remote sensing

MDPI

Article

DSAN: Dual-Scale Aligned Network with Asymmetric Priors and Differentiable Soft-Edge Loss for SAR-to-Optical Image Translation

Yingying Kong * and Dongmin Wang

College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; sx2404078@nuaa.edu.cn

* Correspondence: yayako_zy@nuaa.edu.cn

Highlights

What are the main findings?

- An asymmetric dual-prior Dual-Scale Aligned Network (DSAN) integrating Large Kernel Attention and Collaborative Attention (CoA) is developed to bridge cross-scale semantic gaps in SAR-to-optical translation.
- A continuous differentiable soft-edge loss is formulated with logarithmic dynamic range compression to enforce high-frequency geometric boundary fidelity.

What are the implications of the main findings?

- The proposed DSAN achieves superior structural consistency and texture realism over state-of-the-art GAN and Transformer paradigms while maintaining real-time inference capability.
- The framework provides a reliable, high-fidelity data generation pathway for all-weather remote sensing observation and downstream multi-modal earth vision applications.

Abstract

Synthetic aperture radar (SAR) provides all-weather imaging but faces challenges in visual interpretation due to low contrast and coherent speckle noise. To address contrast deficiency and edge blurring in SAR-to-optical translation, we propose a Dual-Scale Aligned Network (DSAN) built upon Pix2PixHD. First, an asymmetric dual-prior architecture is designed: the global generator ingests low-resolution SAR images enhanced by histogram equalization to capture macroscopic structures, while the local generator utilizes original high-resolution SAR images to preserve microscopic details, alleviating the trade-off between contrast and fine textures. Second, a Dual-Scale Fusion Module (DSFM) coupling Large Kernel Attention and Collaborative Attention breaks scale barriers, enabling bidirectional cross-scale alignment and deep fusion. Third, a continuous differentiable soft-edge loss is formulated using logarithmic dynamic range compression to prevent highlights from dominating gradients and enforce boundary consistency across urban areas, water bodies, and farmlands. Experiments on the Nanjing and public SEN1-2 datasets demonstrate that DSAN outperforms state-of-the-art models—including Pix2PixHD, CycleGAN, MSTM-Net, and ICMA—in perceptual distribution realism (FID) with the sharpest geometric boundaries. Ablation studies confirm the effectiveness of the asymmetric dual-prior design, DSFM, and the refined edge loss.

Check for updates

Academic Editor: Dusan Gleich

Received: 27 July 2026
Revised: 29 August 2026
Accepted: 3 September 2026
Published: 5 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: synthetic aperture radar; image-to-image translation; asymmetric dual-prior; differentiable soft-edge loss

1. Introduction

Since its emergence in the 1950s, synthetic aperture radar (SAR) has achieved high-resolution imaging by utilizing the platform motion to synthesize an equivalent long aperture and performing coherent processing on received echo signals. SAR provides all-weather and round-the-clock earth observation capabilities, effectively penetrating clouds without illumination constraints. Consequently, it holds significant value across various domains, such as environmental monitoring, military reconnaissance, and disaster response [1,2]. However, owing to the coherent imaging mechanism, the gray-level features of SAR images strictly depend on the electromagnetic scattering properties of ground objects. Variations in radar wave scattering across distinct land covers yield specific gray-level distributions and texture patterns (e.g., water bodies typically appear as dark regions with low gray values, whereas buildings present as bright regions with high gray values). This specialized imaging mechanism poses substantial challenges to visual interpretation; by contrast, optical remote sensing images align much better with human visual perception habits. Furthermore, most mainstream algorithms for image classification and target recognition are optimized for optical imagery, which remains highly susceptible to cloud coverage and variable lighting. Therefore, converting SAR images into high-fidelity optical images is of paramount importance for improving the visual interpretability of SAR data and expanding its downstream applications.

Image translation and image style transfer share fundamental conceptual similarities. Early methods built upon convolutional neural networks (CNNs) made notable progress in tasks such as image segmentation [3], classification and recognition [4], and super-resolution [5]. Nevertheless, they face severe hurdles when addressing complex mapping tasks like image translation: on the one hand, parameter scales expand dramatically as network depth increases; on the other hand, constrained by traditional loss function designs, these models struggle to effectively recover fine texture details and exhibit limited generalization in complex scenes. In 2014, Goodfellow et al. introduced the generative adversarial network (GAN) [6], offering a new paradigm for image generation through adversarial min-max games between a generator and a discriminator. Building upon this, various GAN variants have been proposed. For instance, CycleGAN established a cycle-consistent framework for mutual translation, effectively mitigating content inconsistency in cross-domain translation without requiring paired training data; it was subsequently applied to SAR image translation to reduce color artifacts. However, unconditional GANs typically rely on random noise inputs, leading to low controllability over the generation process and making it difficult to preserve spatial structural details such as terrain and buildings during SAR-to-optical conversion. The conditional generative adversarial network (CGAN) partially resolved this limitation [7] by incorporating source images as condition constraints, enabling precise guidance of generated content and successfully preserving crucial structural information like building outlines, topography, and spatial layouts [8,9].

Although CGANs demonstrate strong controllability in image translation tasks, they still face critical challenges in the specific SAR-to-optical scenario. First, the gray-level distribution of source SAR images is overly concentrated, making it difficult for the network to distinguish boundaries and fine details of different land cover types during early training stages. Second, SAR images encompass both macro-level global topography and micro-

level local building structures; existing generative networks struggle to balance both during multi-scale feature extraction and fusion. Third, severely affected by inherent multiplicative speckle noise, traditional edge operators (e.g., Canny) involve non-differentiable "hard thresholding" operations. Consequently, they cannot be directly integrated into end-to-end neural network training and tend to produce false edges triggered by noise, leading to blurry boundaries in translated outputs. To address these limitations, taking the CGAN as the base framework, this paper proposes a multi-scale collaborative attention network for SAR-to-optical image translation. Compared with existing approaches, the fundamental innovation of this paper lies in reformulating cross-scale interaction and boundary constraints, with its design rationale summarized in three aspects. First, unlike conventional multi-scale architectures that extract features independently and combine them via simple upsampling or naive concatenation, our Dual-Scale Fusion Module (DSFM) couples Large Kernel Attention (LKA) and Collaborative Attention (CoA) to establish bidirectional cross-scale alignment. This transforms rigid feature concatenation into adaptive soft alignment, breaking scale barriers and enabling direct mutual guidance between global context and local details. Second, this dual-scale alignment is tailored to SAR imaging physics: since SAR penetration and coherent interference mix significant ground scattering with speckle noise, multi-path cross-scale exchange allows the network to effectively decouple structural semantics from noise at appropriate receptive fields, ensuring that the synthesized outputs faithfully capture optical radiometric properties. Third, the accompanying asymmetric dual-prior input architecture feeds complementary contrast-enhanced and fine-grained features into the dual streams, maximizing the representational capacity of the cross-scale alignment mechanism. The main contributions of this work are summarized as follows:

1. To resolve the trade-off between contrast enhancement and high-frequency texture preservation caused by narrow SAR dynamic ranges, an asymmetric dual-prior input architecture is developed. The global generator ingests a histogram-equalized low-resolution SAR image to capture macroscopic topological layouts across a stretched dynamic range, while the local generator receives the original high-resolution SAR image to preserve fine-grained structural details. This multi-prior paradigm enhances boundary discriminability during early training while preventing the distortion of native structural features inherent to monolithic equalization.
2. To address the difficulty of balancing global terrain and local building details during translation, this paper leverages the dual-generator architecture of Pix2PixHD to model global and local features separately, incorporating a Dual-Scale Fusion Module (DSFM) that couples Large Kernel Attention (LKA) and Collaborative Attention (CoA). This module eliminates inter-scale isolation and establishes direct bidirectional cross-scale alignment between complementary representations, thereby significantly enhancing the overall clarity and structural fidelity of generated images.
3. To overcome the non-differentiability and noise sensitivity of traditional edge operators, a continuous differentiable soft-edge loss is formulated. By leveraging the illumination insensitivity and logarithmic dynamic range compression, this loss mitigates gradient domination from specular highlights and fairly captures weak boundaries across shadowed terrain. Furthermore, replacing hard thresholding with continuous differentiable gating ensures smooth gradient backpropagation during end-to-end optimization, significantly enhancing the geometric sharpness and boundary fidelity of diverse land covers, including urban layouts, water bodies, and agricultural fields.

1.1. Related Work

1.1.1. Conditional Generative Adversarial Networks (CGANs)

Although GANs effectively capture data distributions and generate realistic outputs, their generation process lacks controllability, hindering their widespread application in complex practical scenarios. To overcome this, Mirza et al. introduced the Conditional Generative Adversarial Network (CGAN) in 2014 [7]. The core idea lies in introducing additional conditional information simultaneously into both the generator and discriminator. In image translation tasks, the source image to be translated is typically provided as the condition input to guide targeted generation toward the target image domain. The introduction of CGAN broke the limitations of unconditional generation in original GANs. Subsequent research has predominantly evolved along two directions:

1. Network Architecture Optimization: Examples include the Deep Convolutional GAN (DCGAN) proposed by Radford et al. [10], which deeply integrates the feature extraction capacity of CNNs with the generative mechanism of GANs; and E^2GAN introduced by Gong et al. [11], which combines diffusion model distillation with LoRA fine-tuning to achieve efficient GAN training and lightweight inference.
2. Loss Function Design: A prominent example is the perceptual loss proposed by Johnson et al. [12], which effectively mitigates blurriness and lack of realism caused by traditional pixel-level losses.

Among various CGAN variants, Pix2Pix [13] represents a major milestone. This framework established the complete paradigm of "CGAN + paired data + hybrid loss" for the first time, using input images as conditional constraints to ensure pixel-level alignment between generated and real images. As a result, it has been widely adopted in image translation, style transfer, and medical image reconstruction. However, images produced by Pix2Pix are generally limited to lower resolutions. To address this drawback, Wang et al. developed Pix2PixHD [14], incorporating key enhancements such as multi-scale generator and discriminator architectures, instance-level feature matching loss, and support for semantic label-guided generation. The network architecture proposed in this paper takes Pix2PixHD as its baseline framework; since our datasets lack corresponding semantic labels, apart from semantic label guidance, we primarily adapt and optimize the remaining two core strategies.

1.1.2. SAR Image Translation

Prior to the widespread application of deep learning techniques, SAR-to-optical image translation primarily relied on traditional physical model-driven approaches and hand-crafted feature extraction. The fundamental principle of these methods was to construct mapping models across radiation, geometry, and texture based on the physical imaging disparities between the two modalities. For instance, techniques such as speckle filtering and histogram matching were employed to align the statistical properties of SAR and optical images [15]. However, traditional approaches suffer from obvious limitations: on the one hand, inherent speckle noise and non-linear scattering characteristics in SAR images make hand-crafted features insufficient for capturing complex texture and structure [16]; on the other hand, these methods rely heavily on strict physical model assumptions, resulting in poor generalization when encountering complex terrain or multitemporal variations in practical applications. Consequently, generated optical images often suffer from blurry details and distorted textures, failing to meet the requirements of downstream high-level interpretation tasks [17].

In recent years, deep learning-based SAR image translation has emerged as a vibrant research hotspot in remote sensing. Most studies adopt the CGAN as their foundational architecture, making deep improvements in network structures, loss functions, and train-

ing strategies to accommodate the unique characteristics of SAR imagery. For example, Fu et al. [18] proposed a generator with a multi-scale cascaded residual structure, passing downsampled SAR feature maps into intermediate layers of varying depths via residual connections to facilitate feature flow and enhance texture representation. Tan et al. [19] designed a multi-task serial training strategy, deploying two independent CGAN models to sequentially perform SAR denoising and colorization, cascading them to achieve high-quality translation. Yang et al. [20] introduced a dual-branch network structure to extract shallow and deep features independently, incorporating a Gaussian blur chromaticity loss to effectively boost color fidelity. Ji et al. [21] introduced the CycleGAN framework into SAR image translation, leveraging cycle-consistency constraints to substantially reduce color artifacts in generated images. Kong et al. [22] designed a multi-scale pyramid architecture based on Swin Transformers and proposed MSTMNet, extracting image features across different receptive fields using parallel Swin Transformer blocks to significantly enhance global structural quality.

2. Materials and Methods

2.1. Overall Framework

Built upon the classic Pix2PixHD architecture within the conditional generative adversarial network (CGAN) paradigm, this paper constructs an end-to-end generative adversarial network manifold tailored for SAR-to-optical image translation. As illustrated in Figure 1, the network adopts a topological structure defined by the co-evolution of dual generators and multi-scale discriminators.

On the generation side, the system is compounded by a global generator G_g and a local generator G_l. Both are enhanced variants of the U-Net backbone, with their primary innovation lying in the embedding of the proposed Dual-Scale Fusion Module (DSFM). In terms of mathematical operators, this module deeply integrates Large Kernel Attention (LKA) [23] and Collaborative Attention (CoA) [24] mechanisms, jointly driving cross-domain translation and precise alignment of two heterogeneous feature branches within a non-linear latent space.

To achieve parallel complementarity between macro-spatial semantics and micro-radiometric textures, the input feature spaces of the two sub-generators incorporate distinct physical priors. The global generator G_g ingests the contrast-enhanced SAR image preprocessed by adaptive histogram equalization (HE), focusing on capturing continuous macroscopic topological layouts and broad geographic configurations across a stretched dynamic range. Conversely, the local generator G_l accepts the pristine high-resolution SAR image as input, dedicated to preserving native high-frequency radiometric textures and fine-grained structural features along land-water interfaces and complex land cover boundaries.

On the discrimination side, to overcome the classification bottleneck where traditional discriminators easily ignore micro-textures, this paper introduces the fully convolutional PatchGAN architecture [13]. A multi-scale verification mechanism is jointly formed by a global discriminator D_g and a local discriminator D_l. Unlike conventional single-sample discrimination, both sub-discriminators take conditional logarithmic manifolds as joint inputs. Specifically, D_g is responsible for low-resolution macro inspection, while D_l performs original-resolution micro verification. Their inputs comprise both a real sample pair formed by the "source SAR image and real optical image" and a fake sample pair composed of the "source SAR image and generated optical image." By calculating real/fake probability matrices over overlapping local receptive fields (patches) across different spatial resolution scales, they achieve multi-dimensional collaborative evaluation of both geometric and semantic authenticity of the generated results.

During the network optimization phase, the Adam stochastic gradient optimizer is employed to iteratively update the dual-generator family $\{G_g, G_l\}$ and the multi-scale discriminator family $\{D_g, D_l\}$ in an alternating fashion. To ensure multi-dimensional constrained convergence, a composite joint total loss functional is formulated, comprising adversarial loss, perceptual loss, feature matching loss, and continuous differentiable soft-edge loss. Driven by gradient backpropagation through this multi-task loss, the generative and discriminative components undergo dynamic evolution in a minimax game. The adversarial loss provides global driving force for manifold distributions. The perceptual loss and the feature matching loss enforce latent consistency across deep semantic and multi-scale feature manifolds, respectively, while the differentiable soft-edge loss meticulously regulates high-frequency topological structures in the log domain. Under their synergy, translation distortion across heterogeneous modalities is progressively suppressed, ultimately driving generated optical images to approximate the joint distribution of real imagery with high fidelity.

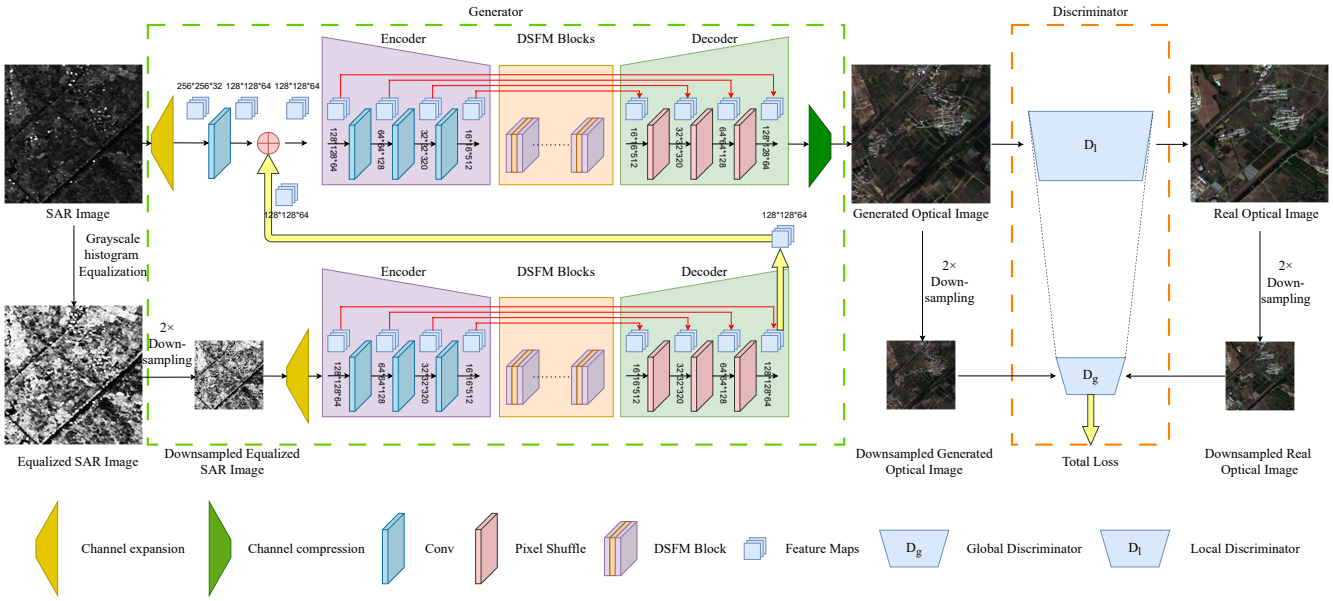


Figure 1. Overall framework of the proposed SAR-to-optical image translation network.

2.2. Data Preprocessing

Gray-level histogram equalization is a classic and efficient algorithm for image contrast enhancement. Its core advantages lie in low computational overhead and significant enhancement effects on low-contrast images, making it widely adopted during the preprocessing stage of diverse image processing tasks [25,26]. Due to the overly concentrated gray-level distributions in original SAR images from the RADARSAT-2 and SEN1-2 satellite public datasets used in this study, overall image contrast remains low, which severely constrains subsequent feature extraction by the network. Therefore, an adaptive histogram equalization preprocessing strategy is introduced to significantly boost image contrast, thereby improving the network's initial perception of land cover geometry and texture details.

The fundamental principle of histogram equalization is to shift and stretch gray-level probability measures concentrated in narrow intervals into an approximately uniform distribution via a non-linear mapping functional, thereby maximizing the global information entropy of the image. Let a discrete input SAR image be defined as a two-dimensional mapping $I : \Omega \to \mathcal{L}$, where $\Omega \subset \mathbb{Z}^2$ denotes the spatial pixel domain, and $\mathcal{L} = \{0, 1, \ldots, L-1\}$ represents the dynamic gray space ($L = 256$ in this study). For any given discrete gray

level $r_k \in \mathcal{L}$, its marginal probability mass function (PMF) $p(r_k)$ is strictly defined via the statistical properties of spatial pixel sets as

$$p(r_k) = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \delta(I(\mathbf{x}) - r_k), \quad \forall k \in [0, L-1] \cap \mathbb{Z} \tag{1}$$

where $|\Omega|$ represents the cardinality of the spatial domain (i.e., the total number of pixels N), $\mathbf{x}$ denotes the two-dimensional spatial coordinate vector, and $\delta(\cdot)$ is the discrete Kronecker delta function (which equals 1 if and only if its input is 0).

To achieve equal-probability mapping toward the maximum entropy state, a monotonically increasing non-linear transformation operator $\mathcal{T} : \mathcal{L} \to \mathcal{L}$ is constructed to map source gray level r_k to target gray level y_k. This transformation operator depends on the empirical cumulative distribution function (ECDF) $F_R(r_k)$ of input gray levels, mathematically formulated as

$$y_k = \mathcal{T}(r_k) = \lfloor (L-1) \cdot F_R(r_k) \rfloor = \left\lfloor (L-1) \cdot \sum_{j=0}^{k} p(r_j) \right\rfloor \tag{2}$$

where $\lfloor \cdot \rfloor$ denotes the floor operation to ensure output values remain strictly closed within the discrete gray space $\mathcal{L}$, and $F_R(r_k)$ represents the cumulative probability measure within interval $[0, r_k]$. Multiplying it by the range factor $(L-1)$ projects the normalized distribution full-scale onto the full target dynamic space. Through this discrete mapping operator, gray-level subspaces with high occurrence frequencies are adaptively stretched and expanded, whereas low-frequency subspaces are moderately merged and compressed. Consequently, the global gray-level distribution of the output image becomes more uniform, significantly enhancing visual contrast.

2.3. Generator Architecture

Existing studies demonstrate that multi-scale feature representations effectively balance global macro-structures with local micro-details [14,22]. Inspired by the literature [27] with targeted improvements, and deeply integrating the design concepts of Pix2PixHD, the core network architecture adopts a dual U-Net topology featuring parallel global generator G_g and local generator G_l. Both sub-generators hierarchically process input images at two resolution scales: 256×256 and 128×128.

Within each individual U-Net branch, the forward propagation is divided into three core phases: feature encoding, feature translation, and feature decoding. The encoding phase comprises an initial channel expansion followed by four downsampling stages, with cumulative downsampling factors of 2, 4, 8, and 16, and corresponding output channel numbers set to 64, 128, 320, and 512, respectively. Features extracted at each downsampling stage are preserved as multi-level encoding features for that specific scale. Subsequently, these four multi-scale encoding features are fed into several proposed Dual-Scale Fusion Modules (DSFMs) for cross-domain translation, producing deep translated features at corresponding scales. Finally, during the decoding phase, the network restores spatial resolution and performs layer-wise fusion of translated features across scales via upsampling. To minimize spatial information loss, the upsampling process utilizes pixel shuffle technology [28], ultimately yielding translation outputs matching the spatial resolution of original input images.

2.3.1. Global Generator and Local Generator

Following Pix2PixHD design principles, the generative network is decoupled into a global generator G_g and a local generator G_l, expressed as

$$G = \{G_g, G_l\}. \tag{3}$$

to model macro-global structures and micro-local details of images, respectively. Given original image spatial dimensions of 256 × 256, the specific feature generation workflow is as follows: first, the original SAR image undergoes gray-level histogram equalization preprocessing and is downsampled to 128 × 128 spatial resolution to serve as input for global generator G_g, capturing global context. The local generator G_l accepts the original 256 × 256 SAR image as input and performs layer-by-layer fusion with corresponding feature maps from G_g at specific encoder stages, thereby introducing global priors to guide local detail generation. Through this cascaded architecture, the network maintains overall topological consistency while effectively enhancing the clarity of local textures.

Conventional U-Net architectures consist of an encoder, a bottleneck, and a decoder. However, multi-stage downsampling in encoders causes severe spatial dimension reduction, inducing feature extraction bottlenecks [29] characterized by two major drawbacks:

1. Excessive resolution decay leads to local feature fragmentation and global semantic inconsistency. In SAR image translation, this manifests as divergent responses to land cover types across feature maps of different scales, making it difficult for subsequent decoding stages to fuse logically consistent images.
2. Downsampling causes irreversible loss of high-frequency details. Even when decoders restore spatial dimensions via upsampling, reconstructing fine textures and edge structures remains challenging, resulting in blurry generated images with missing details.

To mitigate these issues, an improved skip connection mechanism is introduced into U-Net, embedding Dual-Scale Fusion Modules along connection paths to achieve deep cross-scale feature alignment and representation enhancement.

2.3.2. Dual-Scale Fusion Module

As illustrated in Figure 2, the input feature map is denoted as $F_{\text{in}} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the spatial height, width, and channel depth at the current scale, respectively. The proposed module first applies a split operator mapping along the orthogonal channel basis of F_{in}, taking the first half of the channels to construct the global feature branch F_g and the remaining half for the local feature branch F_l. The mathematical decoupling process can be expressed via tensor slicing operations as

$$F_g = F_{\text{in}}\left[:, :, 1 : \frac{C}{2}\right], \quad F_l = F_{\text{in}}\left[:, :, \frac{C}{2} + 1 : C\right] \tag{4}$$

where $F_g, F_l \in \mathbb{R}^{H\times W\times \frac{C}{2}}$ carry independent feature manifold information.

To construct a cross-scale global reference, the global branch is downsampled to 16 × 16 spatial resolution (i.e., $F_g \in \mathbb{R}^{H_{16}\times W_{16}\times \frac{C}{2}}$), whereas the local branch retains its original spatial resolution. To enhance the representation capabilities of both branches for their respective scales, Large Kernel Attention (LKA) modules are introduced to refine features independently:

$$F_g = \text{LKA}(F_g), \quad F_l = \text{LKA}(F_l) \tag{5}$$

For the local feature branch, the LKA module captures subtle textures and high-frequency edge details precisely by leveraging its broad receptive field; for the global feature branch, LKA integrates long-range dependencies from a macroscopic view, extracting structural patterns with global semantic context.

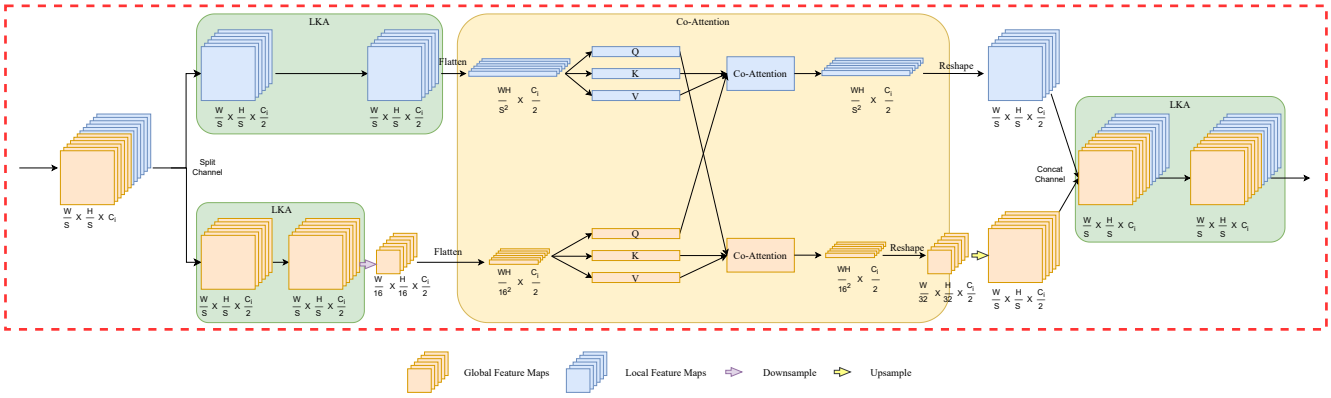


Figure 2. Structure of the Dual-Scale Fusion Module (DSFM).

In multi-scale feature fusion, a naive paradigm upsamples low-resolution global features via bilinear interpolation and concatenates them directly with local features along the spatial dimension. However, since dual-scale features operate under different semantic receptive fields, direct concatenation leads to severe spatial misalignment. To explicitly uncover the intrinsic relationships between both representations, a Collaborative Attention (CoA) mechanism is introduced to achieve learnable, dynamic, bidirectional cross-scale alignment and deep coupling.

Adopting Collaborative Attention (CoA) for cross-scale soft alignment rather than naive concatenation is governed by two complementary design rationales. First, SAR backscattering intrinsically couples valid surface semantics with multiplicative speckle noise across heterogeneous receptive fields; through multi-scale, multi-path feature exchange, the network can dynamically decouple salient structural semantics from noise distributions at appropriate spatial scales, extracting robust land-cover representations required for optical synthesis. Second, Large Kernel Attention (LKA) strengthens long-range contextual dependencies in the global stream while refining high-frequency textural cues in the local stream, providing highly discriminative inputs for cross-scale interaction; Collaborative Attention (CoA) then establishes soft semantic alignment via learnable bidirectional cross-attention maps. This breaks the representational isolation between global and local hierarchies and facilitates direct mutual guidance, effectively mitigating spatial distortion and scale mismatch inherent to rigid upsampling-concatenation schemes.

First, spatial flattening operations are performed on the processed feature tensors to reshape them into sequence manifolds. Let the local sequence length be $N_l = H \times W$, and the global sequence reference length be $N_g = H_{16} \times W_{16}$. The flattened feature sequences are represented as $\mathbf{f}_l \in \mathbb{R}^{N_l \times \frac{C}{2}}$ and $\mathbf{f}_g \in \mathbb{R}^{N_g \times \frac{C}{2}}$, respectively. Subsequently, parameterized linear projection layers map the feature sequences into latent subspaces, generating Query (Q), Key (K), and Value (V) matrices for corresponding branches:

$$\begin{aligned} Q_l &= \mathbf{f}_l W_Q^l, \quad K_l = \mathbf{f}_l W_K^l, \quad V_l = \mathbf{f}_l W_V^l \\ Q_g &= \mathbf{f}_g W_Q^g, \quad K_g = \mathbf{f}_g W_K^g, \quad V_g = \mathbf{f}_g W_V^g \end{aligned} \tag{6}$$

where $W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{\frac{C}{2} \times d_k}$ $(i \in \{l, g\})$ denote learnable linear projection weight matrices, and d_k represents the dimension of the mapped subspace.

Based on the projection matrices, scaled dot-product operations across branches are computed to measure bidirectional similarities between local and global tokens. Normalized cross-scale attention weight matrices are then generated using Softmax activation functions:

$$S_{g\to l} = \text{Softmax}\left(\frac{Q_l K_g^\top}{\sqrt{d_k}}\right), \quad S_{l\to g} = \text{Softmax}\left(\frac{Q_g K_l^\top}{\sqrt{d_k}}\right) \tag{7}$$

where $S_{g\to l} \in \mathbb{R}^{N_l \times N_g}$ characterizes the importance of each global feature token to the corresponding local feature region, guiding local features toward global context alignment; similarly, $S_{l\to g} \in \mathbb{R}^{N_g \times N_l}$ enables global features to adaptively absorb high-frequency local details. Dynamically aligned hybrid feature representations are aggregated via attention weighting:

$$\tilde{\mathbf{f}}_l = S_{g\to l} V_g, \quad \tilde{\mathbf{f}}_g = S_{l\to g} V_l \tag{8}$$

After completing soft alignment, the hybrid features $\tilde{\mathbf{f}}_l \in \mathbb{R}^{N_l \times d_k}$ and $\tilde{\mathbf{f}}_g \in \mathbb{R}^{N_g \times d_k}$ are each passed through a 1×1 convolution layer to restore the initial channel dimension $\frac{C}{2}$, and inversely reshaped into 2D grid topologies to yield aligned feature maps $\tilde{F}_l \in \mathbb{R}^{H \times W \times \frac{C}{2}}$ and $\tilde{F}_g \in \mathbb{R}^{H_{16} \times W_{16} \times \frac{C}{2}}$. Next, spatial upsampling (Up) is applied to the low-resolution global representation $\tilde{F}_g$ to match the spatial dimensions of $\tilde{F}_l$, followed by channel-wise concatenation (Concat). Finally, the concatenated features are fed into an LKA module to complete the ultimate cross-scale integration:

$$F_{\text{out}} = \text{LKA}\Big(\text{Concat}\big(\tilde{F}_l, \text{Up}(\tilde{F}_g)\big)\Big) \tag{9}$$

Through this adaptive closed-loop information interaction, the DSFM overcomes pixel misalignment bottlenecks caused by static fusion strategies and breaks the information barrier between scales. This allows the generative network to maintain the macro-structural framework of SAR images while reproducing local optical textures with high fidelity, significantly boosting semantic consistency and visual naturalness.

2.4. Discriminator Architecture and Adversarial Game

Substantial physical disparities exist between SAR and optical imaging mechanisms, with optical imagery harboring richer spatial details and textural information. Consequently, SAR-to-optical image translation requires generative networks not only to preserve manifold distribution consistency with target optical images in global visual content, but also to achieve high fidelity in detail features and texture rendering. This imposes stricter topological constraints on the discrimination capability of discriminator networks.

Conventional generative adversarial network (GAN) discriminators map an input image to a single scalar value, relying solely on the final layer output for global binary classification. This paradigm overemphasizes macro-distributions while easily ignoring high-frequency texture details. To overcome this limitation, this paper adopts a discriminator network based on the PatchGAN architecture [13]. Breaking away from global mapping mechanisms, this discriminator maps the input image space Ω to a series of highly overlapping local receptive fields (patches).

Let the joint input feature pair to the discriminator be $\mathbf{x}$ (defined as the channel-wise concatenation of the condition SAR image with either the real optical target or the synthesized optical output), with the corresponding source SAR condition denoted as $\mathbf{x}_{\text{sar}}$ and the real optical image as $\mathbf{y}$. Mathematically, PatchGAN functions as a fully convolutional network mapping $\mathcal{D} : \Omega \to \mathbb{R}^{H_D \times W_D}$, outputting a probability matrix of size $H_D \times W_D$. Each element $D(\mathbf{x})_{i,j}$ strictly represents the confidence score that the (i, j)-th local patch belongs to real optical image features. Adopting the least-squares objective of the Least Squares GAN (LSGAN) for stabilized adversarial gradients, the objective functions for the multi-scale discriminators and generator are formulated as

$$\mathcal{L}_{\text{cGAN}}(D_k) = \frac{1}{2H_D W_D} \sum_{i=1}^{H_D} \sum_{j=1}^{W_D} \left[\left(D_k(\mathbf{x}_{\text{sar}}, \mathbf{y})_{i,j} - 1 \right)^2 + \left(D_k(\mathbf{x}_{\text{sar}}, G(\mathbf{x}_{\text{sar}}))_{i,j} \right)^2 \right] \tag{10}$$

$$\mathcal{L}_{\text{cGAN}}(G) = \sum_{k=1}^{2} \frac{1}{H_D W_D} \sum_{i=1}^{H_D} \sum_{j=1}^{W_D} \left(D_k(\mathbf{x}_{\text{sar}}, G(\mathbf{x}_{\text{sar}}))_{i,j} - 1 \right)^2 \tag{11}$$

where $D_k \in \{D_l, D_g\}$ denotes the multi-scale sub-discriminators operating at the fine and coarse scales, respectively, and $G(\mathbf{x}_{\text{sar}})$ represents the full-resolution synthesized optical image. The key advantage of this design lies in balancing macro-distributions with micro-details: spatial continuity constraints between adjacent patches safeguard the global semantic consistency of generated images, while independent patch-level discrimination captures microscopic variations in high-frequency local textures and edges, preventing generated results from falling into "globally plausible but locally distorted" optimization traps.

Furthermore, to enhance discrimination accuracy across multiple physical resolutions, a multi-scale evaluation strategy is integrated into the PatchGAN pipeline. Two independent sub-discriminators with identical network topologies but different input spatial scales are constructed, denoted as the global sub-discriminator D_g and the local sub-discriminator D_l, as illustrated in Figure 3. Specifically, local sub-discriminator D_l focuses on verifying high-frequency edges and fine land-cover textures at the original resolution scale; global sub-discriminator D_g inspects macro-geometric layouts and spatial continuity across large background regions at a downsampled lower resolution scale.

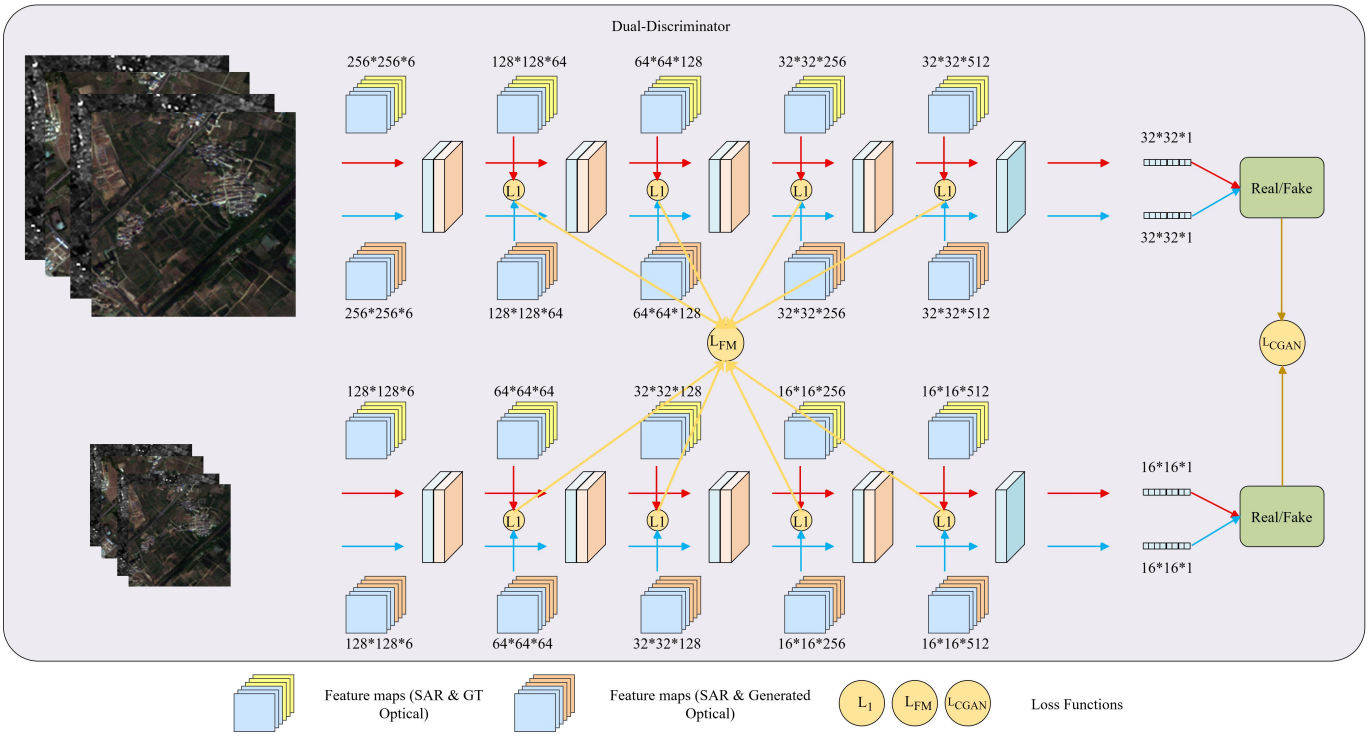


Figure 3. Structure of the proposed multi-scale discriminators.

During network training, the Adam stochastic gradient optimizer alternates updates between the dual-generator family $\{G_g, G_l\}$ and the multi-scale discriminator family $\{D_g, D_l\}$. Through dynamic min-max adversarial play, the system approaches a Nash equilibrium, driving visual quality and multi-scale physical feature fidelity toward optimal convergence.

2.5. Loss Functions

The joint loss functional constructed in this paper comprises four components: adversarial loss [6], perceptual loss [12], feature matching loss [30], and the continuous differentiable soft-edge loss proposed in this paper. Specifically, the adversarial loss adopts the least squares objective from LSGAN; the feature matching loss measures distribution discrepancies across intermediate feature manifolds of multi-scale discriminators; and the perceptual loss extracts and maps deep semantic features from Visual Geometry Group (VGG) networks. Existing studies demonstrate that under a baseline adversarial loss weight of 1, setting feature matching and perceptual loss weights to 10 yields optimal network convergence [22]. Subsequent loss weight ablation experiments in this paper are conducted using this hyperparameter manifold baseline.

However, applying the classic Canny edge detection algorithm to end-to-end deep learning optimization poses severe theoretical flaws: on the one hand, non-maximum suppression (NMS) and double-threshold hysteresis operations act as discrete "hard truncation" operators that are mathematically non-differentiable, severely blocking continuous backpropagation of error gradients; on the other hand, its built-in Sobel spatial differential operator assumes additive Gaussian noise in optical images. When facing the unique physical multiplicative speckle noise in SAR imagery, gradient variances in bright regions diverge abnormally, triggering numerous false edges.

To address these challenges, this paper dispenses with traditional hard-threshold edge operators and constructs a differentiable soft-edge perceptual loss that is higher-order continuous and fully differentiable in the spatial domain. The design of this loss is motivated by the following physical insight: illumination and surface reflectance vary dramatically across optical remote sensing scenes, where highly reflective structures (e.g., building roofs) tend to dominate gradient backpropagation, submerging the structural cues of low-contrast regions such as shadows and water bodies. To overcome this gradient imbalance, the image signals are first mapped into the logarithmic domain, as illustrated in Figure 4:

$$I_{\log}(\mathbf{x}) = \ln\left(I(\mathbf{x}) + \epsilon\right), \quad \forall \mathbf{x} \in \Omega \tag{12}$$

where $I(\mathbf{x})$ denotes the image mapping defined on two-dimensional discrete spatial domain $\Omega \subset \mathbb{Z}^2$, spatial coordinate vector $\mathbf{x} = [x, y]^\top \in \Omega$, and $\epsilon = 10^{-5}$ is a small regularization constant to prevent log numerical singularity. In accordance with the Weber–Fechner Law of human perception and Retinex illumination-reflectance decomposition, logarithmic transformation effectively compresses the high dynamic range of radiometric intensities. This renders the differential operator invariant to absolute illumination scaling, thereby equitably capturing structural boundaries across both specular highlight regions and weak-contrast areas such as shaded terrain and inland water bodies. Within this log domain, spatial filtering with an isotropic Gaussian kernel suppresses high-frequency noise interference, deriving an adaptive gradient magnitude operator $M_{\mathrm{SAR}}(\mathbf{x})$:

$$M_{\mathrm{SAR}}(\mathbf{x}) = \left\| \left(\nabla K_\sigma\right) * I_{\log}(\mathbf{x}) \right\|_2 \tag{13}$$

where K_σ represents a 2D Gaussian kernel with scale parameter σ, $\nabla = [\partial/\partial x, \partial/\partial y]^\top$ denotes the spatial gradient differential operator, $*$ represents 2D spatial convolution, and $\|\cdot\|_2$ is the L_2 norm of a vector. This operator is mathematically equivalent in expected value to the local Ratio of Averages (ROA) operator, further eliminating sensitivity to absolute radiometric intensity and substantially boosting the gradient responses of structural boundaries across low-contrast land covers.

To break non-differentiable truncation barriers in classic algorithms, a continuous differentiable parameterized Logistic activation functional is introduced to smoothly approximate ideal double-threshold step boundaries, generating a higher-order continuous Soft-Edge Probability Map (SEPM):

$$E(\mathbf{x}) = \sigma_{\beta,\tau}\left(M_{\mathrm{SAR}}(\mathbf{x})\right) = \left[1 + \exp\left(-\beta \cdot \left(M_{\mathrm{SAR}}(\mathbf{x}) - \tau\right)\right)\right]^{-1} \tag{14}$$

where τ is an adaptive shift parameter defining the edge decision center, and β is a scaling factor controlling the sharpness of the probability transition band. To grant dynamic optimization capabilities to the network, τ and β are jointly parameterized as learnable parameters, enabling them to automatically converge toward optimal discrimination scales during minimax adversarial play.

Based on these mathematical derivations, the proposed differentiable soft-edge perceptual loss function is defined as the expected L_1 distance between generated and real optical images across soft-edge probability topological spaces:

$$\mathcal{L}_{\mathrm{edge}} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \left\| E_{\mathrm{gan}}(\mathbf{x}) - E_{\mathrm{GT}}(\mathbf{x}) \right\|_1 \tag{15}$$

where $|\Omega|$ represents the total pixel cardinality of the spatial domain (equivalent to $W \times H$), and E_{gan} and E_{GT} map the soft-edge probability spatial distributions extracted from generated and real target images, respectively. This loss functional guarantees smooth gradient flow during end-to-end optimization and explicitly enforces geometric alignment along semantic boundaries such as urban building footprints, water-land interfaces, and agricultural field boundaries, effectively driving translation networks across modal distribution gaps to substantially boost geometric realism and structural fidelity along land–water boundaries and complex land cover edges.

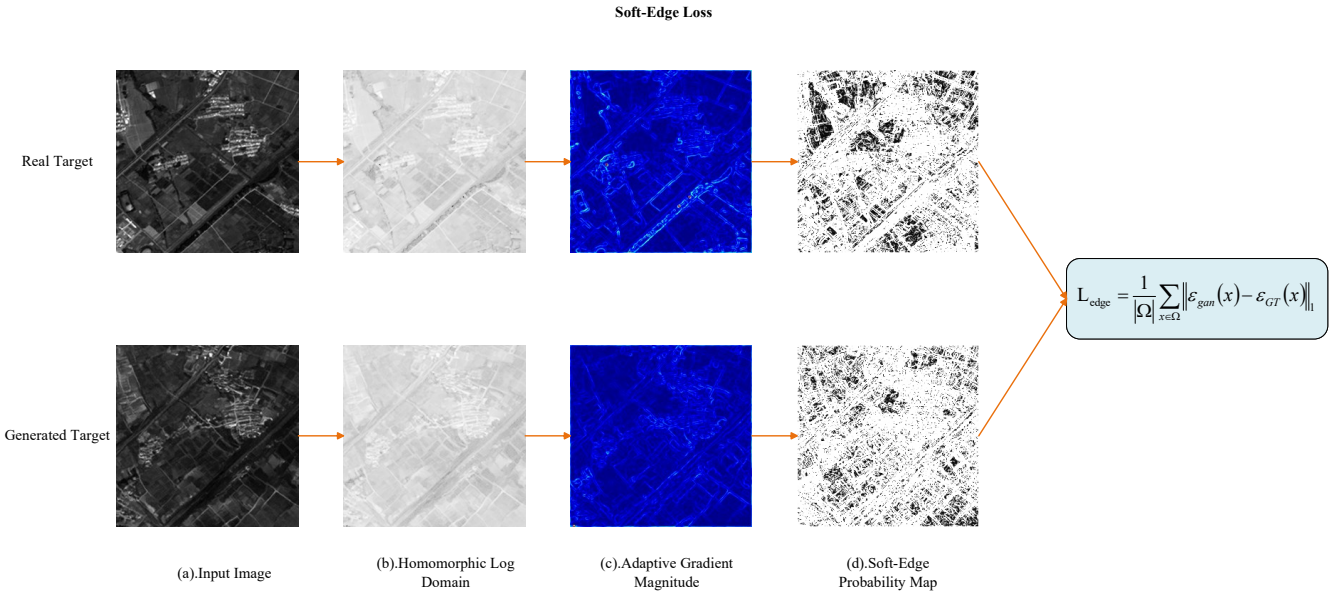


Figure 4. Pipeline of the continuous differentiable soft-edge loss function ($\mathcal{L}_{\mathrm{edge}}$): Both real target images (I_{GT}) and generated fake images (I_{gan}) are transformed into the log domain to compress the dynamic range and eliminate illumination differences, followed by adaptive gradient magnitude extraction. Parameterized Sigmoid functions are mapped to continuous Soft-Edge Probability Maps (E_{GT} and E_{gan}), where an L_1 distance is computed to enforce structural boundary consistency.

Total Loss

Combining the individual loss constraints described above, the total loss function for network optimization is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_0 \mathcal{L}_{\text{cGAN}} + \lambda_1 \mathcal{L}_{\text{FM}} + \lambda_2 \mathcal{L}_{\text{VGG}} + \lambda_3 \mathcal{L}_{\text{edge}} \tag{16}$$

where $\lambda_0, \lambda_1, \lambda_2$, and λ_3 denote hyperparameter weights for adversarial loss, feature matching loss, perceptual loss, and edge loss, respectively. Following the baseline literature and experimental configurations, these four weight terms are set to 1, 10, 10, and 15, respectively.

3. Results

3.1. Experimental Parameters and Setup

This paper utilizes paired SAR and optical remote sensing images acquired over Nanjing, Jiangsu Province, China, as the primary experimental dataset. The SAR imagery was acquired by the RADARSAT-2 satellite, and the optical imagery was obtained from the RapidEye satellite, both featuring a spatial resolution of 5 m. Original images were cropped into uniform 256 × 256 pixel patches and augmented via random flipping and rotation, resulting in a standardized benchmark comprising 8452 training pairs, 1500 validation pairs, and 1500 test pairs. Regarding the partitioning protocol, the Nanjing dataset was partitioned strictly at the level of original large-scale satellite scenes prior to slicing. Non-overlapping 256 × 256 patches cropped from a specific large scene were exclusively allocated to one of the training, validation, or testing subsets. Consequently, the geographic extents of all test samples remained entirely disjoint from the training distribution, fundamentally preventing spatial autocorrelation and cross-set data leakage. To evaluate model generalization across diverse geographic regions, the public SEN1-2 dataset was introduced as a supplementary benchmark. Since the standard SEN1-2 distribution utilizes a 128-pixel sliding stride that introduces a 50% spatial overlap between adjacent patches, standard random splitting would inevitably incur severe data leakage and inflated evaluation metrics. To eliminate this spatial overlap, we reconstructed continuous spatial scenes by mosaicking overlapping patches and re-cropped mutually independent, non-overlapping patches. Geographic regions allocated to the test subset were strictly isolated from the training partition. Following this rigorous cleaning protocol and identical data augmentation strategies, 5696 image pairs were selected for training, with 824 pairs reserved for validation and testing, respectively.

Regarding hyperparameter configurations, the generator accepts 256 × 256 × 3 SAR image tensors, while the discriminator processes concatenated 256 × 256 × 6 image pairs. Weights for individual terms in the joint loss functional were set to 1, 10, 10, and 15, respectively. The network was optimized using the Adam optimizer with a batch size of 8 and an initial learning rate of 0.0002. Training lasted for 150 epochs, maintaining the initial learning rate for the first 50 epochs and linearly decaying it to zero over the remaining 100 epochs. All experiments were conducted on a deep learning workstation equipped with dual NVIDIA GeForce RTX 4070 Ti Super GPUs (NVIDIA Corporation, Santa Clara, CA, USA).

3.2. Evaluation Metrics

In image-to-image translation, mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) serve as standard metrics. Operating on pixel-level physical differences, these metrics offer computational simplicity. However, they often fail to capture high-level semantic structures and fine texture details. For instance, localized spatial smoothing may induce minimal MSE variations while

severely degrading subjective visual quality. Consequently, deep feature-driven perceptual metrics are essential for comprehensive performance evaluation.

Fréchet Inception Distance (FID) [31] evaluates generative quality by measuring feature distribution discrepancies in deep semantic space. Leveraging a pre-trained Inception-v3 network, 2048-dimensional feature vectors are extracted from real optical and generated images. Multivariate Gaussian distributions fitted to these features are then compared by calculating the Fréchet distance between their mean vectors and covariance matrices. Unlike pixel-level metrics, FID assesses global semantic consistency and data distribution alignment, correlating closely with human visual perception. Thus, FID was designated as the primary evaluation metric in this work.

Learned Perceptual Image Patch Similarity (LPIPS) [32] is another deep perceptual metric. Operating similarly to perceptual loss mechanisms, LPIPS utilizes pre-trained deep neural networks to extract multi-scale feature maps, calculating weighted feature-block distances between generated outputs and corresponding ground truth images. Compared with FID, LPIPS emphasizes paired perceptual similarity, providing fine-grained quantification of local texture and structural fidelity aligned with human vision.

In summary, a multi-dimensional evaluation system was established, anchored by FID and LPIPS as primary metrics and complemented by PSNR, SSIM, and MSE. By combining global distribution alignment (FID) with localized perceptual similarity (LPIPS), quantitative evaluation was conducted across macro-semantic and micro-visual dimensions.

3.3. Dataset Preprocessing

In the Nanjing dataset, original SAR gray values were heavily concentrated in the 0–75 range, causing severe visual radiometric compression (dark appearance). This gray-level collapse severely hinders the extraction of global topological and structural features. Therefore, adaptive histogram equalization was applied to preprocess the raw data. As shown in Figures 5 and 6, preprocessed images exhibited significantly broadened, uniform gray distributions, making land cover boundaries (e.g., roads and rivers) visually distinct. Quantitative evaluation is summarized in Table 1: prior to preprocessing, SAR images in the Nanjing dataset yielded an average contrast of 33.6001 and an average information entropy of 6.6041; post-processing, average contrast increased to 50.805 and average information entropy reached 7.4996, corresponding to a 51.2% contrast gain and a 13.6% entropy enhancement. Notably, the information entropy increases rather than decreases following equalization. In the raw SAR imagery, the radiometric distribution suffers from severe dynamic range compression, where a vast majority of pixels are congested within a very narrow gray-level band, resulting in significant underutilization of the available quantization space. By nonlinearly redistributing and stretching the probability density function across the full dynamic range, equalization disperses clustered pixel intensities into previously vacant states, thereby enriching the distinguishable gray-level hierarchy. This entropy increment objectively confirms that the preprocessing expands the perceptible radiometric information capacity while enhancing contrast, providing superior discriminative features for subsequent boundary delineation and textural synthesis. Similarly, on the SEN1-2 dataset, histogram equalization achieved a 130.27% contrast gain with a minor 2.38% entropy loss, validating the robustness and generalization of the preprocessing strategy across datasets.

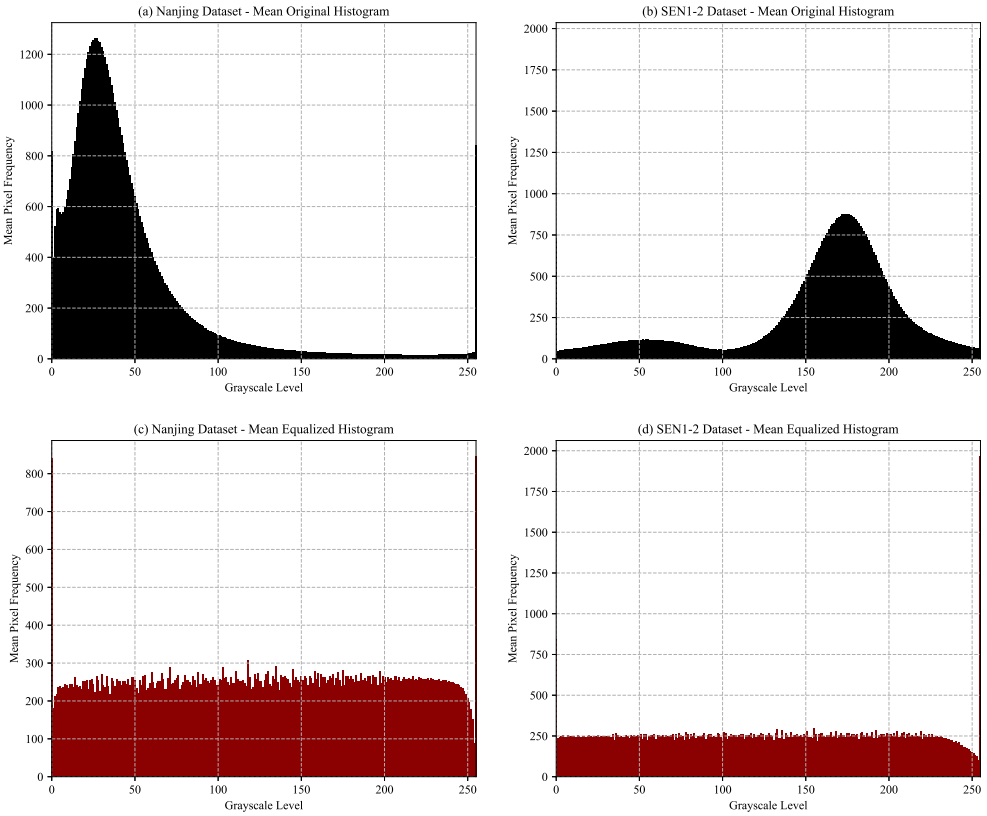


Figure 5. (**a**) Original mean gray-level histogram of the Nanjing dataset. (**b**) Original mean gray-level histogram of the SEN1-2 dataset. (**c**) Equalized mean gray-level histogram of the Nanjing dataset. (**d**) Equalized mean gray-level histogram of the SEN1-2 dataset. Black bars represent original distributions; dark red bars denote equalized distributions.

Table 1. Comparison of evaluation metrics before and after histogram equalization across different datasets.

Metric	Nanjing Dataset			SEN1-2 Dataset		
	Before	**After**	**Change (%)**	**Before**	**After**	**Change (%)**
Contrast	33.6001	50.805	+51.2	29.8218	68.6704	+130.27
Entropy	6.6041	7.4996	+13.6	6.7417	6.5813	−2.38

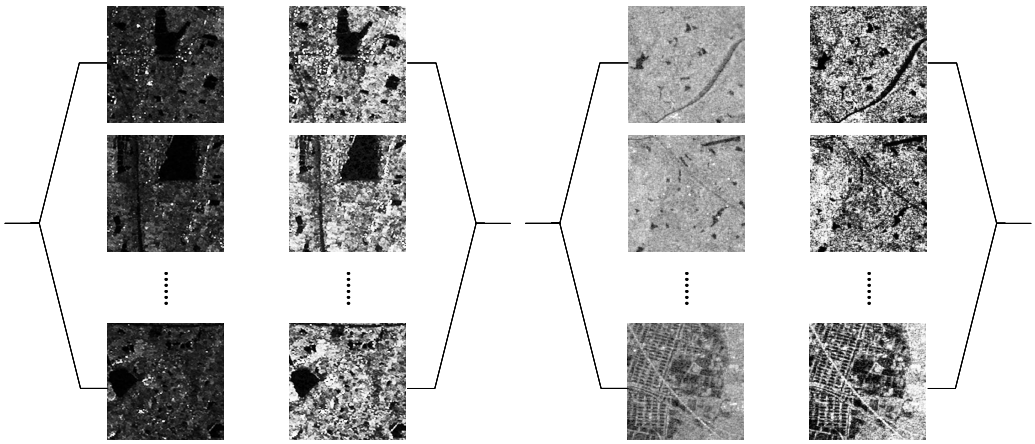


Figure 6. Schematic diagram of gray-level histogram equalization, showing equalized results for the Nanjing dataset (**left**) and the SEN1-2 dataset (**right**).

Comparison of Preprocessing Strategies

To evaluate the rationale behind employing adaptive histogram equalization (HE) as the macroscopic prior, we systematically compared it against two representative contrast-enhancement paradigms: Contrast Limited Adaptive Histogram Equalization (CLAHE) and a Markov Random Field (MRF)-based spatial regularization method. The radiometric statistics of the preprocessed inputs and the resulting downstream translation performance are reported in Table 2.

Table 2. Comparison of radiometric statistics and downstream translation performance under distinct preprocessing strategies on the Nanjing dataset.

Strategy	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MSE ↓	Contrast	Entropy
MRF	63.471	0.5692	17.287	**0.2269**	0.5582	49.0295	7.4038
CLAHE	56.039	**0.5594**	**17.514**	0.2241	**0.5428**	48.7210	7.2641
HE (Ours)	**53.121**	0.5605	17.485	0.2193	0.5471	**50.805**	**7.4996**

Note: Bold values indicate the best results. ↓ and ↑ denote that lower and higher values are better, respectively.

From the perspective of input radiometric characteristics, HE achieves the highest contrast of 50.805, compared with 48.7210 for CLAHE and 49.0295 for MRF, and the largest information entropy of 7.4996, compared with 7.2641 for CLAHE and 7.4038 for MRF. This confirms that global dynamic range stretching effectively broadens the quantization space and enriches radiometric gradations without suppressing macro-scale variations. Regarding translation fidelity, HE achieves the lowest FID of 53.121, improving by 2.918 over CLAHE and 10.350 over MRF, while its LPIPS of 0.5605 remains virtually on par with the 0.5594 of CLAHE and notably outperforms the 0.5692 of MRF. Although CLAHE achieves marginally higher scores in low-level pixel alignment metrics such as PSNR, SSIM, and MSE, pixel-wise averaging objectives inherently penalize high-frequency structural variations and favor over-smoothed outputs. In contrast, HE provides stronger macro-topological contrast that empowers the generator to synthesize sharper land-cover boundaries and more realistic textural distributions, yielding superior global perceptual alignment as reflected by the pronounced FID advantage. These empirical validations substantiate that the combination of optimal contrast expansion and maximal information entropy under HE provides the most effective conditioning distribution for the global generation stream.

3.4. Experimental Results

3.4.1. Comparative Experiments

To comprehensively evaluate the translation performance of the proposed network, comparative experiments were conducted against baseline algorithms, including Pix2Pix [13], CycleGAN [33], Pix2PixHD [14], MSTMNet [22], as well as the recent state-of-the-art SAR-to-optical translation network based on InternImage and cascaded multi-head attention (ICMA) [34], as presented in Figure 7 and Table 3. Constrained by GPU VRAM capacity, Pix2Pix utilized a U-Net-128 architecture, CycleGAN adopted a ResNet-6Blocks backbone, and Pix2PixHD employed a ResNet-9Blocks setup. Quantitative metrics indicate that among the three general-purpose image-to-image translation networks, CycleGAN achieved relatively superior results; nevertheless, its FID score remained 24.344 higher than our method, its LPIPS was 0.0338 higher, and its pixel-level metrics were consistently inferior to ours. Compared with MSTMNet, a domain-specific algorithm for SAR-to-optical translation, our approach achieved a substantial reduction in FID by 14.471. Although LPIPS experienced a marginal increase of 0.0006, all three pixel-level evaluation metrics demonstrated consistent improvements. Among the recent domain-specific benchmarks, ICMA leverages the large receptive field and deformable convolutions of

InternImage to achieve commendable overall performance over MSTMNet. Nonetheless, our proposed DSAN still outperforms ICMA by 2.389 in FID. While ICMA maintains slight leads in pixel-wise and perceptual distances, with margins of 0.0040, 0.0025, and 0.0066 in LPIPS, SSIM, and MSE, respectively, our method achieves the optimal balance in distribution realism and peak signal fidelity, attaining FID and PSNR values of 53.121 and 17.485, respectively. This reflects superior global structural and radiometric consistency.

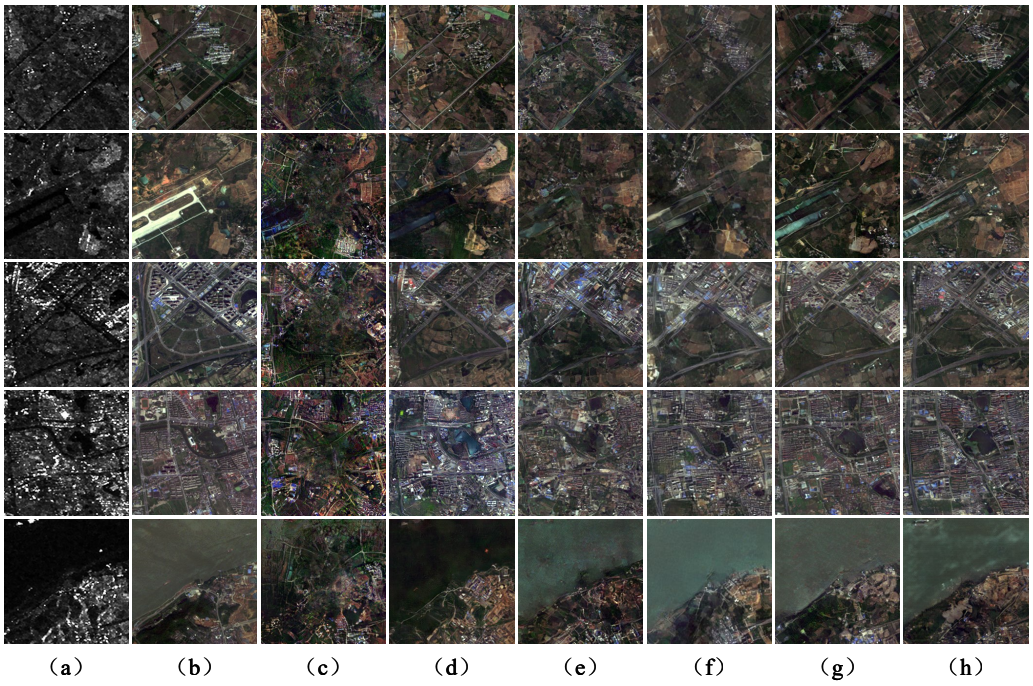


Figure 7. Comparative experiment results on the Nanjing dataset. From left to right: (**a**) SAR image, (**b**) Ground Truth optical image, (**c**) Pix2Pix, (**d**) CycleGAN, (**e**) Pix2PixHD, (**f**) MSTMNet, (**g**) ICMA, and (**h**) Ours.

From a subjective visual perspective, Pix2Pix exhibited severe color distortion and artifacts, yielding the worst performance and misidentifying large water bodies. Pix2PixHD outputs were overly smooth and blurry, struggling to reconstruct fine geometric topologies of roads and buildings while only roughly delineating macroscopic water boundaries. CycleGAN and MSTMNet performed comparatively better, reconstructing basic land-cover categories such as houses, roads, and farmlands; however, their local detail fidelity fell short of our approach. Although ICMA produces enhanced global semantic coherence compared with MSTMNet, its local geometric delineation along structural boundaries (e.g., linear roads and complex building clusters) remains less distinct than our method. For instance, in the first sample set, roads generated by our method exhibited smoother, straight edges matching the Ground Truth optical image most closely. In the second sample set, only our method and ICMA successfully reconstruct the small-scale water body, whereas MSTMNet incurs noticeable chromatic aberration. In the fourth set, CycleGAN misclassified central bare land as a water body, whereas both our method and MSTMNet successfully avoided such semantic confusion, with our approach reconstructing building outlines closer to reality. In the final set involving large-area water body translation, our method alone produced radiometric color attributes consistent with the real optical image.

Table 3. Performance comparison of different models on the Nanjing dataset.

Method	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MSE ↓
Pix2Pix	155.021	0.6171	15.862	0.1461	0.6747
CycleGAN	77.465	0.5943	16.339	0.1854	0.6390
Pix2PixHD	104.921	0.5916	16.915	0.1867	0.5953
MSTMNet	67.592	0.5599	17.048	0.2074	0.5825
ICMA	55.510	**0.5565**	17.480	**0.2218**	**0.5405**
Ours	**53.121**	0.5605	**17.485**	0.2193	0.5471

Note: Bold values indicate the best results. ↓ and ↑ denote that lower and higher values are better, respectively.

To evaluate operational feasibility, the number of parameters, FLOPs, per-image latency, and throughput across all benchmarks are detailed in Table 4. Our model comprises 62.2 M parameters—comparable to the 56.8 M of ICMA—with the parameter footprint primarily attributed to the dual-stream generator and multi-scale discriminator topologies. In terms of computational complexity, our network incurs 103.1 G FLOPs, which is significantly more compact than ICMA at 297.8 G (achieving a 65.4% reduction) and CycleGAN at 113.7 G. Regarding inference efficiency, our model requires 71.37 ms per patch, yielding a throughput of approximately 14 FPS that operates safely within the acceptable real-time threshold of at least 10 FPS for remote sensing batch processing. Regarding training overhead, our model requires 802 s per epoch, totaling approximately 33.4 h across the full 150-epoch schedule. Although this training cost is higher than that of Pix2Pix (42 s), MSTMNet (431 s), and ICMA (477 s), it remains markedly lower than CycleGAN (1107 s) and Pix2PixHD (1369 s). Since training constitutes a one-time offline investment while deployment efficiency is governed by inference latency, this mid-range training overhead represents an acceptable cost. While slightly slower than single-stream baselines due to the unoptimized bidirectional CoA quadratic complexity and dual-branch forward passes, the substantial gains in perceptual fidelity and sharp geometric boundary preservation justify this computational investment. Future deployments can further accelerate inference via operator fusion (e.g., FlashAttention) and structural pruning.

To further verify model generalization across differing data distributions, supplementary validation was conducted on the public SEN1-2 dataset, as illustrated in Figure 8 and Table 5. Quantitatively, while MSTMNet and ICMA maintain slight advantages in pixel-wise MSE/SSIM and PSNR, respectively, our method demonstrates pronounced superiority in perceptual alignment and distribution matching: our FID decreases substantially by 23.582 compared with MSTMNet and by 13.319 against ICMA, while LPIPS achieves the optimum of 0.4758, surpassing MSTMNet and ICMA by 0.0114 and 0.0039, respectively. This confirms that ICMA's performance lies between MSTMNet and our model in macroscopic perceptual realism. Visually, across the first three sample sets, our method alone accurately restored global color distributions matching the target scene: although Pix2PixHD partially mapped colors, its outputs lacked high-frequency textures and appeared blurry; MSTMNet suffered from obvious color shifts (e.g., washed-out colors in the first set and overall dark tones in the second set), and ICMA likewise exhibited chromatic inconsistencies in broad homogeneous regions. In the final two complex scenes, although no algorithm achieved perfect color alignment, our method generated the sharpest geometric boundaries, outperforming baseline methods in translation fidelity for pond and building outlines.

Table 4. Comparison of computational complexity, inference efficiency, and training cost across different models.

Model	Params (M)	FLOPs (G)	Inference Time (ms)	FPS	Training Time (s/Epoch)
Pix2Pix	54.4	36.3	1.72	581	42
CycleGAN	11.4	113.7	6.27	160	1107
Pix2PixHD	11.4	32.1	3.42	292	1369
MSTMNet	44.6	94.1	28.79	35	431
ICMA	56.8	297.8	18.29	55	477
Ours	62.2	103.1	71.37	14	802

Table 5. Performance comparison of different models on the SEN1-2 dataset.

Method	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MSE ↓
Pix2Pix	231.267	0.6860	11.852	0.1046	0.8760
CycleGAN	118.897	0.6181	13.475	0.1349	0.6772
Pix2PixHD	128.252	0.5188	17.331	0.2982	0.4635
MSTMNet	114.393	0.4872	17.688	**0.3287**	**0.4310**
ICMA	104.130	0.4797	**17.720**	0.3101	0.4372
Ours	**90.811**	**0.4758**	17.320	0.2891	0.4561

Note: Bold values indicate the best results. ↓ and ↑ denote that lower and higher values are better, respectively.

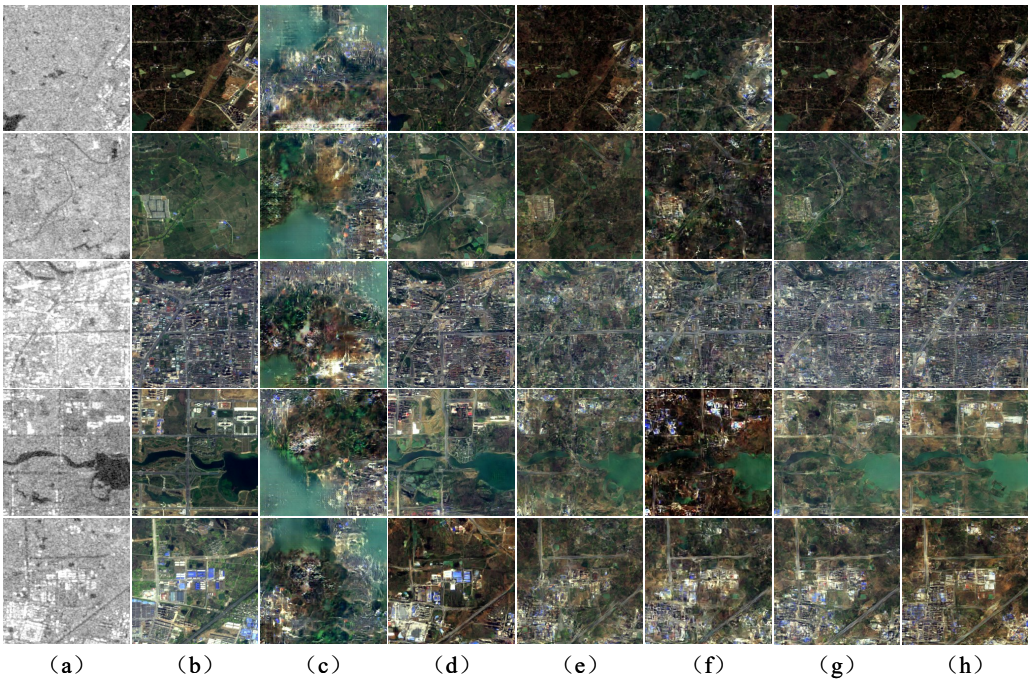


Figure 8. Comparative experiment results on the SEN1-2 dataset. From left to right: (**a**) SAR image, (**b**) Ground Truth optical image, (**c**) Pix2Pix, (**d**) CycleGAN, (**e**) Pix2PixHD, (**f**) MSTMNet, (**g**) ICMA, and (**h**) Ours.

3.4.2. Ablation Study on Network Modules

To verify the effectiveness of the core components proposed in this paper, ablation studies were conducted, with quantitative and qualitative results presented in Figure 9 and Table 6.

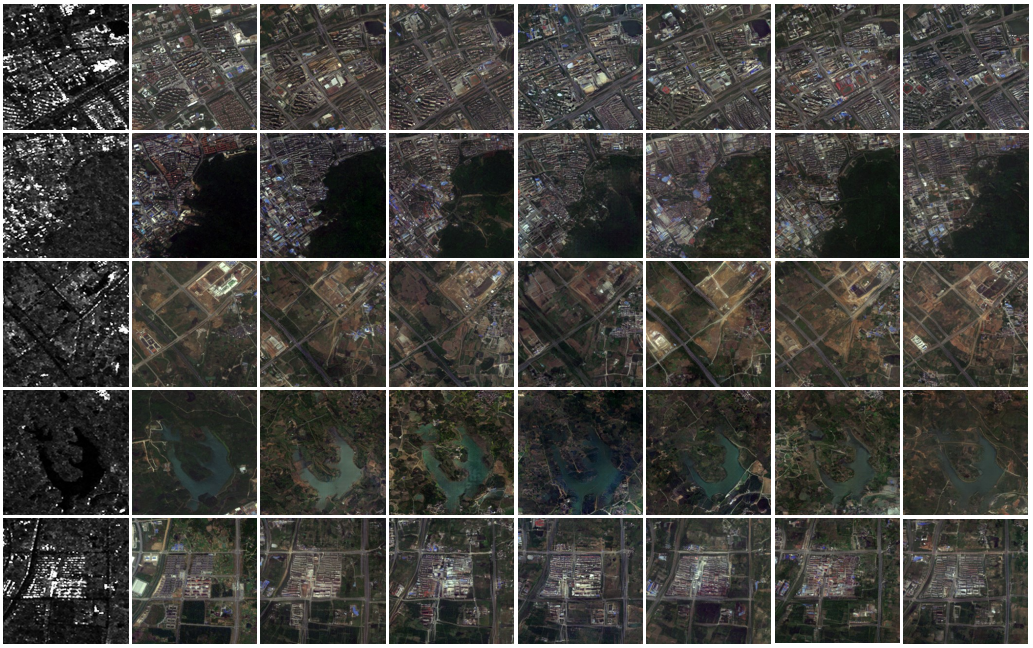


Figure 9. Module ablation experiment results. From left to right: (**a**) SAR image, (**b**) Ground Truth optical image, (**c**) w/o Equalization, (**d**) Dual Equalization, (**e**) w/o Global Generator, (**f**) w/o LKA, (**g**) w/o Soft-Edge Loss, and (**h**) Ours.

Table 6. Performance comparison of different network module variants.

Module Variant	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MSE ↓
w/o Equalization	56.860	0.5615	17.379	0.2118	0.5604
Dual Equalization	59.590	0.5614	17.458	0.2203	0.5467
w/o Global Generator	57.322	0.5688	17.243	0.2134	0.5595
w/o LKA	58.544	0.5638	17.373	0.2186	0.5515
w/o Soft-Edge Loss	56.579	0.5618	**17.636**	**0.2238**	**0.5362**
Ours	**53.121**	**0.5605**	17.485	0.2193	0.5471

Note: Bold values indicate the best results. ↓ and ↑ denote that lower and higher values are better, respectively.

First, an ablation study was conducted on the histogram equalization strategy used in the preprocessing phase. Quantitative metrics indicate that incorporating this strategy reduced FID significantly by 3.739, while improving all three pixel-level metrics compared to the raw input setup, with LPIPS remaining comparable. Second, applying histogram equalization simultaneously to both global and local generator inputs (dual equalization) was investigated. Although dual equalization further boosted contrast, it inevitably introduced structural degradation: its pixel-level metrics varied inconsistently, and FID increased substantially by 6.469 compared to our single-equalization approach. Visually, without equalization, the central building in the first sample set and the water area in the third set exhibited severe radiometric color distortion and distorted road boundaries; conversely, dual equalization generated over-hallucinated redundant textures in the forest area of the second set. These findings confirm the rationality of our asymmetric equalization preprocessing, which enhances contrast for feature extraction while maximally preserving authentic topological structures.

Furthermore, the global generator branch designed under Pix2PixHD principles was ablated. Removing the global generator led to severe degradation in core perceptual metrics despite minimal fluctuations in pixel-level metrics: LPIPS degraded by 0.0083 and

FID increased by 4.201. Visual inspection further revealed severe semantic confusion and color shifts along the left road in the third sample set and around the water body in the fourth set. This underscores the indispensability of the global generator in capturing macro-contextual features.

3.4.3. Sensitivity Analysis on Edge Loss Weight

To validate the weight assigned to the edge loss, a sensitivity analysis was conducted on hyperparameter λ_3, as depicted in Figure 10 and Table 7. Based on the baseline literature and prior experiments, optimal weights for adversarial, feature matching, and perceptual losses were fixed at 1, 10, and 10, respectively. Fixing these three parameters, candidate values of 0, 5, 10, 15, 20, 25, and 30 were evaluated for λ_3.

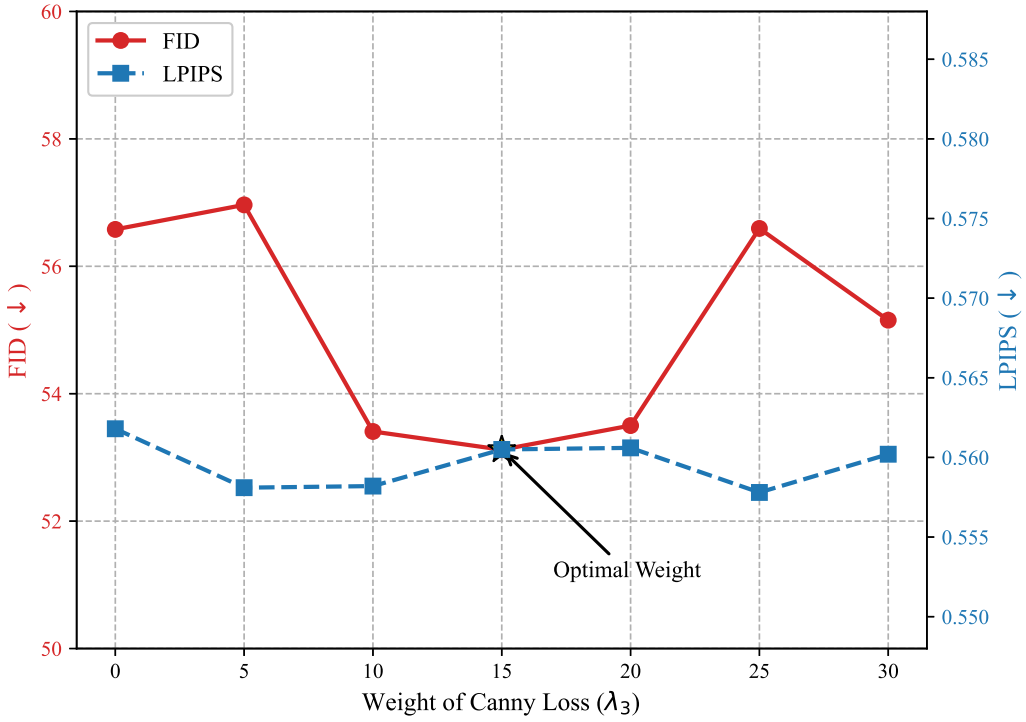


Figure 10. FID and LPIPS curves across different loss weights λ_3 (↓ indicates that lower values are better). The red curve represents FID (left axis), the blue curve represents LPIPS (right axis), and the star marks the optimal weight location.

Table 7. Performance sensitivity analysis across different loss weight values λ_3.

λ_3	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MSE ↓
0	56.579	0.5618	**17.636**	**0.2238**	**0.5362**
5	56.964	0.5581	17.566	0.2225	0.5417
10	53.407	0.5582	17.517	0.2215	0.5437
15	**53.121**	0.5605	17.485	0.2193	0.5471
20	53.500	0.5606	17.383	0.2195	0.5505
25	56.594	**0.5578**	17.544	0.2220	0.5409
30	55.154	0.5602	17.429	0.2198	0.5479

Note: Bold values indicate the best results. ↓ and ↑ denote that lower and higher values are better, respectively.

Quantitative results demonstrate that at $\lambda_3 = 15$, the primary perceptual metric FID achieves optimal performance. Interestingly, pixel-level metrics reach peak values when $\lambda_3 = 0$ (omitting edge loss). This apparent "metric divergence" occurs because,

without high-frequency edge constraints, the network tends to generate overly smooth, blurry outputs to artificially minimize pixel-level error expectations (e.g., MSE). This paradox highlights the necessity of the soft-edge loss in compelling the model to learn sharp boundaries and high-frequency textures. Visual comparisons confirm its effectiveness: at $\lambda_3 = 15$, geometric road straightness in the first set, topological connectivity of small river networks in the second set, and river boundary precision in the fourth set display optimal fidelity.

3.4.4. Ablation Study on Attention Mechanisms

To investigate the specific contribution of the constituent attention mechanisms within the DSFM for dual-scale feature interaction, several ablation and substitution variants were evaluated. These include complete removal of LKA (w/o LKA), applying LKA only during scale separation (LKA Before), applying LKA only during cross-scale fusion (LKA Behind), replacing LKA with Global Attention Mechanism (GAM) [35] or Coordinate Attention (CA) [36], and complete removal of Collaborative Attention (w/o CoA, which reverts cross-scale interaction to naive interpolation-concatenation). Results are illustrated in Figure 11 and Table 8.

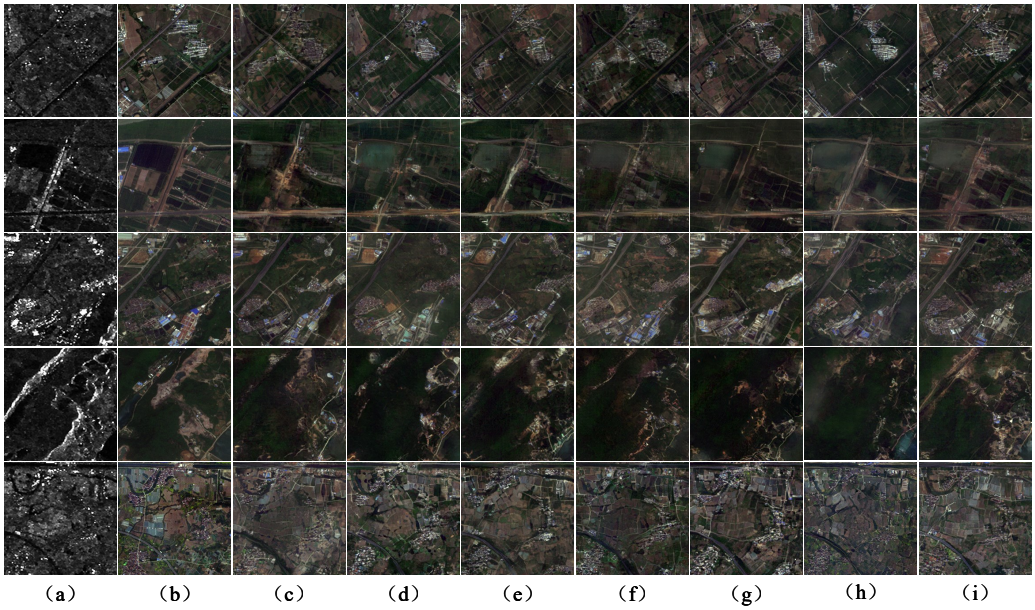


Figure 11. Visual comparison of ablation study on attention mechanism configurations and substitutions. From left to right: (**a**) SAR image, (**b**) Ground Truth optical image, (**c**) w/o LKA, (**d**) LKA Before, (**e**) LKA Behind, (**f**) GAM, (**g**) CA, (**h**) w/o CoA, and (**i**) Ours.

Table 8. Performance comparison of different attention mechanism configurations.

Configuration	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MSE ↓
w/o LKA	58.544	0.5638	17.373	0.2186	0.5515
LKA Before	55.934	0.5644	17.254	0.2165	0.5581
LKA Behind	59.212	0.5667	17.422	0.2186	0.5481
GAM	57.313	0.5638	17.393	0.2202	0.5512
CA	59.532	0.5628	17.438	0.2202	0.5477
w/o CoA	60.657	0.5616	**17.554**	**0.2368**	**0.5415**
Ours	**53.121**	**0.5605**	17.485	0.2193	0.5471

Note: Bold values indicate the best results. ↓ and ↑ denote that lower and higher values are better, respectively.

Quantitatively, the synergistic integration of LKA and CoA in our method achieved optimal scores in both FID and LPIPS, with FID dropping by 2.813 compared to the second-best variant (LKA Before) and by a substantial 7.536 compared to the w/o CoA baseline. Notably, the w/o CoA variant achieves higher numerical scores in low-level pixel alignment metrics such as PSNR, SSIM, and MSE. This apparent "metric divergence" aligns with established generative image synthesis principles: in the absence of explicit cross-scale bidirectional alignment, the generator tends to produce over-smoothed, blurry textures that minimize mean squared pixel deviations. While this artifact suppression artificially inflates pixel-wise metrics, it fundamentally compromises high-frequency structural fidelity and topological coherence, resulting in severe degradation in distribution realism, with FID deteriorating from 53.121 to 60.657. Subjective analysis revealed that omitting LKA at any stage degrades local structural fidelity. For instance, complete LKA removal caused severe color distortion in building areas; retaining LKA solely during scale separation led to an overall lack of visual sharpness; whereas using LKA only during cross-scale fusion reduced translation fidelity in complex textures such as farmlands. Likewise, omitting CoA degrades feature interaction into rigid concatenation, leading to evident semantic boundary dislocations and structural blur across heterogeneous land-cover transitions—such as building edges in urban layouts and riparian contours along river boundaries. These evaluations confirm the optimal configuration of coupling LKA receptive field expansion with CoA bidirectional soft alignment across both feature decomposition and fusion stages.

4. Discussion

The proposed SAR-to-optical translation framework demonstrates consistent performance gains over existing mainstream paradigms across both the Nanjing and SEN1-2 benchmarks. This performance advantage stems from the synergistic orchestration of three complementary designs across the input, intermediate feature, and output supervision stages. At the input stage, the asymmetric dual-prior architecture allocates macroscopic topological perception to the global stream and fine-grained texture preservation to the local stream. Ablation results confirm that removing either prior severely degrades distribution alignment, with FID increasing by 3.739 and 6.469 for the w/o-equalization and dual-equalization variants, respectively, substantiating that complementary conditioning distributions form the bedrock of synthesis fidelity. At the feature stage, the DSFM eliminates inter-scale isolation via LKA receptive field expansion and CoA bidirectional alignment: omitting CoA causes a sharp FID degradation from 53.121 to 60.657, and the concurrent artificial inflation of pixel-level metrics confirms the indispensable role of cross-scale soft alignment in enforcing global semantic consistency. At the output stage, the differentiable soft-edge loss leverages logarithmic dynamic range compression to eliminate specular gradient dominance, equitably constraining structural boundaries across diverse terrain categories (e.g., urban fabrics, water bodies, and agricultural plots). Acting coherently across the input, feature, and output stages, these components establish an end-to-end optimization chain that fundamentally outperforms isolated single-stage modifications.

Compared with the recent domain-specific benchmark ICMA, our method achieves substantial FID reductions of 2.389 and 13.319 on the two benchmarks, respectively, alongside superior PSNR on the Nanjing dataset, whereas ICMA exhibits marginal leads in specific pixel-level metrics. This divergence highlights distinct structural paradigms: ICMA leverages the large receptive field of InternImage to optimize local textures under semantic priors, whereas our DSAN establishes explicit bidirectional alignment between global context and local details, offering superior fidelity in global distribution matching and high-frequency boundary coherence. Regarding computational overhead, our model contains 62.2 M parameters—comparable to the 56.8 M of ICMA—with the memory footprint

primarily governed by the dual-generator and multi-scale discriminator topologies. While requiring 103.1 G FLOPs, our network is substantially more compact than ICMA at 297.8 G, achieving a 65.4% reduction. The per-image inference latency of 71.37 ms, yielding approximately 14 FPS, remains well within the operational threshold of at least 10 FPS for remote sensing batch processing. The 802 s per-epoch training cost, totaling approximately 33.4 h over the full 150-epoch schedule, is likewise lower than that of CycleGAN and Pix2PixHD, representing an acceptable one-time offline overhead. This computational cost represents an advantageous trade-off for pronounced perceptual and boundary gains, and latency can be further streamlined via operator fusion and structural pruning.

Despite these advantages, several limitations warrant consideration. First, constrained by GPU memory overhead during multi-scale adversarial training, the model currently operates at a patch resolution of 256×256, leaving full-scene ultra-high-resolution direct inference for future exploration. Second, the dual-stream topology introduces a larger parameter footprint and latency compared to single-branch baselines, posing potential bottlenecks for resource-constrained edge platforms (e.g., onboard satellite payloads); this can be addressed through knowledge distillation, channel pruning, or lightweight attention kernels. Third, although our evaluations cover two distinct spaceborne SAR sensors (RADARSAT-2 and Sentinel-1) and span diverse land-cover categories across strictly disjoint test scenes, translation robustness across extreme terrain (e.g., arid deserts, polar ice sheets, and rugged mountainous relief) requires further validation.

Accordingly, future research will pursue three key avenues: (1) designing lightweight attention backbones to facilitate onboard deployment; (2) integrating progressive super-resolution modules to enable seamless translation of large-scale gigapixel remote sensing swaths; and (3) incorporating heterogeneous multi-sensor datasets to enhance domain generalization across atypical land-cover distributions.

5. Conclusions

To address low contrast and blurry texture representations in synthetic aperture radar (SAR) images, this paper proposes a multi-scale collaborative attention network for SAR-to-optical image translation. Built upon the Pix2PixHD dual-generator baseline, three core innovations are introduced. First, an asymmetric dual-prior input architecture is developed. The global generator ingests histogram-equalized low-resolution SAR imagery to model macroscopic topological configurations, while the local generator utilizes pristine high-resolution SAR data to retain fine-grained textural cues, resolving the trade-off between contrast enhancement and spatial detail preservation. Second, a Dual-Scale Fusion Module (DSFM) coupling Large Kernel Attention (LKA) with Collaborative Attention (CoA) is designed to eliminate scale barriers and achieve bidirectional soft alignment across global semantics and local details. Third, a continuous differentiable soft-edge loss is formulated to overcome the non-differentiability and noise sensitivity of traditional edge operators. By exploiting logarithmic dynamic range compression, this loss mitigates gradient domination from specular highlights, equitably capturing subtle edge structures across shaded areas and enforcing geometric sharpness along complex land-cover boundaries.

Experiments on the Nanjing dataset and the public SEN1-2 dataset validate the superiority and generalization of our approach. Compared with representative baselines including Pix2Pix, CycleGAN, Pix2PixHD, MSTMNet, and the recent ICMA, the proposed DSAN achieves state-of-the-art performance in core perceptual metrics, attaining FID values of 53.121 and 90.811 on the two datasets, respectively, while faithfully reconstructing optical radiometric distributions and structural boundaries. Future work will focus on structural pruning and operator optimization to reduce computational latency, accelerating the real-time engineering deployment of high-fidelity remote sensing translation.

Author Contributions: Conceptualization, Y.K. and D.W.; methodology, D.W.; software, D.W.; validation, D.W. and Y.K.; formal analysis, D.W.; investigation, D.W.; writing—original draft preparation, D.W.; writing—review and editing, D.W. and Y.K.; supervision, Y.K.; funding acquisition, Y.K.; project administration, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 61501228, No. 62171220); the Natural Science Foundation of Jiangsu (No. BK20140825); the Aeronautical Science Foundation of China (No. 20152052029, No. 20182052012); the Fundamental Research Funds for the Central Universities (No. NS2015040, No. NS2021030); the National Science and Technology Major Project (2017-II-0001-0017); and the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education (NJ20240002).

Data Availability Statement: The data are not publicly available due to privacy.

Acknowledgments: During the preparation of this manuscript, the authors used Gemini 3.1 Pro (Google) for the purposes of language polishing and grammar editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Song, J.; Zhang, S.; Kong, X.; Liao, C.; Li, H.; Che, D. Application of a deep learning framework integrating SAR imagery and terrain slope to flood mapping. *Geocarto Int.* **2025**, *40*, 2596248. [CrossRef]
2. Lindsay, E.; Ganerød, A.J.; Devoli, G.; Reiche, J.; Nordal, S.; Frauenfelder, R. Understanding Landslide Expression in SAR Backscatter Data: Global Study and Disaster Response Application. *Remote Sens.* **2025**, *17*, 3313. [CrossRef]
3. Zhou, H.; Yang, J.; Zhang, T.; Dai, A.; Wu, C. EAS-CNN: Automatic Design of Convolutional Neural Network for Remote Sensing Images Semantic Segmentation. *Int. J. Remote Sens.* **2023**, *44*, 3911–3938. [CrossRef]
4. Manoharan, T.; Basha, S.H.; Murugan, J.S.; Suja, G.P.; Rajkumar, R.; Srimathi, S. A Novel Framework for Classifying Remote Sensing Images using Convolutional Neural Networks. In *Proceedings of the 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*; IEEE: New York, NY, USA, 2024; Volume 1, pp. 1–6. [CrossRef]
5. Vasileiou, C.; Smith, J.; Thiagarajan, S.S.; Nigh, M.; Makris, Y.; Torlak, M. Efficient CNN-Based Super Resolution Algorithms for MMWave Mobile Radar Imaging. In *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2022; pp. 3803–3807. [CrossRef]
6. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
7. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784v1. [CrossRef]
8. Gao, J.; Yuan, Q.; Li, J.; Wei, W.; Zhang, H. Cloud Removal with Fusion of High Resolution Optical and SAR Images Using Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 191. [CrossRef]
9. Fang, J.; Ma, X.; Wang, J.; Zhao, L. A Noisy SAR Image Fusion Method Based on NLM and GAN. *Entropy* **2021**, *23*, 410. [CrossRef] [PubMed]
10. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016. [CrossRef]
11. Gong, Y.; Zhan, Z.; Jin, Q.; Li, Y.; Idelbayev, Y.; Liu, X.; Zharkov, A.; Aberman, K.; Tulyakov, S.; Wang, Y.; et al. E^2GAN: Efficient Training of Efficient GANs for Image-to-Image Translation. *arXiv* **2024**, arXiv:2401.06127.
12. Johnson, J.; Alahi, A.; Li, F.-F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 694–711.
13. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 1125–1134.
14. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 8798–8807.
15. Lopes, A.; Nezry, E.; Touzi, R.; Laur, H. Structure Detection and Statistical Adaptive Speckle Filtering in SAR Images. *Int. J. Remote Sens.* **1993**, *14*, 1735–1758. [CrossRef]
16. Lee, J.S.; Wen, J.H.; Ainsworth, T.L.; Chen, K.S.; Chen, A.J. Improved Sigma Filter for Speckle Filtering of SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 202–213. [CrossRef]

17. Fuentes Reyes, M.; Auer, S.; Merkle, N.; Henry, C.; Schmitt, M. SAR-to-Optical Image Translation Based on Conditional Generative Adversarial Networks—Optimization, Opportunities and Limits. *Remote Sens.* **2019**, *11*, 2067. [CrossRef]
18. Fu, S.; Xu, F.; Jin, Y.Q. Reciprocal translation between SAR and optical remote sensing images with cascaded-residual adversarial networks. *Sci. China Inf. Sci.* **2021**, *64*, 122301. [CrossRef]
19. Tan, D.; Liu, Y.; Li, G.; Yao, L.; Sun, S.; He, Y. Serial GANs: A feature-preserving heterogeneous remote sensing image transformation model. *Remote Sens.* **2021**, *13*, 3968. [CrossRef]
20. Yang, X.; Wang, Z.; Zhao, J.; Yang, D. FG-GAN: A fine-grained generative adversarial network for unsupervised SAR-to-optical image translation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5621211. [CrossRef]
21. Ji, G.; Wang, Z.; Zhou, L.; Xia, Y.; Zhong, S.; Gong, S. SAR image colorization using multidomain cycle-consistency generative adversarial network. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 296–300. [CrossRef]
22. Kong, Y.; Liu, S.; Peng, X. Multi-Scale translation method from SAR to optical remote sensing images based on conditional generative adversarial network. *Int. J. Remote Sens.* **2022**, *43*, 2837–2860. [CrossRef]
23. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual Attention Network. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
24. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 289–297.
25. Dubey, V.; Katarya, R. Adaptive histogram equalization based approach for sar image enhancement: A comparative analysis. In *Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*; IEEE: New York, NY, USA, 2021; pp. 878–883.
26. Kowal, B.; Kwaśnicka, H. Impact of Histogram Equalization on the Classification of Retina Lesions from OCT B-Scans. *Electronics* **2024**, *13*, 4996. [CrossRef]
27. He, X.; Kong, L.; Dong, J.; Wang, Y.; Huang, W.; Zhang, H. Dual-stream Network for Visual Recognition. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34. [CrossRef]
28. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 1874–1883. [CrossRef]
29. Amin, Y.; Trivedi, N.S.; Bhattad, R. A Comparative Study of U-Net Architectures for Change Detection in Satellite Images. *arXiv* **2025**, arXiv:2506.07925.
30. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training GANs. In *Proceedings of the Advances in Neural Information Processing Systems*; Association for Computing Machinery: New York, NY, USA, 2016; Volume 29, pp. 2234–2242.
31. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.
32. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 586–595. [CrossRef]
33. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2223–2232. [CrossRef]
34. Xu, C.; Kong, Y. SAR-to-Optical Remote Sensing Image Translation Method Based on InternImage and Cascaded Multi-Head Attention. *Remote Sens.* **2026**, *18*, 55. [CrossRef]
35. Liu, Y.; Shao, Z.; Han, Z.; Liu, J.; Li, X.; Zhou, J. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561. [CrossRef]
36. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 13713–13722. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.