S^{3}OIL: Semi-Supervised SAR-to-Optical Image Translation via Multi-Scale and Cross-Set Matching

Xi Yang, *Senior Member, IEEE*, Haoyuan Shi, Ziyun Li, Maoying Qiao, *Member, IEEE*, Fei Gao, *Member, IEEE*, and Nannan Wang, *Senior Member, IEEE*

***Abstract*—Image-to-image translation has achieved great success, but still faces the significant challenge of limited paired data, particularly in translating *Synthetic Aperture Radar* (SAR) images to optical images. Furthermore, most existing semi-supervised methods place limited emphasis on leveraging the data distribution. To address those challenges, we propose a *Semi-Supervised SAR-to-Optical Image Translation* (S^{3}OIL) method that achieves high-quality image generation using minimal paired data and extensive unpaired data while strategically exploiting the data distribution. To this end, we first introduce a *Cross-Set Alignment Matching* (CAM) mechanism to create local correspondences between the generated results of paired and unpaired data, ensuring cross-set consistency. In addition, for unpaired data, we apply weak and strong perturbations and establish intra-set *Multi-Scale Matching* (MSM) constraints. For paired data, intra-modal semantic consistency (ISC) is presented to ensure alignment with the ground truth. Finally, we propose local and global cross-modal semantic consistency (CSC) to boost structural identity during translation. We conduct extensive experiments on SAR-to-optical datasets and another sketch-to-anime task, demonstrating that S^{3}OIL delivers competitive performance compared to state-of-the-art unsupervised, supervised, and semi-supervised methods, both quantitatively and qualitatively. Ablation studies further reveal that S^{3}OIL can ensure the preservation of both semantic content and structural integrity of the generated images. Our code is available at: https://github.com/XduShi/SOIL**

***Index Terms*—Semi-supervised learning, cross-set alignment matching (CAM), multi-scale matching (MSM), intra-modal semantic consistency (ISC), cross-modal semantic consistency (CSC).**

Received 22 December 2024; revised 7 July 2025 and 16 August 2025; accepted 24 September 2025. Date of publication 7 October 2025; date of current version 13 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62372348, Grant 62571395, Grant U22A2096, and Grant 62036007; in part by the Key Research and Development Program of Shaanxi under Grant 2024GXZDCYL-02-10; in part by Shaanxi Outstanding Youth Science Fund Project under Grant 2023-JC-JQ-53; in part by the Scientific and Technological Innovation Teams in Shaanxi Province under Grant 2025RS-CXTD-011; in part by Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11; and in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23042. The associate editor coordinating the review of this article and approving it for publication was Prof. Liqiang Nie. *(Corresponding author: Fei Gao.)*

Xi Yang and Nannan Wang are with Xidian University, Xi'an 710071, China (e-mail: yangx@xidian.edu.cn; nnwang@xidian.edu.cn).

Haoyuan Shi and Fei Gao are with Hangzhou Institute of Technology, Xidian University, Hangzhou 311231, China (e-mail: xdshy@stu.xidian.edu.cn; fgao@xidian.edu.cn).

Ziyun Li is with KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: ziyli@kth.se).

Maoying Qiao is with the University of Technology, Sydney, NSW 2007, Australia (e-mail: maoying.qiao@uts.edu.au).

Digital Object Identifier 10.1109/TIP.2025.3616576

I. Introduction

SYNTHETIC Aperture Radar (SAR) and optical sensing are two important means of earth observation. SAR can be used for all-day and all-weather earth observation, but it has the disadvantages of speckle noise and geometric distortion, which are not conducive to human eye recognition. These differences complicate the task of achieving high-fidelity translations. Given the unique challenges posed by SAR and optical sensing in earth observation, applying *image-to-image translation* (I2I) techniques offers a promising approach to bridge the gap between the two modalities.

In the domain of I2I, supervised learning methods have traditionally relied on large-scale paired datasets to establish precise mappings between domains. Methods like conditional generative models [1] and regression-based techniques [2] have demonstrated strong performance in controlled scenarios by leveraging such datasets. However, constructing paired datasets, particularly for SAR and optical data, is both expensive and time-intensive due to the challenges of simultaneous data acquisition.

To address the limitations of supervised approaches, unsupervised and semi-supervised methods have emerged, reducing reliance on paired datasets. Unsupervised techniques, such as cycle consistency [3], enforce reconstruction fidelity across domains but often struggle to preserve structural and semantic details in complex settings like urban landscapes. Advanced approaches, including feature-guided SAR-to-optical translation [4] and thermodynamics-inspired networks [5], improve image quality but may introduce artifacts or compromise local details, particularly in high-frequency regions. Moreover, due to the absence of explicit supervision, these methods often suffer from a distribution mismatch between the source and target domains, especially in cross-modal settings where texture, geometry, and semantics differ significantly.

Semi-supervised methods combine labeled and unlabeled data to enhance generalization and reduce the dependence on paired datasets. Techniques like FixMatch [6], CorrMatch [7], and graph-based methods [8], [9] improve label efficiency and propagate information effectively. However, both unsupervised and semi-supervised methods face challenges in aligning data distributions, as mismatched paired and unpaired distributions can lead to semantic inconsistencies, structural discrepancies, and performance degradation in generative models [10]. This issue becomes more pronounced in remote sensing scenarios

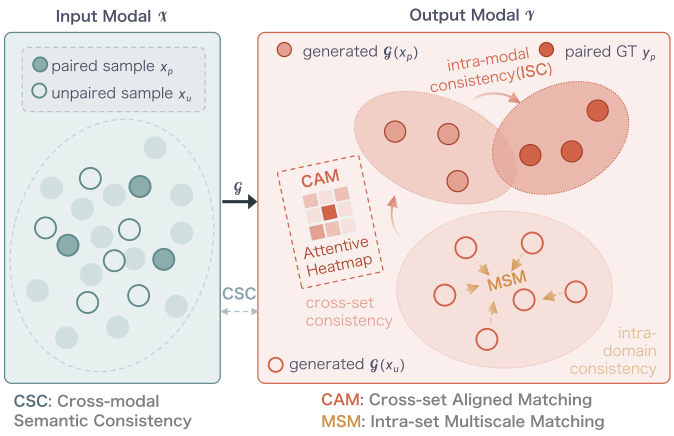


Fig. 1. Motivation of the proposed semi-supervised image-to-image translation (S^{3}OIL) method. Aiming to minimize the distribution distance between paired and unpaired data, we first incorporate Cross-Set Aligned Matching (CAM) to achieve cross-set consistency. Furthermore, to enhance the distribution density of the unpaired data, we introduce the Intra-Set Multi-Scale Matching (MSM) mechanism. Cross-Modal Semantic Consistency (CSC) ensures that the generated content remains consistent with the original semantics across all levels. Additionally, Intra-Modal Semantic Consistency (ISC) ensures that paired data aligns closely with the ground truth within the same modality, contrasting with the intra-domain alignment addressed by MSM, which enhances consistency within unpaired data.

involving SAR and optical modalities, where the visual gap between modalities results in fundamentally different feature spaces. Without strong mechanisms to explicitly bridge these gaps, existing methods risk misalignment between the generative outputs and target semantics, limiting translation fidelity.

Therefore, we address the limitations of both supervised and unsupervised methods by introducing a ***S**emi-**S**upervised **SAR**-to-**O**ptical **I**mage Trans**l**ation* (S^{3}OIL) method. Unlike previous approaches, S^{3}OIL prioritizes the alignment of data distributions across domains, enabling high-quality image translation with minimal paired data. We conceptualize the target domain's data distribution as a low-dimensional manifold embedded in a high-dimensional space, making feature extraction and distribution alignment essential. In this work, our primary focus is on aligning feature distributions to ensure consistency in structure and semantics.

Specifically, S^{3}OIL incorporates four key constraints, as illustrated in Fig. 1. (i) *Cross-Set Aligned Matching* (CAM) aligns the distributions between paired and unpaired data within the target domain. Since paired data inherently offers more accurate and detailed information, aligning these distributions is crucial for preserving the images' structural integrity and content fidelity, ensuring that the generated outputs from the unpaired data remain consistent and reliable. (ii) *Intra-Set Multi-Scale Matching* (MSM) addresses unpaired data using multi-scale matching combined with contrastive learning. By introducing both weak and strong perturbations, MSM ensures the preservation of detailed structures and broad context, yielding a more consistent and compact distribution. (iii) *Intra-Modal Semantic Consistency* (ISC) involves using a Huber loss [11] between the generated optical image and the ground truth, ensuring that paired data aligns with the ground truth and indirectly pulls the unpaired distribution closer to it in the global semantic context. (iv) *Cross-Modal Semantic Consistency* (CSC) is applied to both paired and unpaired data to achieve local and global semantic alignment, ensuring that the generated content remains consistent with the original semantics across all levels.

Our contributions are summarized as follows:

- We propose S^{3}OIL, a semi-supervised image-to-image translation framework that excels in generating high-quality images with minimal paired data and extensive use of unpaired data. Our method creates strong connections between paired and unpaired datasets, ensuring both local and global semantic consistency. By aligning features and preserving meaning at every level, it effectively bridges the gap between limited supervision, resulting in high-quality image translations
- We introduce four data distribution consistency constraints, *Cross-Set Aligned Matching* (CAM), *Intra-Set Multi-Scale Matching* (MSM), *Intra-Modal Semantic Consistency* (ISC), and *Cross-Modal Semantic Consistency* (CSC), from a manifold perspective to ensure the preservation of both semantic content and structural integrity.

The remainder of this paper is organized as follows. Section II reviews related works, analyzing the achievements and limitations in image translation to identify the gaps that S^{3}OIL aims to fill. Building on these insights, Section III introduces the S^{3}OIL framework, providing a detailed explanation of its architecture and the mechanism of its modules. Section IV presents experimental results, using comparisons and ablation studies to prove S^{3}OIL's superiority. Finally, Section V concludes with a summary of the contributions.

II. Related Works

A. Image-to-Image Translation

Recent advancements in image-to-image translation have focused on supervised, unsupervised, and semi-supervised learning methods. CycleGAN-based approaches have shown strong performance in different semantic areas but fall short in gray value consistency [12], [13]. However, supervised methods require large collections of aligned SAR-optical images, which involve substantial engineering effort.

Unsupervised learning methods have also been explored, which do not need paired images. These methods include CycleGAN-based architectures to reduce color distortion [14] and contrastive unpaired translation models aimed at breaking the cycle consistency constraint [15], [16]. Despite these improvements, some challenges like incorrect gray values persist. Refinement approaches that modify predicted target domain images step by step have been proposed to enhance fine detail preservation while maintaining semantic information [17]. Further enhancements in multi-modal registration are achieved through exemplar-based I2I modules for style consistency with transformer-based networks for accurate deformation prediction [18], which ensure style consistency across modalities. Related advances in multimodal fusion, such

as dynamic multimodal fusion via meta-learning for micro-video recommendation [19], demonstrate the effectiveness of adaptively integrating heterogeneous modalities to mitigate distribution gaps [20], [21]. Further enhancements in cross-modal translation have been achieved by exemplar-based image-to-image modules combined with contrastive learning in transformer or diffusion-based architectures [22], which employs prior-guided diffusion with global-local contrastive objectives to ensure structural fidelity and style consistency.

Semi-supervised methods aim to leverage both labeled and unlabeled data to improve model generalization while reducing the dependency on fully labeled datasets. Various approaches have been explored to achieve this goal. Structured generative modeling techniques, such as those proposed in [23], utilize generative frameworks to integrate unlabeled data effectively. Regularization-based methods, including those in [24], focus on enhancing consistency to improve learning stability. In addition to these, graph-based methods have been explored for propagating label information across graph structures, as shown in [8] and [9]. Despite their strengths, these methods often overlook the critical challenge of aligning data distributions effectively. For generative models, a mismatch between the generated and actual data distributions can lead to significant performance degradation, as discussed in [10]. This highlights the need for strategies that address distribution alignment to further improve the robustness and efficacy of semi-supervised learning frameworks.

B. SAR-to-Optical Image Translation

SAR-to-optical image translation is a critical task in remote sensing, with various GAN-based models like cGAN, Pix2Pix, and CycleGAN being commonly used. The release of the SEN1-2 dataset [25] greatly advanced research in this field, particularly by enabling the use of Pix2Pix for SAR-to-optical image translation. However, generating fine details and ensuring accurate color fidelity remain significant challenges [12]. Several optimization strategies, such as parallel feature fusion and multi-scale discriminators, have been proposed to address these issues and improve texture and contour details [26], [27]. More recently, thermodynamics-inspired networks, such as S2O-TDN [5], have introduced pixel-molecule analogies for feature extraction and diffusion regularization to improve the quality of the generated images.

Despite these advancements, most existing methods rely on paired samples, which limits their scalability. Unsupervised methods, such as CycleGAN and combined loss functions, have shown promise in addressing these limitations and have the potential to improve SAR-to-optical translation with minimal supervision [27], [28], [29]. In view of the limitations in the translation process from SAR-to-optical images, Reyes et al. [12] proposed several optimization strategies based on CGAN. In 2020, Turnes et al. [30] introduced a CGAN network architecture utilizing atrous convolution. The proposed generator and discriminator incorporated an Atrous Spatial Pyramid Pooling (ASPP) module, leveraging spatial background to enhance the fineness of generated images at multiple scales.

To address challenges in complex optical scenes and high-frequency speckle noise in SAR images, Zhang et al. [4] proposed a feature-guided SAR-to-optical translation method that extracts multi-layer features and employs a discrete cosine transform-based loss function to reduce noise, significantly enhancing image quality. In the same year, Hwanget al. [27] proposed a strongly constrained genetic algorithm built upon the structural similarity index measure (SSIM) [31] and the L1 norm within the CGAN framework. This method established constraints between generated and target images in the structural information space, resulting in the generation of images with improved architectural detail.

C. Semi-Supervised Learning via Weak-Strong Matching

Semi-supervised learning methods offer a balance between supervised and unsupervised methods by training on both aligned and unaligned images. FixMatch simplifies semi-supervised learning by focusing on consistency and confidence [6]. Transformation consistency regularization ensures invariant predictions for unlabeled data [32], and segmentation-guided frameworks enhance semi-supervised translation [33]. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation has shown significant improvements [34]. CorrMatch leverages label propagation via correlation matching for semi-supervised semantic segmentation, further advancing the field [7], revealing the potential of the matching consistency for image translation tasks.

In the context of SAR-to-optical (S2O) image translation, recent semi-supervised methods have attempted to overcome the challenge of limited paired data by using weakly labeled or unpaired images [35]. However, these methods still struggle with maintaining fidelity and semantic consistency due to the core issue of data distribution misalignment. When paired and unpaired data are misaligned, it disrupts the model's ability to preserve structural integrity and semantic accuracy, limiting the effectiveness of these approaches. Therefore, addressing this alignment gap is crucial for improving SAR-to-optical image translation.

III. Method

A. Overview

Building upon the insights from prior research, which emphasize the critical roles of feature alignment, consistency regularization, and semantic preservation in image-to-image translation, we propose a novel approach that effectively leverages minimal paired data while achieving improved alignment of data distributions. In S^{3}OIL, illustrated in Fig. 2, paired $(\mathcal{X}_p, \mathcal{Y}_p)$ and unpaired $(\mathcal{X}_u)$ data are processed with four key consistency constraints: CAM aligns semantic and structural information across paired and unpaired data; MSM applies multi-scale matching to improve generalization for unpaired data by comparing weakly and strongly perturbed inputs, denoted as $\mathcal{P}_x = \{x_u^w, x_u^s\}$; ISC ensures that generated optical images match the ground truth on the paired data; and CSC maintains semantic integrity during image-to-image translation for both paired and unpaired samples. $\mathcal{G}$ denotes the forward generator that maps SAR images to optical images, while $\mathcal{F}$

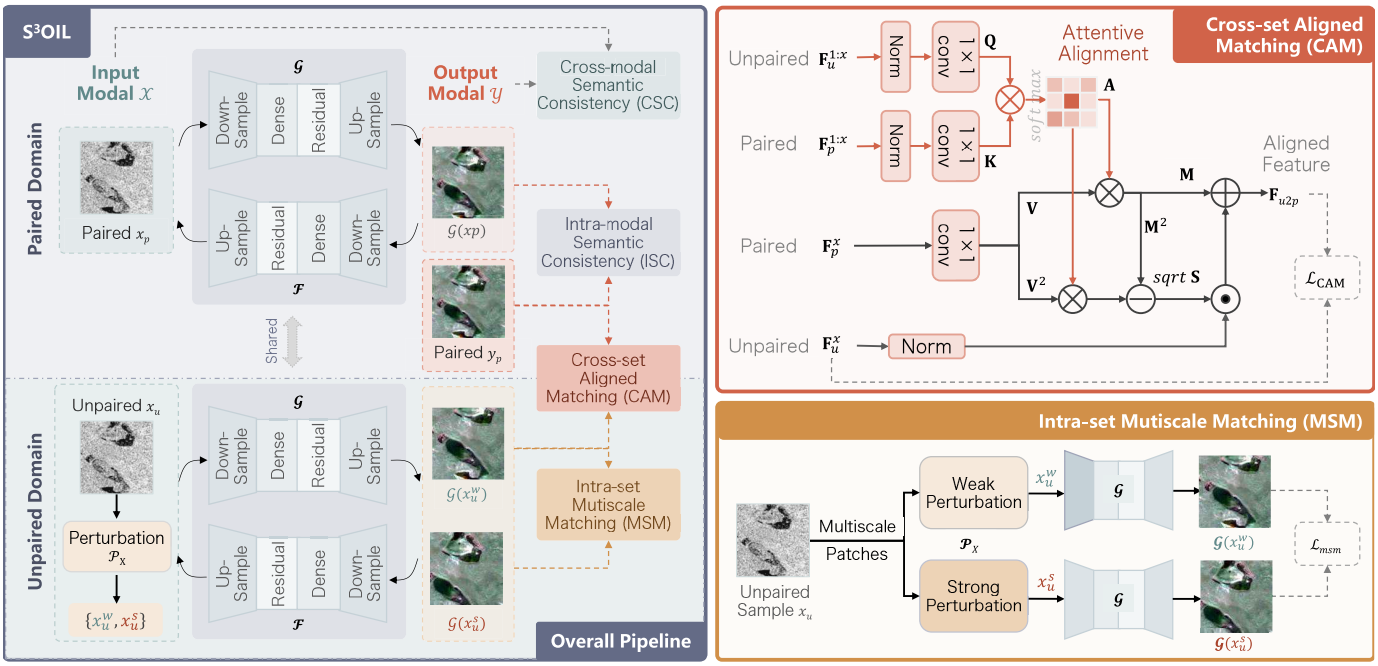


Fig. 2. Overview of the proposed ***S**emi-**S**upervised **S**AR-to-**O**ptical **I**mage Trans**l**ation* (S^{3}OIL) method. In S^{3}OIL, we use both the paired set $\{\mathcal{X}_p, \mathcal{Y}_p\}$ and the unpaired set $\{\mathcal{X}_u\}$ during training. First, we propose the *cross-set aligned matching* (CAM) to constrain the consistency between the paired set and the unpaired set. Second, we propose to constrain the *intra-set consistency* on unpaired samples through *multi-scale matching* (MSM). In addition, we constrain the *cross-modal semantic consistency* (CSC) between the input and the output, and the *intra-modal semantic consistency* (ISC) on the paired sample.

represents the inverse generator that supports the backward process for cyclic reconstruction and semantic consistency evaluation.

B. Cross-Set Aligned Matching (CAM)

To align features across paired and unpaired domains, we introduce the *Cross-Set Aligned Matching* (CAM) mechanism. As shown in Fig. 2, our model, inspired by AdaAttN [36], uses ResNet50 as a feature extractor. We integrate features from the `ReLU3_1`, `ReLU4_1`, and `ReLU5_1` layers of ResNet, denoted as $\mathbf{F}_x^* \in \mathbb{R}^{C\times H\times W}$, representing either unpaired (x_u) or paired (x_p) data:

$$\mathbf{F}_*^{1:i} = \mathbf{D}_x(\mathbf{F}_*^1) \oplus \mathbf{D}_x(\mathbf{F}_*^2) \oplus \cdots \oplus \mathbf{F}_*^i, \tag{1}$$

where $\mathbf{D}_x$ is the bilinear interpolation layer that downsamples the input to match $\mathbf{F}_x^*$, and $\oplus$ denotes concatenation along the channel dimension. $\mathbf{F}_x^*$ denotes the feature representation extracted from a specific of the ResNet backbone for a given sample x (either x_p or x_u). In contrast, $\mathbf{F}_*^{1:i}$ represents the multi-level concatenation of downsampled feature maps from layers 1 through i, used to construct hierarchical attention maps. This distinction enables our model to capture semantic correspondences at varying levels of abstraction during the alignment process. Our approach introduces an enhanced attention mechanism, which incorporates both low-level and high-level layers of paired and unpaired features, in contrast to traditional methods that predominantly rely on deeper features. This allows for a more comprehensive measurement of similarity across different domains. To compute the attention map $\mathbf{A}$ of layer i, we define the query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) as follows:

$$\mathbf{Q} = f\left(\text{Norm}(\mathbf{F}_u^{1:i})\right), \tag{2}$$

$$\mathbf{K} = g\left(\text{Norm}(\mathbf{F}_p^{1:i})\right), \tag{3}$$

$$\mathbf{V} = h\left(\mathbf{F}_p^{i}\right), \tag{4}$$

where f, g, and h are 1×1 learnable convolutional layers, and Norm as channel-wise mean-variance normalization, as used in instance normalization. We choose 1×1 convolutions here for their role as efficient channel-wise projection operators that preserve spatial resolution. This allows us to reduce dimensionality and introduce learnable transformations within each spatial location, enabling the attention mechanism to focus on semantic relationships without introducing additional spatial context. Such a design is commonly employed in attention-based architectures (e.g., AdaAttN, Non-Local Networks) and is particularly suitable in our semi-supervised setting, where computational efficiency and limited supervision must be balanced with representational capacity. Here, $\mathbf{F}_u^{1:i}$ and $\mathbf{F}_p^{1:i}$ represent the unpaired and paired features up to layer i, respectively. $\mathbf{F}_p^i$ denotes the feature map of the paired sample extracted specifically from the i-th layer of the ResNet backbone. And then the attention map $\mathbf{A}$ can be calculated as:

$$\mathbf{A} = \text{Softmax}(\mathbf{Q}^\top \otimes \mathbf{K}). \tag{5}$$

The attention-weighted mean $\mathbf{M}$ is then derived as:

$$\mathbf{M} = \mathbf{V} \otimes \mathbf{A}^\top. \tag{6}$$

To evaluate the efficacy of feature alignment, we achieve our alignment strategy by integrating an attention mechanism.

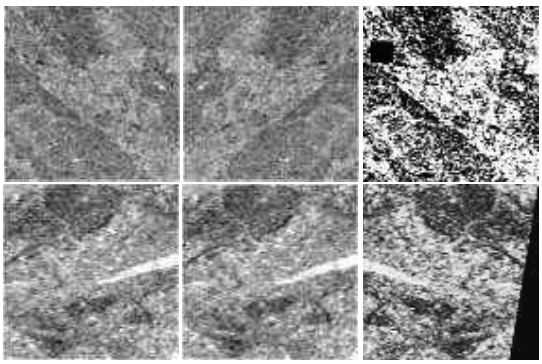


Fig. 3. Illustration of input image perturbations. (a) Input image, (b) weak perturbation of input image, and (c) strong perturbation of input image.

As shown in Fig. 2, the attention-weighted standard deviation is denoted as $\mathbf{S} = \sqrt{(\mathbf{V}^2) \otimes \mathbf{A}^\top - \mathbf{M}^2}$. where $\mathbf{V}^2$ and $\mathbf{M}^2$ denote the element-wise squares of the matrices $\mathbf{V}$ and $\mathbf{M}$, respectively. The scale $\mathbf{S}$ and shift $\mathbf{M}$ are applied to generate the aligned feature map:

$$\mathbf{F}^i_{align} = \mathbf{S} \cdot \text{Norm}(\mathbf{F}^i_u) + \mathbf{M}. \tag{7}$$

To ensure cross-set consistency, we use the $\mathcal{L}_1$ loss:

$$\mathcal{L}_{cam} = \|\mathbf{F}_u - \mathbf{F}_{align}\|_1. \tag{8}$$

This measures the absolute differences between $\mathbf{F}_{align}$ and $\mathbf{F}_u$, promoting cross-set alignment and high-quality image-to-image generation.

C. Intra-Set Multi-Scale Matching (MSM)

In the context of unpaired domains, utilizing data effectively poses significant challenges. To address these, we leverage contrastive learning's concept [16], [37], [38] by introducing both weak and strong perturbations while ensuring consistency between them, as depicted in Fig. 3.

Weak perturbations, such as flip-and-shift, preserve the core structure, allowing the model to refine its understanding of fine details. For instance, we apply random horizontal flipping with a 50% probability on all datasets, to introduce minimal yet effective diversity. Additionally, we perform random translations of the images, shifting them vertically and horizontally.

Strong perturbations, derived from AutoAugment [39] and Cutout [40], introduce more substantial changes, challenging the model to maintain semantic alignment under varying conditions. AutoAugment applies a predefined or learned policy of transformations, including operations such as rotations, brightness adjustments, and color distortions, to create substantially altered versions of the input data. These augmentations simulate a wide range of real-world variabilities, ensuring the model can generalize beyond the training dataset's specific conditions. By combining multiple augmentation strategies in a single policy, AutoAugment introduces rich variations while retaining the core semantics of the original data. Cutout, on the other hand, masks out random regions of the input image by overlaying fixed-size black or gray rectangles. This encourages the model to infer missing information by leveraging the surrounding context, thereby improving its ability to recognize and process incomplete or occluded patterns. By masking different regions in each training iteration, Cutout introduces an additional layer of robustness, preparing the model for scenarios where critical features may be obscured.

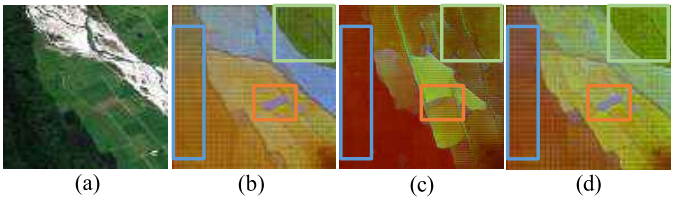


Fig. 4. Semantic visualization of intra-modal semantic consistency. (a) optical image $\mathcal{Y}_p$ from paired dataset, (b) semantic segmentation of $\mathcal{Y}_p$, (c) semantic segmentation of the image generated by methods lacking semantic consistency, and (d) semantic segmentation of the image generated using the proposed semantic consistency modules. The blue, orange and green boxes highlight distinct semantic regions, underscoring the effectiveness of intra-modal semantic consistency in preserving semantic coherence and consistency.

This matching process between different perturbation levels enhances the model's robustness and generalization capabilities. Additionally, to retain both detailed structures and broader information from the unpaired data, we use a multiscale strategy, extracting patches of different scales (16×16, 32×32, and 64×64) from the input images. Based on these two aspects, the multiscale loss $\mathcal{L}_{msm}$ from the unpaired data is designed as:

$$\mathcal{L}_{msm} = \sum_{i=1}^{N} \|\mathcal{G}(x^w_{u^{(i)}}) - \mathcal{G}(x^s_{u^{(i)}})\|_1, \tag{9}$$

where N represents the number of scales, $\mathcal{G}$ denotes the generated model, and $_1$ denotes the $\mathcal{L}_1$ loss, i.e., the sum of the absolute differences between the corresponding pixels of two images.

D. Intra-Modal Semantic Consistency (ISC)

To achieve semantic fidelity within the same modality, we aim to enforce consistency between the generated images and the ground truth. Traditional methods that rely on L1 loss have shown significant limitations, both in terms of quantitative metrics and visual quality, often failing to capture fine-grained details or maintain semantic coherence [3]. As illustrated in Fig. 4 (c), these methods struggle to preserve semantic integrity, resulting in noticeable distortions or structural loss in the generated images. Therefore, a more robust alternative is necessary to ensure better global semantic consistency and higher-quality generation.

For paired data, we apply an intra-modal semantic consistency constraint using a Huber loss [11] between the generated optical image and the ground truth, formulated as follows:

$$\mathcal{L}_{isc} = \begin{cases} \frac{1}{2}(y_p - \mathcal{G}(x_p))^2 & |y_p - \mathcal{G}(x_p)| \le \delta, \\ \delta|y_p - \mathcal{G}(x_p)| - \frac{1}{2}\delta^2 & \text{otherwise}, \end{cases} \tag{10}$$

where y_p is the ground-truth image and δ is the threshold parameter. The Huber loss operates in two distinct regions, each designed to handle specific error magnitudes effectively.

In the quadratic region ($|y_p - \mathcal{G}(x_p)| \leq \delta$), the loss behaves as a squared error, emphasizing precise alignment between the generated and ground-truth images. This region excels at capturing fine-grained variations, thereby preserving intricate local details such as textures and edges. Conversely, in the linear region ($|y_p - \mathcal{G}(x_p)| > \delta$), the loss transitions to a linear form, mitigating the influence of large deviations. This design reduces the risk of destabilizing training or skewing optimization, aligning with physical systems where extreme errors are moderated to ensure stability. Together, as shown in Fig. 4 (d), these regions enable the Huber loss to balance precision and robustness, making it effective for maintaining semantic consistency and enhancing image generation quality.

E. Cross-Modal Semantic Consistency (CSC)

While we have addressed semantic inconsistency within a single modality, establishing a robust semantic connection between two modalities within the paired domain remains a critical issue. Inspired by the feature matching loss employed in Pix2PixHD [41], our approach seeks to enforce both local and global semantic consistency across modalities. By leveraging CSC, we ensure local and global semantic consistency for paired data. Locally, we segment each category in both original and generated images by fine-tuning SAM model, applying a cross-entropy loss to measure and align the semantic discrepancies. The local cross-modal semantic consistency loss $\mathcal{L}_{local\_csc}$ is defined as:

$$\mathcal{L}_{local\_csc} = -\sum_{i=1}^{K}\sum_{j=1}^{C} \mathbf{P}(x_i^j) \log \mathbf{P}(\mathcal{G}(x_i)^j), \tag{11}$$

where K is the number of segments in the original image, C is the number of semantic categories, x_i^j is the original segment for category j, $\mathcal{G}(x_i)$ is the generated segment, and $P(x_i^j)$ is the predicted probability of x_i belonging to category j, computed using the CLIP model [42] with softmax.

For global semantic consistency, we leverage a CLIP loss [43] to capture the overall semantic information. The global cross-modal semantic consistency loss $\mathcal{L}_{global\_csc}$ is formulated as follows:

$$\mathcal{L}_{\text{CLIP}} = \|\text{CLIP}(x) - \text{CLIP}(\mathcal{G}(x))\|. \tag{12}$$

By integrating local segmentation semantic loss and global image semantic loss, we align the semantics between generated and original segmentations (i.e, $\mathcal{L}_{local\_csc}$) and ensure that the generated image maintains the semantic integrity of the original content (i.e., $\mathcal{L}_{\text{global\_csc}}$). The overall cross-modal semantic consistency $\mathcal{L}_{csc}$ is denoted as:

$$\mathcal{L}_{csc} = \mathcal{L}_{local\_csc} + \mathcal{L}_{global\_csc}. \tag{13}$$

F. Overall Loss

We also apply the adversarial loss by adding discriminators to encourage generated images to belong to their respective domains [44], denoted as $\mathcal{L}_{\text{GAN}}$. Finally, the overall loss $\mathcal{L}_{overall}$ in our framework is defined as:

$$\begin{aligned}\mathcal{L}_{overall} = \lambda_{cam}\mathcal{L}_{cam} + \lambda_{\text{msm}}\mathcal{L}_{\text{msm}} + \lambda_{isc}\mathcal{L}_{isc} \\ + \lambda_{csc}\mathcal{L}_{csc} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}},\end{aligned} \tag{14}$$

where we set $\lambda_{cam} = 50$, $\lambda_{csc} = 10$, $\lambda_{isc} = 10$, $\lambda_{msm} = 50$, $\lambda_{\text{GAN}} = 50$ from experimental experience. Additionally, we utilize the unbalanced generator with DenseNet blocks in the middle [14] and incorporate a patch-based discriminator [41] in practice.

IV. Experiments

We conduct a comprehensive evaluation of our proposed framework by providing both qualitative and quantitative comparisons using classic SAR-to-optical datasets. Additionally, we extend our analysis by including comparisons on a sketch-to-anime dataset to demonstrate the robustness and generalizability of our model.

A. Settings

1) Datasets: We utilize two SAR-to-optical datasets for both training and evaluation: SEN12MS and SEN12. The SEN12MS dataset [52] contains 180,662 patch triplets, each consisting of corresponding Sentinel-1 dual-polarization SAR data, Sentinel-2 multispectral images, and MODIS-derived land cover maps. These patches are distributed globally, covering all four meteorological seasons. The Sentinel-1 data in the dataset includes both VV-polarized and VH-polarized images. For our approach, we specifically use the VV polarization channel from Sentinel-1 as the input SAR data. While the Sentinel-2 multispectral data includes 13 bands, we select only the three RGB channels for generating the required optical images. The SEN12 dataset [25] consists of 282,384 SAR-optical patch pairs collected from Sentinel-1 and Sentinel-2. These patches are gathered from diverse locations worldwide, spanning all four seasons. The Sentinel-1 images include VV-polarized data, while the Sentinel-2 images comprise three bands: red, green, and blue.

2) Implementation Details: Our experiments are conducted using a single RTX 4090 GPU with 24GB of memory, and the implementation is done using PyTorch. We train our models for 200 epochs on different datasets, utilizing the Adam optimizer with a batch size of 4. The initial learning rate is set to 0.0002 and gradually reduced using a learning rate scheduler. We set $\lambda_{cam} = 50$, $\lambda_{csc} = 10$, $\lambda_{isc} = 10$, $\lambda_{msm} = 50$, $\lambda_{\text{GAN}} = 50$ based on experimental experience.

In our model architecture, we introduce a generator architecture known as the Unbalanced Generator (UBG), initially proposed by FG-GAN [14]. This architecture is particularly suited for image translation tasks involving modality gaps (e.g., SAR-to-optical), as it features asymmetric encoder-decoder pathways that better accommodate the structural and statistical differences between source and target domains. The encoder in UBG utilizes several convolutional layers with residual blocks to capture fine-grained spatial features, while the decoder progressively upsamples these features using transposed convolutions to reconstruct high-resolution optical images. To enhance the model's performance, we incorporate discriminators in both the supervised and unsupervised branches of the model. In the supervised training, we employ a PatchGAN discriminator [1] that effectively handles local structural information, ensuring the accuracy of generated

TABLE I

QUANTITATIVE COMPARISON WITH SOTA I2IT METHODS IN THE SAR-TO-OPTICAL IMAGE TRANSLATION TASK, ON THE SEN12MS AND SEN12 DATASETS. † INDICATES THE SUPERVISED VARIANT OF OUR MODEL

	SEN12MS				SEN12			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
fully-supervised learning: 6000 paired samples								
Pix2Pix [1]	12.45	0.1545	0.606	145.5	15.84	0.2568	0.673	139.1
CDiffusion [45]	11.60	0.1228	0.637	103.4	16.23	0.5259	0.352	99.0
BBDM [46]	12.35	0.1244	0.618	104.5	15.15	0.3655	0.493	130.6
S^{3}OIL† (Ours)	12.53	0.1438	0.623	127.5	16.21	0.2188	0.655	103.6
unsupervised learning: : 6000 unpaired samples								
FG-GAN [14]	11.22	0.1311	0.602	105.1	11.34	0.1405	0.661	97.4
U-GAT-IT [47]	10.94	0.1220	0.613	106.6	10.71	0.1312	0.661	110.6
DivCo [48]	10.84	0.1183	0.605	183.5	11.05	0.1931	0.665	198.8
CycleGAN [3]	10.82	0.1228	0.609	150.8	11.18	0.1402	0.588	101.8
StegoGAN [49]	11.72	0.1595	0.593	145.7	11.29	0.1189	0.595	131.3
CCM [50]	8.87	0.1587	0.603	195.3	13.06	0.2753	0.563	237.7
semi-supervised learning: 600 paired samples and 5400 unpaired samples								
WS-I2I [51]	10.85	0.1174	0.604	137.8	11.60	0.1539	0.704	374.8
Scenimefy [33]	11.30	0.1263	**0.591**	165.7	10.38	0.1165	**0.605**	177.3
Semi-I2I [35]	12.22	0.1780	0.593	118.1	12.98	**0.1978**	0.651	122.0
S^{3}OIL (Ours)	**12.51**	**0.1862**	**0.591**	**105.6**	**13.10**	0.1812	0.649	**116.8**

image details. In the unsupervised branch, we use multi-scale discriminators [14], [41] to improve the consistency and quality of generated images across multiple scales.

Then we build upon the foundational structure of Semi-I2I [35], modifying its generator to integrate the UBG and utilizing it as our base model. The generator's encoder is initialized with pre-trained ResNet-50 layers for improved feature extraction, while the decoder consists of multiple deconvolutional layers with skip connections to retain high-resolution spatial cues. For optimization, we use the Adam optimizer due to its good convergence properties and stability. During training, we monitor validation loss to avoid overfitting, and apply gradient clipping (max norm = 5) to prevent gradient explosion in early training. During training, we employ a learning rate scheduler to dynamically adjust the learning rate, improving training efficiency and model performance. Specifically, the initial learning rate is set to 0.0002 and gradually reduced during training to allow finer weight adjustments in the later stages.

3) Evaluation Metrics: We conduct evaluations using both reference and no-reference indicators. *Peak Signal-to-Noise Ratio* (PSNR), *Structural Similarity Index* (SSIM), and *Frechet Inception Distance* (FID) [53] are utilized as reference evaluation metrics, while *Learned Perceptual Image Patch Similarity* (LPIPS) [54] serve as no-reference evaluation metrics. Using these metrics together provides a comprehensive evaluation of image quality, covering pixel accuracy, perceptual similarity, distribution alignment, and visual quality.

B. Comparison With the State-of-the-Art Methods

Baselines: We conduct a comprehensive comparison of our proposed method against state-of-the-art approaches, including supervised, unsupervised, and semi-supervised techniques. Specifically, for supervised methods, we evaluate against Pix2Pix [1] CDiffusion [45] and BBDM [46]; for unsupervised methods, we consider FG-GAN [14], U-GAT-IT [47], DivCo [48], CycleGAN [3], StegoGAN [49] and CCM [50]; and for semi-supervised methods, we compare with WS-I2I [51], Scenimefy [33], and Semi-I2I [35]. For supervised methods, we utilize them primarily as performance benchmarks, and they are excluded from qualitative and quantitative comparisons.

1) Quantitative Comparison: In Table I, we compare our model with SOTA I2IT methods in the SAR-to-optical image translation task. The **bold** values indicate the best-performing results, while the underlined values denote the second-best results for each metric.

On the SEN12MS dataset, S^{3}OIL achieves the highest PSNR of 12.51 and the best SSIM score of 0.1862 among semi-supervised learning methods. For comparison, other semi-supervised methods like Semi-I2I deliver slightly lower PSNR (12.22) and SSIM (0.1780), while unsupervised methods such as FG-GAN and CycleGAN achieve even lower SSIM values (0.1311 and 0.1228, respectively). Despite its focus on conditional generation, CCM appears ineffective in preserving either structural or semantic content when no supervision is available. S^{3}OIL's FID score is only second to FG-GAN, indicating that S^{3}OIL effectively balances image quality and perceptual realism. This can be attributed to S^{3}OIL's emphasis on leveraging both paired and unpaired data through mechanisms such as Cross-Set Aligned Matching (CAM) and Intra-Set Multi-Scale Matching (MSM). These components prioritize robust generalization and semantic consistency across datasets, ensuring the preservation of structural

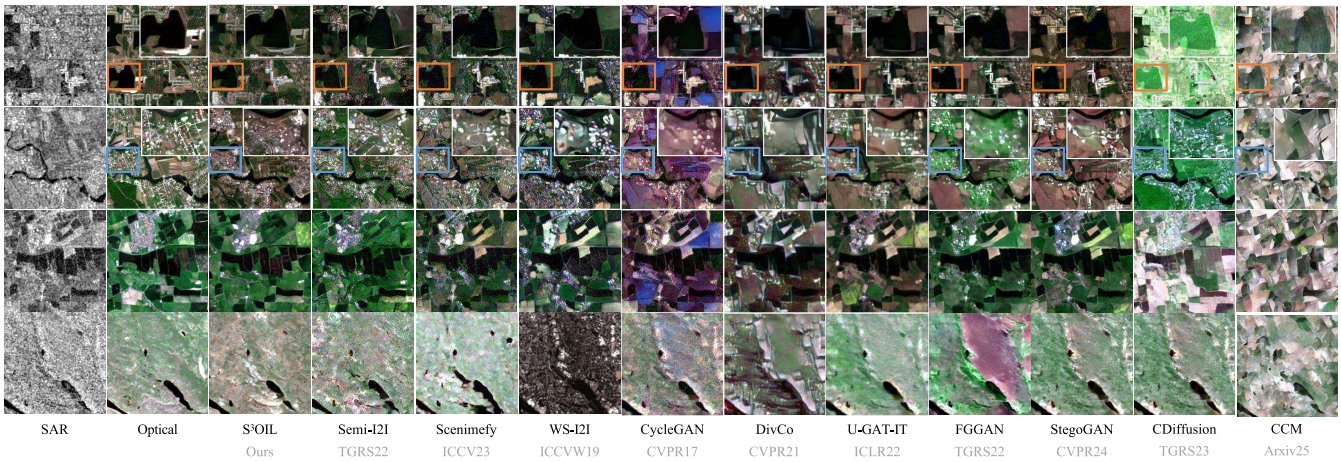


Fig. 5. Qualitative comparison with SOTA method on SEN12MS dataset.

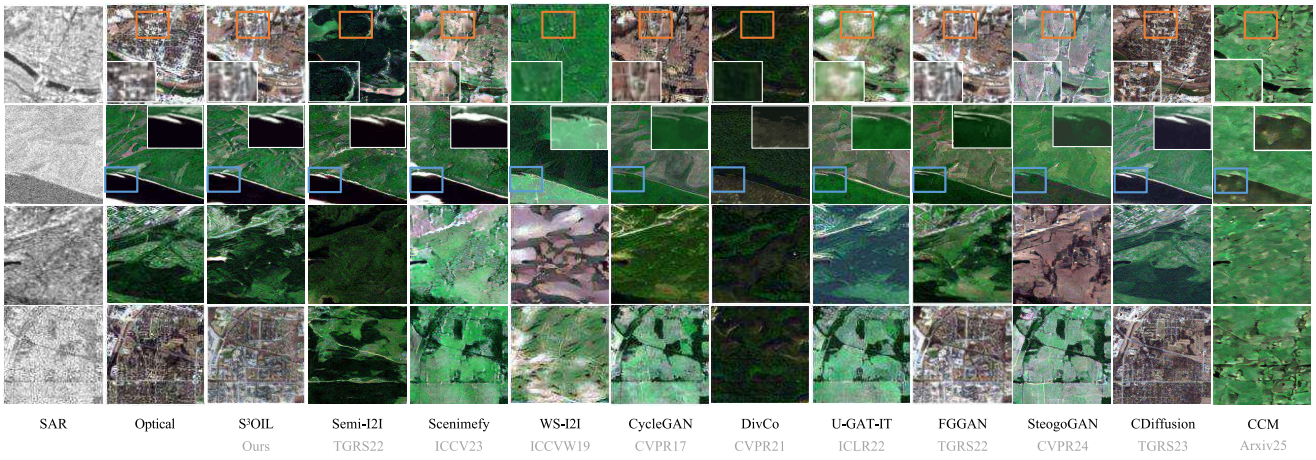


Fig. 6. Qualitative comparison with SOTA method on SEN12 dataset.

integrity and content alignment. However, this focus may inadvertently lead to slight compromises in fine-grained image realism, which is the primary aspect measured by the FID metric. All the metrics imply that S^{3}OIL more accurately reconstructs the key structural, local details, and semantic elements of the original images.

2) Qualitative Comparison: In this section, we present qualitative results to illustrate the advantages of our method, S^{3}OIL, in preserving both structural integrity and semantic consistency across various image domains. The orange box in Fig. 5 underscores S^{3}OIL's capacity to preserve wetland boundaries and maintain their structural integrity reliably preserving linear and angular structures such as roads, river boundaries, and building edges, which are compromised by methods like DivCo, CDiffusion, CycleGAN, or CCM. Similarly, the blue box highlights the method's semantic consistency in urban areas, where competing models exhibit texture distortion and misaligned object boundaries. Notably, S^{3}OIL avoids semantic drift by aligning textures and shapes with realistic classes, unlike models like StegoGAN that generate semantically inconsistent content.

In Fig. 6, the orange box emphasizes S^{3}OIL's strength in maintaining accurate color rendering in semi-urban areas, preserving realistic and natural color tones even in low-texture regions, effectively avoiding the color shifts, over-saturation, or weak-supervision artifacts common in baselines. The blue box further demonstrates S^{3}OIL's ability to recover fine-grained textures along riverbanks and complex terrain boundaries, successfully reconstructing high-frequency features like vegetation textures, rooftops, and water ripple,s yielding sharp, contextually aligned results that outperform supervised (e.g., BBDM) and unsupervised methods. Similarly, CCM [50] exhibits unstable spatial performance due to absent structural guidance and paired supervision, underscoring the necessity of hybrid strategies like those in S^{3}OIL.

Taken together, these observations highlight the superiority of S^{3}OIL in generating outputs that not only exhibit high perceptual realism but also maintain semantic coherence and spatial structure. By leveraging both paired and unpaired data during training, S^{3}OIL achieves more stable domain alignment and better generalization across heterogeneous landscape types

TABLE II
QUANTITATIVE COMPARISON FOR THE ABLATION STUDY WITH MODULE INCLUSION INDICATED

CAM	MSM	ISC	CSC	PSNR↑	SSIM↑	LPIPS↓	FID↓
✓	-	-	-	12.28	0.1778	0.594	**100.1**
-	✓	-	-	12.43	0.1810	0.593	113.5
-	-	✓	-	12.21	0.1804	0.594	135.0
-	-	-	✓	12.36	0.1780	0.593	120.5
✓	✓	-	-	<u>12.48</u>	<u>0.1850</u>	<u>0.592</u>	106.9
✓	-	✓	-	12.30	0.1798	0.593	114.6
✓	-	-	✓	12.28	0.1789	0.593	117.4
-	✓	✓	-	12.32	0.1803	0.593	116.5
-	✓	-	✓	12.30	0.1792	0.593	113.9
-	-	✓	✓	12.20	0.1782	0.594	110.3
✓	✓	✓	✓	**12.51**	**0.1862**	**0.591**	<u>105.6</u>

compared to both traditional unsupervised and even recent supervised baselines.

C. Ablation Study

In the ablation study, the baseline model [14], [35] is defined as the version of the framework with the four key modules removed. By systematically excluding each component, we can quantitatively and qualitatively assess its impact on the overall performance.

1) Quantitative Analysis: As demonstrated in Table II, the inclusion of CAM significantly improves the FID score to 100.1. This suggests a better alignment with real image distributions. The addition of MSM results in the highest PSNR and SSIM scores, demonstrating notable improvements in image quality and structural similarity. When combining both CAM and MSM, the results show further enhancements, achieving a PSNR of 12.48 and an SSIM of 0.1850, while reducing the LPIPS score to 0.592. The FID also decreases to 106.9, reflecting a balanced improvement in quality and perceptual realism. Finally, combining all components yields the best results across all metrics, achieving the highest PSNR, SSIM, and the lowest LPIPS. This suggests that the integration of MSM and CAM provides a balanced enhancement in both quality and perceptual similarity. However, there is a slight drop in FID, which may stem from the trade-off between preserving structural integrity and improving perceptual realism. While the structural consistency is significantly improved, the enhanced focus on fine-grained details might slightly affect the overall perceptual match with real images.

Table II shows that each module contributes positively to performance when applied individually, targeting a specific aspect of the distribution gap or semantic misalignment. However, the combinations of two modules often do not yield additive improvements and may even result in performance degradation compared to single-module settings. This behavior is due to the complementary—not strictly additive—nature of the modules. Each constraint (CAM, MSM, ISC, CSC) is designed to address a distinct dimension of the translation problem: CAM aligns distributions between paired and unpaired domains; MSM promotes intra-set consistency among unpaired samples; ISC preserves ground-truth fidelity for paired data; and CSC ensures cross-modal semantic alignment. Without the full synergy of all modules, partial combinations may cause redundant constraints or interfere with optimization objectives. For instance, CAM and CSC both enforce cross-domain alignment, but lack intra-set compactness without MSM. Conversely, MSM alone enhances generalization but cannot guarantee semantic alignment across domains. This explains why some two-module combinations perform worse than a single module alone. The complete integration of all four modules yields the best results across all metrics, demonstrating their collective effectiveness in preserving semantic integrity and structural fidelity. This confirms that the design of S^{3}OIL requires the full interaction among CAM, MSM, ISC, and CSC to ensure high-quality translation.

TABLE III
COMPARISON BETWEEN MSM AND SINGLE SCALE MATCHING (SSM)

	PSNR↑	SSIM↑	LPIPS↓	FID↓
Baseline	12.22	0.1780	0.593	118.1
+SSM	12.32	0.1805	0.596	**101.3**
+MSM	**12.43**	**0.1810**	**0.593**	113.5

We also use *Davies-Bouldin Index* (DBI) to evaluate the distribution distance between the paired and unpaired set [55]. After applying the CAM module, the DBI metric decreases from 42.44 to 35.97, indicating that the distribution of paired and unpaired samples has become more similar. This reduction suggests that the CAM module effectively reduces the distance differences between the sets.

Additionally, as shown in Table III, the MSM (B+MSM) method outperforms the B and SSM (B+SSM) method greatly in PSNR, SSIM and LPIPS. This indicates that our multi-scale technique contributes to a more refined and detailed learning process in image generation.

2) Qualitative Analysis: Fig. 7 provides a detailed analysis of the contributions of individual modules in our model. The orange box highlights the impact of the MSM module, which significantly enhances the clarity and precision of water boundaries. By addressing multi-scale features with weak and strong perturbations, the MSM module ensures more refined spatial details, preserving the structural integrity of natural features like rivers and lakes. The CAM module further demonstrates its ability to produce fine-grained results by effectively aligning paired and unpaired data distributions within the target domain. This alignment enables better structural and contextual fidelity, ensuring that the generated outputs maintain the original scene's essential details. The CSC and ISC modules excel in achieving semantic consistency across both local and global regions. In particular, as shown in the blue box, these modules reduce blurring and enhance edge definition around the lake boundaries.

To further illustrate the alignment capability of CAM, we present a t-SNE-based feature distribution visualization in Fig. 8. The embeddings are extracted from intermediate encoder layers for both paired and unpaired samples. Without

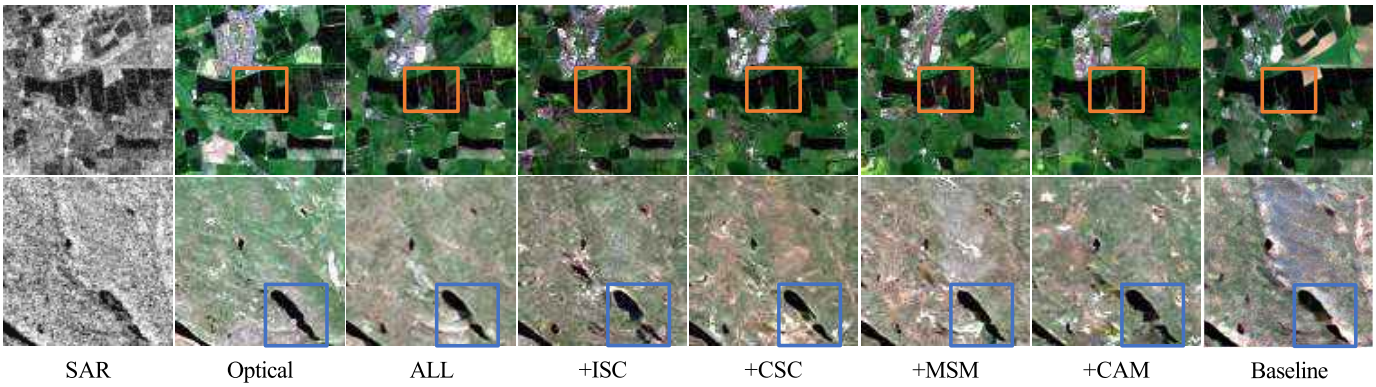


Fig. 7. Qualitative comparison for ablation study.

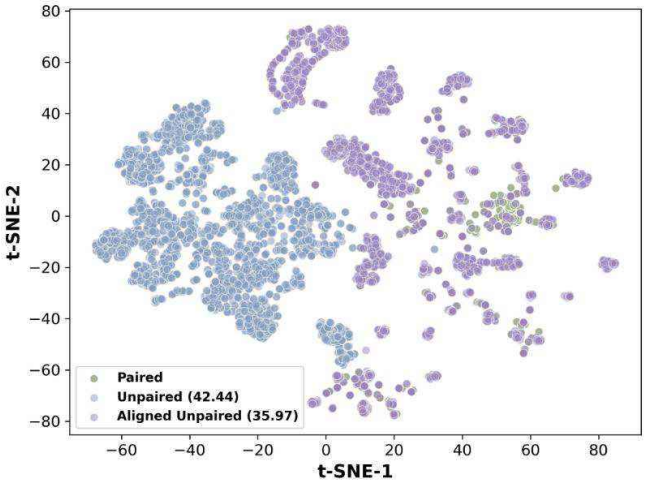


Fig. 8. T-SNE results of CAM module, blue represents the unpaired feature, purple represents the paired feature green represents the aligned feature after CAM. The DBI values (shown in parentheses) reflect the clustering compactness of each group. Compared to the paired features, the aligned features (DBI = 35.97) are significantly closer to the unpaired ones (DBI = 42.44), indicating that CAM effectively narrows the domain gap. These visualizations demonstrate that CAM significantly reduces the domain gap by bringing unpaired and paired features into closer proximity.

CAM, the two domains remain clearly separated in the t-SNE space, indicating poor cross-domain alignment. In contrast, when CAM is applied, the paired and unpaired features form overlapping and compact clusters, reflecting significantly improved semantic consistency across modalities.

The results illustrate the effectiveness of the combined approach (ALL), which integrates MSM, CAM, ISC, and CSC. This comprehensive framework ensures both structural and semantic consistency, particularly in complex areas like lakes, where the model achieves precise detail preservation and maintains the overall scene context. These results underscore the robustness and adaptability of the proposed model in addressing diverse challenges in image translation tasks.

D. Impact of the Number of Samples

To enhance the efficiency and robustness of our proposed semi-supervised framework, we employ various data segmentation strategies during training. We select 5,400 unpaired images and 600 paired images from the SEN12MS dataset and a total of 6,000 unpaired images to train the models for the unsupervised methods. Similarly, for the supervised methods, we incorporate 6,000 paired images in the training process.

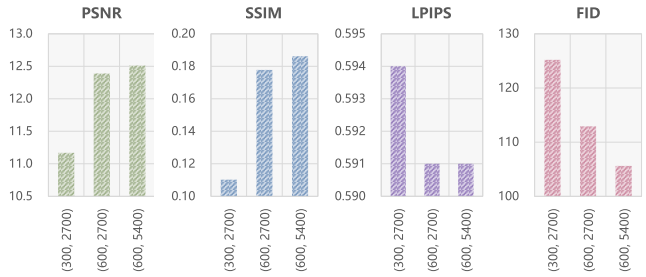


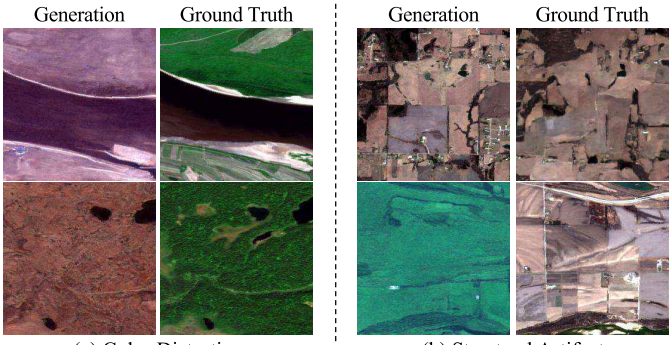
Fig. 9. Impact of the number of samples, i.e. (N_p, N_u).

As shown in Fig. 9, we evaluate dataset segmentation strategies to analyze the impact of sample distribution. The three strategies, (2700:300), (2700:600), and (5400:600), reflect supervised learning data splits. Metrics reveal distinct trends: PSNR and SSIM peak at a 1:9 paired/unpaired ratio (600/5400), achieving 12.51 and 0.1862, respectively, highlighting the benefits of unpaired data for reconstruction and structural similarity. Increasing paired data further boosts PSNR (11.17 → 12.39 → 12.51), SSIM, and FID, indicating enhanced detail reconstruction. LPIPS remains stable, suggesting perceptual quality is less affected. These results demonstrate the critical role of paired data in improving semi-supervised training performance, with the model maintaining robustness even with limited data.

E. Failure Examples

While S^{3}OIL demonstrates strong performance across both qualitative and quantitative evaluations, certain limitations remain, particularly in edge cases involving complex textures or ambiguous semantics. Fig. 10 presents two representative failure cases that highlight the model's current challenges.

1) Color Distortion: In the left example of Fig. 10, the model generates an optical image with incorrect greenish hues in areas that correspond to built-up regions. This color



(a) Color Distortion (b) Structural Artifact

Fig. 10. Failure cases of S^{3}OIL. (a) color distortion in urban textures. (b) artifacts along structural edges.

distortion typically occurs when the radar backscatter pattern in SAR images strongly deviates from those seen in the paired data, leading to incorrect mapping in the color space. Since the model relies on feature-level alignment across paired and unpaired domains, extreme distribution shifts can still result in mismatched style inference. The absence of global appearance priors or color regularization exacerbates this issue.

2) Structural Artifacts: In the right example of Fig. 10, artifacts appear near object boundaries, such as roads and building edges. These artifacts are often caused by inconsistencies between multi-scale features or by imperfect cross-modal matching when the semantic structure is subtle or noisy in SAR. Although our CAM and CSC modules aim to align features and preserve structure, they can fail in low-contrast or occluded regions, resulting in blurry or deformed outputs.

These cases underscore the need for further enhancement in handling extreme domain shifts and ambiguous feature representations. In future work, we plan to explore adaptive attention, uncertainty-aware matching, and external semantic priors to improve robustness under such conditions.

F. Generalization to Other I2IT Tasks

We broaden our analysis by evaluating our model on other datasets like Sketch2Anime to prove its robustness and generalizability. The Sketch2Anime dataset comprises anime images sourced from Danbooru2018 [56], with corresponding sketches generated using [57]. All anime images and sketches are aligned, resulting in a total of 135,509 paired images. Following the segmentation strategy proposed in our paper, we select 2,700 unpaired images and 300 paired images from the Sketch2Anime dataset. For unpaired methods, we employ a total of 3,000 unpaired images to train the models. Similarly, for the supervised methods, we incorporate 3,000 paired images into the training process.

As illustrated in Fig. 11, our model consistently outperforms Semi-I2I, WS-I2I, and Scenimefy. The superior performance of our model can be attributed to its advanced architecture and novel components that significantly enhance feature extraction and image generation capabilities. Unlike Semi-I2I and WS-I2I, our model incorporates mechanisms for semantic consistency and improved structural detail. Compared to Scenimefy and CycleGAN, our model exhibits more accurate and realistic image synthesis due to enhanced texture and detail preservation. Additionally, our approach surpasses DivCo, U-GAT-IT, StegoGAN, and FG-GAN by achieving higher fidelity and consistency inthe generated image. In summary, S^{3}OIL exhibits significant advantages in the task of Sketch2Anime image translation. In Table IV, our model performs significantly in Sketch2Anime datasets. These results highlight the robustness and generalization capabilities of our approach, showcasing its effectiveness in various image translation tasks and confirming its adaptability to different domains.

TABLE IV
QUANTITATIVE RESULTS ON SKETCH2ANIME DATASET. † INDICATES THE SUPERVISED VARIANT OF OUR MODEL

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓
Pix2Pix	14.71	0.5232	0.660	64.8
S^{3}OIL† (Ours)	14.77	0.6326	0.633	47.4
CDiffusion	15.40	0.6450	0.615	45.8
FG-GAN	12.44	0.5734	0.578	58.1
U-GAT-IT	11.18	0.5439	0.596	65.4
DivCo	10.12	0.3674	0.589	100.3
CycleGAN	12.05	0.5232	0.575	75.6
StegoGAN	11.08	0.5111	0.588	108.8
WS-I2I	11.35	0.4319	0.590	<u>90.2</u>
Scenimefy	10.98	0.5310	<u>0.590</u>	97.8
Semi-I2I	12.09	<u>0.5790</u>	0.591	57.1
S^{3}OIL (Ours)	**13.74**	**0.6176**	**0.587**	**54.5**

G. Hyperparameter Analysis

In our final model configuration, we empirically set the weights of different loss components as follows: $\lambda_{\text{CAM}} = 50$, $\lambda_{\text{CSC}} = 10$, $\lambda_{\text{ISC}} = 10$, $\lambda_{\text{MSM}} = 50$, and $\lambda_{\text{GAN}} = 50$. These values were determined based on experimental validation across multiple held-out scenes, with the goal of balancing semantic preservation, structural consistency, and perceptual realism. A higher weight for λ_{CAM} emphasizes cross-set alignment, which is particularly crucial in our semi-supervised setting where unpaired data may introduce significant domain gaps. A similar high value is used for λ_{MSM} to enforce strong intra-set consistency via multi-scale perturbation, which improves generalization. In contrast, λ_{CSC} and λ_{ISC} are set lower, as these modules operate on more localized constraints (semantic matching and paired reconstruction) and are already guided by structural priors through CAM. The adversarial loss weight λ_{GAN} is set moderately high to maintain texture realism without overwhelming the consistency constraints.

We observed that lowering λ_{CAM} or λ_{MSM} weakens the cross-domain alignment and results in color inconsistencies and structural drift in unpaired scenes. Conversely, excessively increasing λ_{GAN} tends to produce sharper textures but introduces hallucinated details or unstable training. A detailed sensitivity analysis of each hyperparameter, shown in Table V, confirms that the model performs stably within a reasonable

Fig. 11. Qualitative results on Sketch2Anime dataset.

TABLE V
SENSITIVITY ANALYSIS OF LOSS WEIGHTS ON SEN12MS DATASET

Loss Weight	Range Tested	LPIPS ↓
λ_{CAM}	10 – 80	0.660 – 0.630
λ_{MSM}	10 – 80	0.655 – 0.620
λ_{GAN}	10 – 100	0.640 – 0.648
λ_{CSC}	5 – 20	0.643 – 0.635
λ_{ISC}	5 – 20	0.641 – 0.636

TABLE VI
COMPARISON OF PARAMETERS AND INFERENCE TIME OF DIFFERENT METHODS

Method	Type	Paras/M	Inf. Time /H
Ours	Semi-supervised	30.6	36.7
Pix2Pix	Supervised	40.1	48.1
StegoGAN	Unsupervised	24.5	29.4
Semi-I2I	Semi-supervised	38.3	46.0

range, but extreme values for λ_{CAM} or λ_{MSM} significantly affect LPIPS.

In terms of network architecture, we adopt an unbalanced generator with DenseNet blocks in the middle layers, following [14], to improve feature propagation and gradient flow in deep latent spaces. This is particularly beneficial for maintaining global structure while enhancing fine details. Furthermore, we use a patch-based discriminator [41], which provides localized feedback and helps the generator produce spatially coherent textures, especially in high-frequency regions.

Together, these hyperparameter settings and architectural choices contribute to the observed robustness and effectiveness of S^{3}OIL across a variety of remote sensing scenarios.

H. Efficiency Comparison of Different Models

To comprehensively evaluate the efficiency of our proposed S3OIL model, we compare its computational complexity with representative baseline methods in terms of model size (number of parameters) and inference speed. As summarized in Table VI, S3OIL comprises 30.6 million parameters, which is fewer than Pix2Pix (40.1M) and Semi-I2I (38.3M), while moderately larger than StegoGAN (24.5M). This indicates that S^{3}OIL maintains a compact architecture compared to most supervised and semi-supervised models. Across 200 training epochs S^{3}OIL achieves a total inference time of 36.7 hours, which is substantially faster than Pix2Pix (48.1h) and Semi-I2I (46.0h), despite delivering comparable or superior translation quality. Although StegoGAN requires less time (29.4h) due to its smaller parameter count, it suffers from noticeable degradation in structural consistency and visual fidelity. These results demonstrate that S^{3}OIL strikes a favorable balance between model compactness and computational efficiency, making it suitable for scenarios requiring both high performance and moderate resource consumption.

V. CONCLUSION

In this paper, we propose S^{3}OIL method to address the challenge of limited paired data in image-to-image translation. Our approach integrates a CAM mechanism to ensure local correspondences and uses MSM constraints with weak and strong perturbations to improve intra-set consistency. We also incorporate local and global cross-modal semantic consistency to enhance structural identity. S^{3}OIL outperforms state-of-the-art methods in both quantitative metrics and qualitative assessments across various datasets, such as SAR-to-optical and sketch-to-anime tasks. Ablation studies confirm that CAM and MSM significantly enhance the stylistic realism and structural consistency of the generated images. While S^{3}OIL achieves impressive results, it does require careful management of computational resources due to the complexity of the cross-set and multi-scale mechanisms. However, the benefits in translation quality and consistency outweigh these computational demands, making S^{3}OIL a robust solution for image translation with limited paired data.

REFERENCES

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.
[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[4] J. Zhang, J. Zhou, and X. Lu, "Feature-guided SAR-to-optical image translation," *IEEE Access*, vol. 8, pp. 70925–70937, 2020.
[5] M. Zhang, J. Xu, C. He, W. Shang, Y. Li, and X. Gao, "SAR-to-optical image translation via thermodynamics-inspired network," 2023, *arXiv:2305.13839*.

[6] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 596–608.

[7] B. Sun, Y. Yang, L. Zhang, M.-M. Cheng, and Q. Hou, "CorrMatch: Label propagation via correlation matching for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3097–3107.

[8] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2008.

[9] H. Wang, C. Zhou, X. Chen, J. Wu, S. Pan, and J. Wang, "Graph stochastic neural networks for semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 19839–19848.

[10] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8934–8954, Sep. 2022.

[11] J. T. Barron, "A general and adaptive robust loss function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4331–4339.

[12] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial networks—Optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, Sep. 2019.

[13] W.-L. Du, Y. Zhou, J. Zhao, and X. Tian, "K-means clustering guided generative adversarial networks for SAR-optical image matching," *IEEE Access*, vol. 8, pp. 217554–217572, 2020.

[14] X. Yang, Z. Wang, J. Zhao, and D. Yang, "FG-GAN: A fine-grained generative adversarial network for unsupervised SAR-to-optical image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621211.

[15] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 752–762.

[16] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 319–345.

[17] B. Kim, G. Kwon, K. Kim, and J. Chul Ye, "Unpaired image-to-image translation via neural Schrödinger bridge," 2023, *arXiv:2305.15086*.

[18] M. Ye, M. Kanski, D. Yang, L. Axel, and D. Metaxas, "Unsupervised exemplar-based image-to-image translation and cascaded vision transformers for tagged and untagged cardiac cine MRI registration," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 7629–7639.

[19] H. Liu, Y. Wei, F. Liu, W. Wang, L. Nie, and T.-S. Chua, "Dynamic multimodal fusion via meta-learning towards micro-video recommendation," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, pp. 1–26, Mar. 2024.

[20] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[21] J.-Y. Zhu et al., "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[22] G. Zhong, Y. Guo, J. Yuan, Q. Zhang, W. Guan, and L. Chen, "PROMOTE: Prior-guided diffusion model with global-local contrastive learning for exemplar-based image translation," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 3313–3322.

[23] Z. Deng et al., "Structured generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3902–3912.

[24] Y. Liu, G. Deng, X. Zeng, S. Wu, Z. Yu, and H.-S. Wong, "Regularizing discriminative capability of CGANs for semi-supervised generative learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5719–5728.

[25] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1–2 dataset for deep learning in SAR-optical data fusion," 2018, *arXiv:1807.01569*.

[26] X. Yang, J. Zhao, Z. Wei, N. Wang, and X. Gao, "SAR-to-optical image translation based on improved CGAN," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108208.

[27] J. Hwang, C. Yu, and Y. Shin, "SAR-to-optical image translation using SSIM and perceptual loss based cycle-consistent GAN," in *Proc. Inf. Commun. Technol. Conf.*, 2020, pp. 191–194.

[28] L. Liu and B. Lei, "Can SAR images and optical images transfer with each other?," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 7019–7022.

[29] Z.-G. Liu, Z.-W. Zhang, Q. Pan, and L.-B. Ning, "Unsupervised change detection from heterogeneous data based on image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403413.

[30] J. N. Turnes, J. D. B. Castro, D. L. Torres, P. J. S. Vega, R. Q. Feitosa, and P. N. Happ, "Atrous CGAN for SAR-to-optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 4003905.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[32] A. Mustafa and R. K. Mantiuk, "Transformation consistency regularization—A semi-supervised paradigm for image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 599–615.

[33] Y. Jiang, L. Jiang, S. Yang, and C. C. Loy, "Scenimefy: Learning to craft anime scene via semi-supervised image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7323–7333.

[34] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7236–7246.

[35] W.-L. Du, Y. Zhou, H. Zhu, J. Zhao, Z. Shao, and X. Tian, "A semi-supervised image-to-image translation framework for SAR–optical image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[36] S. Liu et al., "AdaAttN: Revisit attention mechanism in arbitrary neural style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6649–6658.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1597–1607.

[38] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[39] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[40] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[41] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[42] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[43] C. Chan, F. Durand, and P. Isola, "Learning to generate line drawings that convey geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7915–7925.

[44] R. Labaca-Castro, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 73–76.

[45] X. Bai, X. Pu, and F. Xu, "Conditional diffusion for SAR to optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[46] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "BBDM: Image-to-image translation with Brownian bridge diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1952–1961.

[47] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*.

[48] R. Liu, Y. Ge, C. L. Choi, X. Wang, and H. Li, "DivCo: Diverse conditional image synthesis via contrastive generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16377–16386.

[49] S. Wu et al., "StegoGAN: Leveraging steganography for non-bijective image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 7922–7931.

[50] A. Bhagat, M. Jain, and A. V. Subramanyam, "Conditional consistency guided image translation and enhancement," 2025, *arXiv:2501.01223*.

[51] S. Shukla, L. Van Gool, and R. Timofte, "Extremely weak supervised image-to-image translation for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3368–3377.

[52] M. Schmitt, L. Haydn Hughes, C. Qiu, and X. Xiang Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," 2019, *arXiv:1906.07789*.

[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.

[54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[55] S. Petrovic, "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters," in *Proc. 11th Nordic Workshop Secure IT Syst.*, 2006, pp. 53–64.
[56] B. Gwernand G. Aaron. (Jan. 2019). *Danbooru2018: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset*. [Online]. Available: https://www.gwern.net/Danbooru2018
[57] X. Xiang, D. Liu, X. Yang, Y. Zhu, and X. Shen. (2021). *Anime2sketch: A Sketch Extractor for Anime Arts With Deep Networks*. [Online]. Available: https://github.com/Mukosame/Anime2Sketch

Xi Yang (Senior Member, IEEE) received the B.Eng. degree in electronic information engineering and the Ph.D. degree in pattern recognition and intelligence system from Xidian University, Xi'an, China, in 2010 and 2015, respectively. From 2013 to 2014, she was a Visiting Ph.D. Student with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA. In 2015, she joined the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, where she is currently a Professor in communications and information systems. She has published over 60 articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, CVPR, ICCV, and ACM MM. Her current research interests include image/video processing, computer vision, and machine learning.

Haoyuan Shi received the B.Eng. degree in communication engineering from Hangzhou Dianzi University, Hangzhou, China, in 2023. He is currently pursuing the M.S. degree in electronic and information engineering with Hangzhou Institute of Technology, Xidian University, Hangzhou. His current research interests include deep learning and image translation.

Ziyun Li received the B.Eng. degree in communication engineering from Hangzhou Dianzi University, Hangzhou, China, in 2017, the M.Sc. degree in computer science from the City University of Hong Kong, Hong Kong SAR, China, in 2018, and the Dr.rer.nat. degree in IT-systems engineering from the Hasso Plattner Institute, University of Potsdam, Germany, in 2024. She is currently a Post-doctoral Researcher with the KTH Royal Institute of Technology, Stockholm, Sweden, funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP). Her recent research has been published in venues such as CVPR, ICCV, AAAI, NeurIPS and TMLR. Her current research interests include generative AI, flow matching, diffusion models, generalized class discovery, out-of-distribution detection, and knowledge distillation. She has served as a reviewer for leading journals and conferences, including IJCV, TMLR, NeurIPS, ICML, ICLR, CVPR, and ICCV.

Maoying Qiao (Member, IEEE) received the Ph.D. degree from the University of Technology Sydney (UTS), Ultimo, NSW, Australia, in 2016. She is currently a Senior Lecturer with the School of Computer Science, FEIT UTS. She has published more than 15 papers in high-quality journals and conferences. Her research interests include machine learning, especially in diversity-promoting modeling. She has actively served as a PC Member for conferences, for example, NeurIPS, ICML, UAI, and IJCNN, and a Reviewer for journals, for example, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, and *ACM Transactions on Knowledge Discovery from Data*.

Fei Gao (Member, IEEE) received the bachelor's degree in electronic engineering and the Ph.D. degree in information and communication engineering from Xidian University, Xi'an, China, in 2009 and 2015, respectively. From October 2012 to September 2013, he was a Visiting Ph.D. Candidate with the University of Technology, Sydney (UTS), Australia. From July 2015 to June 2023, he was with Hangzhou Dianzi University. He mainly applies machine learning techniques to computer vision problems. His research results have expounded in more than 30 publications at prestigious journals and conferences. He served for a number of journals and conferences. His research interests include visual quality assessment and enhancement, intelligent visual arts generation, and biomedical image analysis.

Nannan Wang (Senior Member, IEEE) received the B.Sc. degree in information and computation science from Xi'an University of Posts and Telecommunications in 2009 and the Ph.D. degree in information and telecommunications engineering from Xidian University in 2015. He is currently a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published over 150 papers in refereed journals and proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV, CVPR, and ICCV. His current research interests include computer vision and machine learning.