



ADTransGAN: detail-enhanced and cross-modality alignment-optimized GAN for SAR-to-optical image translation

Man Li, Yiyang Tan & Kai Xu

To cite this article: Man Li, Yiyang Tan & Kai Xu (2025) ADTransGAN: detail-enhanced and cross-modality alignment-optimized GAN for SAR-to-optical image translation, International Journal of Remote Sensing, 46:22, 8711-8736, DOI: [10.1080/01431161.2025.2571236](https://doi.org/10.1080/01431161.2025.2571236)

To link to this article: <https://doi.org/10.1080/01431161.2025.2571236>



Published online: 08 Oct 2025.



Submit your article to this journal [↗](#)



Article views: 44



View related articles [↗](#)



View Crossmark data [↗](#)



ADTransGAN: detail-enhanced and cross-modality alignment-optimized GAN for SAR-to-optical image translation

Man Li , Yiyang Tan  and Kai Xu 

School of Internet, Anhui University, Hefei, China

ABSTRACT

Synthetic aperture radar enables consistent data acquisition with all-weather and all-time observation capabilities. However, synthetic aperture radar images suffer from speckle noise and geometric distortions, affecting interpretation and analysis. Optical images are more intuitive but are limited by weather and illumination conditions. To integrate the advantages of both and fill the gaps in optical data, a typical method is to employ generative adversarial network for the translation of synthetic aperture radar images into optical images. Nevertheless, traditional methods struggle to accommodate synthetic aperture radar's distinctive characteristics, resulting in insufficient retention of fine details and poor cross-modality alignment accuracy in the translated optical images. To address these challenges, we propose ADTransGAN, featuring innovative designs in three aspects. First, an autoencoder-based discriminator is designed to embed modality prototype orientation constraints, guiding the generator in approximating the feature distribution of the target modality. Second, a local-global synchronization module is designed, serially fusing CNN and Transformer. This ensures the preservation of detail integrity while maintaining overall scene consistency. Third, an adaptive dynamic convolution module is developed, dynamically adjusting convolution kernel parameters based on local synthetic aperture radar features. This suppresses speckle noise and enhances the mapping accuracy of key details. Comprehensive experiments on the SEN1-2 and SAR2OPT datasets demonstrate that ADTransGAN outperforms state-of-the-art methods across multiple evaluation metrics. Specifically, it achieves 17.1160 and 0.5277 in PSNR and LPIPS, which are 1.3898 higher in PSNR and 0.0085 lower in LPIPS than those of the second-best methods.

ARTICLE HISTORY

Received 1 August 2025

Accepted 30 September 2025

KEYWORDS

dynamic convolution;
Autoencoder; CNN-
Transformer; prototype;
Synthetic aperture radar
(SAR)

1. Introduction

Optical sensors, as common passive imaging devices, operate within the visible spectrum and capture solar radiation reflected from ground objects. The resulting images contain rich spectral information that reveals the physical and chemical properties of surface materials (Persson, Duckett, and Lilienthal 2007). However, their imaging quality is highly dependent on weather and illumination conditions. Cloud cover and precipitation often

lead to interrupted observations, thereby limiting the long-term usability of optical datasets (H. Liu et al. 2024).

In contrast, Synthetic aperture radar (SAR) is an active microwave imaging system operating in the centimetre-to-millimetre wavelength range. It can penetrate clouds, smoke, and precipitation, enabling stable all-time, all-weather observations (Kulkarni and Rege 2020). SAR imagery also provides abundant structural and scattering features, which have demonstrated significant application value in tasks such as image classification (Sharifzadeh, Akbarizadeh, and Kavian 2019), object detection (Mahmoudi, Shokouhi, and Akbarizadeh 2022; Samadi, Akbarizadeh, and Kaabi 2019), image registration (Norouzi, Akbarizadeh, and Eftekhari 2018) and image segmentation (Modava, Akbarizadeh, and Soroosh 2019). Although these studies show the value of SAR in specific applications, the inherent differences in imaging mechanisms, physical properties, and image characteristics compared with optical imagery result in a lack of intuitive visual representation, making it difficult for non-experts to interpret SAR data directly (Xu and Jin 2024).

Therefore, translating SAR images into optical representations has become a research focus, as it not only assists non-specialists in understanding SAR data but also helps to overcome the spatiotemporal limitations of optical observations (Enomoto et al. 2018; Turnes et al. 2022; H. X. Wang et al. 2022; L. Wang et al. 2019; J. X. Zhang, Zhou, and Lu 2020; M. J. Zhang et al. 2024).

In recent years, deep learning has achieved remarkable breakthroughs in classification, segmentation, and detection tasks. For the SAR-to-optical image translation (S2OIT) problem, several approaches have been proposed, such as wavelet decomposition (H. H. Li et al. 2022), multitemporal information (Bermudez et al. 2019) and multi-scale fusion (Y. K. Chen et al. 2024). However, these methods still fall short in complex scenarios. By contrast, Generative Adversarial Networks (GAN) (Goodfellow et al. 2014) can model data distributions through adversarial learning and generate images that approximate real distributions, demonstrating unique advantages in cross-modal generation tasks. Representative methods such as Pix2Pix (Isola et al. 2016) and CycleGAN (J. Y. Zhu et al. 2017) have achieved significant success in image-to-image translation, providing feasible pathways for S2OIT (Turnes et al. 2022; L. Wang et al. 2019). Yet, remote sensing scenes often involve diverse land-cover types with large feature variations, while SAR images are frequently accompanied by speckle noise and blurred edges. These factors lead existing models to suffer from insufficient global consistency, low colour fidelity, detail loss, and difficulties in cross-modal feature alignment.

From the above analysis, it can be seen that S2OIT imposes more stringent requirements on representation learning. First, speckle noise in SAR imagery is often misinterpreted as texture features during translation, resulting in noisy outputs. To address this, we design an adaptive dynamic convolution module (ADCM), which dynamically adjusts convolution kernels according to local features, effectively suppressing noise while enhancing critical detail mapping. Second, due to inherent differences in imaging mechanisms, there exists a significant distribution gap between SAR and optical modalities. Conventional discriminators relying solely on binary classification cannot provide effective cross-modal feature alignment. To overcome this limitation, we introduce an autoencoder-based discriminator, which leverages feature reconstruction to guide the generator towards approximating the optical modality, thereby improving cross-modal

consistency. Finally, since remote sensing imagery contains diverse land-cover types, a single network struggles to learn multi-scale features effectively (Pan, Khan, and Meng 2023). To this end, we propose a local-global synchronous module (LGSM), which integrates CNN and Transformer in a complementary manner, enabling the network to capture fine-grained local scattering details while modelling global spatial dependencies, thus achieving a balance between local detail and global consistency.

Based on these designs, we propose ADTransGAN, which effectively mitigates the limitations of existing methods in terms of detail loss, noise interference, and insufficient modality alignment. The main contributions of this work are summarized as follows:

- (1) We introduce an autoencoder-based discriminator embedded with modality prototype orientation constraints (MPOC), which guides the generator in approximating the feature distribution of the target modality.
- (2) We propose LGSM that integrates CNN and Transformer, achieving a balance between local detail fidelity and global scene consistency.
- (3) We design ADCM that effectively suppresses speckle noise while preserving critical structural details.
- (4) Comprehensive experiments on the SEN1-2 and SAR2OPT datasets validate that ADTransGAN outperforms SOTA methods in metrics including PSNR and SSIM, confirming its promising application potential.

The rest of this paper is organized as follows. Section II provides a review of related research. Section III elaborates on the architecture of ADTransGAN and its key technical modules. Section IV presents experimental findings and discussions. Section V concludes the paper.

2. Related works

2.1. GAN in SAR-to-optical translation

Since Goodfellow et al. (2014) introduced the GAN in 2014, notable breakthroughs have been made in cross-modal image translation. For instance, CycleGAN (J. Y. Zhu et al. 2017) introduces cycle consistency loss to achieve cross-modality conversion, preserving content consistency and generating structurally coherent images. MUNIT (Huang et al. 2018) further decouples content and style, enabling more flexible image translation. SPADE (Park et al. 2019) employs spatially adaptive normalization and multi-task learning strategies to enhance the detail preservation and structural consistency. CUT (Park et al. 2020) introduces a contrastive learning framework based on InfoNCE loss, maximizing mutual information between input-output patches to guide the encoder in preserving source-domain structural features while integrating but provide limited guidance but provide limited guidance target-domain appearance traits, thus enhancing unpaired cross-modality translation performance.

Despite these advancements, such methods still face challenges in S2OIT (Nie et al. 2024; Xiong et al. 2023; M. J. Zhang et al. 2022). With the increasing demands of remote sensing applications, GAN-based frameworks have been widely adopted for S2OIT, and several variants tailored for SAR characteristics have been proposed. Parallel-GAN

(H. X. Wang et al. 2022) separates content and style through parallel generators and leverages adversarial learning to strengthen the mapping between SAR scattering features and optical styles. GFTT (Liang et al. 2025) introduces a Geographic Imaging Tokenizer (GIT) and combines self-supervised tasks with contrastive loss to improve semantic correspondence and scene consistency. EDCGAN (Chouhan et al. 2022) employs a multi-scale attention mechanism to refine feature representations and enhance mapping accuracy. ICGAN (Yang et al. 2021) fuses high and low-level features through parallel branches, improving the colour fidelity of optical images while retaining SAR contours.

However, most of these methods lack explicit mechanisms for modality alignment. Their discriminators primarily constrain generators through authenticity judgements, but provide limited guidance regarding optical modality characteristics such as colour distribution and texture patterns, leading to suboptimal cross-modal alignment. To address this issue, our model introduces an autoencoder-based discriminator framework.

2.2. CNN-transformer hybrid architectures

In recent times, the Transformer (Vaswani et al. 2017) architecture has demonstrated outstanding performance in image recognition and generation, motivating researchers to explore hybrid architectures (CNN) (X. P. Li and Li 2022; Z. Li et al. 2022; Yuan et al. 2023), aiming to combine CNN's strengths in local feature perception with Transformer's advantages in global context modelling. Due to its inherent local receptive field and efficient computing structure, CNN is often used to extract low-level local features in image translation tasks. However, CNN has inherent limitations in modelling long-distance semantic dependencies and is difficult to capture global context information in images. In contrast, Transformer can effectively model the dependencies between non-neighbouring pixels in images through the self-attention mechanism, achieving remarkable results in multiple tasks, including image generation, object detection, and semantic segmentation.

In the exploration of fusion architectures, TransUNet (J. Chen et al. 2021) first verified its value in medical image segmentation tasks. It uses CNN to extract local structural features and Transformer to capture global semantic associations, achieving accurate fusion of multi-scale features through skip connections. CvT (Wu et al. 2021) improves model performance and computational efficiency through designs such as convolutional token embedding and convolutional projection, providing ideas for the engineering application of fusion architectures. This fusion idea has also been introduced into cross-modality translation tasks. Si-CTFNet (J. Zhang et al. 2024) adopts a dual-path architecture, realizes the effective aggregation of CNN local features and Vision Transformer global features through the BIA fusion module, and improves decoding accuracy by combining multi-scale context analysis of the MS-DAM module. HVT-cGAN (Zhao et al. 2025) performs patch embedding through a convolutional stem, extracts local details and global information through parallel CNN and vision Transformer branches, respectively, and adaptively aggregates features through the Convolutional Attention Fusion Module, significantly improving the detail retention and colour fidelity of generated images.

However, existing methods still face challenges in the image translation task from SAR to optical images. One is the insufficient granularity of local-global interaction. Existing fusion mostly stays at the level of feature concatenation or simple weighting, making it

difficult to generate globally semantically coherent images that conform to optical imaging laws while retaining key SAR scattering features. The other is the lack of adaptability to SAR characteristics. No adaptive modelling mechanisms have been designed for SAR's unique attributes, such as speckle noise and inhomogeneous scattering, which limit the generalization ability of the model in complex ground object scenes and make it difficult to balance the contradiction between noise suppression and detail retention.

2.3. Autoencoder

The autoencoder, a classic framework for unsupervised feature learning, projects high-dimensional inputs into a low-dimensional latent space via an encoder and reconstructs inputs from latent representations via a decoder. It is widely used in image reconstruction (Makhzani et al. 2015), feature extraction (Hinton and Salakhutdinov 2006; Ng 2011), and cross-modality alignment (Karpathy and Li 2017).

Recent advancements have expanded the autoencoder's capabilities through structural innovations and loss function design. Amtrano et al. (2024) proposed the sparse autoencoder, introducing sparsity constraints to mimic biological neuron activation and enhance feature discriminability. Makhzani et al. (2015) integrated GAN adversarial loss into the autoencoder, using discriminators to supervise reconstruction authenticity and significantly improving representation learning and image generation quality. Han et al. (2022) designed the style-based autoencoder to learn style features of GAN-generated images, providing effective tools for GAN inversion and style transfer.

The discriminator proposed in this paper builds on the autoencoder, introducing dual functionalities of authenticity judgement and feature guidance via MPOC. This constraint constructs feature prototypes of the target modality to guide the generator's output towards aligning with the target modality's feature distribution.

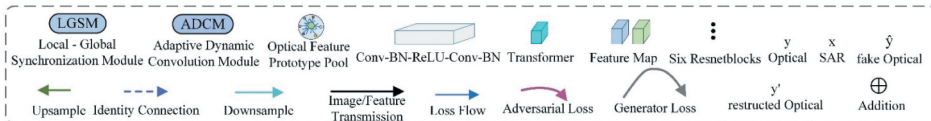
3. Proposed method

This section starts by outlining ADTransGAN, then details the generator, autoencoder-based discriminator, and the loss functions used for network training.

3.1. Overview of network

ADTransGAN is trained on paired SAR and optical images, as depicted in Figure 1, consisting of a generator (G) and an autoencoder-based discriminator (D). G uses LGSM in its encoder, which combines CNN and Transformer to process local and global features. The decoder incorporates ADCM to adjust features from the encoder and maps SAR images (x) to optical-like outputs (\hat{y}). D performs authenticity judgement and feature guidance, using an autoencoder with MPOC.

In the workflow, paired SAR images (x) and optical images (y) are input into the network. G converts SAR images to generated images (\hat{y}), while the encoder of the autoencoder encodes real optical images (y) into latent features. The decoder reconstructs images (y') from these features. The autoencoder's encoder, also serving as D, distinguishes between real and generated images based on feature guidance. During training, D classifies (x, y) as real and (x, \hat{y}) as fake, while G is trained to fool the discriminator. Through alternating



ing both authenticity judgement and feature guidance.

G independently performs translation tasks.

cross-modal features through multi-scale extraction and fusion.

L-GSM serially fuses CNN and T

global scene dependencies, achieving jointly ensures of local detail integrity and global

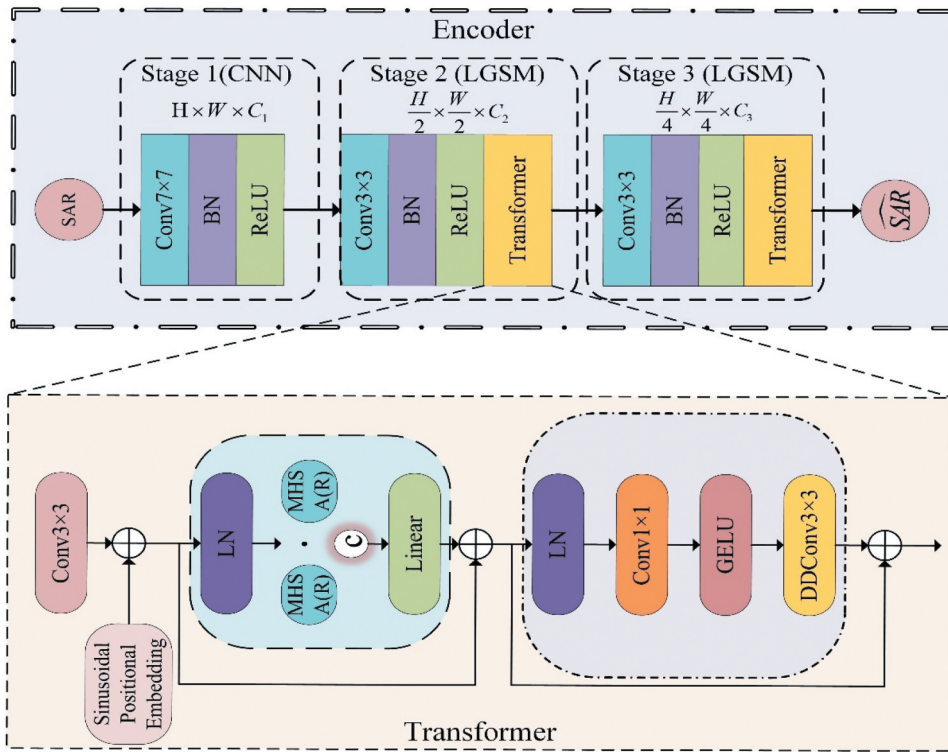


Figure 2. Details of the encoder. The abbreviations used in the figure are as follows: conv: convolution, BN: BatchNorm. MHSA: multi-head self-attention, R: relative position bias, DDConv: depth-wise convolution, c: concat.

scene consistency during downsampling. Key improvements focus on adapting the Transformer (as illustrated in Figure 2) to SAR characteristics, addressing issues such as loss of fine-grained spatial details caused by patch segmentation and poor spatial awareness in traditional ViT for SAR translation.

3.2.1.1. Convolution-enhanced feature embedding. Conventional ViT (Dosovitskiy et al. 2020) splits images into patches, which can disrupt local details in SAR images. The Transformer in LGSM employs an embedding approach of convolutional preprocessing and sequence conversion. It first receives multi-scale local feature maps from the CNN encoder. Through a 3×3 convolution, it enhances pixel adjacency to preserve subtle feature continuity. Then, it flattens the processed feature maps into sequences. This approach avoids information loss caused by patch-based operations (Z. Chen et al. 2021; L. Zhu, Jiang, and Wu 2023) and provides a detailed foundation for global modelling.

3.2.1.2. Position-aware multi-head self-attention (MHSA). To strengthen spatial topology modelling, we design a position-aware MHSA. The matrices of Query (Q), Key (K), and Value (V) are derived from input features, with attention weights dynamically assigned via similarity calculation:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \quad (1)$$

where $Q \in R^{n \times d}$, $V \in R^{n \times d}$, $K \in R^{n \times d}$, $d_h = 64$ (dimension per attention head). Four parallel attention heads enable multi-subspace feature learning, with outputs concatenated and linearly transformed into 256-dimensional features.

To enhance position sensitivity, learnable ‘relative position biases’ are introduced before attention calculation, adding position difference encoding to the similarity matrix:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T + R}{\sqrt{d_h}}\right)V \quad (2)$$

where R denotes the learnable relative position bias matrix. This improves the model’s ability to capture spatial dependencies in SAR images

3.2.1.3. Convolutional feed-forward and residual connections. The conventional Transformer feed-forward network (FFN) uses fully connected layers, which struggles to handle complex local dependencies in SAR features. We design a convolutional FFN with the structure ‘ 1×1 convolution \rightarrow GELU activation $\rightarrow 3 \times 3$ depth-wise convolution’, enhancing local feature interaction while reducing computational complexity.

Additionally, layer normalization (LN) combined with residual connections is introduced to alleviate gradient vanishing, thereby enhancing training stability and feature expressiveness:

$$X_1 = \text{MHSA}(\text{LN}(X)) + X \quad (3)$$

$$X_2 = \text{FFN}(\text{LN}(X_1)) + X_1 \quad (4)$$

where X denotes input features and X_2 denotes output features.

The encoder workflow incorporating LGSM is illustrated in [Figure 2](#). The single-channel SAR input image $\in R^{256 \times 256 \times 1}$ is first mapped to the local feature space via a 7×7 convolution layer with padding = 3 and stride = 1 to expand the input channel to 64 dimensions as a preprocessing step. This is followed by two downsampling layers for local information encoding: two 3×3 convolutions with padding = 1 and stride = 2 sequentially increase the feature dimension to 128 and 256, while downsampling the spatial size to 1/2 and 1/4 of the original, respectively. After each downsampling step, the Transformer module is introduced for global spatial dependency modeling, with output features saved as $y_1 \in R^{128 \times 128 \times 128}$ and $y_2 \in R^{64 \times 64 \times 256}$.

3.2.2. Residual-enhanced feature interaction module

Features from the encoder are fed into a latent feature interaction layer consisting of 6 residual blocks (shown as ResnetBlocks in [Figure 1](#)). Each residual block adopts the structure ‘Conv-BN-ReLU-Conv-BN’ with two 3×3 convolution layers for feature transformation and non-linear activation. Residual connections enhance feature propagation across scales, improving training stability of deep networks and preserving semantic fidelity.

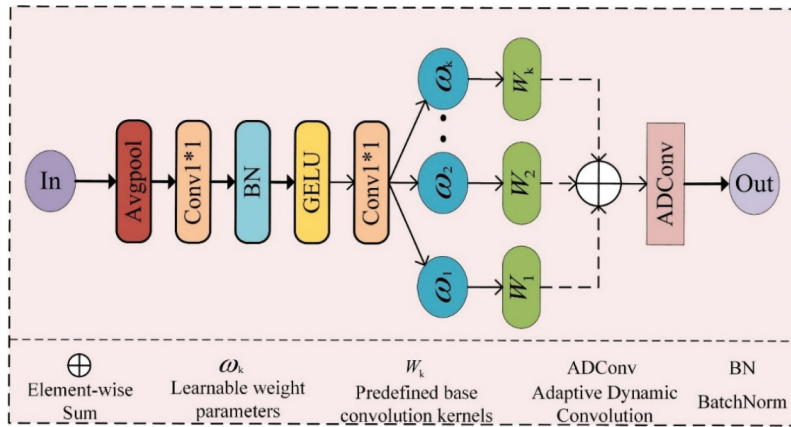


Figure 3. Structure of the adaptive dynamic convolution module (ADCM). The upper part illustrates the workflow of the ADCM. The lower part provides explanations of the symbols used in the figure.

3.2.3. ADCM-driven decoder

Due to differences in imaging mechanisms between SAR and optical images (Yu et al. 2021), cross-modality translation requires the accurate identification and conversion of key structural features in SAR data. Traditional methods rely on fixed feature aggregation (Hughes et al. 2020), which fails to adapt to modality differences. Thus, we propose the ADCM (detailed in Figure 3) to adaptively extract critical scattering features from SAR images and map them to the texture feature space of optical images. ADCM can also specifically suppress speckle noise in SAR images while maintaining structural integrity during the denoising process. To clarify its working mechanism, we first review the calculation of basic depth-wise convolution:

$$\text{DepthwiseConv}(i, j) = \sum_{p, q} W(p, q) \cdot X(i + p, j + q) \quad (5)$$

where (i, j) are the coordinates of the output feature, (p, q) traverse the local region of the SAR feature map X , and W denotes the fixed convolution kernel weights. In the ADCM, the convolution kernel weights are no longer fixed but dynamically generated based on the SAR feature map X :

$$W(X) = \sum_{k=1}^K \omega_k(X) \cdot W \quad (6)$$

The key improvement here is that W adapts in real time to the scattering characteristics of the SAR image.

In the decoder, two upsampling operations are first employed to progressively restore spatial resolution, and subsequently, ADCM is used to establish skip connections with the encoder features, enabling effective noise suppression while preserving critical structural information. Finally, the decoder maps the features to the three-channel RGB space and normalizes them to the range $[-1, 1]$ through the Tanh activation function, ensuring consistency with the characteristics of optical images.

3.3. Dual-function modality alignment discriminator

The discriminator is trained on target modality data to learn feature representations in an autoencoder configuration, which shares the same architecture as the generator and embeds the MPOC. The autoencoder's encoder (\mathcal{A}_{enc}) maps target modality images (y) to features:

$$f = \mathcal{A}_{\text{enc}}(y) \quad (7)$$

while the decoder reconstructs images from these features:

$$y' = \mathcal{A}_{\text{dec}}(f) \quad (8)$$

This process fully captures the target modality's feature distribution.

Simultaneously, the MPOC conducts statistical learning on the encoding features of a large number of target-modality images to obtain the feature prototype (μ_{opt}) representing the core features of this modality. This makes the features of the target modality present an aggregated distribution in the latent space and constructs an accurate modality feature distribution model. On this basis, the encoder of the autoencoder (\mathcal{A}_{enc}) is directly used as the core module of the discriminator. In the adversarial training stage, this encoder not only needs to distinguish the authenticity of the input image but also provides an implicit constraint for the generator through the target-modality feature distribution learned by its pre-training and the MPOC, which involves calculating the distance between the encoding feature of the generated image (\hat{y}) and the target-modality prototype (μ_{opt}), that is:

$$\|W \cdot \mathcal{A}_{\text{enc}}(\hat{y}) - \mu_{\text{opt}}\|_2^2 \quad (9)$$

where W denotes the feature mapping matrix.

This design enables the discriminator to have the dual functions of adversarial supervision and feature guidance simultaneously. The former ensures the visual realism of the generated images through the adversarial process, while the latter relies on the target modality features learned by the autoencoder and the directional constraints of MPOC to make the generated results conform to the inherent properties of the target modality at the feature level.

3.4. Loss function

To achieve high-quality S2OIT, the model enhances the generation quality through the collaborative optimization of four loss functions, which respectively improve from four dimensions: feature alignment, adversarial supervision, modality constraints, and pixel fidelity. The specific design is as follows.

3.4.1. Hierarchical feature collaboration loss

To enable the generator to learn the inherent feature distribution of optical images, we leverage the autoencoder's self-representation capability for the target modality, providing supervision via hierarchical feature alignment. The autoencoder is pre-trained on optical images, enabling its decoder to output optical-consistent hierarchical features.

During synthesis, the generator's decoder is constrained to match the feature distribution of corresponding layers in the autoencoder. The loss is defined as:

$$\mathcal{L}_{\text{ffc}} = \sum_{i=1}^L \text{JS}(\mathcal{G}_{\text{dec}}^i(z) || \mathcal{A}_{\text{dec}}^i(a)) \quad (10)$$

where L is the number of convolutional layers of the decoder, i is the output feature of the potential layer of the generator, $\mathcal{G}_{\text{dec}}^i(z)$ represents the feature output by the i -th layer of the generator decoder, and $\mathcal{A}_{\text{dec}}^i(a)$ represents the feature output by the i -th layer of the autoencoder decoder. JS is the Jensen-Shannon divergence, which is used to measure the difference in the feature distributions of the two layers.

This loss guides the generator to progressively learn the hierarchical structure of optical images by minimizing distribution differences across layers.

3.4.2. Dual-function adversarial loss

The discriminator, composed of the autoencoder's encoder, not only distinguishes the authenticity of input images but also guides the generator via the pre-trained target modality feature distribution. Specifically, the discriminator outputs a high-confidence 'real' label for the real optical image (y), and high-confidence 'fake' labels for the generated image (\hat{y}) and the reconstructed image of the autoencoder (y'). The loss is defined as:

$$\mathcal{L}_{\text{dfl}} = \mathbb{E}_{y \sim p_y} [\log D(y)] + \mathbb{E}_{\hat{y} \sim p_{\hat{y}}} [\log(1 - D(\hat{y}))] + \mathbb{E}_{y' \sim p_{y'}} [\log(1 - D(y'))] \quad (11)$$

where $D(\cdot) = \sigma(\mathcal{A}_{\text{enc}}(\cdot))$ is the output probability of the discriminator (σ denotes the Sigmoid function), and y' is the reconstruction result of the autoencoder for the real image.

Through this design, the discriminator not only improves the overall fidelity of the generated image through an adversarial game but also enhances the sensitivity to the local details of the optical image with the help of the reconstruction error feedback of the autoencoder.

3.4.3. Modality prototype distance loss

Based on the MPOC, this loss strengthens intra-modality compactness by measuring the distance between generated features and optical prototypes. First, it is learned through the optical image dataset. In the training stage, the distance between the feature representation of the generated image and is forced to be minimized. The loss is defined as:

$$\mathcal{L}_{\text{mpd}} = \exp\left(-\frac{\|W \cdot \mathcal{G}_{\text{enc}}(\hat{y}) - \mu_{\text{opt}}\|_2^2}{2\tau^2}\right) \quad (12)$$

where \mathcal{G}_{enc} denotes the encoder output of generated images, W is the feature mapping matrix, and is a learnable scaling parameter.

3.4.4. Generation loss

The generation loss supervises the output accuracy of the generator and autoencoder via pixel-level constraints, ensuring generated images are visually close to real optical images

while enhancing the autoencoder's feature learning capability for the target modality. It is defined as:

$$\mathcal{L}_{gen} = \mathbb{E}_{x \sim p_{SAR}} [G(x) - y_1] + \mathbb{E}_{y \sim p_y} [\mathcal{A}(y) - y_1] \quad (13)$$

where $G(x)$ denotes the generated image and $\mathcal{A}(y)$ denotes the autoencoder's reconstructed image.

3.4.5. Total loss function

The overall optimization objective of ADTransGAN is a weighted combination of the four losses:

$$\mathcal{L}_{total} = \lambda_{lfc} \cdot \mathcal{L}_{lfc} + 1 \cdot \mathcal{L}_{dfl} + \lambda_{BCE} \cdot \mathcal{L}_{mpd} + \lambda_{L1} \cdot \mathcal{L}_{gen} \quad (14)$$

where these parameters are determined empirically through ablation experiments:

$$\lambda_{lfc} = 0.4, \lambda_{BCE} = 5, \lambda_{L1} = 100$$

4. Experiment results

In this section, we introduce the implementation details of ADTransGAN, comparative analysis with existing methods, hyperparameter analysis experiments, ablation experiments, and multi-channel spectral feature distribution matching experiments.

4.1. Implementation details

4.1.1. Datasets

We selected two representative paired SAR-optical datasets (SEN1-2 (Schmitt, Hughes, and Zhu 2018) and SAR2OPT (Y. Wang and Zhu 2018)) to validate ADTransGAN's performance in cross-modality conversion.

The SEN1-2 dataset comprises 282,384 pairs of SAR and optical image patches, which were acquired by the Sentinel-1 and Sentinel-2 satellites, respectively. These patches have a 10-metre ground sampling distance, covering global land areas and encompassing all four seasons. Sentinel-1 provides SAR data with VV polarization, while Sentinel-2's optical images include three bands: red, green, and blue. To balance computational efficiency and data diversity, 12 typical scenes were selected, including 520 SAR images for testing and 1,400 for training.

The SAR2OPT dataset covers remote sensing images of 10 cities across Asia, North America, Oceania, and Europe from 2007 to 2013, including 1,450 training samples and 627 test pairs. SAR data are sourced from the TerraSAR-X satellite with a 1-metre ground sampling resolution, and corresponding optical images are obtained from Google Earth Engine, with image patches uniformly sized at 600×600 pixels. For experiments, these images are randomly cropped to 256×256 patches to fit the network input. Due to its moderate scale and broad scene coverage, the full dataset is used directly for model training and testing.

In the following experiments, Ground Truth (GT) refers to the paired optical images provided by the datasets. Using these sensor-based references ensures reliable validation and prevents the subjectivity associated with manual labelling.

4.1.2. Experimental setup

All experiments were implemented on a computing platform equipped with an NVIDIA RTX 3080 GPU, based on the PyTorch framework version 2.2.1 and accelerated by CUDA 12.2. The model was trained for 100 epochs on both datasets. The AdamW optimizer was utilized to update network parameters with preset configurations, where the momentum and weight decay were set to 0.5 and 0.999, and a fixed learning rate of 0.0002 was adopted throughout the training process.

4.1.3. Evaluation metrics

Image quality evaluation metrics serve to enable quantitative comparisons between different methods. Here, we select four metrics to assess various approaches from multiple angles, including the structural similarity index measure (SSIM), feature similarity index (FSIM), peak signal-to-noise ratio (PSNR), and learned perceptual image patch similarity (LPIPS).

Each of these metrics focuses on distinct aspects of image quality. PSNR centres on pixel-level discrepancies between generated images and reference images. It calculates the ratio between the maximum possible signal power and the noise power, with higher PSNR values indicating smaller pixel differences and thus better image quality. SSIM evaluates the similarity in three key aspects: brightness, contrast, and structural information between generated and reference images. This metric aligns closely with human visual perception, as it mimics how humans perceive image similarity, making it a reliable indicator of perceptual consistency. A higher SSIM score signifies a closer match in structural characteristics. FSIM is a feature-based metric that leverages phase congruency and gradient magnitude for evaluation. It emphasizes the importance of salient visual features, which are critical for human perception. A higher FSIM metric indicates that the generated image retains more vital structural and gradient features, showing consistency with the reference image. LPIPS, a learning-based metric, assesses perceptual differences by comparing high-level feature representations extracted from pre-trained deep neural networks. Unlike pixel-based metrics, it captures semantic and perceptual dissimilarities that are more aligned with human subjective judgement. A lower LPIPS score means the generated image is more perceptually similar to the reference image, reflecting better quality in terms of high-level visual consistency.

4.1.4. Comparative methods

For S2OIT comparative experiments, six methods are chosen. To guarantee the fairness of the comparison, only those methods with open-source implementations or that are improved based on existing open-source frameworks are taken into consideration. The specific details of each method are as follows:

- (a) CycleGAN (J. Y. Zhu et al. 2017): It accomplishes image translation by leveraging adversarial loss to drive the generator and discriminator competitively, along with cycle consistency loss to ensure that the translated image can be converted back to the original input, thus maintaining content consistency.
- (b) MUNIT (Huang et al. 2018): It decouples the content and style of images through separate encoders and decoders, enabling more flexible translation between

different modalities by capturing and recombining content features and style features.

- (c) CUT (Park et al. 2020): A contrastive learning-based approach for unpaired image translation. It enhances mutual information between input and output image patches and utilizes the InfoNCE loss function, enabling the encoder to focus on cross-domain commonalities. This design retains the input image's structural features while integrating the target domain's appearance characteristics.
- (d) DivCo-DRIT (R. Liu et al. 2021): A method for diverse conditional image synthesis. It introduces contrastive learning into the generative adversarial network framework, aiming to generate more diverse and realistic images under conditional constraints, which applies to cross-modality image translation tasks.
- (e) FG-GAN (Yang et al. 2022): A fine-grained generative adversarial network designed for unsupervised SAR-to-optical image translation. It introduces regional information as conditional input, effectively enhancing the accuracy and interpretability of the generated images and addressing the issue of SAR images lacking colour information.
- (f) Parallel-GAN (H. X. Wang et al. 2022): Comprises an optical image reconstruction subnetwork and S2OIT subnetwork. The optical image reconstruction subnetwork provides information constraints on the latent features during the S2OIT process, leveraging the adversarial characteristics of GAN to strengthen the matching between ground object structures and optical styles, thus effectively retaining key scattering features in SAR images.

4.2. Comparison with State-of-the-Art (SOTA) methods

We conducted a comparative analysis between our proposed ADTransGAN and other SOTA approaches. To verify the superiority of ADTransGAN in SAR-to-optical image translation, we evaluated and contrasted different methods from multiple perspectives, including qualitative visualization results and quantitative evaluation metrics. These comprehensive comparisons collectively highlight the advantages of our proposed method in this cross-modality translation task.

4.2.1. Experimental analysis on the SAR2OPT dataset

We evaluated different SAR-to-optical image translation methods using the evaluation metrics aforementioned.

Table 1. Quantitative results of various comparison methods on the SAR2OPT dataset (red: optimal, blue: suboptimal).

Method	PSNR↑	SSIM↑	FSIM↑	LPIPS↓
CycleGAN	12.6103	0.1543	0.6259	0.5362
MUNIT	13.8373	0.1702	0.5902	0.6292
CUT	15.7262	0.2650	0.5373	0.5627
DivCo-DRIT	14.9542	0.2276	0.4981	0.7982
Parallel-GAN	14.0605	0.2079	0.6605	0.5563
FG-GAN	14.5708	0.2478	0.6043	0.5394
ADTransGAN	17.1160	0.2610	0.6449	0.5277

Table 1 presents the numerical results of all comparative methods. MUNIT and CycleGAN perform poorly across nearly all evaluation metrics. MUNIT struggles to extract sufficient information for translation due to speckle noise in SAR images and fails to accurately capture key scattering features. While CycleGAN can maintain the overall structure, it is limited by its single network architecture and cannot fully model cross-modal mapping relationships. DivCo-DRIT performs poorly in PSNR (14.9542), LPIPS (0.7982), and FSIM (0.4981) because its latent enhancement-based code generation neglects the target domain's imaging rules, hindering translation of complex surface categories. CUT performs prominently in SSIM (0.2650), reflecting the advantage of contrastive learning in modelling structural consistency. Parallel-GAN achieves the optimal FSIM (0.6605), indicating that it has certain effects in restoring local visual features. The proposed ADTransGAN demonstrates excellent comprehensive performance. It ranks first in both PSNR (17.1160) and LPIPS (0.5277), which indicates that the generated images have the smallest pixel error and the lowest perceptual difference, and possess the strongest pixel matching degree and high-level semantic consistency with the reference images. Meanwhile, its SSIM (0.2610) and FSIM (0.6449) are also at the top level, showing that it performs excellently in terms of structural consistency and key visual feature fidelity, thus verifying the effectiveness of ADTransGAN in the SAR-to-optical image translation task.

The qualitative visualization results on the SAR2OPT dataset ([Figure 4](#)) further confirm these findings. In SAR-to-optical image translation tasks, the core criteria for evaluating translation quality lie in two aspects: accurately preserving the inherent ground object structural details in SAR images and restoring the unique colour authenticity of optical images. From a visual perspective, the quality of generated images is significantly correlated with the complexity of surface coverage types. For simple coverage types, such as open farmland and single water areas, most methods can achieve acceptable translation

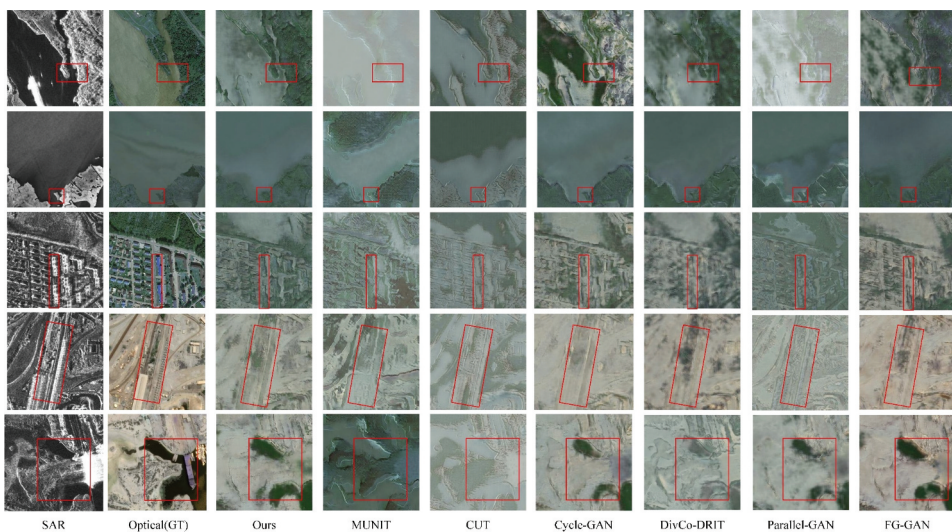


Figure 4. Visualization results of different methods on the SAR2OPT dataset. The red box indicates the key comparison area.

results. However, when dealing with more complex coverage types, such as dense residential areas and urban–rural junctions, the translation quality generally declines. In these cases, prominent issues such as detail loss and colour distortion are frequently observed.

In Row 1, red boxes highlight water-land boundaries. CycleGAN preserves contours but loses shoal textures. MUNIT blends colours and obscures edges. CUT keeps the layout but produces flat water surfaces. Parallel-GAN and FG-GAN generate unnatural block-like water bodies. By contrast, ADTransGAN shows natural gradients, realistic reflections, and clear shoreline details such as shadows and ripples, closely resembling real optical images. In Row 2, CycleGAN distorts mountain textures, MUNIT causes colour mismatches with water, CUT fails to capture details, and Parallel-GAN and FG-GAN blur terrain hierarchies. ADTransGAN restores snow-rock patterns and erosion landforms with natural vegetation – water transitions. In Row 3, red boxes cover residential clusters. CycleGAN misses roof textures, MUNIT mixes roof materials, CUT keeps layouts but with flat textures, and Parallel-GAN and FG-GAN blur edges. ADTransGAN distinguishes materials, restores tile patterns and shadows, and improves feature clarity. In Row 4, CycleGAN blurs infrastructure boundaries, MUNIT produces colour mismatches, CUT omits details such as runway markings, and Parallel-GAN and FG-GAN distort structures. ADTransGAN generates clear boundaries, rich textures, and realistic gradients, capturing runway variations and building shadows. In Row 5, CycleGAN obscures object boundaries, MUNIT creates colour conflicts, CUT struggles with complex layouts, and Parallel-GAN and FG-GAN misalign elements. ADTransGAN separates cover types, enables smooth colour transitions, and reproduces details such as wetland vegetation and building – water interactions, supporting downstream classification tasks.

In summary, through multi-scale feature fusion, ADTransGAN adapts to diverse surface types, producing results closer to real optical images and showing robustness in complex remote sensing scenes.

In addition, we further analyse the model complexity and computational costs, as summarized in Table 2. Our ADTransGAN adopts a lightweight generator with only 8.048 M parameters, the smallest among all methods, which highlights the efficiency of its generator design. The ADCM is integrated into the generator, and although the dynamic generation of convolution kernels inevitably introduces additional computations, the overhead remains limited thanks to the compact generator design. However, the

Table 2. Comparative analysis of the different S2OIT methods in terms of number of parameters, FLOPs, and time costs (red: optimal).

Method	Params(M)			FLOPs(G)			Training Time (s/per epoch)	Test Time (s/per 100 images)
	Generetor	Discriminator	Total	Generetor	Discriminator	Total		
CycleGAN	22.756	5.530	28.268	113.728	6.298	120.026	216	8.630
MUNIT	30.052	16.542	46.594	154.650	4.360	159.01	226	24.130
CUT	11.378	2.765	14.143	80.985	6.425	87.410	201	7.891
DivCo-DRIT	21.271	43.768	65.039	60.576	10.838	71.414	233	10.707
Parallel-GAN	68.153	2.769	70.922	45.185	3.203	48.388	145	5.518
FG-GAN	54.41	5.53	59.94	96.14	3.20	99.34	280	10.000
Ours	8.048	143.383	151.431	43.814	140.864	184.678	254	2.803

autoencoder-based discriminator contributes 143.383 M parameters, leading to the largest total parameter count of 151.431 M and FLOPs of 184.678 G. This complexity results in relatively high training time per epoch (254s), which is slightly faster than FG-GAN (280s). Importantly, ADTransGAN achieves the fastest inference speed, requiring only 2.803 s to process 100 test images, far surpassing other methods. These results show that while our model imposes a higher computational burden during training, it ensures excellent efficiency during inference. Future work will aim to streamline the discriminator design to reduce redundancy while preserving the inference advantage.

4.2.2. Experimental analysis on the SEN1-2 Dataset

To further validate the model's generalization ability across diverse scenarios, we conducted supplementary experiments on the SEN1-2 dataset, which covers global land areas with multi-seasonal variations and a 10-metre ground sampling distance.

As shown in Table 3, ADTransGAN achieves the best results on three perceptual metrics, namely SSIM (0.1911), FSIM (0.5953), and LPIPS (0.5685). Regarding PSNR, ADTransGAN reaches 13.0601, which is slightly lower than the highest value of 13.8284 obtained by Parallel-GAN. This is mainly because ADTransGAN emphasizes perceptual fidelity and structural consistency guided by Transformer modelling and prototype constraints, thereby sacrificing part of the pixel-level accuracy. In future work, we plan to further refine the generator design to improve pixel-wise reconstruction accuracy and achieve a better balance between PSNR and perceptual quality. Overall, ADTransGAN maintains stable performance across datasets with varying resolutions, scene complexities, and seasonal characteristics, fully verifying its robust generalization ability in practical remote sensing applications.

Figure 5 presents visual comparison results on the SEN1-2 dataset across multiple scenes. From the results, it can be observed that most baseline methods exhibit varying degrees of translation errors. Specifically, MUNIT and FG-GAN tend to generate blurry images with oversmoothed details, where roads and vegetation boundaries are difficult to distinguish. CUT and CycleGAN often introduce colour distortions and fail to accurately preserve structural textures. DivCo-DRIT and Parallel-GAN also show limited capability in handling complex land-cover categories, leading to local artefacts and inaccurate tones. In contrast, the proposed ADTransGAN produces translation results that are more consistent with the ground truth in both colour and structure. As highlighted in the red boxes, our method better reconstructs fine linear structures. Overall, ADTransGAN demonstrates superior visual quality, delivering clearer details and more realistic colour representation across diverse scene categories.

Table 3. Quantitative results of various comparison methods on the SEN1-2 dataset (red: optimal, blue: suboptimal).

Method	PSNR↑	SSIM↑	FSIM↑	LPIPS↓
CycleGAN	12.4790	0.0953	0.5316	0.6187
MUNIT	11.9324	0.1469	0.5114	0.8486
CUT	12.7153	0.1899	0.4719	0.6218
DivCo-DRIT	10.3755	0.0177	0.3879	0.8830
Parallel-GAN	13.8284	0.1069	0.5233	0.6694
FG-GAN	12.0500	0.1343	0.4962	0.5976
ADTransGAN	13.0601	0.1911	0.5953	0.5685

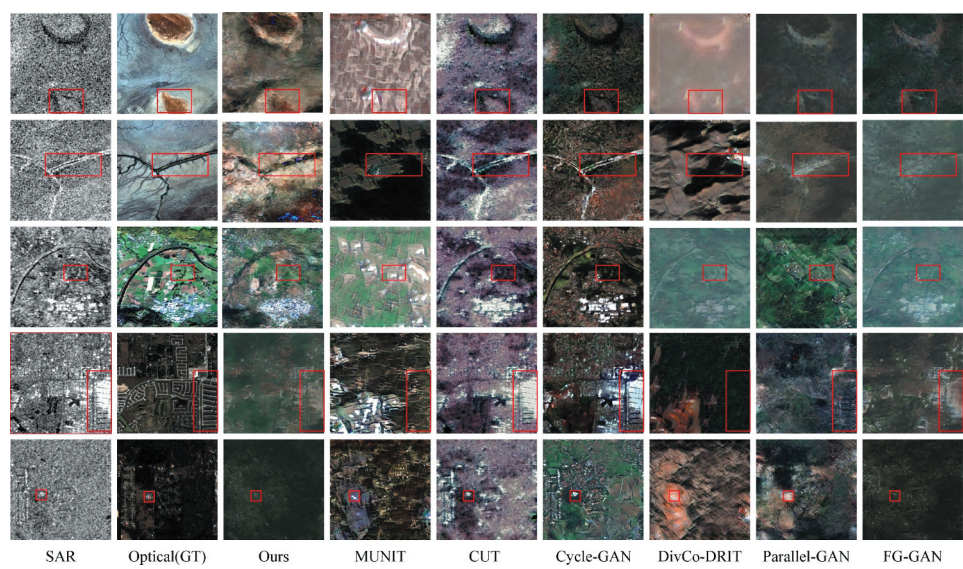


Figure 5. Visualization results of different methods on the SEN1-2 dataset. The red box indicates the key comparison area.

Overall, experiments on both datasets demonstrate ADTransGAN’s superior performance and robustness in S2OIT.

4.3. Ablation experiments

Ablation experiments are conducted on the SAR2OPT dataset to validate the effectiveness of key components, with the evaluation results of different module combinations listed in Table 4. As mentioned earlier, these evaluation metrics focus on different aspects: PSNR emphasizes pixel-level differences, SSIM and FSIM focus on structural and feature similarity, and LPIPS reflects perceptual consistency.

First, the baseline model (Group 1) shows relatively poor performance across all metrics, indicating that the absence of these key modules limits the model’s ability to capture details and align cross-modal features. Second, when only the autoencoder-based discriminator (AED) is introduced (Group 2), all metrics are improved compared to the baseline. This demonstrates that the AED, with its dual functions of authenticity judgement and feature guidance, effectively enhances cross-modality alignment, thereby improving the overall quality of generated images. Third, incorporating only the ADCM

Table 4. Quantitative results of ablation study of proposed ADTransGAN on the SAR2OPT dataset (red: optimal, blue: suboptimal).

	AED	ADCM	LGSM	PSNR↑	SSIM↑	FSIM↑	LPIPS↓
Group 1	×	×	×	16.8227	0.2413	0.5997	0.6438
Group 2	√	×	×	16.9420	0.2495	0.6125	0.6410
Group 3	×	√	×	16.7153	0.2361	0.6258	0.5418
Group 4	√	√	×	17.0305	0.2419	0.5765	0.6315
Group 5	√	√	√	17.1160	0.2610	0.6449	0.5277

(Group 3) leads to significant improvements in LPIPS and FSIM, though there is a slight drop in PSNR. This indicates that the ADCM, which dynamically adjusts convolution kernels based on local features, plays a crucial role in suppressing speckle noise and enhancing key detail mapping, especially in improving perceptual quality and feature fidelity. Fourth, combining the AED and ADCM (Group 4) results in a higher PSNR than Groups 2 and 3, but lags in FSIM. This suggests that while the AED helps optimize pixel-level consistency, the lack of the LGSM limits the model's ability to capture global scene dependencies, leading to insufficient feature integrity in complex scenes. Finally, when all three modules are integrated (Group 5), the model achieves the best performance across all metrics.

This confirms that the synergistic effect of the AED, ADCM, and LGSM is critical for performance enhancement. The AED ensures cross-modality feature alignment, the ADCM enhances local detail mapping and noise suppression, and the LGSM balances local detail integrity and global scene consistency. Their combined application enables the model to achieve optimal overall performance in SAR-to-optical image translations.

The visualization of translation results for different groups (Figure 6) further corroborates these findings, with distinct performance variations observed in diverse scenarios. In ROW 1 (scattered small-building scene), Group 1 yields blurred building contours, Group 2 renders visible outlines via AED but lacks fine details, Group 3 accentuates local features through ADCM yet suffers from misalignments, Group 4 achieves clear structural definitions post-fusion but with abrupt edges and Group 5, by integrating all modules, presents sharp building edges and natural textures, accurately reconstructing the scene with high fidelity to real images. In ROW 2 (rural road scene), Group 1 fails to distinguish road edges and markings, Group 2 clarifies road contours using AED but lacks pavement details, Group 3 enhances local features via ADCM

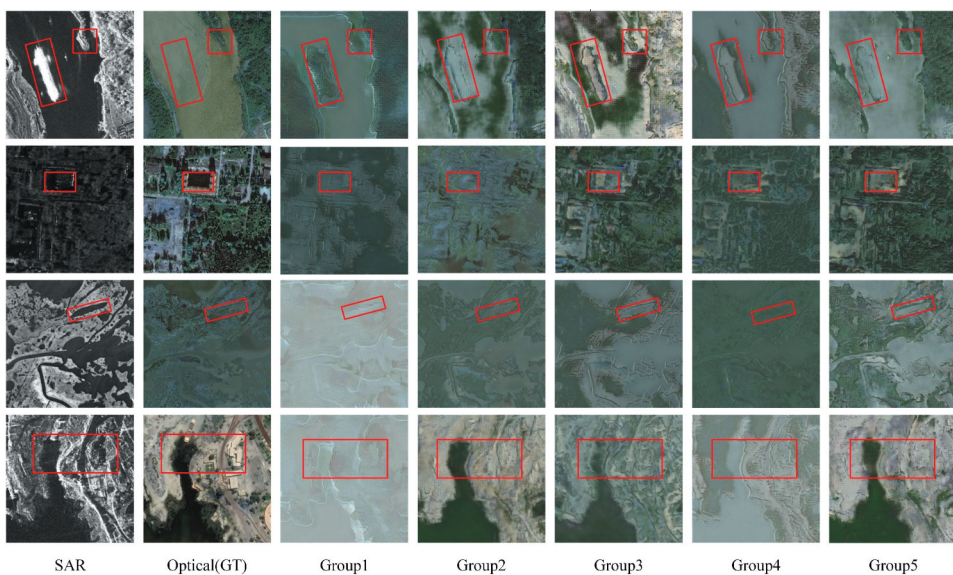


Figure 6. Visualization results of the ablation studies of the proposed ADTransGAN on the SAR2OPT dataset. The red box indicates the key comparison area.

while exhibiting misalignments, Group 4 delivers prominent road structures after fusion but with chaotic layouts at complex intersections and Group 5, through module collaboration, exhibits sharp road edges and clear lane divisions, faithfully restoring the real rural road scenario. In ROW 3 (forest-grass transition zone), Group 1 shows indistinct forest-grass boundaries and poor leaf texturing, Group 2 clarifies spatial distribution via AED but lacks hierarchical details, Group 3 enriches fine features through ADCM yet presents disordered transitions, Group 4 improves colour consistency but with unnatural transitions and Group 5, via full-module integration, demonstrates natural layering, realistic textures, and smooth transitions, conforming to the characteristics of real optical images. In ROW 4 (large-scale complex terrain scene), Group 1 loses terrain contours entirely, Group 2 captures general topographic trends but lacks details, Group 3 retains local features but with chaotic layouts, Group 4 achieves moderately clearer structures post-fusion but with collapsed large-scale hierarchies and Group 5, by integrating all modules, presents rich hierarchical details and realistic textures, fully restoring terrain morphology with high alignment to real images.

4.4. Hyperparameter analysis experiments

We conducted experiments on the SAR2OPT dataset to select appropriate values for the two hyperparameters corresponding to the generation loss (λ_L) and the modality prototype distance loss (λ_{BCE}) primarily constrains the pixel-level similarity between generated images and real optical images, ensuring basic visual fidelity. λ_{BCE} guides the generated features to align with the prototype distribution of the target modality, enhancing cross-modal feature consistency. λ_L is tested at $\{1, 50, 100, 150, 200\}$, and λ_{BCE} at $\{1, 3, 5, 7, 10\}$, resulting in 25 parameter combinations. Figure 7 shows metrics vary across these combinations.

From the trend of change, the PSNR index shows non-monotonic characteristics. When λ_L is in the middle interval (around 100), the heights of the bar charts are generally higher. This indicates that moderate generation constraints are critical for preserving pixel-level fidelity. Excessive constraints may force the model to overly focus on pixel matching, leading to rigid textures, while insufficient constraints result in blurred details and pixel-level deviations. The high-value regions of the FSIM and SSIM sub-graphs highly overlap, which confirms that there is a synergistic effect between feature alignment and structural similarity, and an overly large λ_{BCE} is likely to break this balance. It may prioritize modality prototype alignment at the cost of local structural consistency, causing mismatches between fine-grained features and the global scene. The LPIPS index shows a downward trend as λ_{BCE} increases within a certain range, reflecting that strengthening modality prototype constraints helps reduce perceptual differences between generated images and real optical images. However, a large λ_{BCE} (exceeding 7) will cause the perceptual restoration to rebound, as excessive emphasis on prototype alignment may suppress the diversity of local textures, resulting in unnatural visual effects. Considering the overlap of high-value regions of Figure 7, λ_L in $[80, 120]$ and λ_{BCE} in $[3, 7]$ are the optimal parameter intervals. To further pinpoint the best combination within these ranges, we analyzed the metric performance at key points within the intervals and found that $\lambda_L = 100$ and $\lambda_{BCE} = 5$ achieved the most balanced results across all evaluation

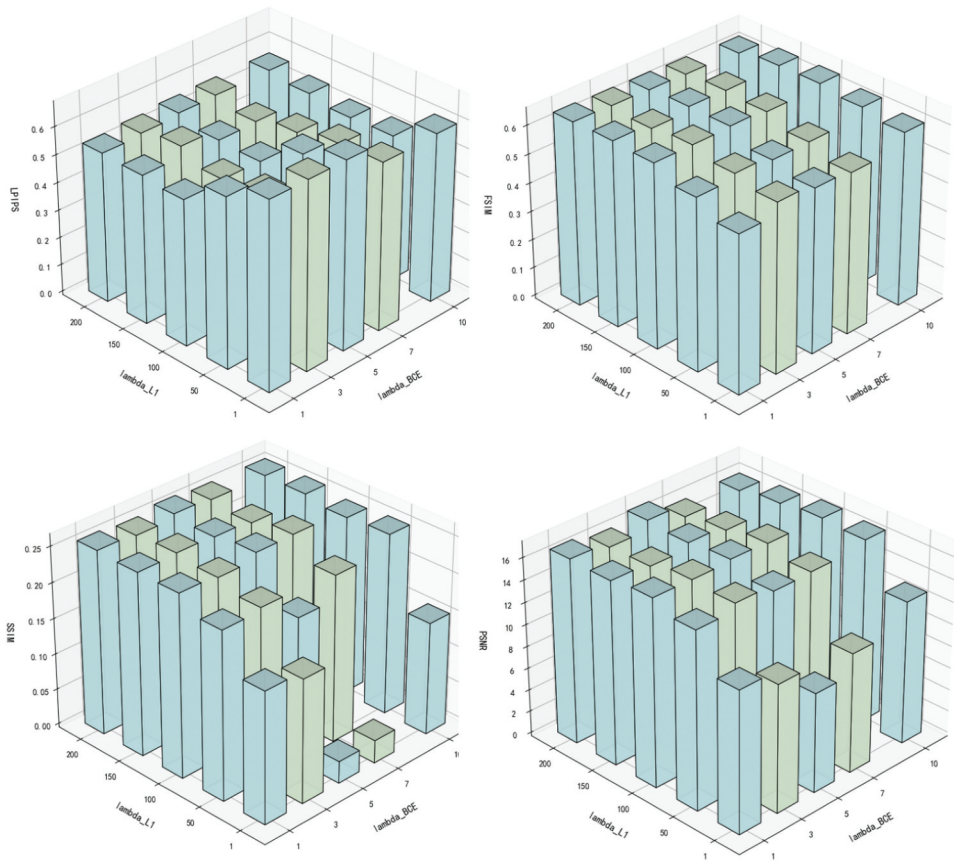


Figure 7. Parametric sensitivity analysis of multi-metrics: λ_{LI} and λ_{BCE} .

indicators. This combination achieves the best trade-off between pixel fidelity, structural consistency, feature alignment, and perceptual quality, further validating the rationality of the loss function design.

4.5. Multi-channel spectral alignment validation for cross-modality translation

As illustrated in Figure 8, this experiment selected multiple pairs of real optical images and ADTransGAN-generated counterparts. For each image pair, pixel values of the Red, Green, and Blue channels are extracted separately. Then, scatter density plots of generated values versus real values are constructed. In these plots, the horizontal axis denotes the channel pixel values of the generated images, the vertical axis represents the corresponding channel pixel values of the real optical images, and the colour coding indicates the distribution density of samples with the same (generated value, real value) combination. These visualizations enable quantitative comparison of spectral feature alignment between generated and real images.

As shown in Figure 8, across scenarios like urban, suburban, and water areas, the scatter points of the red, green, and blue channels all form high-density clusters along the diagonal where the generated values equal the real values. Qualitatively, the brightness of

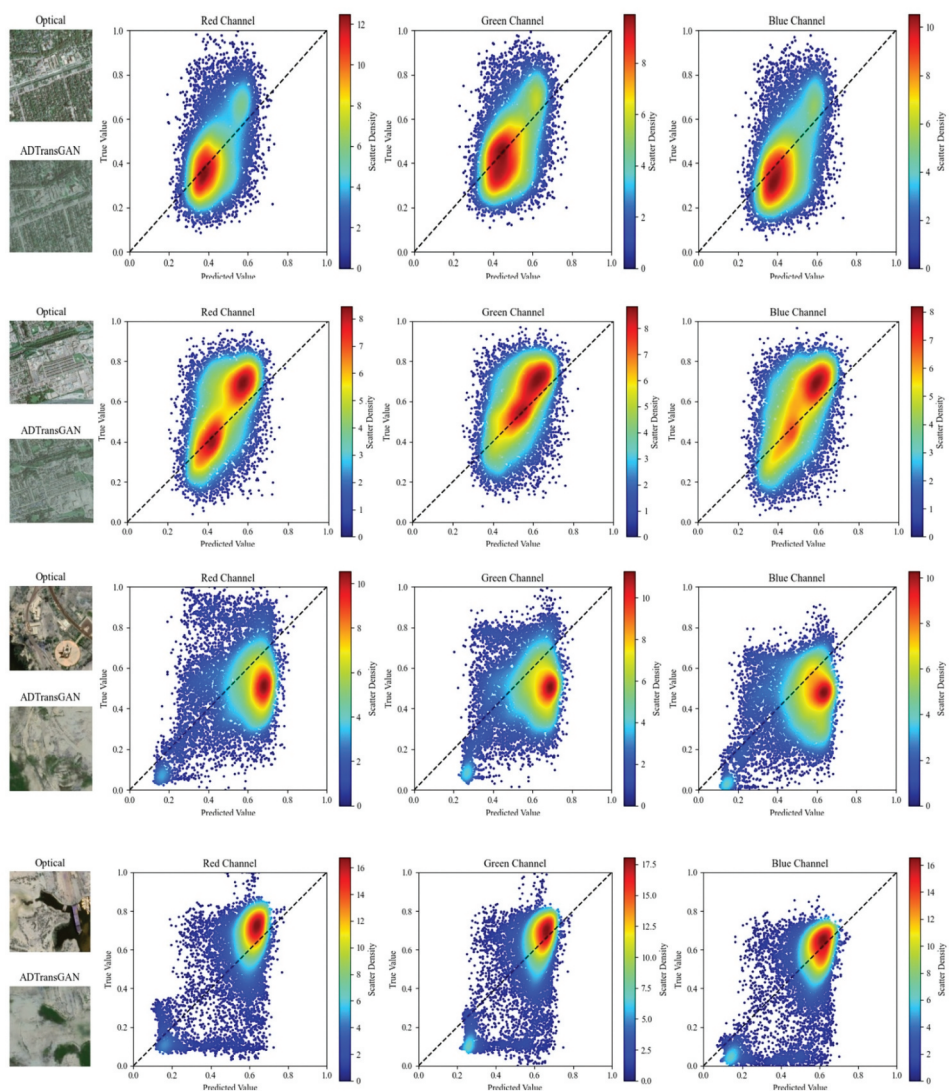


Figure 8. Scatter density plots of channel pixel values between images generated by ADTransGAN and real optical images.

the colour in the dense scatter regions directly reflects how close they are to the ideal spectral matching. Regions with brighter clusters tightly around the diagonal, meaning most generated pixel values highly match the spectra of real optical images. Even in complex urban scenes with diverse land covers or homogeneous water surfaces, this density distribution pattern along the diagonal remains stable, demonstrating the model's robustness.

Quantitatively, by counting scatter points within a narrow deviation band from the diagonal relative to the total samples, consistent patterns emerge. In urban scenes, over 72% of the points in the red, green, and blue channels fall within this range, with the red channel density peaking at 75%, suburban scenes show an even higher concentration –

an average of 75% of the points across channels cluster near the diagonal, and the green channel even reaches 77%, for water areas, about 73% of the samples align closely with the diagonal, maintaining spectral consistency despite uniform textures.

To further verify, we calculated the linear correlation between the generated and real channel values. The red channel shows a strong positive correlation, with values ranging from 0.87 to 0.91 across scenarios. The green channel follows, with coefficients between 0.86 and 0.90, and the blue channel maintains a solid 0.85–0.89 correlation. These metrics, combined with the visual density patterns, confirm that ADTransGAN-generated images achieve high-fidelity spectral matching with real optical data, regardless of scene complexity.

In conclusion, the model effectively learns the spectral conversion rules from SAR to optical images, resolves cross-modality alignment challenges, and ensures the spectral authenticity of the generated image.

5. Conclusion

In this paper, a new S2OIT method, ADTransGAN, is proposed to generate high-quality optical images and effectively address the issues of detail loss, speckle noise interference, and cross-modality inconsistency. The method consists of three key modules. The autoencoder-based discriminator with MPOC leverages the autoencoder structure to learn the target modality distribution during reconstruction and introduces prototype constraints in the feature space to achieve precise cross-modality alignment. LGSM connects CNN and Transformer in sequence to integrate fine-grained local texture modelling with global dependency learning, thereby ensuring structural integrity and detail fidelity. ADCM dynamically generates convolution kernels according to local SAR scattering characteristics to effectively suppress speckle noise and enhance critical texture representation. Extensive experiments on SAR2OPT and SEN1-2 datasets demonstrate that our model achieves superior performance over existing methods in four key evaluation metrics, namely PSNR, SSIM, FSIM and LPIPS, confirming both its quantitative advantages and visual quality improvements.

In future work, we plan to extend ADTransGAN to other remote sensing modalities to further validate its generalization ability across diverse imaging scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 42401455.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Code and data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Man Li  <http://orcid.org/0009-0003-7197-9754>
 Yiyang Tan  <http://orcid.org/0009-0008-9024-7694>
 Kai Xu  <http://orcid.org/0000-0001-6310-2977>

References

- Amitrano, D., G. Di Martino, A. Di Simone, and P. Imperatore. 2024. "Flood Detection with SAR: A Review of Techniques and Datasets." *Remote Sensing* 16 (4). <https://doi.org/10.3390/rs16040656>.
- Bermudez, J. D., P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira. 2019. "Synthesis of Multispectral Optical Images from SAR/Optical Multitemporal Data Using Conditional Generative Adversarial Networks." *IEEE Geoscience & Remote Sensing Letters* 16 (8): 1220–1224. <https://doi.org/10.1109/lgrs.2019.2894734>.
- Chen, J., Y. Lu, Q. Yu, X. Luo, and Y. Zhou. 2021. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation."
- Chen, Y. K., Z. G. Zhu, Y. Huang, P. Wang, B. Huang, and M. Dalla Mura. 2024. "MSF: A Multi-Scale Fusion Generative Adversarial Network for SAR-to-Optical Image Translation." Paper presented at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Athens, Greece, July 07–12.
- Chen, Z., Y. Zhu, C. Zhao, G. Hu, and M. Tang. 2021. "DPT: Deformable Patch-Based Transformer for Visual Recognition." In *29th ACM International Conference on Multimedia (MM)*, 2899–2907.
- Chouhan, A., N. Jindal, A. Sur, D. Chutia, and S. Aggarwal. 2022. "EDCGAN: Encoder Decoder Based Conditional GAN for SAR to Optical Image Translation." In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing*.
- Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby. 2020. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale."
- Enomoto, K., K. Sakurada, W. Wang, N. Kawaguchi, M. Matsuoka, and R. Nakamura. 2018. "Image Translation Between SAR and Optical Imagery with Generative Adversarial Nets." Paper presented at the 38th IEEE International Geoscience and Remote Sensing Symposium (IGARSS), July 22–27.
- Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Nets." MIT Press.
- Han, L., S. H. Musunuri, M. R. Min, R. Gao, Y. Tian, and D. Metaxas. 2022. "AE-StyleGAN: Improved Training of Style-Based Auto-Encoders." Rutgers The State University of New Jersey.
- Hinton, G. E., and R. R. Salakhutdinov. 2006. "Reducing the Dimensionality of Data with Neural Networks." *Science* 313 (5786): 504–507. <https://doi.org/10.1126/science.1127647>.
- Huang, X., M. Y. Liu, S. Belongie, and J. Kautz. 2018. "Multimodal Unsupervised Image-to-Image Translation." In *Proceedings of the European conference on computer vision (ECCV)*. Cham: Springer.
- Hughes, L. H., D. Marcos, S. Lobry, D. Tuia, and M. Schmitt. 2020. "A Deep Learning Framework for Matching of SAR and Optical Imagery." *ISPRS Journal of Photogrammetry & Remote Sensing* 169:166–179. <https://doi.org/10.1016/j.isprsjprs.2020.09.012>.
- Isola, P., J. Y. Zhu, T. Zhou, and A. A. Efros. 2016. "Image-to-Image Translation with Conditional Adversarial Networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE.
- Karpathy, A., and F. F. Li. 2017. "Deep Visual-Semantic Alignments for Generating Image Descriptions." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39 (4): 664–676. <https://doi.org/10.1109/tpami.2016.2598339>.
- Kulkarni, S. C., and P. P. Rege. 2020. "Pixel Level Fusion Techniques for SAR and Optical Images: A Review." *Information Fusion* 59:13–29. <https://doi.org/10.1016/j.inffus.2020.01.003>.
- Li, H. H., C. Gu, D. Q. Wu, G. Cheng, L. Guo, and H. Liu. 2022. "Multiscale Generative Adversarial Network Based on Wavelet Feature Learning for SAR-to-Optical Image Translation." *IEEE Transactions on Geoscience & Remote Sensing* 60:15. <https://doi.org/10.1109/tgrs.2022.3211415>.

- Li, X. P., and S. Q. Li. 2022. "Transformer Help CNN See Better: A Lightweight Hybrid Apple Disease Identification Model Based on Transformers." *Agriculture-Basel* 12 (6): 16. <https://doi.org/10.3390/agriculture12060884>.
- Li, Z., D. Li, C. Xu, W. Wang, Q. Hong, Q. Li, and J. Tian. 2022. "TFCNs: A CNN-Transformer Hybrid Network for Medical Image Segmentation." *International Conference on Artificial Neural Networks* 13532:781–792.
- Liang, H., X. Yang, X. Yang, J. Luo, and J. Zhu. 2025. "GFTT: Geographical Feature Tokenization Transformer for SAR-to-Optical Image Translation." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 18:2975–2989.
- Liu, H., F. Wang, Y. Jin, X. Ma, S. Li, Y. Bian, and G. Situ. 2024. "Learning-Based Real-Time Imaging through Dynamic Scattering Media." *Light: Science & Applications* 13 (1).
- Liu, R., Y. Ge, C. L. Choi, X. Wang, and H. Li. 2021. "DIVCO: Diverse Conditional Image Synthesis via Contrastive Generative Adversarial Network." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16372–16381.
- Mahmoudi, F., S. B. Shokouhi, and G. Akbarizadeh. 2022. "A New Technique for Segmentation of the Oil Spills from Synthetic-Aperture Radar Images Using Convolutional Neural Network." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 15:8834–8844. <https://doi.org/10.1109/jstars.2022.3213768>.
- Makhzani, A., J. Shlens, N. Jaitly, and I. Goodfellow. 2015. "Adversarial Autoencoders." *Computerence*.
- Modava, M., G. Akbarizadeh, and M. Soroosh. 2019. "Hierarchical Coastline Detection in SAR Images Based on Spectral-Textural Features and Global-Local Information." *IET Radar Sonar & Navigation* 13 (12): 2183–2195. <https://doi.org/10.1049/iet-rsn.2019.0063>.
- Ng, A. 2011. "Sparse Autoencoder."
- Nie, H., B. Luo, J. Liu, Z. Fu, W. Liu, C. Wang, and X. Su. 2024. "A Novel Rotation and Scale Equivariant Network for Optical-SAR Image Matching." In *IEEE Transactions on Geoscience and Remote Sensing*.
- Norouzi, M., G. Akbarizadeh, and F. Eftekhari. 2018. "A Hybrid Feature Extraction Method for SAR Image Registration." *Signal, Image and Video Processing* 12 (8): 1559–1566.
- Pan, Y., I. A. Khan, and H. Meng. 2023. "SAR-to-Optical Image Translation Using Multi-Stream Deep ResCNN of Information Reconstruction." *Expert Systems with Application* 224.
- Park, T., A. A. Efros, R. Zhang, and J. Y. Zhu. 2020. "Contrastive Learning for Unpaired Image-to-Image Translation."
- Park, T., M. Y. Liu, T. C. Wang, and J. Y. Zhu. 2019. "Semantic Image Synthesis with Spatially-Adaptive Normalization." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Persson, M., T. Duckett, and A. Lilienthal. 2007. "Improved Mapping and Image Segmentation by Using Semantic Information to Link Aerial Images and Ground-Level Information." Paper presented at the 13th International Conference on Advanced Robotics, August 22–25.
- Samadi, F., G. Akbarizadeh, and H. Kaabi. 2019. "Change Detection in SAR Images Using Deep Belief Network: A New Training Approach Based on Morphological Images." *IET Image Processing* 13 (12): 2255–2264. <https://doi.org/10.1049/iet-ipr.2018.6248>.
- Schmitt, M., L. H. Hughes, and X. X. Zhu. 2018. "The SEN1-2 Dataset for Deep Learning in SAR-Optical Data Fusion." *ISPRS TC I Mid-Term Symposium on Innovative Sensing - From Sensors to Methods and Applications* 4-1:141–146.
- Sharifzadeh, F., G. Akbarizadeh, and Y. S. Kavian. 2019. "Ship Classification in SAR Images Using a New Hybrid CNN-MLP Classifier." *The Journal of the Indian Society of Remote Sensing* 47 (4): 551–562. <https://doi.org/10.1007/s12524-018-0891-y>.
- Turnes, J. N., J. D. B. Castro, D. L. Torres, P. J. S. Vega, R. Q. Feitosa, and P. N. Happ. 2022. "Atrous CGAN for SAR to Optical Image Translation." *IEEE Geoscience & Remote Sensing Letters* 19:5. <https://doi.org/10.1109/lgrs.2020.3031199>.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. "Attention Is All You Need." Paper presented at the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, December 04–09.

- Wang, H. X., Z. G. Zhang, Z. Y. Hu, and Q. L. Dong. 2022. "SAR-to-Optical Image Translation with Hierarchical Latent Features." *IEEE Transactions on Geoscience & Remote Sensing* 60:12. <https://doi.org/10.1109/tgrs.2022.3200996>.
- Wang, L., X. Xu, Y. Yu, R. Yang, R. Gui, Z. Z. Xu, and F. L. Pu. 2019. "SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks." *IEEE Access* 7:129136–129149. <https://doi.org/10.1109/access.2019.2939649>.
- Wang, Y., and X. X. Zhu. 2018. "The SARptical Dataset for Joint Analysis of SAR and Optical Image in Dense Urban Area." In *38th IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 6840–6843.
- Wu, H. P., B. Xiao, N. Codella, M. C. Liu, X. Y. Dai, L. Yuan, and L. Zhang. 2021. "CVT: Introducing Convolutions to Vision Transformers." Paper presented at the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Electr Network, October 11–17.
- Xiong, Q., G. Q. Li, X. C. Yao, and X. D. Zhang. 2023. "SAR-to-Optical Image Translation and Cloud Removal Based on Conditional Generative Adversarial Networks: Literature Survey, Taxonomy, Evaluation Indicators, Limits and Future Directions." *Remote Sensing* 15 (4): 20. <https://doi.org/10.3390/rs15041137>.
- Xu, F., and Y. Jin. 2024. "Microwave Vision and Intelligent Perception of Radar Imagery." *Journal of Radars* 13 (2): 285–306. <https://doi.org/10.12000/JR23225>.
- Yang, X., Z. H. Wang, J. Y. Zhao, and D. Yang. 2022. "FG-GAN: A Fine-Grained Generative Adversarial Network for Unsupervised SAR-to-Optical Image Translation." *IEEE Transactions on Geoscience & Remote Sensing* 60:11. <https://doi.org/10.1109/tgrs.2022.3165371>.
- Yang, X., J. Zhao, Z. Wei, N. Wang, and X. Gao. 2021. "SAR-to-Optical Image Translation Based on Improved CGAN." *Pattern Recognition*.
- Yu, Q. Z., D. W. Ni, Y. X. Jiang, Y. X. Yan, J. C. An, and T. Sun. 2021. "Universal SAR and Optical Image Registration via a Novel SIFT Framework Based on Nonlinear Diffusion and a Polar Spatial-Frequency Descriptor." *ISPRS Journal of Photogrammetry & Remote Sensing* 171:1–17. <https://doi.org/10.1016/j.isprsjprs.2020.10.019>.
- Yuan, J., F. Zhou, Z. Guo, X. Li, and H. Yu. 2023. "HCformer: Hybrid CNN-Transformer for LDCT Image Denoising." *Journal of Digital Imaging* 36 (5): 16.
- Zhang, J., W. Zhang, X. Zhou, Q. Chu, X. Yin, G. Li, X. Dai, S. Hu, and F. Jin. 2024. "CNN and Transformer Fusion Network for Sea Ice Classification Using GaoFen-3 Polarimetric SAR Images." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 17:18898–18914.
- Zhang, J. X., J. J. Zhou, and X. W. Lu. 2020. "Feature-Guided SAR-to-Optical Image Translation." *IEEE Access* 8:70925–70937. <https://doi.org/10.1109/access.2020.2987105>.
- Zhang, M. J., C. Y. He, J. Zhang, Y. X. Yang, X. Q. Peng, and J. Guo. 2022. "SAR-to-Optical Image Translation via Neural Partial Differential Equations." Paper presented at the 31st International Joint Conference on Artificial Intelligence (IJCAI), Vienna, Austria, July 23–29.
- Zhang, M. J., P. Zhang, Y. H. Zhang, M. H. Yang, X. F. Li, X. G. Dong, and L. C. Yang. 2024. "SAR-to-Optical Image Translation via an Interpretable Network." *Remote Sensing* 16 (2): 19. <https://doi.org/10.3390/rs16020242>.
- Zhao, W., N. Jiang, X. Liao, and J. Zhu. 2025. "HVT-cGAN: Hybrid Vision Transformer CGAN for SAR-to-Optical Image Translation." In *IEEE Transactions on Geoscience and Remote Sensing*.
- Zhu, J. Y., T. Park, P. Isola, and A. A. Efros. 2017. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." In *Proceedings of the IEEE international conference on computer vision*. IEEE.
- Zhu, L., C. Jiang, and M. Wu. 2023. "A Patch Information Supplement Transformer for Person Re-Identification." *Electronics* 12 (9): 14.