中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)08-3045-15

论文引用格式:Du W L, Guo B, Zhao J Q, Yao R and Zhou Y. 2026. Semantic-guided contrastive learning for SAR and optical image translation. Journal of Image and Graphics, 31(8):3045-3059(杜文亮, 郭波, 赵佳琦, 姚睿, 周勇. 2026. 语义引导对比学习的SAR与光学图像转换. 中国图象图形学报, 31(8):3045-3059)[DOI:10.11834/jig.250526]

语义引导对比学习的SAR与光学图像转换

杜文亮,郭波,赵佳琦,姚睿,周勇*

1. 中国矿业大学计算机科学与技术学院/人工智能学院,徐州　221116; 2. 矿山数字化教育部工程研究中心,徐州　221116;
3. 地下空间智能感知与应急物联江苏省产业技术工程化中心,徐州　221116

摘　要:**目的**　合成孔径雷达(synthetic aperture radar, SAR)与光学图像转换能够融合两种模态数据的优势,提供全天时、全天候与高分辨率的观测能力。然而,当前基于循环一致性生成对抗网络的方法主要侧重于图像结构的宏观重建,未能充分利用跨模态间的深层语义信息来指导图像生成,限制了生成图像的语义保真度和在下游任务中的性能。同时,现有基于对比学习的转换方法在处理遥感图像时,因同类地物特征高度自相关导致正负样本难以区分,造成对比机制失效。针对上述问题,提出了一种语义引导对比学习的SAR与光学图像转换方法。**方法**　提出基于语义分割的特征提取模块,利用预训练的SAR与光学语义分割模型提取像素级语义信息;提出了语义引导的对比学习模块,利用先验的语义分割信息,在对比学习空间中显式构建基于类别一致性的正负样本筛选机制,有效解决了遥感图像特征同质化导致的传统对比学习失效问题;设计了融合循环生成结构与对比学习的联合优化框架,通过引入循环语义分割损失与生成对抗损失,约束生成图像在结构、纹理和语义层面的一致性。**结果**　实验在WHU-OPT-SAR(Wuhan University-optical-SAR)和DDHRNet(dual-stream deep high-resolution network)两个公开数据集上进行。实验结果表明,与当前优秀方法相比,在SAR到光学及光学到SAR的图像转换任务中,生成质量指标分别最高提升了11.9%和3.8%;在下游任务中,语义分割准确率分别提升了49.26%和10.60%,特征匹配的正确内点比例最高提升了1%。消融实验研究表明,语义引导对比学习模块与循环语义分割损失对提升模型性能均起到关键作用。**结论**　本文提出的语义引导对比学习的SAR与光学图像转换方法,能够有效解决传统对比学习在遥感图像转换中的失效问题,显著提升了生成图像的语义保真度与跨模态特征对齐能力,在下游语义分割和图像匹配任务中取得了最优的综合性能,为无监督SAR与光学图像转换提供了新的解决思路。本文代码开源在链接:https://www.scidb.cn/s/VVVBnu。

关键词:图像转换;合成孔径雷达(SAR);遥感图像;对比学习;语义分割

Semantic-guided contrastive learning for SAR and optical image translation

Du Wenliang, Guo Bo, Zhao Jiaqi, Yao Rui, Zhou Yong*

1. School of Computer Science and Technology / School of Artificial Intelligence, China University of Mining and Technology, Xuzhou　221116, China; 2. Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou　221116, China; 3. Jiangsu Provincial Industrial Technology Engineering Center for Intelligent Sensing and Emergency IoT in Underground Space, Xuzhou　221116, China

收稿日期:2025-11-07;**修回日期**:2026-01-08;**预印本日期**:2026-01-15

*通信作者:周勇　yzhou@cumt.edu.cn

基金项目:国家自然科学基金项目(62272461,62172417,62002360);中国矿业大学"双一流"建设提升自主创新能力项目(2022ZZCX06);澳门科学技术发展基金项目(0020/2024/RIA1)

Supported by: National Natural Science Foundation of China(62272461, 62172417, 62002360); "Double First-Class" Project of China University of Mining and Technology for Independent Innovation and Social Service(2022ZZCX06); Science and Technology Development Fund of Macau(0020/2024/RIA1)

Abstract: Objective Synthetic aperture radar (SAR) and optical remote sensing represent two complementary Earth observation modalities. SAR imagery enables all-weather, all-day observation due to its active imaging mechanism, while optical imagery provides higher spatial resolution and more intuitive visual details. Cross-modal translation between SAR and optical images not only compensates for missing data in either modality but also enhances downstream applications, such as image matching and semantic segmentation, through multi-modal information fusion. However, current methods encounter notable challenges. Approaches based on cycle-consistent generative adversarial networks (CycleGAN) predominantly emphasize macroscopic structural reconstruction and fail to exploit deep semantic correlations across source and target domains. In particular, traditional contrastive learning techniques are limited in remote sensing due to the high spatial autocorrelation of similar land-cover features, which blurs the distinction between positive and negative samples and causes the contrastive mechanism to underperform. This study addresses these challenges by proposing a semantic-guided contrastive learning method for SAR and optical image translation, which effectively mitigates feature homogenization in remote sensing image translation and significantly improves semantic fidelity and downstream task performance. **Method** The proposed framework comprises three core components: 1) a semantic feature extraction module, 2) a semantic-guided contrastive learning module, and 3) a joint optimization scheme combining cyclic generation and contrastive learning. Initially, this approach utilizes pre-trained SAR and optical semantic segmentation models (DeepLabV3) to extract pixel-level semantic features from real and reconstructed images. Subsequently, the semantic-guided contrastive learning module implements a class-consistency-based positive-negative sample selection strategy: In each query patch in the generated image, patches sharing the same semantic class in the real image serve as positive samples, while those from differing classes are treated as negatives. This strategy effectively mitigates the feature homogenization problem that challenges traditional contrastive learning in remote sensing contexts. Contrastive loss is computed within a shared feature space projected by a lightweight perceptron. Finally, the joint optimization framework integrates cyclic consistency losses at pixel and semantic levels with adversarial losses. The cyclic semantic segmentation loss enforces consistency in structure, texture, and semantics between generated and real images, while adversarial losses enhance image realism. The overall loss function balances these components via weighted hyperparameters. **Result** Comprehensive experiments on two public datasets—WHU-OPT-SAR and DDHRNet—evaluate the proposed method against state-of-the-art approaches, including CycleGAN, contrastive unpaired translation, query-selected attention, and conditional diffusion, across image translation quality and downstream tasks. Regarding image translation, our method consistently achieves superior performance. On WHU-OPT-SAR, SAR-to-optical translation yielded a peak signal-to-noise ratio (PSNR) improvement of 11.9% and a mean absolute error (MAE) reduction of 31.1% over the second-best method; optical-to-SAR translation achieved gains of 3.8% in PSNR, 5.6% in structural similarity index measure (SSIM), and a 5.2% reduction in MAE. On DDHRNet, our method sustained leading performance across diverse geographical contexts. In downstream tasks, the semantic segmentation and feature matching results confirm marked gains. The semantic segmentation pixel accuracy improved by 49.26% for optical image generation and 10.60% for SAR image generation on WHU-OPT-SAR, outperforming all baselines. In feature matching tasks, the inlier ratio (IR) and Euclidean distance-based inlier ratio improved by up to 1% and 0.49%, respectively. Qualitative comparisons reveal superior visual quality with enhanced detail preservation and grayscale consistency. Ablation studies demonstrate the necessity of each component: removal of the semantic-guided contrastive learning module caused a 10.75% drop in pixel accuracy for optical images, while omitting the cyclic semantic segmentation loss reduced PSNR by 7.0% in optical-to-SAR translation. These findings validate the critical role of our innovations in overcoming remote sensing translation challenges. **Conclusion** The proposed semantic-guided contrastive learning method for SAR and optical image translation effectively addresses the prevalent issue of feature homogenization in remote sensing image translation. Our method substantially enhances the semantic fidelity and downstream utility of generated images by integrating semantic segmentation guidance with contrastive learning within a cyclic generative framework, offering a novel solution for unsupervised SAR and optical image translation. Extensive experiments demonstrate that this approach provides a robust and effective solution for unpaired SAR-optical image translation with notable advantages in semantic consistency and cross-modal feature alignment. Our main contributions include: 1) identifying the feature homogenization problem inherent in traditional contrastive learning for remote sensing translation; 2) proposing a novel semantic-guided contrastive learning frame-

work with category-consistent sample selection; and 3) developing a unified architecture combining cyclic generation and contrastive learning. Future work will explore self-supervised semantic guidance to improve applicability in scenarios lacking semantic annotations, thereby enhancing generalizability and robustness across diverse remote sensing applications. Our codes are available at https://www.scidb.cn/s/VVVBnu.

Key words: image-to-image translation; synthetic aperture radar (SAR); remote sensing image; contrastive learning; semantic segmentation

0 引　言

合成孔径雷达(synthetic aperture radar, SAR)与光学遥感是两种互补的对地观测手段。SAR图像凭借其主动成像机制,能够穿透云雾、不受光照限制,提供全天时、全天候的观测能力。与之相比,光学图像则具有更高的空间分辨率和更符合人类视觉感知的物理特性,信息更直观。因此,研究SAR与光学图像之间的跨模态转换技术,不仅能在任一模态数据缺失时进行补充,更能通过融合多模态信息,为图像匹配、语义分割、图像去雾(Wang等,2025)和目标检测(Hu等,2025)等下游应用提供更鲁棒、更丰富的数据支持。

循环一致性生成对抗网络(cycle-consistent generative adversarial network, CycleGAN)(Zhu 等, 2017)首次引入循环一致性损失,解决了非配对数据下的域适应问题。然而,该方法在处理跨模态等风格差异显著的转换任务时,存在模式崩溃的风险。为提升跨模态转换稳定性并支持多属性协同转换,双重对偶生成对抗网络(吴柳玮 等,2020)提出通过4个生成器和判别器的对偶架构实现风格与年龄双属性协同转换,并引入专属损失函数保障语义一致性。Torbunov等人(2023)在CycleGAN架构中嵌入视觉Transformer瓶颈层,结合梯度惩罚与自监督预训练缓解模式崩溃问题。在遥感领域,融合残差网络与密集连接卷积网络的并行生成器网络(余佩伦等,2021)提出通过双分支分别处理输入图像,结合基于通道阈值分割的线性插值融合策略优化输出,为跨模态转换的网络结构设计提供了多分支融合的有效思路;Yang等人(2022)以CycleGAN为基础,针对无监督SAR到光学图像转换中生成图像细节不足的问题,提出了包含非平衡生成器、多尺度判别器和综合归一化组的细粒度网络,显著提升了生成图像的细节质量与物理表征一致性;Guo等人(2024)则针对现有方法中场景记忆不足导致的区域地貌变形和斑点噪声干扰问题,设计了多尺度表示生成器以实现SAR图像场景特征的多尺度融合与利用,构建多感受野判别器以增强不同地貌场景的记忆与生成质量,并引入子带收缩去噪模块有效抑制斑点噪声的影响。然而,当前基于CycleGAN的方法主要侧重于图像结构的宏观重建,未能充分挖掘原始域与目标域之间的深层语义关联。

鉴于此,基于对比无监督转换方法(contrastive unpaired translation, CUT)被提出(Park等,2020),通过引入对比学习机制,充分利用原始图像与生成图像的局部语义信息,通过最大化图像块(Patch)间的互信息实现细节一致性优化,在自然图像转换任务中取得性能提升。基于CUT框架,Kang和Park(2020)结合对比损失与条件判别策略,提升转换质量与训练稳定性。为进一步增强对比效果,Wang等人(2021)提出实例级困难负样本生成方法,有效提升模型性能;Han等人(2021)通过双编码器结构增强非配对图像间的映射能力。后续研究进一步关注负样本质量与语义结构建模,如查询选择注意力(query-selected attention, QS-Attn)(Hu 等,2022)选择关键锚点进行语义对齐;Zhan等人(2022)基于图像对比度动态调整负样本权重,缓解伪影问题;Jung等人(2022)结合语义关系正则化与解耦对比学习,有效提升图像内部异质语义的表达能力与转换多样性。对比学习的核心目标在于最大化正样本对之间的互信息,同时最小化负样本对之间的互信息。然而,遥感图像的特殊性在于:同类地物在特征空间中的分布常高度一致(例如水体和森林等),导致同一语义类别的负样本与正样本在特征空间中的相似度极高甚至趋同,从而使传统对比学习在处理相同语义特征的样本时失效,严重制约了其在遥感图像转换任务中的效能。如图1所示,蓝、绿和红框分别代表查询、正样本和负样本图像块。在以CUT为代表的传统对比学习方法中,查询图像块与正样本图像

块在对比空间中的特征应尽可能相似，而与负样本的特征则应尽可能区分。但在图1中，同一语义类别的正样本与负样本的特征极为接近，若继续采用传统对比学习策略，将导致对比机制失效。

因此，本文提出语义引导对比学习框架，利用先验的语义分割信息（Jiang等，2025），在对比学习空间中显式构建基于类别一致性的正负样本筛选机制（如图1），以突破遥感图像特征同质化对传统对比学习的约束。同时，为进一步有效利用语义分割信息，在语义引导对比学习框架中加入了CycleGAN模型的循环生成结构，并计算循环重构图像与原始图像之间的语义分割损失。

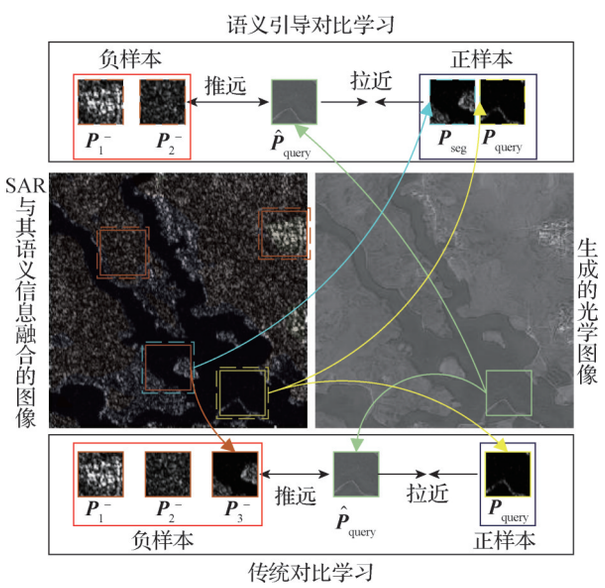


图1　传统对比学习与语义引导对比学习示意图

Fig. 1　Schematic diagram of traditional contrastive learning and semantic-guided contrastive learning

综上，本文主要贡献如下：1）揭示了传统对比学习应用在多源遥感图像转换过程中遇到的特征同质性问题；2）提出语义引导对比学习框架，利用语义分割标签引导查询图像块选择同类区域为正样本，跨类区域为负样本，有效避免遥感图像的同质性问题；3）构建了融合循环生成结构与对比学习的图像转换框架，为非配对遥感图像转换提供了新思路。

1　方　法

本文提出的语义分割引导对比学习的SAR与光学图像转换方法的整体框架如图2所示。该框架主要包括两个核心模块：语义特征提取模块和语义引导的对比学习模块。

如图2所示，本文方法同时训练了两个生成器：SAR到光学图像生成器$G_{S\to O}$和光学到SAR图像生成器$G_{O\to S}$。本文方法将原始的SAR图像和光学图像（$\boldsymbol{real}_{sar}$和$\boldsymbol{real}_{opt}$）分别输入至生成器$G_{S\to O}$和$G_{O\to S}$，得到生成的光学和SAR图像：$\boldsymbol{fake}_{opt}$和$\boldsymbol{fake}_{sar}$，然后将$\boldsymbol{fake}_{opt}$和$\boldsymbol{fake}_{sar}$输入至生成器$G_{O\to S}$和$G_{S\to O}$，得到重建的SAR和光学图像：$\boldsymbol{rec}_{sar}$和$\boldsymbol{rec}_{opt}$。在生成器训练过程中，原始和重建的图像需输入至语义特征提取模块，获得相应的语义分割图。最后，根据原始和重建的SAR与光学图像及其语义分割图，求解分割对比损失、循环语义分割损失、循环损失以及生成对抗损失，以训练生成器和判别器的参数。

1.1　语义特征提取模块

语义特征提取模块由预训练的SAR与光学语义分割模型组成。本文方法的语义分割模型使用的是DeepLabV3（Chen等，2017）。其中，在SAR图像及其语义分割标注训练集中训练获得的DeepLabV3模型用于获得SAR图像的语义分割特征，命名为SARSeg；在光学及其语义分割标注训练集中训练获得的DeepLabV3模型用于获得光学图像的语义分割特征，命名为OPTSeg。

真实的SAR与光学图像（$\boldsymbol{real}_{sar}$和$\boldsymbol{real}_{opt}$）以及重建的SAR与光学图像（$\boldsymbol{rec}_{sar}$和$\boldsymbol{rec}_{opt}$）都需在语义特征提取模块中获得其相应的语义特征。

真实以及重构的SAR语义分割特征由相应图像经SARSeg分割模型获得，并命名为$\boldsymbol{Seg}_{sar}^{real}$和$\boldsymbol{Seg}_{sar}^{rec}$；真实以及重构的光学语义分割特征由相应图像经OPTSeg分割模型获得，并命名为$\boldsymbol{Seg}_{opt}^{real}$和$\boldsymbol{Seg}_{opt}^{rec}$。其中，真实的语义分割特征输入至语义引导的对比学习模块，用于判断正确的正负样本；真实和重构的SAR与光学语义分割特征用于计算生成器的语义损失L_{seg}，约束生成器图像生成的语义一致性，计算式为

$$L_{seg}=L_{ce}\left(\boldsymbol{Seg}_{sar}^{rec},\operatorname{argmax}\left(\boldsymbol{Seg}_{sar}^{real}\right)\right)+L_{ce}\left(\boldsymbol{Seg}_{opt}^{rec},\operatorname{argmax}\left(\boldsymbol{Seg}_{opt}^{real}\right)\right)\tag{1}$$

式中，L_{ce}代表交叉熵损失。

1.2　语义引导的对比学习模块

由于$G_{S\to O}$和$G_{O\to S}$两个生成器的训练过程都使用了语义引导的对比学习模块，且使用过程一致，因

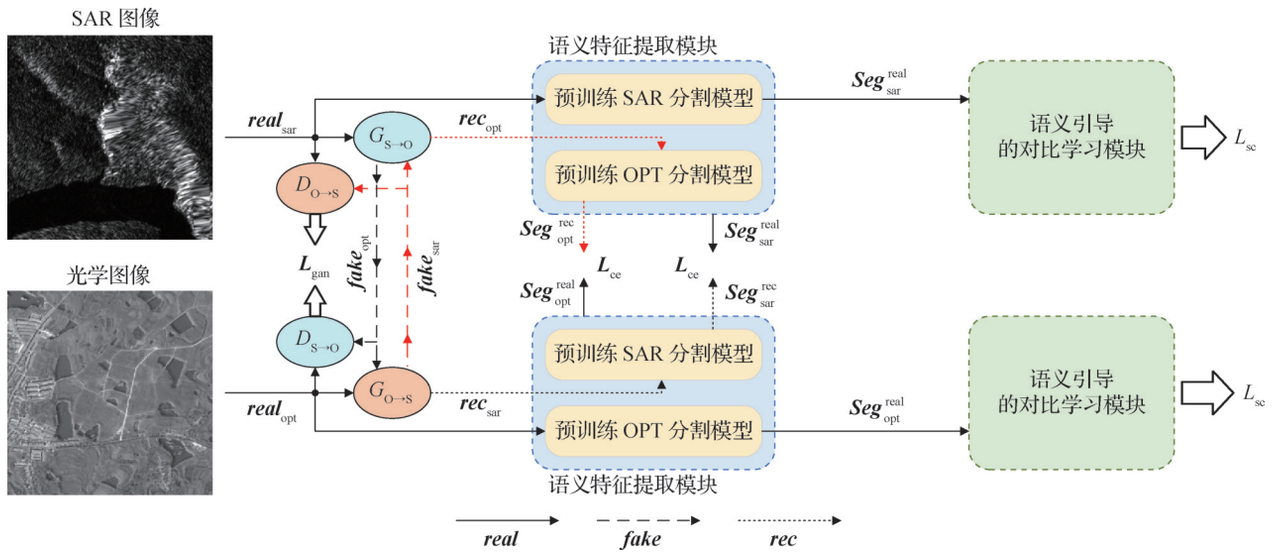


图2　本文提出方法的架构图

Fig. 2　Architecture diagram of the method proposed in this article

此，以SAR图像到光学图像生成为例，介绍语义引导的对比学习模块。语义引导的对比学习模块示意图如图3所示，主要包含一个特征编码器（$E^l_{S\to O}$）和一个特征映射感知机（$MLP^l_{S\to O}$）。其中，$\hat{\boldsymbol{p}}_{query}$为生成查询向量，$\boldsymbol{p}_{query}$为真实查询向量，$\boldsymbol{p}_{seg}$为与查询向量相同类别的特征向量，$\boldsymbol{p}_{-}$为负样本特征向量。特征编码器$E_{S\to O}$是$G_{S\to O}$的一部分，用于提取真实SAR图像与生成光学图像的特征。$E^l_{S\to O}$表示编码器$E_{S\to O}$第l层的输出。其中，$l\in\{1,2,3,\cdots,L\}$，L表示$E_{S\to O}$为从$G_{S\to O}$提取的L层网络。在网络架构层面，本模块沿用了CUT模型（Park等，2020）中提出的图像块级对比学习框架，通过特征编码器（$E_{S\to O}$）提取L层特征，并利用与CUT相同结构的多层感知机（multilayer perceptron，MLP）将特征映射至对比空间。

然而，传统的CUT方法在对比学习策略上仅基于空间位置划分正负样本，即同一位置的图像块为正样本其余为负样本。这种策略在处理具有高纹理重复性的遥感图像时容易失效。与之不同，本文提出的语义引导对比学习模块在对比学习架构的基础上，设计并引入了语义一致性筛选机制，以确保正负样本的语义一致性，其具体过程如下：

在构建语义引导的对比学习正样本与负样本过程中，首先将在生成图像（$\boldsymbol{fake}_{opt}$或$\boldsymbol{fake}_{sar}$）中选取的

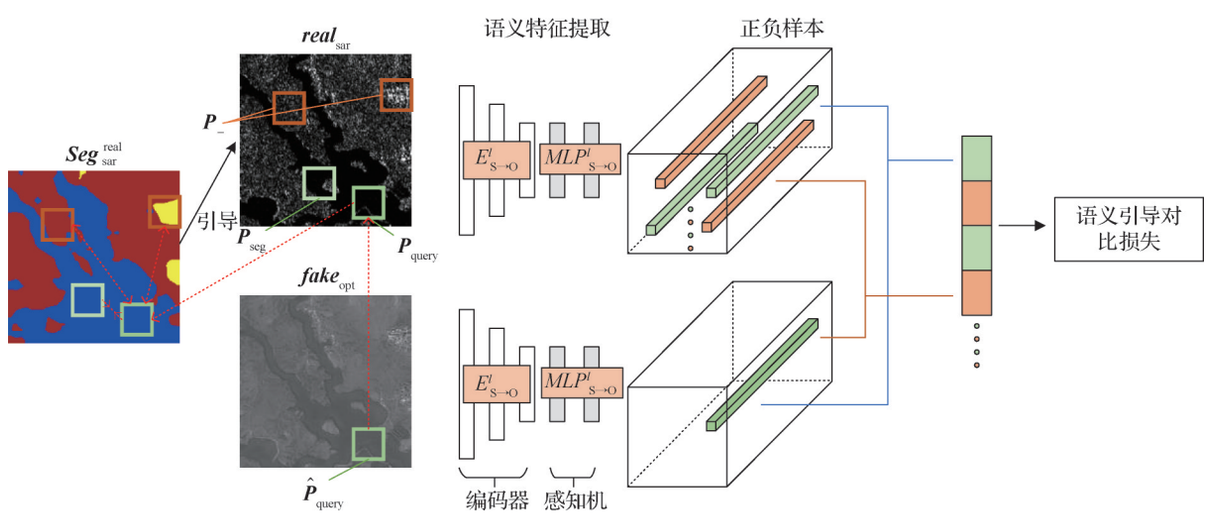


图3　语义引导的对比学习模块

Fig. 3　Semantic-guided contrastive learning module

一个图像块(Patch)作为生成查询块$\hat{\boldsymbol{p}}_{\text{query}}$,并将真实图像中对应位置的图像块作为真实查询块$\boldsymbol{p}_{\text{query}}$。然后,根据$\boldsymbol{p}_{\text{query}}$在真实图像语义分割特征$\boldsymbol{Seg}_{\text{sar}}^{\text{real}}$中对应的类别,获取真实图像中同类别的图像块$\boldsymbol{p}_{\text{seg}}$,并将不同类别的图像块作为负样本图像块$\boldsymbol{p}_{-}$。正样本图像块$\boldsymbol{p}_{+}$则为同类别图像块$\boldsymbol{p}_{\text{seg}}$与真实查询块$\boldsymbol{p}_{\text{query}}$的并集,即$\boldsymbol{p}_{+}=\{\boldsymbol{p}_{\text{query}},\boldsymbol{p}_{\text{seg}}\}$。传统对比学习方法通常采用$1\times1$像素作为图像块的尺寸,此时$\boldsymbol{p}_{+}$中图像块的个数与$\boldsymbol{p}_{-}$中图像块个数之和为图像的像素个数。

接下来,利用特征编码器$E_{\text{S}\to\text{O}}^{l}$和特征映射感知机$MLP_{\text{S}\to\text{O}}^{l}$将$\hat{\boldsymbol{p}}_{\text{query}}$、$\boldsymbol{p}_{\text{query}}$、$\boldsymbol{p}_{\text{seg}}$、$\boldsymbol{p}_{+}$和$\boldsymbol{p}_{-}$映射到对比空间中,得到相应的$C_l$维的特征向量:$\hat{\boldsymbol{v}}_{\text{query}}^{l}\in\mathbf{R}^{C_l}$、$\boldsymbol{v}_{\text{query}}^{l}\in\mathbf{R}^{C_l}$、$\boldsymbol{v}_{\text{seg}}^{l}\in\mathbf{R}^{P\times C_l}$、$\boldsymbol{v}_{+}^{l}\in\mathbf{R}^{(P+1)\times C_l}$和$\boldsymbol{v}_{-}^{l}\in\mathbf{R}^{N\times C_l}$。其中,$P$为相同类别块的个数,$N$为不同类别图像块的个数,$C_l$为特征编码器第$l$层特征维度。设置语义引导的交叉熵损失,令$\hat{\boldsymbol{v}}_{\text{query}}^{l}$与$\boldsymbol{v}_{+}^{l}$更相似、与$\boldsymbol{v}_{-}^{l}$更不同。语义引导的交叉熵损失为

$$\ell(\hat{\boldsymbol{v}}_{\text{query}}^{l},\boldsymbol{v}_{+}^{l},\boldsymbol{v}_{-}^{l})=-\log\left[\frac{\sum\limits_{\boldsymbol{v}_{+}^{l}\in\boldsymbol{Posi}}\exp\left(\hat{\boldsymbol{v}}_{\text{query}}^{l}\cdot\frac{\boldsymbol{v}_{+}^{l}}{\tau}\right)}{\sum\limits_{\boldsymbol{v}_{+}^{l}\in\boldsymbol{Posi}}\exp\left(\hat{\boldsymbol{v}}_{\text{query}}^{l}\cdot\frac{\boldsymbol{v}_{+}^{l}}{\tau}\right)+\sum\limits_{\boldsymbol{v}_{+}^{l}\in\boldsymbol{Neg}}\exp\left(\hat{\boldsymbol{v}}_{\text{query}}^{l}\cdot\frac{\boldsymbol{v}_{-}^{l}}{\tau}\right)}\right]\tag{2}$$

式中,$\boldsymbol{Posi}$和$\boldsymbol{Neg}$分别代表正样本和负样本特征向量的集合,$\boldsymbol{Posi}=\left\{\boldsymbol{v}_{\text{query}}^{l},\boldsymbol{v}_{\text{seg}}^{l(1)},\cdots,\boldsymbol{v}_{\text{seg}}^{l(S)}\right\}$,$\boldsymbol{Neg}=\left\{\boldsymbol{v}_{-}^{l(1)},\cdots,\boldsymbol{v}_{-}^{l(N)}\right\}$,$\tau$表示缩放系数。

由于SAR到光学和光学到SAR生成器的训练过程都使用了语义引导的对比学习模块训练,因此编码器$E_{\text{S}\to\text{O}}$和$E_{\text{O}\to\text{S}}$在选取的L个特征层上都计算了两个模态在对比空间中的对比损失,以确保两个模态特征在对比空间的语义一致性。同时,还需在两个编码器对应的L个特征层计算各模态的一致性损失,以约束各模态在对比空间中语义的不变性。其中,对于编码器的第l层,查询块共有Q_l个。综上,语义引导的对比学习损失L_{sc}的计算过程为

$$\begin{aligned}L_{\text{sc}}=&\text{E}_{s\sim S}\sum_{l=1}^{L}\sum_{q=1}^{Q_l}\ell\left(\hat{\boldsymbol{v}}_{\text{query}}^{l(q)},\boldsymbol{v}_{+}^{l(q)},\boldsymbol{v}_{-}^{l(q)}\right)+\\&\text{E}_{o\sim O}\sum_{l=1}^{L}\sum_{q=1}^{Q_l}\ell\left(\hat{\boldsymbol{v}}_{\text{query}}^{l(q)},\boldsymbol{v}_{+}^{l(q)},\boldsymbol{v}_{-}^{l(q)}\right)+\\&\text{E}_{s\sim S}\sum_{l=1}^{L}\sum_{q=1}^{Q_l}\ell\left(\bar{\boldsymbol{v}}_{\text{query}}^{l(q)},\boldsymbol{v}_{+}^{l(q)},\boldsymbol{v}_{-}^{l(q)}\right)+\\&\text{E}_{o\sim O}\sum_{l=1}^{L}\sum_{q=1}^{Q_l}\ell\left(\bar{\boldsymbol{v}}_{\text{query}}^{l(q)},\boldsymbol{v}_{+}^{l(q)},\boldsymbol{v}_{-}^{l(q)}\right)\end{aligned}\tag{3}$$

式中,E表示数学期望,$\boldsymbol{S}$和$\boldsymbol{O}$分别代表真实的SAR和光学图像的数据集,$\boldsymbol{s}$和$\boldsymbol{o}$分别代表数据集中真实的SAR和光学图像,$\bar{\boldsymbol{v}}_{\text{query}}^{l}$表示一致性生成图像($\boldsymbol{idt}_{\text{sar}}$或$\boldsymbol{idt}_{\text{opt}}$)查询块对应的特征向量,$\boldsymbol{idt}_{\text{sar}}$和$\boldsymbol{idt}_{\text{opt}}$由$G_{\text{O}\to\text{S}}(\boldsymbol{real}_{\text{sar}})$和$G_{\text{S}\to\text{O}}(\boldsymbol{real}_{\text{opt}})$得到。式(3)的前半部分为两模态间在对比空间的对比损失,后半部分为各模态内在对比空间的一致性损失。

1.3 损失函数

本文方法除了使用前两小节提到的语义损失L_{seg}和语义引导的对比学习损失L_{sc},还使用了生成对抗损失和循环损失以保证SAR与光学图像转换的基础生成质量。

生成对抗损失的计算式为

$$\begin{aligned}L_{\text{gan}}=&\text{E}_{o\sim O}\log D_{\text{S}\to\text{O}}\left(\boldsymbol{real}_{\text{opt}}\right)+\\&\text{E}_{s\sim S}\log\left(1-D_{\text{S}\to\text{O}}\left(G_{\text{S}\to\text{O}}\left(\boldsymbol{real}_{\text{sar}}\right)\right)\right)+\\&\text{E}_{s\sim S}\log D_{\text{O}\to\text{S}}\left(\boldsymbol{real}_{\text{sar}}\right)+\\&\text{E}_{o\sim O}\log\left(1-D_{\text{O}\to\text{S}}\left(G_{\text{O}\to\text{S}}\left(\boldsymbol{real}_{\text{opt}}\right)\right)\right)\end{aligned}\tag{4}$$

循环损失函数计算过程为

$$\begin{aligned}L_{\text{cyc}}=&\lambda_1\left(\left\|\boldsymbol{real}_{\text{sar}}-\boldsymbol{rec}_{\text{sar}}\right\|_1+\left\|\boldsymbol{real}_{\text{opt}}-\boldsymbol{rec}_{\text{opt}}\right\|_1\right)+\\&\lambda_2\left(\left\|\boldsymbol{real}_{\text{sar}}-\boldsymbol{idt}_{\text{sar}}\right\|_1+\left\|\boldsymbol{real}_{\text{opt}}-\boldsymbol{idt}_{\text{opt}}\right\|_1\right)\end{aligned}\tag{5}$$

式中,λ_1和λ_2为超参数,分别设置为10和0.5。

本文总损失函数为

$$L_{\text{all}}=L_{\text{cyc}}+\lambda_3L_{\text{gan}}+\lambda_4L_{\text{seg}}+\lambda_5L_{\text{sc}}\tag{6}$$

式中,λ_3、λ_4和λ_5为超参数,统一设置为1。

2 实验结果及分析

2.1 对比方法与实验设置

本文对比方法包括两个基线方法CycleGAN和CUT以及当前最优的基于对比学习的图像转换方法QS-Attn和基于扩散模型的图像转换方法条件扩散(conditional diffusion, Con-Diffusion)(Bai等,2024)。

本文实验均在包含1个NVIDIA GeForce RTX 4090D GPU的服务器上完成。批次大小设置为2,训练epoch数设置为100,优化器采用Adam。学习率在前50个epoch训练过程中设置为2×10^{-4},在后50个epoch以线性衰减到0。式(2)中τ设置为0.07。

2.2 数据集

实验采用两个SAR与光学图像语义分割数据

集：WHU-OPT-SAR（Wuhan University-optical-SAR）（Li 等，2022）和 DDHRNet（dual-stream deep high-resolution network）（Ren 等，2022）。WHU-OPT-SAR 数据集由高分 1 号拍摄的光学图像与高分 3 号拍摄的 SAR 图像构成。该数据集的训练集、验证集和测试集分别包含 20 580 对、5 880 对和 2 940 对分辨率为 256 × 256 像素的 SAR 和光学图像及其对应的语义分割标注。由于现有图像转换模型训练速度较慢，本文从训练集中随机选取 2 000 对图像，用于训练本文模型及 4 种对比方法。

DDHRNet 数据集由高分 2 号拍摄的光学图像与高分 3 号拍摄的 SAR 图像构成，覆盖中国西安、山东及韩国区域，本文实验仅使用山东区域数据。该数据集经随机采样后按 8∶1∶1 的比例划分为训练集、验证集和测试集，分别包含 6 264 对、783 对和 784 对分辨率为 256 × 256 像素的 SAR 与光学图像及其语义分割标注。

两个数据集中的 SAR 与光学图像均为严格配对（paired），可用于有效评估图像转换质量。然而，为模拟仅有非配对数据时的训练场景，本文在训练图像转换模型时，对 SAR 与光学图像分别进行随机采样，使得每个批次中的图像均为非配对形式。

2.3 评估指标

本文对从图像转换质量与下游任务性能两个维度对比模型的 SAR 到光学图像转换质量进行评价。其中，图像转换质量评估采用峰值信噪比（peak signal-to-noise ratio，PSNR）、结构相似性指标（structural similarity index measure，SSIM）以及平均绝对误差（mean absolute error，MAE）作为评估指标。

下游任务评估包括语义分割与图像匹配两项任务。在语义分割任务中，本文采用在语义特征提取模块中针对 SAR 和光学图像分别训练的 DeepLabV3 模型（SARSeg（SAR segmentation）和 OPTSeg（optical segmentation）），对转换后的图像进行语义分割，并使用像素准确率（pixel accuracy，PA）（Badrinarayanan 等，2017）对分割结果进行评估。

在特征匹配任务中，使用在两个数据集上训练的特征匹配模型（Yagmur 等，2024）进行特征点提取与粗匹配，并将粗匹配点数量与提取特征点数量的比值（putative matches ratio，PMR）作为第 1 个评价指标。然后使用随机抽样一致（random sample consensus，RANSAC）算法筛选出正确匹配点，以正确匹配点数量占粗匹配点数量的比例（inlier ratio，IR）作为第 2 个评估指标。此外，由于两个数据集中的 SAR 与光学图像均为严格配对，可通过计算匹配特征点的欧氏距离直接判定正确匹配点，将欧氏距离小于 3 个像素的匹配点定义为欧氏距离正确匹配，并将其数量与粗匹配点数量的比值（inlier ratio of Euclidean distance，IR-ED）作为第 3 个评估指标。

2.4 对比实验和分析

2.4.1 定量实验结果与分析

首先对各类方法在图像转换质量方面的定量实验结果进行对比分析，进而评估其在下游任务中的表现。各方法在 WHU-OPT-SAR 和 DDHRNet 数据集上的 SAR 与光学图像跨模态转换任务的图像质量定量实验结果如表 1 和表 2 所示。在两个数据集上，本文方法与 CycleGAN 在 SAR 到光学图像生成及光学到 SAR 图像生成的双向任务中，均展现出优于其他对比方法的图像生成质量。因此，若仅从传统的图像生成质量评估指标来看，基于循环生成方法在 SAR 与光学图像的跨模态生成中，与真实图像具有更高的像素相似度（如 PSNR、MAE）和结构相似性（如 SSIM）。相比之下，基于对比学习的方法 CUT 和 QS-Attn 在图像生成质量评估指标上的表现弱于 CycleGAN 和本文方法。此外，采用查询选择注意力模块的 QS-Attn 方法相比基线方法 CUT，在 SAR 与光学图像转换任务中并未表现出显著优势。而基于扩散模型的 Con-Diffusion 方法在两个数据集上的图像生成质量评估指标显著弱于其他方法，表明扩散模型在 SAR 与光学图像转换任务的图像质量评估方面不具备优势。

具体而言，在 WHU-OPT-SAR 数据集上的 SAR 到光学图像生成任务中，本文方法在 PSNR 和 MAE 两项指标上均达到最优性能，相较于次优方法 CycleGAN 和 CUT，PSNR 提升了 2.246 4 dB（11.9%），MAE（越低越好）降低了 0.034 5（31.1%）；而在 SSIM 指标上，本文方法较 CycleGAN 降低 0.045（10.6%）。在光学图像到 SAR 图像生成任务中，本文方法在所有 3 项图像质量评价指标上均取得最优表现，相较次优方法在 PSNR 和 SSIM 分别提升了 0.477 1 dB（3.8%）和 0.003（5.6%），在 MAE 降低了 0.009 4（5.2%）。

在 DDHRNet 数据集上，本文方法在 SAR 到光学图像生成任务中的 3 项图像质量评价指标（PSNR、

SSIM、MAE)上均取得最优表现,相较于次优方法CycleGAN,PSNR提升了0.252 4 dB(1.6%),SSIM提升了0.021 6(8.9%),MAE降低了0.007 1(5.5%)。然而,在光学到SAR图像生成任务中,CycleGAN在3项指标上均表现最优,而本文方法未能达到次优水平。这表明,在CycleGAN基础上引入语义引导的对比学习机制,在DDHRNet数据集上生成的SAR图像在图像质量方面并未产生预期提升效果。

表3展示了各类方法在WHU-OPT-SAR数据集上,图像生成结果在语义分割(PA)和特征匹配(PMR、IR和IR-ED)两个下游任务的实验结果。本文方法在两个图像转换任务上都取得了最优结果。具体而言,本文方法在光学图像生成与SAR图像生成的语义分割结果比次优方法分别高出0.162 9(49.26%)和0.048 5(10.60%),表明本文方法在语义分割这一下游任务中具备显著优势。相比之下,同为循环生成架构的CycleGAN虽然在此前图像质量评估中表现优异,却在该任务中性能最差。另一方面,基于扩散模型的Con-Diffusion方法在语义分割任务中并未像其在图像质量评估中那样表现出明显劣势,甚至在SAR生成图像的语义分割结果中取得了次优的性能。

在特征匹配下游任务中,此前在图像生成质量评估中表现最弱的Con-Diffusion方法,在PMR指标上却取得了最优结果。这表明,尽管其生成图像在像素级相似度上与真实图像存在较大差距,但在特征级相似度上表现较强,从而使其生成的59.87%特征点能够构成粗匹配。基于对比学习的方法同样取得了优异的特征匹配结果:CUT方法虽然在PMR指标上比Con-Diffusion方法略低1.78%,但其精匹配比例比Con-Diffusion方法高2.73%(IR)和0.49%(IR-ED)。与之相对,在图像生成质量评估中表现优异的CycleGAN,在特征匹配实验中表现最差,仅有12%的特征点可作为粗匹配,且在这些粗匹配中,仅有2%(IR)和0.2%(IR-ED)的特征点能够通过RANSAC方法与欧氏距离筛选成为精匹配。本文方法在PMR指标上取得了次优结果,并在精匹配中均取得了最优性能,表明本文方法有效结合了

表1　各方法在WHU-OPT-SAR数据集上进行图像转换的定量实验结果

Table 1　Quantitative experimental results of image translation using various methods on the WHU-OPT-SAR dataset

方法	SAR→OPT			OPT→SAR		
	PSNR/dB ↑	SSIM ↑	MAE ↓	PSNR/dB ↑	SSIM ↑	MAE ↓
CycleGAN(Zhu等,2017)	<u>18.883 5</u>	**0.424 3**	0.113 5	<u>12.439 4</u>	<u>0.053 2</u>	<u>0.179 9</u>
CUT(Park等,2020)	18.594 0	0.359 9	<u>0.111 0</u>	11.741 8	0.041 6	0.193 3
QS-Attn(Hu等,2022)	18.434 0	0.368 2	0.112 5	11.780 3	0.041 4	0.192 5
Con-Diffusion(Bai等,2024)	14.945 9	0.245 5	0.192 2	8.152 9	0.020 8	0.330 6
本文	**21.129 9**	<u>0.379 3</u>	**0.076 5**	**12.916 5**	**0.056 2**	**0.170 5**

注:加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好,↓表示值越小越好。

表2　各方法在DDHRNet数据集上进行图像转换的定量实验结果

Table 2　Quantitative experimental results of image translation using various methods on the DDHRNet dataset

方法	SAR→OPT			OPT→SAR		
	PSNR/dB ↑	SSIM ↑	MAE ↓	PSNR/dB ↑	SSIM ↑	MAE ↓
CycleGAN(Zhu等,2017)	<u>16.250 1</u>	<u>0.242 3</u>	<u>0.128 5</u>	**11.800 8**	**0.099 8**	**0.198 1**
CUT(Park等,2020)	15.385 5	0.222 2	0.135 8	11.556 5	<u>0.091 4</u>	0.203 9
QS-Attn(Hu等,2022)	15.435 6	0.225 1	0.134 6	<u>11.622 7</u>	0.090 6	<u>0.202 1</u>
Con-Diffusion(Bai等,2024)	10.188 8	0.121 8	0.293 5	9.862 4	0.081 4	0.259 5
本文	**16.502 5**	**0.263 9**	**0.121 4**	11.533 7	0.083 7	0.202 5

注:加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好,↓表示值越小越好。

CycleGAN的图像生成质量与CUT的特征生成质量

2.4.2　定性实验结果与分析

本节首先对各类方法的图像转换案例进行定性分析，随后对下游语义分割任务和特征匹配任务的案例进行定性分析。所有定性分析均基于WHU-OPT-SAR数据集的案例展开。

图4(a)(b)分别展示了在WHU-OPT-SAR数据集上，各方法在SAR到光学图像和光学图像到SAR转换任务中的4组生成结果，其中左起第1列分别为真实的SAR和光学图像。图像转换结果进一步印证了定量实验的结论：从像素级图像转换结果来看，基于循环生成的方法（CycleGAN和本文方法）的转换效果优于基于对比学习的方法（CUT和QS-Attn）以及基于扩散模型的方法（Con-Diffusion）。

具体而言，Con-Diffusion生成的图像在像素级上与真实图像存在显著差异，尤其在光学图像转换优势。

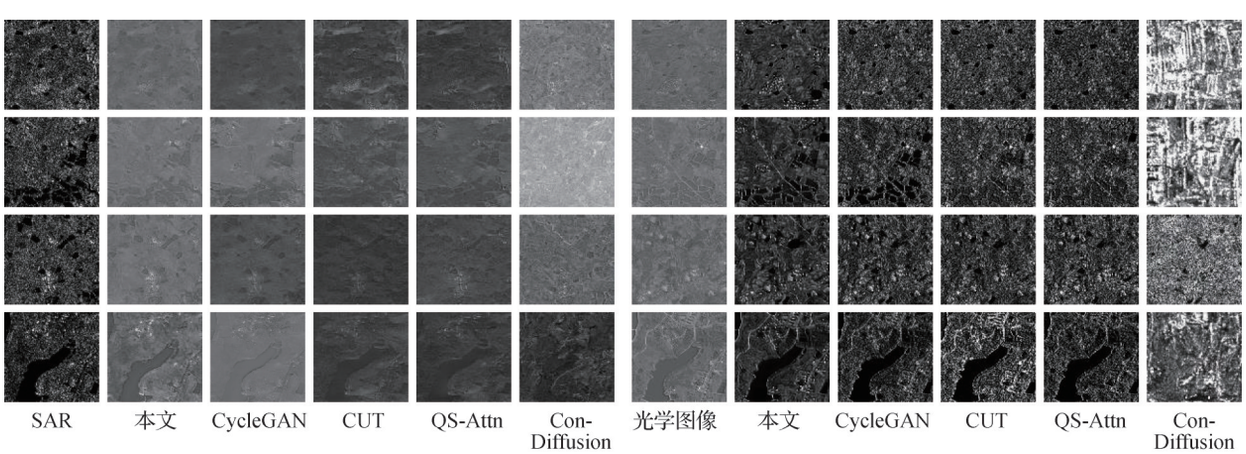
SAR图像方面，生成SAR图像的灰度值与真实SAR图像存在明显偏差。例如，图4(b)第1、2行光学到SAR图像生成结果中，Con-Diffusion生成的SAR图像整体呈现灰白色调，而真实SAR图像则以灰黑色调为主。CUT和QS-Attn方法生成的图像整体质量优于Con-Diffusion，但在SAR到光学图像生成任务中，部分区域仍存在模糊现象。例如，在图4(a)的第4行SAR到光学图像生成案例中，CUT和QS-Attn方法生成的光学图像上半部分的地表纹理与建筑边缘细节生成结果较为模糊。相比之下，本文方法与CycleGAN在该区域生成的图像细节更为丰富。此外，本文方法生成图像在与真实图像的灰度一致性方面，显著优于其他对比方法。例如，图4第1、3行SAR到光学图像生成案例中，本文方法生成的光学图像与真实图像在灰度上具有高度一致性，而其他方法生成的图像整体灰度值较真实图像

表3　各方法在WHU-OPT-SAR数据集的语义分割和特征匹配实验结果

Table 3　Experimental results of semantic segmentation and feature matching for various methods on the WHU-OPT-SAR dataset

方法	PA(SAR→OPT) ↑	PA(OPT→SAR) ↑	PMR ↑	IR ↑	IR-ED ↑
CycleGAN(Zhu等,2017)	0.247 7	0.404 2	0.120 0	0.020 6	0.002 0
CUT(Park等,2020)	0.268 8	0.436 7	0.580 9	<u>0.317 2</u>	<u>0.135 0</u>
QS-Attn(Hu等,2022)	<u>0.330 7</u>	0.441 0	0.508 6	0.313 7	0.077 5
Con-Diffusion(Bai等,2024)	0.184 9	<u>0.457 6</u>	**0.598 7**	0.289 9	0.130 1
本文	**0.493 6**	**0.506 1**	<u>0.588 6</u>	**0.327 1**	**0.139 0**

注：加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好，↓表示值越小越好。



(a) SAR转换为光学图像　　(b) 光学图像转换为SAR

图4　各方法在WHU-OPT-SAR数据集上的转换效果

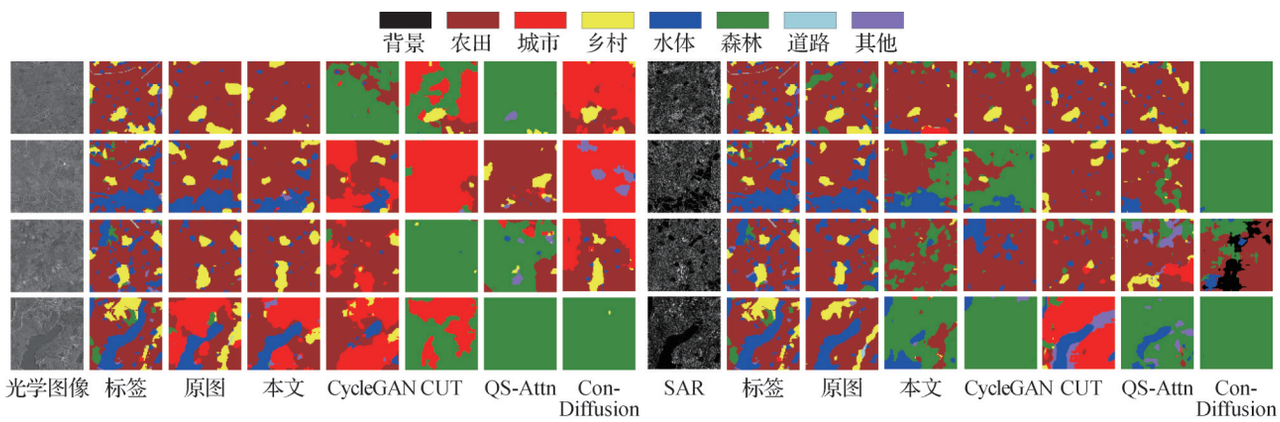
Fig. 4　Qualitative comparison of different methods on the WHU-OPT-SAR dataset

((a) SAR-to-optical image translation results; (b) optical-to-SAR image translation results)

偏暗。

图5(a)(b)分别展示了各方法在图4(a)(b)上生成的4组图像对应的语义分割结果,其中,左起前3列依次为真实图像、语义分割标注,以及分割器OPTSeg和SARSeg对真实图像的分割结果。与定量实验结论一致,在语义分割引导的加持下,本文提出的方法在语义分割下游任务中的表现显著优于其他方法。

具体而言,在图5(a)所示的生成光学图像的语义分割结果中,基于对比学习的方法中仅有QS-Attn方法在第2行中生成的农田特征,以及CUT方法在第1行中生成的部分乡村特征与真实图像较为一致,使其语义分割结果与标注较为接近。其余生成光学图像在OPTSeg分割器下的表现与真实图像分割结果以及分割标注的差异较为显著。例如,CUT方法在4组图像中,将原本为农田的特征生成为城市或森林;QS-Attn方法则将第1、3、4行中的农田及水体特征生成为森林,导致语义分割出现大面积错误。相比之下,基于扩散模型的Con-Diffusion和基于循环生成的CycleGAN的生成效果略优,尤其在第3行中,两种方法生成的农田特征与真实图像较为一致。然而,CycleGAN和Con-Diffusion在第1行中分别将农田特征生成为森林和城市,且Con-Diffusion在第4行中生成的光学图像特征更接近森林,导致整幅图像分割错误。与此相对,本文方法在4组光学图像生成中的特征与真实图像更为接近。尽管其在第4行中的分割结果将农田和乡村的组合误分割为城市,这主要是由于第4行中的特征更为复杂,导致OPTSeg分割器将大片农田与乡村的组合识别为城市。



(a) SAR转换到光学图像　　(b) 光学图像转换到SAR

图5　各方法在WHU-OPT-SAR数据集转换后的分割效果

Fig. 5　Segmentation results of different methods after image translation on the WHU-OPT-SAR dataset

((a) SAR-to-optical image translation; (b) optical-to-SAR image translation)

图6展示了SAR图像转换至光学图像后与真实图像的特征匹配可视化结果。其中,绿色和红色线段分别表示RANSAC识别的正确匹配和错误匹配。该结果与前述定量分析的结论相吻合。具体而言,CycleGAN虽然在像素级的图像转换上取得了较好的视觉效果,但在特征匹配任务中表现不佳,在全部4个测试案例中,仅有一对实现了正确匹配。与之形成对比的是,Con-Diffusion方法,尽管其图像转换质量相对较低,却在特征匹配上表现出色,其生成的粗匹配对数量(图中绿色与红色线的总和)在所有对比方法中达到最多。基于对比学习的CUT和QS-Attn方法,虽然粗匹配总数的绝对值不高,但其中正确匹配(图中绿色直线)的占比较高。本文所提方法则兼具二者优势,不仅获得了数量可观的粗匹配对,同时保持了较高的正确匹配比例。综上,本文方法在像素级图像生成与特征级匹配两个任务上,均展

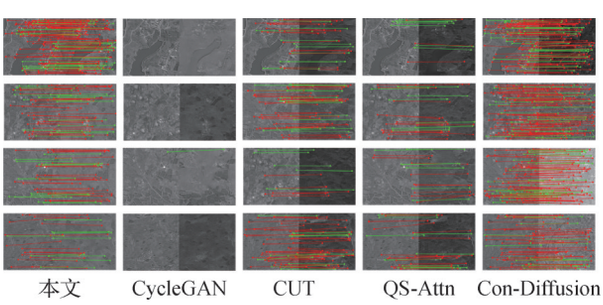


图6　各方法在WHU-OPT-SAR数据集的配准效果

Fig. 6　Registration performance of various methods on the WHU-OPT-SAR dataset

现出卓越的综合性能。

2.5 消融实验

本文提出方法相较仅采用生成对抗损失 L_{gan} 的基线模型，主要引入循环一致性架构及其对应损失 L_{cyc}、语义特征提取模块及其语义分割损失 L_{seg} 和语义引导的对比学习模块及其对应损失 L_{sc}。为验证各模块的有效性及其损失函数对模型性能的独立贡献，在 WHU-OPT-SAR 数据集设计了逐步引入上述损失函数的消融实验。该实验从图像转换质量及下游语义分割准确性两个维度评估模型表现。表 4 展示了不同损失函数组合下的实验结果。

实验结果表明，在仅使用生成对抗损失 L_{gan} 时，由于缺乏有效的约束，SAR 与光学图像的转换质量较差，各项指标均处于最低水平。在此基础上引入循环一致性损失 L_{cyc} 后（即等同于 CycleGAN 基线模型），模型的结构保持能力显著增强，SAR 到光学转换的 PSNR 提升至 18.242 5 dB，初步实现了跨模态转换。

为了验证语义信息的引入对转换性能的提升，在基线模型上逐步叠加了不同的语义相关损失项。当增加语义分割损失（$L_{gan} + L_{cyc} + L_{seg}$）后，模型在下游分割任务中的表现大幅提升。与基线相比，SAR 到光学的 PA 指标升至 0.448 0，表明显式的语义约束能有效指导生成器恢复地物的语义类别信息，减少语义模糊。为验证语义引导对比学习损失 L_{sc} 的有效性，设置了将其替换为原始 CUT 方法中对比损失 L_{nce} 的实验。实验结果表明，使用传统对比学习损失（$L_{gan} + L_{cyc} + L_{nce}$）虽然在一定程度上提升了图像纹理质量，但在下游任务的 PA 指标（0.384 8）上明显低于使用语义对比损失（$L_{gan} + L_{cyc} + L_{sc}$）取得的 PA 指标（0.492 3）。验证了语义引导策略通过类别一致性筛选正负样本，能更精准地实现特征对齐。

本文提出的完整模型（L_{all}）在 PSNR、SSIM 及 MAE 等图像质量指标上均取得了最优结果（PSNR 达到 21.129 9 dB），同时在下游分割任务中保持了最高的像素准确率（PA 为 0.493 6）。尽管在光学到 SAR 的反向转换中，去除语义引导对比损失（$L_{gan} + L_{cyc} + L_{seg}$）的模型在 PSNR 上略有优势，但完整模型在双向转换任务中展现出了最佳的综合性能平衡，验证了融合循环生成结构与语义引导对比学习框架的有效性。

为了验证本文提出的联合优化框架的收敛性，图 7 展示了最优超参设置下（$\lambda_1 = 10, \lambda_2 = 0.5, \lambda_3 = 1, \lambda_4 = 1, \lambda_5 = 1$）各项损失在训练过程中的变化趋势。其中，$L_{cyc}$（红线）初始值最高，且随训练推进呈现持续且稳定的下降趋势，表明反向图像生成的一致性逐渐收敛。L_{seg}（橙线）与 L_{sc}（绿线）均呈现显著下降趋势，并在约 60 Epoch 后收敛至稳定水平，说明模型成功学习并优化了跨模态语义一致性和特征对齐能力。L_{gan}（蓝线）在整个训练过程中保持在相对稳定的区间，没有出现消失或激增。这符合生成对抗网络博弈均衡的典型特征，说明生成器与判别器的性能协同提升并趋于稳定。以上所有损失项均能良好收敛，验证了本文所提联合优化框架的训练稳定性与整体有效性。

同时，为验证不同语义分割模型、分割损失权重、语义引导对比损失权重及图像块大小对本文方法的影响，在 WHU-OPT-SAR 数据集上进行了消融实验。表 5 展示了本文方法基于 3 类经典语义分割

表 4 消融实验结果

Table 4 Module ablation experiment results

损失组成	SAR→OPT				OPT→SAR			
	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑
L_{gan}	17.418 7	0.357 9	0.138 4	0.116 5	11.441 5	0.021 4	0.205 3	0.064 5
$L_{gan} + L_{cyc}$	18.242 5	0.362 9	0.118 5	0.270 1	11.855 6	0.029 0	0.193 1	0.302 6
$L_{gan} + L_{cyc} + L_{seg}$	20.876 5	0.372 0	0.080 6	0.448 0	**13.228 4**	**0.057 8**	**0.165 8**	0.469 4
$L_{gan} + L_{cyc} + L_{sc}$	20.592 1	0.370 4	0.083 1	0.492 3	12.076 8	0.040 1	0.187 3	0.437 3
$L_{gan} + L_{cyc} + L_{nce}$	19.242 2	0.367 9	0.102 6	0.384 8	12.038 3	0.044 0	0.186 7	0.412 8
L_{all}	**21.129 9**	**0.379 3**	**0.076 5**	**0.493 6**	12.916 5	0.056 2	0.170 5	**0.506 1**

注：加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好，↓表示值越小越好。

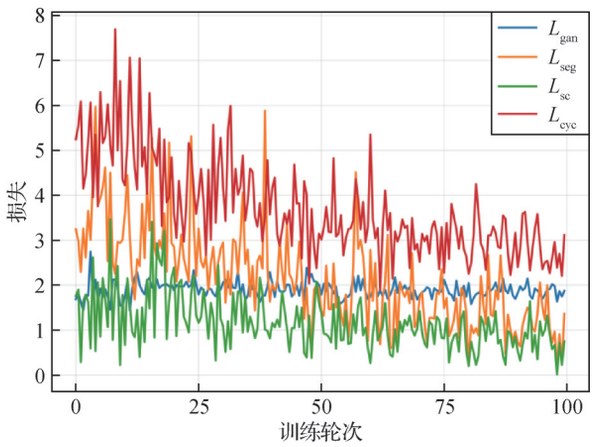


图7　所提方法各项损失在训练过程中的变化曲线

Fig. 7　Training loss curves of the proposed method

模型（U-Net、DeepLabV3、Segformer）的定量实验结果，其中PA（real）表示各模型对真实SAR或光学图像的分割像素准确率。实验结果表明，在真实SAR图像上：3个模型的分割效果差异较小，其中DeepLabV3表现最优。本文方法基于DeepLabV3在图像生成质量和下游语义分割结果上均达到最优。在真实光学图像上：最优模型Segformer与最弱模型U-Net的分割精度相差约6.68%。然而，基于Segformer的本文方法并未在图像生成质量和下游分割结果上取得最优；相对的，基于次优模型DeepLabV3的本文方法在PSNR和下游分割结果上均最优。综上，不同分割模型及其结果对本文方法有一定影响，但当分割效果达到一定阈值后，该影响趋于有限。因此，选择较易实现且常用的DeepLabV3模型为本文方法提供语义分割结果。

表6展示了不同语义分割损失权重（λ_4）下的定量实验结果。在光学到SAR转换任务中，本文方法在$\lambda_4=1$时表现出最佳的综合性能，其PSNR、MAE和PA指标均达到最优，SSIM亦保持次优水平；权重过低（0.5）或过高（5）均会导致各项指标出现不同程度的下降。在SAR到光学转换任务中，本文方法在$\lambda_4=1$时生成图像的PSNR和MAE达到最优；虽然将权重增加至2能略微提升SSIM和PA，但代价是PSNR值出现了明显下降（从21.129 9 dB降至19.915 4 dB）。为在双向转换任务中取得最佳的性能均衡，最终将语义分割权重设置为1。

表7展示了不同语义引导对比损失权重（λ_5）下的定量实验结果。在光学到SAR转换任务中，模型在$\lambda_5=1$时取得最佳性能，权重过低（0.5）或过高（5）均会对图像生成质量及下游语义分割任务的性能造成损害。在SAR到光学转换任务中，$\lambda_5=1$时PSNR值达到最优；继续增加λ_5会改善SSIM和PA，但代价是PSNR和MAE会更差。为在双向转换任务中取得最佳的性能均衡与稳定性，最终将语义引导对比损失权重设置为1。

为了进一步验证语义分割损失和语义引导对比损失在不同权重设置下的训练稳定性，图8展示了不同权重设置下两个训练损失收敛过程。图中曲线已进行移动平均平滑处理，窗口大小设置为5。实验结果表明，语义分割损失和语义引导对比损失在不同的权重设置下均能取得较好的训练收敛效果，表明这两个损失函数与提出方法有很好的适配性。

表8展示了语义引导对比学习模型在不同图像块大小配置下的定量实验结果。当图像块大小设为1×1时，模型能够利用语义信息最精准地筛选正负样本，从而在光学到SAR转换任务上取得了最优的图像生成质量与语义分割性能，并在SAR到光学转换任务上取得了最优的PSNR与MAE值及次优的PA值。该结果验证了1×1配置的有效性。结合传统对比学习也普遍采用此尺寸的实践，最终将图像块大小确定为1×1。

表5　不同分割模型的消融实验结果

Table 5　Ablation experiment results of different segmentation models

模型	SAR→OPT					OPT→SAR				
	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑	PA(real) ↑	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑	PA (real) ↑
U-Net(Ronneberger等,2015)	<u>20.673 1</u>	0.352 3	**0.074 4**	0.347 4	0.686 1	<u>12.394 3</u>	<u>0.053 2</u>	<u>0.181 1</u>	<u>0.377 4</u>	0.782 2
Segformer(Xie等,2021)	19.581 4	**0.446 3**	0.095 5	<u>0.399 1</u>	**0.731 9**	12.028 2	0.032 5	0.189 2	0.309 0	<u>0.785 3</u>
DeepLabV3(Chen等,2017)	**21.129 9**	<u>0.379 3</u>	<u>0.076 5</u>	**0.493 6**	<u>0.709 6</u>	**12.916 5**	**0.056 2**	**0.170 5**	**0.506 1**	**0.786 8**

注：加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好，↓表示值越小越好。

表6　语义分割权重λ_4的消融实验结果

Table 6　Ablation results of the semantic segmentation weight λ_4

λ_4	SAR→OPT				OPT→SAR			
	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑
0.5	20.426 1	0.389 1	0.086 5	0.486 1	12.158 2	0.046 3	0.185 7	0.337 3
1	**21.129 9**	0.379 3	**0.076 5**	0.493 6	**12.916 5**	0.056 2	**0.170 5**	**0.506 1**
2	19.915 4	**0.403 2**	0.089 1	**0.511 0**	12.736 4	**0.056 5**	0.173 2	0.378 2
5	19.795 1	0.400 9	0.082 0	0.476 7	12.627 1	0.054 4	0.172 0	0.378 3

注：加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好，↓表示值越小越好。

表7　语义引导对比损失权重λ_5的消融实验结果

Table 7　Ablation results of the semantic-guided contrastive loss weight λ_5

λ_5	SAR→OPT				OPT→SAR			
	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑
0.5	21.084 9	0.409 8	**0.074 0**	0.433 2	12.256 7	0.050 8	0.184 5	0.383 7
1	**21.129 9**	0.379 3	0.076 5	0.493 6	**12.916 5**	**0.056 2**	**0.170 5**	**0.506 1**
2	20.210 1	**0.423 9**	0.084 7	**0.520 6**	12.176 8	0.039 9	0.186 2	0.309 5
5	21.028 7	0.417 6	0.074 7	0.512 4	11.966 2	0.038 7	0.190 1	0.323 2

注：加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好，↓表示值越小越好。

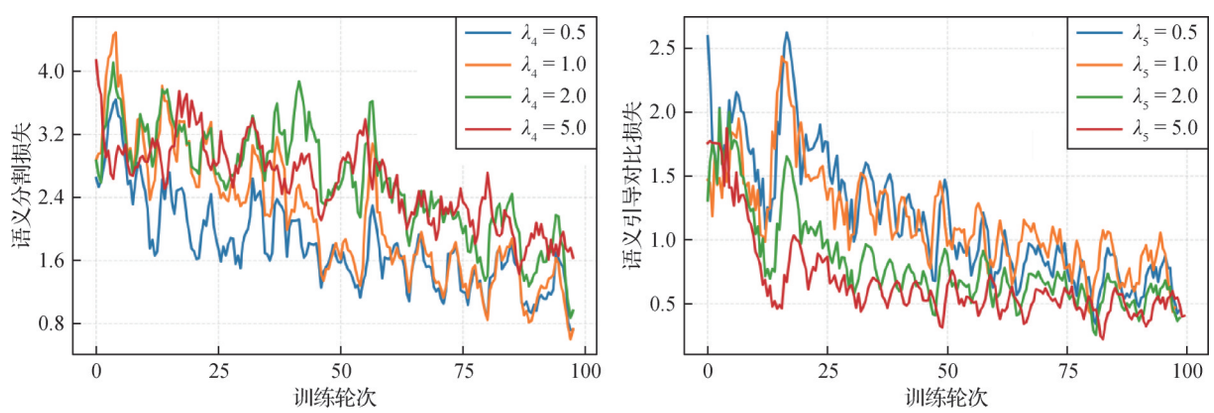


图8　语义分割损失和语义引导对比损失在不同权重设置下训练的变化曲线

Fig. 8　Training curves of the semantic segmentation loss and the semantic-guided contrastive loss under different weight settings

表8　图像块不同大小的消融实验结果

Table 8　Ablation experiment results with different patch sizes

块大小	SAR→OPT				OPT→SAR			
	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑	PSNR/dB ↑	SSIM ↑	MAE ↓	PA ↑
5 × 5	20.637 9	0.433 1	0.078 5	**0.525 3**	11.970 1	0.034 3	0.190 9	0.326 7
3 × 3	19.970 0	**0.435 0**	0.088 0	0.450 5	11.954 0	0.034 0	0.191 1	0.294 5
1 × 1	**21.129 9**	0.379 3	**0.076 5**	0.493 6	**12.916 5**	**0.056 2**	**0.170 5**	**0.506 1**

注：加粗、下划线字体表示各列最优、次优结果。↑表示值越大越好，↓表示值越小越好。

3 结 论

本文提出了一种语义分割引导对比学习的SAR与光学图像转换方法，设计了融合循环生成与对比学习的联合优化框架，有效解决了传统对比学习在遥感图像转换中的特征同质化问题。该方法设计了基于语义分割的类别一致性正负样本筛选策略，构建了语义引导的对比学习模块，有效提升了跨模态特征对齐的准确性；同时，引入循环语义分割损失，约束了生成图像的结构、纹理与语义的一致性，从而提升了生成图像的质量与实用性。在WHU-OPT-SAR和DDHRNet两个公开数据集上的实验结果表明，与相关方法相比，本文方法不仅能生成高质量、高保真度的跨模态图像，还能显著提升生成图像在语义分割和图像匹配等下游任务中的性能。然而，本文方法依赖于预训练的语义分割模型来提供先验知识，这限制了方法在缺乏语义标注数据场景下的应用，且分割模型的精度会直接影响最终的转换效果。同时，当前方法中正负样本的构建严格依赖于预定义的类别标签，这种强监督设定在跨场景或完全无标注的环境中适应性有限，难以支持真正的无监督泛化需求。因此，下一步将研究如何在无需预训练语义分割模型的情况下，实现自监督的语义引导对比学习，并探索不依赖显式类别标签的正负样本构建机制，以增强方法的通用性和鲁棒性。

参考文献(References)

Badrinarayanan V, Kendall A and Cipolla R. 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495 [DOI: 10.1109/TPAMI.2016.2644615]

Bai X Y, Pu X Y and Xu F. 2024. Conditional diffusion for SAR to optical image translation. IEEE Geoscience and Remote Sensing Letters, 21: #4000605 [DOI: 10.1109/LGRS.2023.3337143]

Chen L C, Papandreou G, Schroff F and Adam H. 2017. Rethinking atrous convolution for semantic image segmentation [EB/OL]. [2025-09-03]. https://arxiv.org/pdf/1706.05587.pdf

Guo Z, Zhang Z B, Cai Q L, Liu J Y, Fan Y Y and Mei S H. 2024. MS-GAN: learn to memorize scene for unpaired SAR-to-optical image translation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 11467-11484 [DOI: 10.1109/JSTARS.2024.3411691]

Han J L, Shoeiby M, Petersson L and Armin M A. 2021. Dual contrastive learning for unsupervised image-to-image translation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE: 746-755 [DOI: 10.1109/CVPRW53098.2021.00084]

Hu J M, Li Y L, Zhi X Y, Shi T J and Zhang W. 2025. Complementarity-aware feature fusion for aircraft detection via unpaired Opt2SAR image translation. IEEE Transactions on Geoscience and Remote Sensing, 63: #5628019 [DOI: 10.1109/TGRS.2025.3578876]

Hu X Q, Zhou X Y, Huang Q S, Shi Z Y, Sun L and Li Q L. 2022. QS-Attn: query-selected attention for contrastive learning in I2I translation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 18270-18279 [DOI: 10.1109/CVPR52688.2022.01775]

Jiang W D, Sun Y L, Lei L, Kuang G Y and Ji K F. 2025. AdaptVFMs-RSCD: advancing remote sensing change detection from binary to semantic with SAM and CLIP. ISPRS Journal of Photogrammetry and Remote Sensing, 230: 304-317 [DOI: 10.1016/j.isprsjprs.2025.09.010]

Jung C, Kwon G and Ye J C. 2022. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 18239-18248 [DOI: 10.1109/CVPR52688.2022.01772]

Kang M and Park J. 2020. ContraGAN: contrastive learning for conditional image generation//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1793

Li X, Zhang G, Cui H, Hou S S, Wang S Y, Li X, et al. 2022. MCANet: a joint semantic segmentation framework of optical and SAR images for land use classification. International Journal of Applied Earth Observation and Geoinformation, 106: #102638 [DOI: 10.1016/j.jag.2021.102638]

Park T, Efros A A, Zhang R and Zhu J Y. 2020. Contrastive learning for unpaired image-to-image translation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 319-345 [DOI: 10.1007/978-3-030-58545-7_19]

Ren B, Ma S B, Hou B, Hong D F, Chanussot J, Wang J L, et al. 2022. A dual-stream high resolution network: deep fusion of GF-2 and GF-3 data for land cover classification. International Journal of Applied Earth Observation and Geoinformation, 112: #102896 [DOI: 10.1016/j.jag.2022.102896]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Torbunov D, Huang Y, Yu H W, Huang J, Yoo S, Lin M F, et al.

2023. UVCGAN: UNet vision Transformer cycle-consistent GAN for unpaired image-to-image translation//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 702-712 [DOI: 10.1109/WACV56688.2023.00077]

Wang P, Chen Y K, Huang B, Zhu D Y, Lu T W, Dalla Mura M, et al. 2025. MT_GAN: a SAR-to-optical image translation method for cloud removal. ISPRS Journal of Photogrammetry and Remote Sensing, 225: 180-195 [DOI: 10.1016/j.isprsjprs.2025.04.011]

Wang W L, Zhou W G, Bao J M, Chen D and Li H Q. 2021. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 14000-14009 [DOI: 10.1109/ICCV48922.2021.01376]

Wu L W, Sun R, Kan J S and Gao J. 2020. Double dual generative adversarial networks for cross-age sketch-to-photo translation. Journal of Image and Graphics, 25(4): 732-744 (吴柳玮, 孙锐, 阚俊松, 高隽. 2020. 双重对偶生成对抗网络的跨年龄素描—照片转换. 中国图象图形学报, 25(4): 732-744) [DOI: 10.11834/jig.190329]

Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M and Luo P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #924

Yagmur I C, Ates H F and Gunturk B K. 2024. XPoint: a self-supervised visual-state-space based architecture for multispectral image registration [EB/OL]. [2025-09-03]. https://arxiv.org/pdf/2411.07430.pdf

Yang X, Wang Z H, Zhao J Y and Yang D. 2022. FG-GAN: a fine-grained generative adversarial network for unsupervised SAR-to-optical image translation. IEEE Transactions on Geoscience and Remote Sensing, 60: #5621211 [DOI: 10.1109/TGRS.2022.3165371]

Yu P L, Shi Q and Wang H. 2021. Infrared-to-visible image translation based on parallel generator network. Journal of Image and Graphics, 26(10): 2346-2356 (余佩伦, 施佺, 王晗. 2021. 并行生成网络的红外—可见光图像转换. 中国图象图形学报, 26(10): 2346-2356) [DOI: 10.11834/jig.200113]

Zhan F N, Zhang J H, Yu Y C, Wu R L and Lu S J. 2022. Modulated contrast for versatile image synthesis//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 18259-18269 [DOI: 10.1109/CVPR52688.2022.01774]

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

作者简介

杜文亮，男，讲师，主要研究方向为遥感图像处理。
E-mail: wldu@cumt.edu.cn
周勇，通信作者，男，教授，主要研究方向为机器学习和人工智能。E-mail: yzhou@cumt.edu.cn
郭波，男，硕士研究生，主要研究方向为遥感图像转换。
E-mail: boguo@cumt.edu.cn
赵佳琦，男，副教授，主要研究方向为人工智能和计算机视觉。E-mail: jiaqizhao@cumt.edu.cn
姚睿，男，教授，主要研究方向为人工智能和计算机视觉。
E-mail: ruiyao@cumt.edu.cn