]// Proceedings of the 16th European Conference on Computer Vision. Glasgow: UK, 2020: 319-345.

[16] HAN J, SHOEIBY M, PETERSSON L, et al. Dual contrastive learning for unsupervised image-to-image translation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: USA, 2021: 746-755.

[17] LI B, XUE K T, LIU B, et al. BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models [C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 1952-1961.

[18] XIA M F, ZHOU Y, YI R, et al. A Diffusion Model Translator for Efficient Image-to-Image Translation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10272-10283.

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

[20] HU X Q, ZHOU X Y, HUANG Q S, et al. Qs-attn: query-selected attention for contrastive learning in i2i translation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: USA, 2022: 18291-18300.

[21] JUNG C, KWON G, YE J C. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: USA, 2022: 18260-18269.

[22] HAN Z H, ZHANG Z Y, ZHANG S, et al. Aerial visible-to-infrared image translation: dataset, evaluation, and baseline [J]. Journal of Remote Sensing, 2023, 3(1): 96.

CHENG Zimeng, born in 2002, postgraduate, is a member of CCF (No. 02769G). Her main research interests include computer vision and image-to-image translation.

AI Haojun, born in 1972. Ph.D, associate professor, is a senior member of CCF (No. 06059S). His main research interests include computer vision, artificial intelligence and deepfake detection.

(责任编辑:李亚辉)