# A Frequency-Coupled Latent Space Facilitates Synthetic-to-Measured SAR Image Translation

Jinrui Liao, *Student Member, IEEE*, Qingsong Wang, Yikui Zhai, *Senior Member, IEEE*, Xingwang Hu, Nengcai Li, Haifeng Huang, *Member, IEEE*, C. L. Philip Chen, *Life Fellow, IEEE*, Pasquale Coscia, *Senior Member, IEEE*, and Angelo Genovese, *Senior Member, IEEE*

***Abstract*—Deep learning (DL)-based synthetic aperture radar (SAR) target recognition depends on measured data that are costly to collect, and synthetic images produced by electromagnetic simulation offer an economical substitute whose direct use is limited by the synthetic-to-measured (S2M) domain gap. Owing to the coherent imaging mechanism of SAR, this gap manifests not only in spatial appearance but also in the distribution of image energy across spatial-frequency bands and orientations, a structure that latent representations learned solely for reconstruction encode only implicitly and therefore cannot align selectively. To reduce this deep-feature gap, we propose LFCS2M, a latent diffusion framework for S2M SAR image translation in a frequency-coupled latent space. A frequency domain feature refinement module (FDFRM) imposes a learnable two-dimensional spectral mask on the latent representation, allowing bandwise and orientationwise feature adjustment without assuming a predefined spectral discrepancy. A bidirectional diffusion bridge guided by Measured Information Guided Cross Attention (MIGCA) further maps synthetic latent features toward the measured distribution. Qualitative and quantitative experiments show that LFCS2M generates high-fidelity SAR images that closely resemble measured data, supporting DL-based SAR target recognition by narrowing the domain gap and reducing data acquisition cost. The code of LFCS2M is accessible at https://github.com/Jordan-Liao/LFCS2M.**

***Index Terms*—Synthetic Aperture Radar Automatic Target Recognition (SAR ATR), Deep learning (DL), image-to-image (Img2Img) translation, Diffusion model.**

## I. Introduction

CONSIDERING that synthetic aperture radar (SAR) images can be collected under a wide range of lighting and weather conditions, SAR has been extensively applied to terrestrial and maritime surveillance and reconnaissance. Recently, SAR automatic target recognition (ATR) has advanced rapidly with the integration of deep learning (DL) techniques [1]–[4]. Despite these advances, data availability remains a major obstacle [5], [6]. Collecting and labeling large-scale SAR datasets, especially those containing diverse measured target images, is prohibitively expensive [4]. This scarcity restricts the development and generalization of DL models; consequently, sufficiently diverse and domain-consistent SAR data are essential for reliable recognition [7]–[9].

Two strategies are commonly adopted to enlarge SAR training sets. The first employs generative adversarial networks (GANs) to synthesize or augment SAR target images [10]–[12]. Recent variants improve controllability by constraining azimuth-dependent features [13] or improve physical fidelity by incorporating attributed scattering center (ASC) features into adversarial translation [14]. Nevertheless, their output diversity remains bounded by the support of the measured training distribution and by the prescribed conditioning information; targets at insufficiently represented aspect angles therefore remain difficult to generate reliably (Fig. 1(a)). The second strategy renders synthetic targets from computer-aided design (CAD) models through electromagnetic simulation [15], producing images of arbitrary classes at controlled aspect angles at low marginal cost. Its obstacle is the domain gap: networks trained on synthetic images degrade sharply when applied to measured targets (Fig. 1(b)), because synthetic images may not fully reflect the real-world features required for effective model training [16]–[18]. Therefore, synthetic data cannot be directly used for DL training.

SAR images are formed by coherent processing in range and azimuth, and their responses are dominated by scattering centers and multiplicative speckle [19]. As a result, domain differences are not limited to spatial appearance, but also appear as structured spectral discrepancies. Electromagnetic simulation can model deterministic target scattering, whereas stochastic clutter, sensor responses, and processing effects in measured collections are only approximated. The discrepancy is therefore more likely to concentrate in specific frequency bands and orientations than to spread uniformly, which is consistent with Fig. 1(c). The importance of high-frequency components to SAR diffusion generation has also been demonstrated by DiffuSAR [20]. However, generic diffusion formulations developed for natural imagery [21]–[23]

This work was supported in part by the National Natural Science Foundation of China under Grant 62273365; in part by the Science and Technology Planning Project of the Key Laboratory of Advanced IntelliSense Technology, Guangdong Science and Technology Department, under Grant 2023B1212060024; in part by the Xiaomi Young Talents Program; and in part by the High-Performance Computing Public Platform (Shenzhen Campus) of Sun Yat-sen University. *(Corresponding author: Qingsong Wang.)*

Jinrui Liao, Qingsong Wang, Xingwang Hu, Nengcai Li, and Haifeng Huang are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China (e-mail: liaojr23@mail2.sysu.edu.cn; wangqs5@mail.sysu.edu.cn; huxw35@mail2.sysu.edu.cn; linengcai@foxmail.com; huanghaifeng@mail.sysu.edu.cn).

Yikui Zhai is with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China (e-mail: yikuizhai@163.com).

C. L. Philip Chen is with the College of Computer Science and Engineering, South China University of Technology, Guangzhou, China (e-mail: philip.chen@ieee.org).

Pasquale Coscia and Angelo Genovese are with the Dipartimento di Computer Science and the Dipartimento di Informatica, Universita degli Studi di Milano, 20133 Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it).

do not explicitly parameterize the bandwise and orientation-wise discrepancies between synthetic and measured SAR domains. Addressing the domain discrepancy between synthetic and measured SAR data remains difficult. Recent domain-adaptation methods progressively align simulated and measured samples at the pixel, domain, and class levels [24], while transfer-based methods fuse domain-invariant and domain-specific representations across SAR configurations [25]. These methods improve target-domain recognition, but they operate primarily in the classifier feature space and do not produce measured-like images that can be reused by arbitrary downstream recognition models. This motivates a synthetic-to-measured (S2M) domain translation strategy: translating images from the synthetic domain to the measured domain may reduce the sensitivity of different neural networks to the intrinsic discrepancy between the two domains, as illustrated in Fig. 1(d). S2M translation can be viewed as a distribution adaptation problem, which is more tractable in a compact latent space than in the original pixel space; therefore, latent diffusion provides a suitable solution [26]. However, a latent encoder trained only for reconstruction cannot reweight individual frequency bands or orientations and thus cannot selectively correct spectral-structure deviations. We therefore couple a learnable frequency-domain operator with the latent representation to construct a frequency-coupled latent space that is both compact and spectrally controllable. Since the spectral adjustment is learned rather than predefined, this representation can adapt to the discrepancies exhibited by measured SAR data.

Based on this formulation, we propose LFCS2M, a latent diffusion framework for S2M SAR image translation in a frequency-coupled latent space. As shown in Fig. 2, paired synthetic and measured images are encoded by the FDFRM-equipped encoder $f_{EF}$. FDFRM applies a learnable two-dimensional spectral mask $m_{ijk}$ that adaptively reweights the latent spectrum across frequency bands and orientations , producing frequency-coupled latent codes without prespecifying which components should be enhanced or suppressed . In this latent space, a forward diffusion bridge [27] interpolates from the measured target latent $r_0$ to the synthetic terminal latent $r_T = s$, whereas the learned reverse process performs the desired S2M translation from $s$ to $r_0$. A U-Net denoiser equipped with the auxiliary MIGCA mechanism recovers the translated latent, which is decoded into the output $\widehat{I}_R$ . The model is optimized with a variational lower-bound objective to produce translations whose scattering patterns and spectral statistics are closer to those of their measured counterparts. We evaluate LFCS2M using perceptual similarity metrics, target recognition experiments, and frequency-domain analysis that verifies the convergence of the translated spectrum toward the measured one. The contributions of this work are summarized as follows.

1) We formulate S2M SAR translation as distribution alignment within a latent space coupled with frequency information. This is realized through FDFRM, which uses a learnable two-dimensional spectral mask to adaptively amplify or attenuate frequency bands and directional components without assuming the location of the spectral discrepancy.
2) We introduced the MIGCA mechanism to the training objective optimization to enhance the focus on scattering information in translated images. Experiments verify the advantages of the mechanism in capturing detailed scattering features such as cross-scatter spots and deformed scatterers, which directly strengthens the model's translation accuracy.
3) We conduct comprehensive evaluations using similarity metrics, real-world recognition scenarios, and frequency-domain analysis. Both qualitative and quantitative results demonstrate that LFCS2M method achieves competitive performance. The synthesized data demonstrate strong similarity to measured SAR images and are well-suited for training DL models in ATR.

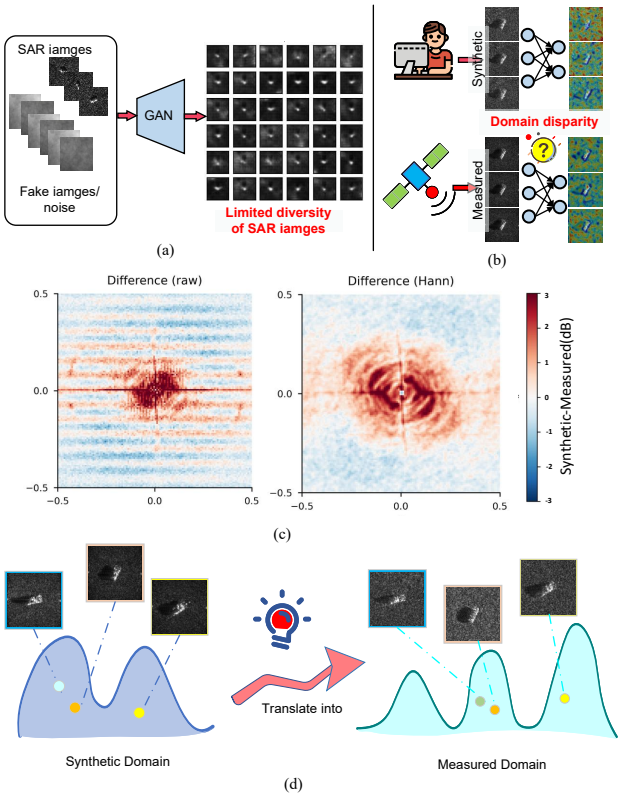



Fig. 1. (a) GAN-based methods tend to produce low-quality and redundant samples when trained on SAR datasets lacking sufficient angular diversity or data volume. (b) There is a pronounced discrepancy at the feature representation level between synthetic SAR images and their measured counterparts. (c) Two-dimensional power spectrum difference between synthetic and measured chips, showing excess power at low and intermediate frequencies. (d) Our motivation: enabling S2M SAR translation by explicitly addressing the domain discrepancy.

The subsequent sections were meticulously arranged according to this structure: Section II reviews advancements in Img2Img research and discusses the utilization of synthetic SAR images in DL applications. Section III explores the design and components of the LFCS2M model in detail. Section IV delineated the experimental details. Section V discusses the results and limitations. Section VI presented conclusions and future research directions.

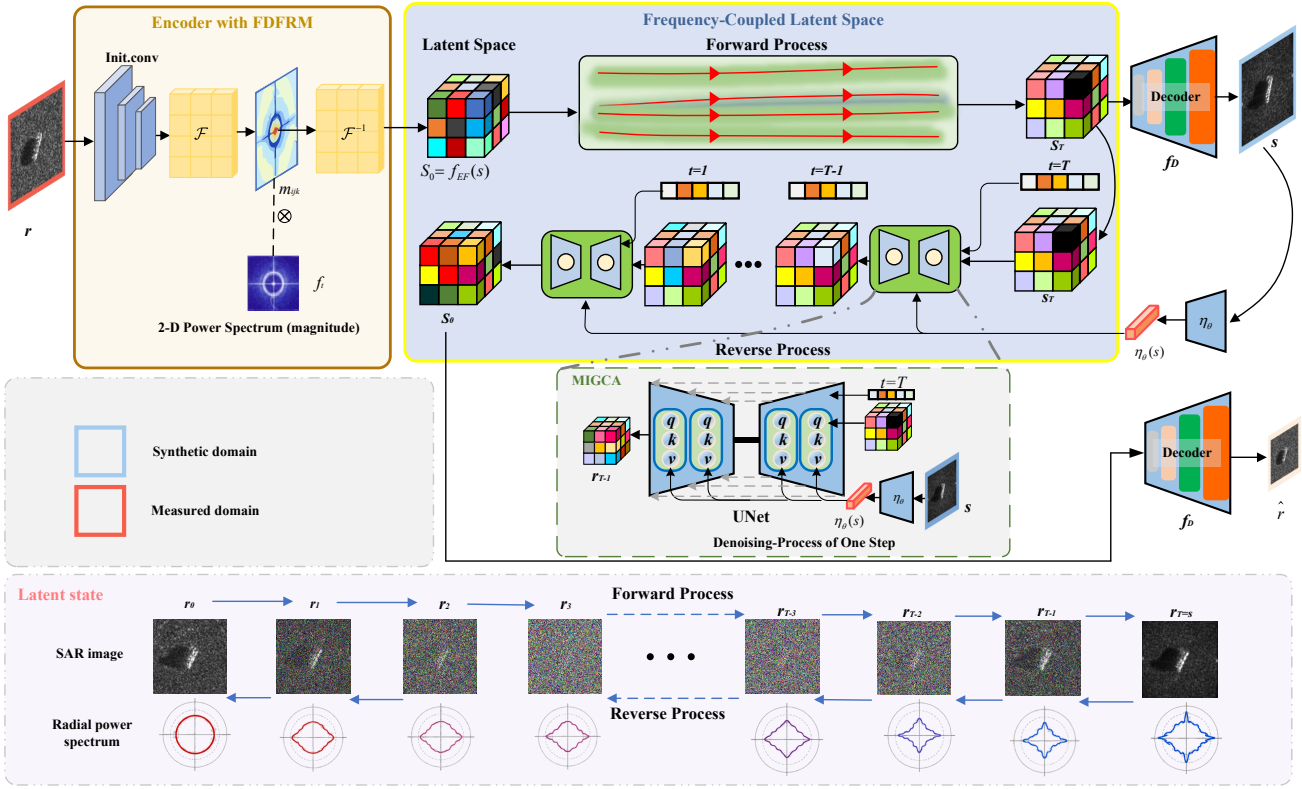



Fig. 2. Overall framework of LFCS2M. The encoder $f_{EF}$ embeds frequency-domain information into the latent representation via the FDFRM. Within the resulting frequency-coupled latent space, a forward bridge runs from the measured target latent $r_0$ to the synthetic terminal latent $r_T = s$, while the learned reverse process (guided by MIGCA) performs the S2M translation. The bottom panel shows the SAR image and its radial power spectrum evolving along the bridge.

## II. RELATED WORK

### A. Image-to-Image Translation

Image-to-Image (Img2Img) translation seeks to convert an image from one domain or style to another [28], [29]. As a computer vision task, this process involves transforming one type of image to visually match the target domain's characteristics or style. Examples include transforming a sketch image into a measured image or colorizing a black-and-white image. Image translation typically leverages DL models, especially GANs or Conditional GANs (cGANs), by learning the mapping relationship between different domains. Image translation can be implemented in various fields, including style transfer, image restoration, and image colorization, which can be trained regardless of whether there are paired data or not.

Img2Img translation involves the transformation of images between domains, with the primary goal of preserving the core features and organization of the original images during conversion. Recent advances in Img2Img translation have been significantly driven by the application of GANs, which have led to considerable enhancements in both the quality and efficiency of Img2Img translation [30], [31], [32]. Pix2Pix [33] is one of the first GAN-based methods that not only implements Img2Img translation, but also theoretically generalizes to any dataset given enough training data. It is also capable of modeling tasks like image restoration and image colorization without modifying the model. CycleGAN [34] can conduct the training of Img2Img GANs without paired data. Although the Img2Img field has been revitalized thanks to the development of GANs, GANs are still confronting a big challenge to ensure high-precision translation with small-scale datasets. Recently, diffusion models have provided an alternative approach to Img2Img [35], [36], [37], [38], showing promising performance even when working with small-scale datasets.

### B. Integration of Synthetic SAR Images to DL

It has been an obstacle for SAR intelligent image interpretation to fully use synthetic SAR data and mitigate the impact of measured SAR data shortage for DL-based target recognition. Therefore, some related studies are reviewed here.

Malmgren-Hansen [39] and Arnold [40] used synthetic SAR data for pre-training, enabling transfer learning in SAR target classification networks. In this framework, the network is first trained on synthetic SAR data to gain generalizable characteristics, and then fine-tuned with a minimal quantity of measured chips to adapt to real-world conditions. While this method improves recognition performance under limited measured data, the inherent domain gap between synthetic and measured SAR images remains. As more measured data are incorporated during fine-tuning, the network's performance gradually converges to that of models trained entirely on measured data.

An alternative strategy seeks to reduce the need for measured data during training by exploiting synthetic SAR images. Sellers et al. [41] used GAN-based augmentation and attained a 95% recognition rate with 99% augmented synthetic data and 1% measured data. Lewis et al. [42] and Inkawhich et al. [16] applied adversarial perturbations and data augmentation to improve the generalization of classifiers trained predominantly or entirely on synthetic data. Du et al. [43] introduced an adversarial encoding framework that mimics physical feature extraction to align simulated and measured representations. More recently, Xiang et al. [13] constrained image generation with two-dimensional azimuth-angle features to improve scattering consistency across viewing angles. These approaches enrich or align training samples for recognition, whereas LFCS2M addresses paired S2M image translation and explicitly adjusts the spectral distribution of each source image before diffusion-based domain mapping.

A third direction translates synthetic SAR images into the measured domain before training a recognition model. Youk et al. [44] used a Transformer to learn representational mappings between paired synthetic and measured targets. Guan et al. [14] instead introduced an ASC-based scattering-feature similarity loss into a CycleGAN-based generator. The former emphasizes spatial representational learning, whereas the latter relies on explicitly extracted scattering-center parameters and adversarial optimization. In contrast, LFCS2M learns a dense two-dimensional spectral mask without prescribing particular scattering centers and performs distribution transport through a source-conditioned latent diffusion bridge . Its MIGCA mechanism conditions each reverse state on the available synthetic terminal latent, while measured-domain information is learned from the paired bridge target during supervised optimization . It therefore differs from both spatial Transformer mapping and ASC-constrained adversarial translation in its representation, optimization process, and mechanism for incorporating measured-domain information.

### C. Diffusion Models for SAR

Recently, diffusion models have offered a more stable and expressive alternative for SAR generation and translation. DiffuSAR [20] separates and adjusts components at high and low frequencies during denoising to improve frequency-aware SAR target generation. Liao et al. [45] developed a complex-valued conditional diffusion model that injects physical scattering information, using generated samples to improve target recognition. Both methods focus on synthesizing or augmenting SAR targets rather than translating a given synthetic target into its measured-domain counterpart. For cross-modal translation, complementary SAR–optical information has also been exploited for registration [46] and for SAR-aided cloud and shadow removal [47]. Within diffusion-based frameworks, Du et al. [48] decomposed SAR-to-optical mapping into noise-domain modality conversion and reverse-diffusion reconstruction, with wavelet alignment used to preserve shared high-frequency details. Their target distribution is optical imagery and therefore differs fundamentally from the same-modality S2M setting considered here. LFCS2M instead transports a source-conditioned synthetic latent toward a paired measured-SAR distribution through a diffusion bridge; its learnable spectral mask provides bandwise and orientationwise control, while MIGCA uses the same synthetic terminal guide during training and inference . Thus, the proposed method is distinct from frequency-aware SAR generation, physical-condition-guided augmentation, and cross-modal SAR-to-optical diffusion.

## III. Methodology

Translating synthetic SAR imagery into its measured counterpart is particularly demanding. Unlike the spatially continuous structure of optical images, SAR scenes comprise discrete point-like returns whose statistics differ markedly from natural imagery; together with multiplicative speckle, they leave a characteristic imprint on the image spectrum. Conventional Img2Img frameworks supervise the mapping with spatial-domain objectives that neither model speckle nor account for the distribution of scattering energy across frequency bands and orientations. By exposing the frequency content within the latent space, the model gains explicit and selective control over exactly the components in which the two domains differ. Building on this basis, we introduce LFCS2M, a framework that aligns the synthetic and measured domains through a diffusion bridge in a frequency-coupled latent space, performing translation in a representation that exposes rather than averages over the bands and orientations where the domains differ. Modeling the global spectral structure of SAR chips in this way reduces the network's reliance on local spatial neighborhoods and improves cross-domain generalization. The following sections detail its architecture and operation.

### A. Latent Space Mapping Process

Given two datasets, $X_R$ and $X_S$, derived from measured and synthetic SAR images, respectively, the measured ones represent real electromagnetic scattering of targets, while the synthetic ones are synthesized through simulation software to predict the electromagnetic characteristics of targets. These two datasets are denoted as domains $R$ and $S$, with the goal of establishing a mapping function from domain $S$ to domain $R$ by SAR image translation. Before constructing the mapping, the measured SAR images are normalized as defined below:

$$\tilde{x}_R = \begin{cases} -1, & x_R \le 0 \\ 1, & x_R \ge \bar{x}_R \\ \frac{2x_R}{\bar{x}_R} - 1, & \text{otherwise} \end{cases} \tag{1}$$

Here, $X_R$ and $\tilde{x}_R$ denote the pixel values of the SAR image prior and subsequent to normalization, respectively, and $X_R$ is the average pixel magnitude of the image $X_R$. This normalization guarantees the intensity values of SAR images lie within the interval $[-1, 1]$, refining the data for better model training.

For a sample image $I_S$ drawn from domain $S$, its latent description is obtained through the encoder, yielding $L_S = f_E(S)$. The encoder $f_E$ compresses $S$ into a compact latent code, while the decoder $f_D$ recovers the pixel-level image from this code.

### B. Frequency Domain Feature Refinement Module

The discrepancy between synthetic and measured SAR images is structured in the frequency domain and may be localized in particular bands and orientations. Conventional latent diffusion models do not explicitly control the frequency content of their latent features, which can reduce fidelity when translating synthetic SAR images into measured ones. We therefore introduce the frequency-domain feature refinement module (FDFRM) immediately after the first convolutional layer. FDFRM learns a two-dimensional response over the full nonredundant spectrum, allowing the training objective to determine which frequency coordinates and directional components should be amplified or attenuated. For an input feature map $x \in \mathbb{R}^{B\times C\times H\times W}$, we define the shared adaptive spectral modulation and the complete FDFRM operator as

$$\begin{aligned} \mathcal{S}_M(x) &= \mathcal{F}_{\mathbb{R}}^{-1}[\mathcal{F}_{\mathbb{R}}(x) \odot \mathcal{R}_{H,W}(M)], \\ \mathrm{FDFRM}(x;\Theta_F) &= x + \gamma \mathcal{S}_M(x), \\ \Theta_F &= \{M, \gamma\}. \end{aligned} \tag{2}$$

Here, $\mathcal{F}_{\mathbb{R}}$ and $\mathcal{F}_{\mathbb{R}}^{-1}$ denote the real-input spectral analysis and synthesis operators, respectively. The resolution-adaptation operator $\mathcal{R}_{H,W}$ bilinearly maps the real-valued base spectral response $M \in \mathbb{R}^{1\times C\times 64\times 64}$ to the nonredundant half-spectrum resolution $H \times (\lfloor W/2 \rfloor + 1)$. The response is initialized once as $M \leftarrow \mathbf{1} + \alpha\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, and is then registered as an `nn.Parameter`; the random perturbation is not resampled during subsequent forward passes. Because $M$ assigns a trainable coefficient to every channel and two-dimensional frequency coordinate, components at the same radial frequency but different orientations can receive different weights. Together with the residual path, the effective response is $1 + \gamma\mathcal{R}_{H,W}(M)$ and can therefore increase or decrease individual spectral components. The learnable scalar gate $\gamma$, initialized to $0.1$, controls the magnitude of this adaptive correction, and $\Theta_F$ collects the trainable FDFRM parameters.

In implementation, $\mathcal{F}_{\mathbb{R}}$ and $\mathcal{F}_{\mathbb{R}}^{-1}$ are realized by the rFFT2/irFFT2 pair. The synthesis operator interprets the nonredundant half-spectrum under Hermitian symmetry and returns a real-valued feature, so no additional complex constraint or projection is required. A single parameter set $\Theta_F$ is shared by the synthetic and measured domains through their common encoder $f_{EF}$. Rather than serving as a fixed Fourier preprocessing step, this shared trainable operator couples the two domains through the same adaptive frequency response and places their latent representations in a common, spectrally controllable coordinate system while preserving domain-specific content for the diffusion bridge. We refer to the resulting representation as the frequency-coupled latent space.

### C. Forward Bridge Process From Measured to Synthetic Domain

Unknown reflective signals and speckle noise pose significant difficulties in directly translating synthetic SAR images into measured ones. We therefore construct a latent diffusion bridge between the two domains . It is important to distinguish the direction of this forward bridge from the direction of image translation: the forward bridge is anchored at the measured target and terminates at the synthetic condition, whereas the learned reverse process performs the desired synthetic-to-measured (S2M) translation . We denote a paired synthetic–measured training sample by $(I_S, I_R)$, where $I_S \in X_S$ and $I_R \in X_R$. Their FDFRM-refined latent representations are denoted by $s = f_{\mathrm{EF}}(I_S)$ and $r = f_{\mathrm{EF}}(I_R)$, respectively. We set the measured latent as the bridge origin, $r_0 = r$, use $r_t$ for the intermediate bridge state, and set the synthetic latent as the terminal condition, $r_T = s$ . The forward bridge is defined as

$$Q_{r\to s}\left(r_t \mid r_0, s\right) = \mathcal{N}\left(r_t; (1-m_t)r_0 + m_t s, \delta_t E\right) \tag{3}$$

$$r_0 = r, \qquad m_t = \frac{t}{T}, \qquad \delta_t = 2\delta_{\max} m_t(1-m_t). \tag{4}$$

Here , $T = 1000$ is the total number of bridge steps, $m_t \in [0,1]$ is the interpolation coefficient, $\delta_{\max} = 1$, and $E$ is the identity matrix in the latent space. Equation (3) therefore describes the marginal distribution of the intermediate state $r_t$ conditioned on the measured target latent $r_0$ and the synthetic terminal latent $s$. As $m_t$ increases from 0 to 1, the mean moves from $r_0$ toward $s$. The variance follows a parabolic, rather than monotonically decreasing, schedule: $\delta_0 = \delta_T = 0$, and it reaches its maximum $\delta_t = 0.5$ at $m_t = 0.5$. The zero endpoint variances pin the bridge to $r_0 = r$ and $r_T = s$, whereas the larger interior variance permits stochastic intermediate states. For numerical stability, the implementation evaluates the $T$ scheduled coefficients over $m_t \in [0.001, 0.999]$; hence, the first and last scheduled variances are both $1.998\times 10^{-3}$, while the two endpoint states themselves are imposed exactly.

The intermediate state and its preceding state can be sampled in discrete form as

$$r_t = (1-m_t)r_0 + m_t s + \sqrt{\delta_t}\,\epsilon_t \tag{5}$$

$$r_{t-1} = (1-m_{t-1})r_0 + m_{t-1} s + \sqrt{\delta_{t-1}}\,\epsilon_{t-1} \tag{6}$$

with $\epsilon_t, \epsilon_{t-1} \sim \mathcal{N}(0, I)$. Here, $r_0$ is the measured target latent, $s$ is the synthetic condition latent, and $r_t$ is the noisy bridge state at step $t$. The interpolation coefficient $m_t$ controls the transition of the mean from the measured endpoint to the synthetic endpoint, while $\delta_t$ controls the uncertainty of the intermediate state.

Substituting Eq. (6) into Eq. (5) yields the one-step transition kernel

$$\begin{aligned} &Q_{r\to s}\left(r_t \mid r_{t-1}, s\right) = \\ &\mathcal{N}\left(r_t; \tfrac{1-m_t}{1-m_{t-1}} r_{t-1} + \left(m_t - \tfrac{1-m_t}{1-m_{t-1}} m_{t-1}\right) s, \delta_{t|t-1} E\right) \end{aligned} \tag{7}$$

$$\delta_{t|t-1} = \delta_t - \delta_{t-1}\frac{(1-m_t)^2}{(1-m_{t-1})^2} \tag{8}$$

For the analytical schedule in Eq. (4), substitution into Eq. (8) gives $\delta_{t|t-1} = 2\delta_{\max}(T-t)/[T(T-t+1)] > 0$ for every stochastic intermediate transition $1 \le t \le T-1$; the zero

variance at $t = T$ is the deterministic terminal constraint. Direct evaluation of the numerically clipped schedule gives $\min_t \delta_{t|t-1} = 9.995 \times 10^{-4} > 0$. Therefore, each intermediate transition in Eq. (7) has a valid covariance. The reverse sampler further uses the posterior variance $\widetilde{\delta}_t = \delta_{t|t-1}\delta_{t-1}/\delta_t$ in Eq. (13), as shown in Algorithm 2. At $t = 0$, $m_0 = 0$ and the bridge starts from the measured latent $r_0 = r$. At $t = T$, $m_T = 1$ and the terminal state satisfies $r_T = s$. Thus, the forward bridge runs from the measured target to the synthetic condition; its learned reverse process realizes the required S2M translation.

### *D. Reverse Process From the Synthetic to the Measured SAR Domain*

During training, the paired measured latent $r_0$ defines the target endpoint of the learned reverse process, while the synthetic latent $s$ defines the terminal state of the forward bridge. At inference , the measured target is unavailable: sampling starts from the available synthetic latent, $r_T = s = f_{\mathrm{EF}}(I_S)$, and predicts $r_{t-1}$ from the current state $r_t$ until the measured-domain estimate $r_0$ is recovered :

$$p_\theta\left(r_{t-1} \mid r_t, s\right) = \mathcal{N}\left(r_{t-1}; \mu_\theta(r_t, t; s), \widetilde{\delta}_t E\right) \tag{9}$$

Here, $p_\theta(r_{t-1} \mid r_t, s)$ is the learned reverse transition conditioned on the available synthetic endpoint $s$. The neural network parameterized by $\theta$ predicts the mean $\mu_\theta(r_t, t; s)$, and $\widetilde{\delta}_t E$ is the reverse-process covariance.

### *E. Optimization Objectives*

We use a U-Net to estimate the bridge objective at each reverse step. During training, paired measured and synthetic latents are available to construct the bridge state and its supervision target. During inference, the trained model starts from the synthetic terminal latent and progressively recovers a measured-domain latent .

During optimisation, the network seeks to align the translated output with the measured image by maximising the evidence lower bound (ELBO):

$$\begin{aligned}&\mathrm{ELBO} = -\mathbb{E}\left[D_{\mathrm{KL}}\left(Q_{r\to s}\left(r_T \mid r_0, s\right) \| p\left(r_T \mid s\right)\right)\right] + \\ &\sum_{t=2}^{T} D_{\mathrm{KL}}\left(Q_{r\to s}\left(r_{t-1} \mid r_t, r_0, s\right) \| p_\theta\left(r_{t-1} \mid r_t, s\right)\right) - \\ &\log p_\theta\left(r_0 \mid r_1, s\right)\end{aligned} \tag{10}$$

The first term aligns the terminal bridge distribution with the synthetic condition, the summation matches the learned reverse transitions to the analytical bridge posterior, and the final term measures reconstruction of the measured endpoint $r_0$. Because $r_T = s$, the terminal term is constant with respect to the trainable reverse model .

Maximizing the ELBO trains the reverse process to recover the measured endpoint from noisy bridge states. After reverse sampling, the translated measured-like image is obtained as $\widehat{I}_R = f_{\mathrm{dec}}(r_0)$. Combining Eqs. (3) and (7), the analytical posterior in the second term is

$$\begin{aligned}&Q_{r\to s}(r_{t-1} \mid r_t, r_0, s) = \frac{Q_{r\to s}(r_t|r_{t-1},s)Q_{r\to s}(r_{t-1}|r_0,s)}{Q_{r\to s}(r_t|r_0,s)} \\ &= \mathcal{N}\left(r_{t-1}; \widetilde{\mu}_t(r_t, r_0, s), \widetilde{\delta}_t E\right)\end{aligned} \tag{11}$$

The mean value term is given by:

$$\begin{aligned}&\widetilde{\mu}_t(r_t, r_0, s) = \tfrac{\delta_{t-1}}{\delta_t}\tfrac{1-m_t}{1-m_{t-1}} r_t + (1-m_{t-1})\tfrac{\delta_{t|t-1}}{\delta_t} r_0 \\ &+ \left(m_{t-1} - m_t \tfrac{1-m_t}{1-m_{t-1}}\tfrac{\delta_{t-1}}{\delta_t}\right) s\end{aligned} \tag{12}$$

Below shows the variance term:

$$\tilde{\delta}_t = \tfrac{\delta_{t|t-1}\cdot\delta_{t-1}}{\delta_t} \tag{13}$$

Since the measured endpoint $r_0$ is unknown during inference, we use the standard reparameterization of the bridge objective. Combining Eqs. (3) and (12), the posterior mean can be written as

$$\widetilde{\mu}_t(r_t, s) = c_{rt} r_t + c_{st} s - c_{\epsilon t}\left[m_t(s - r_0) + \sqrt{\delta_t}\,\epsilon\right] \tag{14}$$

$$c_{rt} = \frac{\delta_{t-1}}{\delta_t}\frac{1-m_t}{1-m_{t-1}} + \frac{\delta_{t|t-1}}{\delta_t}(1-m_{t-1}) \tag{15}$$

$$c_{st} = m_{t-1} - m_t \frac{1-m_t}{1-m_{t-1}}\frac{\delta_{t-1}}{\delta_t} \tag{16}$$

$$c_{\epsilon t} = (1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t} \tag{17}$$

Instead of predicting the complete posterior mean, the U-Net estimates the bridge objective $m_t(s-r_0)+\sqrt{\delta_t}\epsilon$. Accordingly , Eq. (14) is parameterized as

$$\mu_\theta(r_t, s, t) = c_{rt} r_t + c_{st} s - c_{\epsilon t}\epsilon_\theta(r_t, t) \tag{18}$$

Thus, we can break down the training goal ELBO in equation 10 into:

$$\mathbb{E}_{r_0,s,t,\epsilon}\left[c_{\epsilon t}\left\|m_t(s-r_0) + \sqrt{\delta_t}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(r_t, t)\right\|^2\right] \tag{19}$$

### *F. MIGCA Mechanism*

During image translation, the uncertainty of reflective signals and speckle noise in the measured SAR image should be given more attention. Therefore, we introduce MIGCA. MIGCA is applied when optimizing the reverse predictor. Both training and inference use the same synthetic terminal latent $s$ as the attention guide.

Let $H_t^{(i)} = \varphi_i(r_t) \in \mathbb{R}^{N\times d_i}$ denote the flattened feature of the current bridge state at the $i$th U-Net layer. The synthetic terminal latent is encoded as $G_S^{(i)} = \eta_\theta^{(i)}(s) \in \mathbb{R}^{M\times d_\eta}$ and retained as the source-side guide throughout the reverse process. At the $i$th layer, the query, key, and value projections are $Q_t^{(i)} = H_t^{(i)} W_q^{(i)}$, $K_S^{(i)} = G_S^{(i)} W_k^{(i)}$, and $V_S^{(i)} = G_S^{(i)} W_v^{(i)}$, respectively. The cross-attention response is then calculated as $\mathcal{A}_\theta^{(i)}(r_t, s) = \mathrm{softmax}\big(Q_t^{(i)}(K_S^{(i)})^{\mathrm{T}}/\sqrt{d}\big)V_S^{(i)}$.

**Algorithm 1** Training

**Require:** Synthetic–measured SAR image pairs $P = \{(I_S^n, I_R^n)\}_{n=1}^N$; bridge steps $T$

**Repeat**

  **FDFRM encoding:**

    $s \leftarrow f_{\mathrm{EF}}(I_S), \quad r_0 \leftarrow f_{\mathrm{EF}}(I_R)$

  Sample forward diffusion index $t \sim \mathcal{U}\{1, \ldots, T\}$

  Sample Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$

  **Form the noisy latent state:**

$$r_t \leftarrow (1 - m_t)r_0 + m_t s + \sqrt{\delta_t}\,\epsilon$$

  **MIGCA-enhanced reverse prediction :**

    $\widetilde{r}_t \leftarrow \mathrm{MIGCA}(r_t, \mathrm{guide} = s)$

    $\widehat{\epsilon} \leftarrow \epsilon_\theta(\widetilde{r}_t, t)$

  **Update** the network parameters by descending

$$\nabla_\theta \left\| m_t(s - r_0) + \sqrt{\delta_t}\,\epsilon - \widehat{\epsilon} \right\|^2$$

**Until converged**

**Algorithm 2** Sampling

**Require:** Synthetic SAR image $I_S$ ; trained LFCS2M parameters $\theta$; bridge steps $T$

**Ensure:** Translated measured-like SAR image $\widehat{I}_R$

$s \leftarrow f_{\mathrm{EF}}(I_S)$

$r_T \leftarrow s$

$g_S \leftarrow s$ *// synthetic-source guide*

**for** $t = T, T-1, \ldots, 1$ **do**

  $\widetilde{r}_t \leftarrow \mathrm{MIGCA}(r_t, \mathrm{guide} = g_S)$

  $\widehat{\epsilon} \leftarrow \epsilon_\theta(\widetilde{r}_t, t)$

  **if** $t > 1$ **then**

    $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

  **else**

    $z \leftarrow \mathbf{0}$

  **end if**

  $r_{t-1} \leftarrow c_{rt} r_t + c_{st} s - c_{\epsilon t}\widehat{\epsilon} + \sqrt{\widetilde{\delta}_t}\, z$

**end for**

$\widehat{I}_R \leftarrow f_{\mathrm{dec}}(r_0)$

**return** $\widehat{I}_R$

Here, $W_q^{(i)} \in \mathbb{R}^{d_i \times d}$, $W_k^{(i)} \in \mathbb{R}^{d_\eta \times d}$, and $W_v^{(i)} \in \mathbb{R}^{d_\eta \times d_i}$ are learnable projections. The response is fused with the current bridge feature through the output projection $\mathcal{P}_o^{(i)}$ and a residual connection, yielding $\widetilde{H}_t^{(i)} = H_t^{(i)} + \mathcal{P}_o^{(i)}\big(\mathcal{A}_\theta^{(i)}(r_t, s)\big)$. For compactness, the resulting enhanced bridge state is denoted by $\widetilde{r}_t = \mathrm{MIGCA}(r_t, s)$, after which the reverse predictor produces $\widehat{\epsilon}_t = \epsilon_\theta(\widetilde{r}_t, t)$.

Specifically, $r_0$ is used only to construct $r_t$ and its supervised objective during training. The MIGCA condition is $s$ in both training and inference. Consequently, sampling starts from $r_T = s = f_{\mathrm{EF}}(I_S)$ and requires neither a measured image nor any target-derived feature.

Let $b_t = m_t(s - r_0) + \sqrt{\delta_t}\boldsymbol{\epsilon}$ denote the analytical bridge target. Accordingly, the objective in Eq. (19) becomes

$$\mathcal{L}_{\mathrm{MIGCA}} = \mathbb{E}_{r_0, s, t, \epsilon}\Big[c_{\epsilon t}\, \|b_t - \epsilon_\theta(\mathrm{MIGCA}(r_t, s), t)\|_2^2\Big]. \tag{20}$$

Equation (20) minimizes the difference between the predicted bridge objective and $b_t$. Since $s$ is available at both stages, the reverse predictor has the same MIGCA input during optimization and sampling. The complete procedures are summarized in Algorithms 1 and 2, respectively.

## IV. EXPERIMENTS

### A. Dataset

The performance of LFCS2M is evaluated using the Synthetically Measured Paired Labeled Experiment (SAMPLE) dataset [18]. The SAMPLE dataset includes the same target categories as the MSTAR dataset [49], encompassing 10 distinct vehicle types. This dataset is created through the use of CAD models and advanced ray tracing techniques. All vehicles are imaged using SAR at a pitch angle between 14° and 18°. The dataset employs SAR images generated by the polar coordinate format algorithm with a 0.3 meters resolution, a 128×128 pixels image size with horizontal polarization (HH), at 10°–80° azimuth angles. Synthetic SAR data are generated through the use of electromagnetic simulation software, with vehicle targets predicted at various azimuth angles.

TABLE I
OVERVIEW OF SAMPLE DATASETS

| Vehicle | Serial | Elevation angles 14°-16° | Elevation angles 17° | Total |
|---|---|---|---|---|
| 2S1 | B01 | 116 | 58 | 174 |
| BMP2 | 9563 | 55 | 52 | 107 |
| BTR70 | C71 | 43 | 49 | 92 |
| M1 | 0AP00N | 78 | 51 | 129 |
| M2 | MV02GX | 75 | 53 | 128 |
| M35 | T839 | 76 | 53 | 129 |
| M548 | C245HAB | 75 | 53 | 128 |
| M60 | 3336 | 116 | 60 | 176 |
| T72 | 812 | 56 | 52 | 108 |
| ZSU23 | D08 | 116 | 58 | 174 |
| Total | | 806 | 539 | 1,345 |

The SAMPLE dataset contains 806 training samples in total and 80 samples on average in each vehicle category. An illustrative example from the dataset is presented in Fig. 3, while the dataset's characteristics are summarized in Table I.

### B. Implementation Details

The training process of LFCS2M consists of 1000 time steps, with the application of linear noise scheduling. With regard to the specifics of the U-Net architecture design, the model employs a multi-head attention mechanism on feature maps with resolutions of 32, 16, and 8 to capture global information. The latent space is of a size of 64×64 after encoding. Then, 200,000 iterations were performed with a batch size of 4, while setting the initial learning rate at $4 \times 10^{-5}$, with an EMA decay strategy starting from iteration 30,000. The AdamW optimizer was employed with $\beta_1 = 0.9$, $\beta_2 = 0.999$. All Img2Img methods discussed were re-trained from scratch on the SAMPLE training dataset.

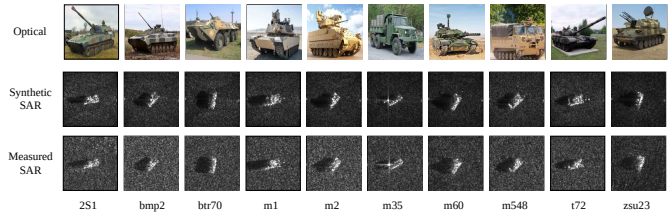



Fig. 3. Examples of SAMPLE dataset, listing 10 targets leftward to rightward, in order from top: optical image of the target; synthetic SAR image; measured SAR image.

### C. Evaluation Protocol

Both qualitative and quantitative analyses are applied to appraise the translated outputs, as outlined in Fig. 4. The qualitative assessment includes visual comparison of the translated images and t-SNE analysis of domain distributions. The quantitative evaluation proceeds from three complementary perspectives. First, a frequency-domain analysis compares the power spectra of simulated, translated, and measured chips, providing a physics-grounded assessment of whether the translation closes the spectral gap identified in the introduction. Second, perceptual similarity to the measured ground truth is quantified by FID [50], DISTS [51], and LPIPS [52]. FID employs a pretrained Inception v3 network to extract high-dimensional features from both generated and real images, treats them as multivariate Gaussian distributions, and computes the Fréchet distance between the two distributions. LPIPS computes the difference between two sets of images in the feature space of a deep neural network. DISTS simultaneously measures structural and texture similarity, providing an overall perceptual similarity score. Third, the translated images are used as training data for SAR target classification networks, and recognition accuracy on measured test images is compared against measured and synthetic bounds. To evaluate recognition performance we employ three classifiers that represent distinct architectural paradigms: a deep convolutional network VGG16 [53], denoted $\mathcal{R}_1$; a residual network ResNet18 [54], denoted $\mathcal{R}_2$; and the Vision Transformer ViT [55], denoted $\mathcal{R}_3$.

### D. Qualitative evaluation

For experimental convenience, we trained the entire SAMPLE dataset by directly pairing measured target slices with synthetic target slices, distributing them into three sets: training set(60%), test set(30%), and validation set(10%) respectively. We employed CycleGAN [34], Pix2Pix [33], CRN [31], AdaIN [56], SPADE [32], and ASAPNet [30] as benchmarks, as depicted in Fig.5. It was observed that while GANs like CycleGAN can produce backgrounds that resemble the actual SAR image distributions, they still blur the finer details of the targets. The other five methods also perform suboptimally in both background distribution and detailed target reconstruction. In contrast, our method convincingly reproduces not only the intricate target details but also the background clutter of the measured images, resulting in more natural-looking images. For comparison, we further evaluated three

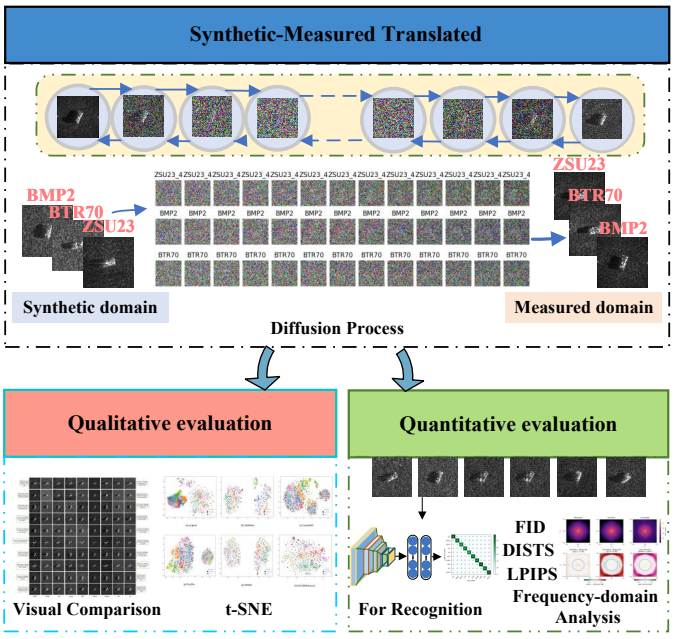



Fig. 4. Comprehensive qualitative and quantitative evaluation of synthetic-to-measured images.

diffusion-based baselines: conditional DDPM (cond.DDPM) [57], MIDMs [58], and LDM [26]. As illustrated in Fig. 6, images generated by LFCS2M remain visually closest to the measured SAR data.

The alignment of distributions between generated and real data is analyzed using t-SNE [59]. Each image is vectorized (flattened) and projected to 2D, with colors indicating target classes and shapes indicating source domain. Fig.7 shows that, initially, the synthetic and measured samples form distinct clusters, reflecting a domain gap. After translation by LFCS2M, the two clusters overlap significantly. This indicates that LFCS2M effectively narrows the distribution gap between synthetic and measured SAR images, which is beneficial for using synthetic data to improve classification on real measured targets.

### E. Quantitative evaluation

*1) Frequency-Domain Characterization:* To test whether LFCS2M reduces the spectral component of the S2M gap, we compare the azimuthally averaged power spectra of paired synthetic and measured chips and report the discrepancy in decibels relative to the measured reference. Before the FFT, each chip is apodized with a two-dimensional Hann window to suppress spectral leakage. Averaged over all SAMPLE image pairs, the synthetic SAR chips exhibit systematic excess power at low and lower intermediate spatial frequencies, with a peak S2M power ratio of 2.20 at $f \approx 0.035$ cyc/px. (Fig. 8). The two-dimensional difference map places this excess in a central lobe associated with low spatial frequencies, the band that contains most of the scene energy and the target's discriminative structure. This bias is consistent across targets: every SAMPLE vehicle shows the same excess power at low frequencies, ranging from about 2 dB for the worst reproduced targets to 0.8 dB for zsu23.

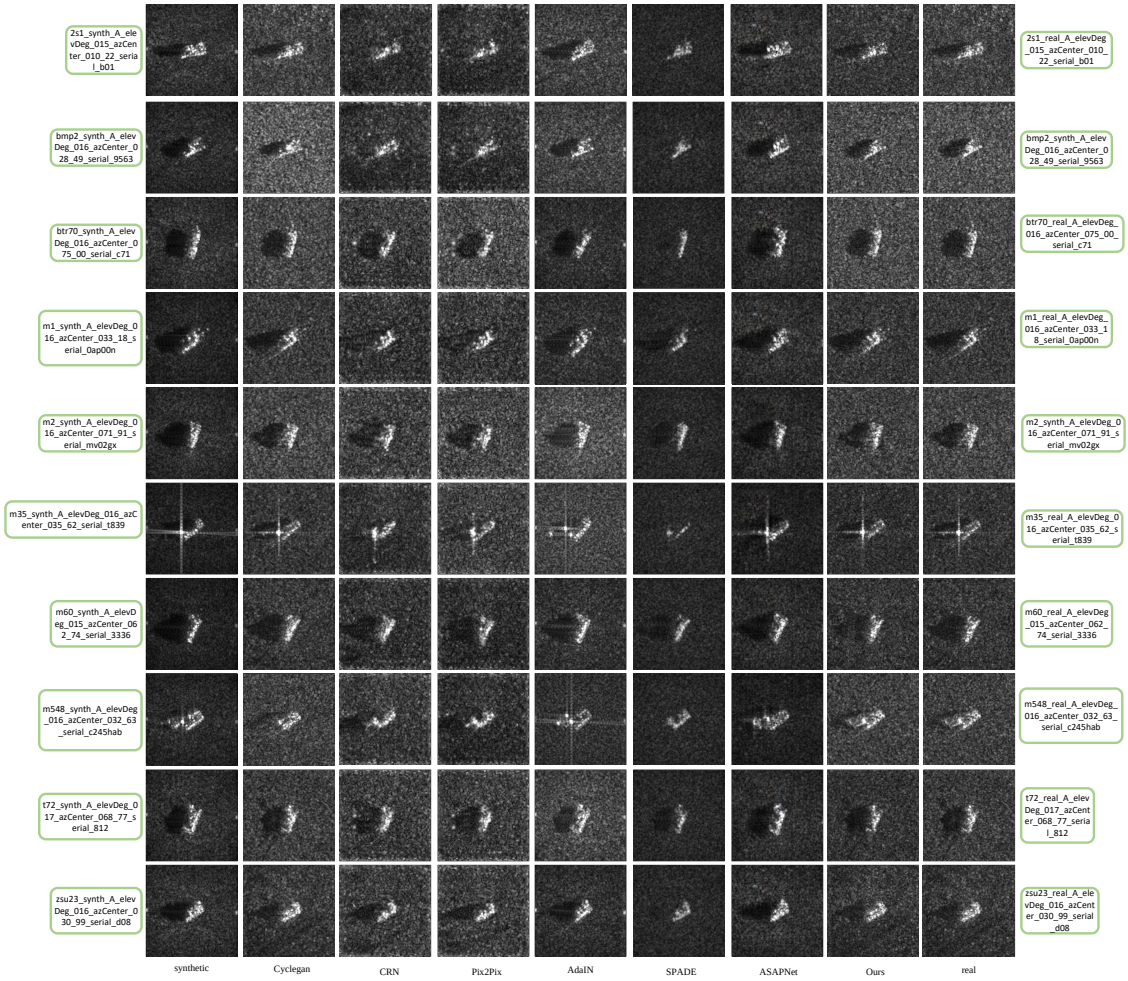



Fig. 5. Visual comparison between LFCS2M and GAN-based Img2Img translation methods on the SAMPLE dataset. LFCS2M yields images that most closely resemble the measured SAR data.

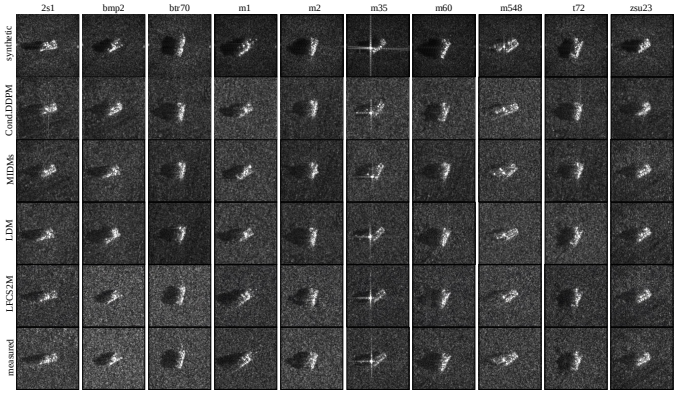



Fig. 6. Visual comparison of LFCS2M and diffusion-based methods (cond. DDPM, MIDMs, and LDM) for SAR image translation on the SAMPLE dataset. LFCS2M yields images that most closely resemble the measured SAR data.

After translation by LFCS2M, the spectral gap is largely closed. Over the structural band $|f| \leq 0.35$ cyc/px, which contains $99.99\%$ of the measured fluctuation energy, the mean gap across classes decreases from $1.71$ dB to $0.19$ dB, an $88.6\%$ reduction (Fig. 9). The two-dimensional difference maps in Fig. 10 show that this improvement is consistent across all ten vehicle classes, with the lobe associated with low spatial frequencies largely removed after translation. The gap computed with energy weighting does not require a band cutoff and decreases from $0.308$ dB to $0.005$ dB, representing a reduction of $98.3\%$. The correction is concentrated at low and lower intermediate frequencies. This distribution is learned from the domain discrepancy in SAMPLE rather than imposed by a predefined spectral prior in FDFRM. These results indicate that LFCS2M substantially reduces the structured and repeatable frequency-domain gap between synthetic and measured SAR chips in the band most relevant to recognition.

*2) Similarity evaluation:* The similarity evaluation of the translated and measured images first requires an introduction of standard Img2Img translation evaluation metrics because translated images are generated through deep neural networks. Therefore, three metrics are launched to comprehensively measure the quality of translated data: including FID [50], LPIPS [52], and DISTS [51]. FID is broadly applied to appraise the generated models. It works with a pre-trained Inception v3 network, which can withdraw high-dimensional characteristics from images, and then treats these features as a multivariate Gaussian distribution. It can calculate the Fréchet distance (i.e.

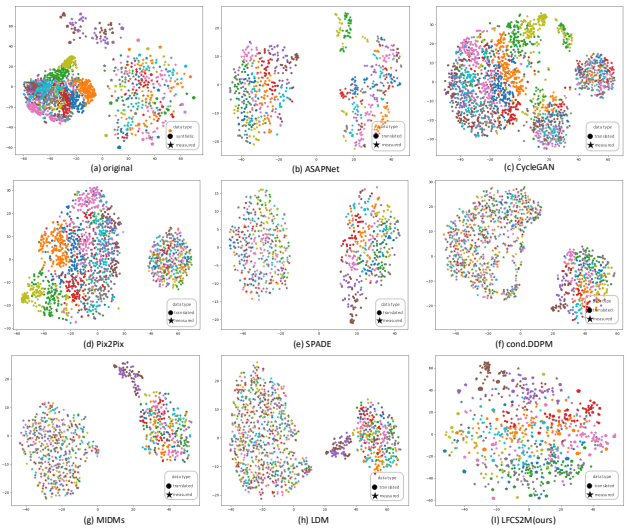



Fig. 7. t-SNE 2D Visualization. (a) Original SAMPLE dataset; (b) SAR images translated by ASAPNet and the measured dataset; (c) SAR images translated by CycleGAN and the measured dataset; (d) SAR images translated by Pix2Pix and the measured dataset; (e) SAR images translated by SPADE and the measured dataset; (f) SAR images translated by cond.DDPM and the measured dataset; (g) SAR images translated by MIDMs and the measured dataset; (h) SAR images translated by LDM and the measured dataset; (i) SAR images translated by LFCS2M and the measured dataset.

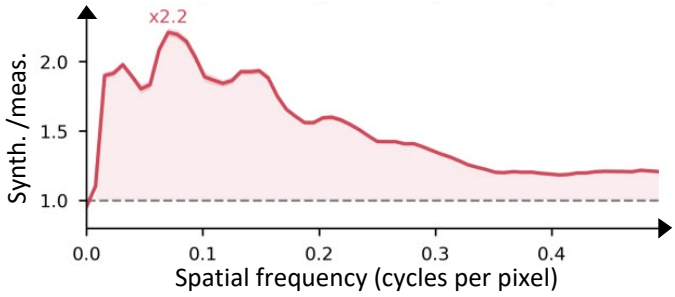



Fig. 8. Azimuthally-averaged power spectrum of synthetic and measured chips for SAMPLE.

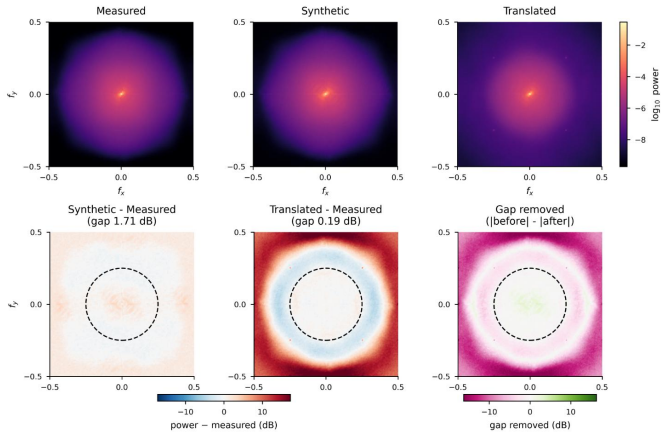



Fig. 9. Azimuthally-averaged power spectrum gap between synthetic and measured chips before and after LFCS2M translation.

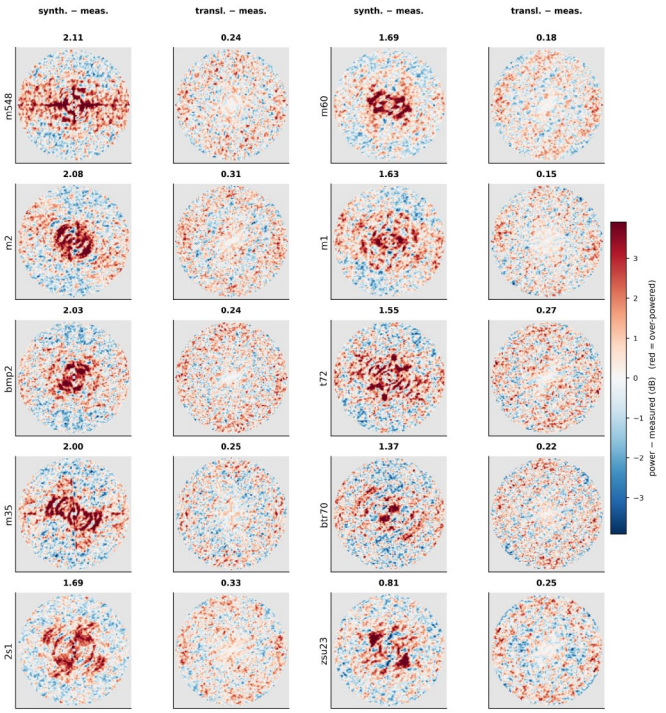



Fig. 10. Classwise two dimensional spectral gap before and after LFCS2M translation, evaluated in the structural band $|f| \leq 0.35$. Smaller values indicate lower absolute spectral gaps in dB. synth.: synthetic; meas.: measured.

Wasserstein-2 distance) between the feature distributions of these two types of images, for the purpose of quantifying the difference between them. LPIPS is designed to compute the difference between two sets of images in feature space, while using features extracted by deep neural networks. DISTS is proposed to concurrently measure the structural and texture similarity of images, reflecting the perceptual differences between them and providing an overall similarity measure. Three metrics comprise the image recognition assessment of this paper, which evaluates the generated images against the measured images. The experimental results are shown on Tab.II. BBDM [27] is additionally included as the bridge-only baseline, allowing the contributions of FDFRM and MIGCA to be evaluated beyond the underlying bridge framework.

*3) Analysis of Practical Scene Recognition:* We trained SAR target classification networks on LFCS2M-translated images and evaluated them on measured test images. The expected performance lies between an upper measured bound (training on all measured images) and a lower synthetic bound (training on synthetic data). If LFCS2M translations are accurate, classification accuracy should approach the measured bound; otherwise any discrepancy indicates domain differences. We defined five experimental scenarios (Table III), following the framework proposed in [44], to systematically assess LFCS2M under limited measured-data conditions. Each scenario varies which target classes and azimuth angles are seen during translation. For recognition, we used three representative networks that represent distinct architectural paradigms: a deep convolutional network VGG16 [53], denoted $\mathcal{R}_1$; a residual network ResNet18 [54], denoted $\mathcal{R}_2$;

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS USING FID, DISTS, AND LPIPS METRICS

| Methods | FID ↓ | DISTS ↓ | LPIPS ↓ |
|---|---|---|---|
| CycleGAN | 57.23 | 0.413 | 1.226 |
| Pix2Pix | 38.98 | 0.460 | 1.033 |
| CRN | 155.90 | 0.839 | 1.561 |
| AdaIN | 55.39 | 0.681 | 0.910 |
| SPADE | 45.41 | 0.720 | 1.128 |
| ASAPNet | 47.26 | 0.356 | 0.831 |
| cond.DDPM | 28.32 | 0.343 | 0.789 |
| MIDMs | 24.87 | 0.085 | 0.731 |
| BBDM | 25.95 | 0.884 | 0.620 |
| LDM | 33.51 | 0.523 | 0.785 |
| LFCS2M (Ours) | 19.57 | 0.115 | 0.477 |

and the Vision Transformer ViT [55], denoted $\mathcal{R}_3$. In all subsequent tables and figures, $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ refer to these three networks, respectively.

**With seen Classes and Azimuth angles Scenario:** This scenario serves as a baseline to verify that LFCS2M accurately translates images across the full range of trained classes and angles. Table IV compares classification accuracy across methods. All networks show a large gap between the measured and synthetic bounds: VGG16 and ResNet18 have gaps of 55.12% and 50.97% respectively, while ViT's gap is smaller (though its absolute measured-bound accuracy is lower). Notably, training on LFCS2M-translated images yields the highest accuracy on measured test images among all methods, approaching the measured-bound. This underscores that LFCS2M outputs closely match the distribution of measured SAR targets. In ResNet18's confusion matrix (Fig.11(a)), almost all classes are correctly recognized; only a few M1 targets are mistaken for BMP2, likely due to their similar scattering signatures after translation.

**Unseen Proximate Azimuth angles Scenario:** This scenario tests LFCS2M's ability to interpolate to unseen angles close to those in the training set. Table IV shows a slight accuracy drop for all networks (due to the smaller training set), but the LFCS2M-trained models again achieve accuracy near the measured-bound. ResNet18 yields 95.72% accuracy and ViT 92.08% (the latter nearly matching the measured-bound of 92.25%). In ResNet18's confusion matrix (Fig.11(b)), almost all classes are predicted correctly, with only a few M2→BMP2 and T72→M60 errors. These minor confusions reflect the remaining similarities in LFCS2M-translated images for those classes.

**Unseen Azimuth angles Scenario:** This examines whether LFCS2M can extrapolate to new azimuths where target scattering patterns differ significantly. TableIV compares methods for Scenario3. The measured and synthetic bounds are similar to Scenarios 1–2 (comparable sample size). However, LFCS2M induces only modest accuracy drops (4.12%, 3.46%, and 3.00% for VGG16, ResNet18, and ViT, respectively) compared to Scenario2, whereas Pix2Pix and CycleGAN suffer much larger losses (especially VGG16). In ResNet18's confusion matrix (Fig.11(c)), most classes remain distinct, with slight confusions in M2, M548, and T72. These findings indicate that LFCS2M successfully synthesizes measured-like SAR images at new angles, enabling effective classification even without measured training data at those angles.

**Unseen Classes Scenario:** This evaluates LFCS2M's ability to translate entirely new target categories. Table V summarizes results. With fewer classes, the synthetic-bound improves (higher baseline accuracy) while the measured-bound stays similar. ResNet18 trained on LFCS2M outputs achieves 93.92% accuracy, 25.03% above the synthetic-bound and 18.40% higher than Pix2Pix. It is only 4.81% below the measured-bound, confirming strong generalization to new classes. The ResNet18 confusion matrix (Fig.11(d)) shows mostly correct classifications; class M2 has the most misclassifications. Overall, other classes are well separated, suggesting that LFCS2M translations of unseen classes closely match the statistics of measured SAR images.

**Unseen Categories and Azimuth Angles:** This scenario rigorously tests LFCS2M's generalization under both novel classes and viewing conditions. TableV shows very large performance gaps: the synthetic-bound is far below the measured-bound (e.g. 56.29% and 53.58% gaps for VGG16 and ResNet18). Under baseline translations (Pix2Pix, CycleGAN), ResNet18's accuracy remains in the 40–60% range. In contrast, LFCS2M-trained ResNet18 achieves about 80% accuracy (up to 88.79%), demonstrating strong performance in this hardest setting. The confusion matrix (Fig.11(e)) shows that M2 is the most challenging class: some M2 samples are misclassified as M35, M60, or ZSU23. Errors among the remaining classes are limited, with some T72 samples misclassified as M35 or M60. Overall, Scenario 5 confirms that LFCS2M can produce realistic SAR images even for entirely new target classes and azimuth angles.

### *F. Ablation Studies*

We conduct ablation experiments to assess the contributions of the FDFRM and MIGCA modules in LFCS2M, first considering perceptual similarity metrics and target recognition performance under Scenario 1. As reported in Table VI, the bridge-only BBDM [27] is used as the baseline, and MIGCA complements FDFRM by conditioning each current bridge feature on the fixed synthetic terminal latent during both optimization and sampling. Relative to BBDM, the full LFCS2M reduces FID from 25.95 to 19.57, DISTS from 0.884 to 0.115, and LPIPS from 0.620 to 0.477, while increasing ResNet18 accuracy from 93.45% to 98.12%. This controlled comparison shows that the improvements arise from FDFRM and MIGCA rather than from the bridge framework alone.

*1) Effectiveness of FDFRM:* To evaluate the impact of the frequency-domain feature refinement module (FDFRM), we compare the training losses of LFCS2M with and without FDFRM in Fig. 12. The FDFRM variant exhibits larger fluctuations during early training but subsequently converges to a lower loss (approximately 0.3) than the variant without

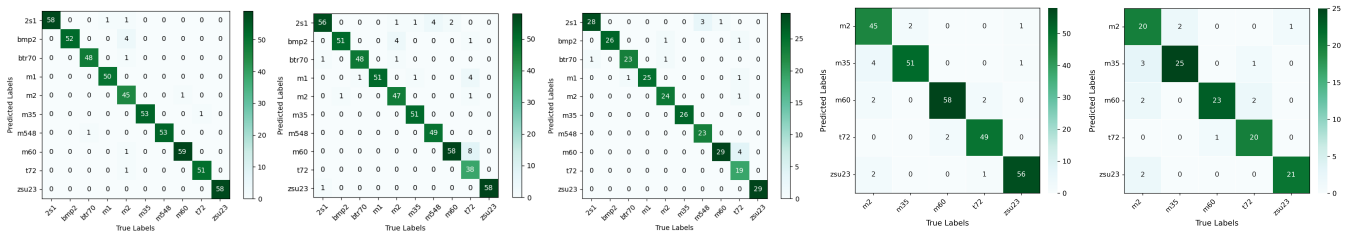
TABLE III
EXPERIMENTAL SETUP FOR SAR TARGET TRANSLATION AND RECOGNITION PHASES ACROSS FIVE SCENARIOS

| Scenario | S2M SAR Target translation Phase | | | | | Translated SAR Target Images for Recognition Phase | | | | | Purpose |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Type | Category | Elevation | Azimuth | Number of Images | Type | Category | Elevation | Azimuth | Number of Images | |
| 1 | Train | 10 | 14°–16° | 10°–80° | 806 | Train/translated | 10 | 14°–16° | 10°–80° | 806 | With seen classes and azimuth angles |
| | Test | 10 | 14°–16° | 10°–80° | 806 | Test/measured | 10 | 17° | 10°–80° | 539 | |
| 2 | Train | 10 | 14°–16° | 2/3 of 10°–80° | 538 | Train/translated | 10 | 14°–16° | 1/3 held out | 268 | Unseen proximate azimuth angles |
| | Test | 10 | 14°–16° | 1/3 held out | 268 | Test/measured | 10 | 17° | 10°–80° | 539 | |
| 3 | Train | 10 | 14°–16° | 10°–45° | 516 | Train/translated | 10 | 14°–16° | 46°–80° | 290 | Unseen azimuth angles |
| | Test | 10 | 14°–16° | 46°–80° | 290 | Test/measured | 10 | 17° | 46°–80° | 247 | |
| 4 | Train | Seen 5 | 14°–16° | 10°–80° | 367 | Train/translated | 10 | 14°–16° | 10°–80° | 439 | Unseen classes |
| | Test | Held-out 5 | 14°–16° | 10°–80° | 439 | Test/measured | 10 | 17° | 46°–80° | 276 | |
| 5 | Train | Seen 5 | 14°–16° | 10°–45° | 239 | Train/translated | 10 | 14°–16° | 46°–80° | 162 | Unseen categories and azimuth angles |
| | Test | Held-out 5 | 14°–16° | 46°–80° | 162 | Test/measured | 10 | 17° | 46°–80° | 123 | |

In Scenarios 4 and 5, the translation network is trained on five seen classes (2S1, BMP2, BTR70, M1, and M548), while M2, M35, M60, T72, and ZSU23 are held out.

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT METHODS IN SCENARIOS 1 TO 3. FOR REFERENCE, MEASURED AND SYNTHETIC BOUNDARIES ARE USED DIRECTLY FOR TARGET RECOGNITION WITH BOTH MEASURED AND SYNTHETIC DATA, FACILITATING A COMPARISON OF THE EFFECTIVENESS OF USING TRANSLATED SAR IMAGES IN ATR

| Methods | Training Data Types | Scenario 1 | | | Scenario 2 | | | Scenario 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathcal{R}_1$ | $\mathcal{R}_2$ | $\mathcal{R}_3$ | $\mathcal{R}_1$ | $\mathcal{R}_2$ | $\mathcal{R}_3$ | $\mathcal{R}_1$ | $\mathcal{R}_2$ | $\mathcal{R}_3$ |
| Mea-Bound | Measured | 98.70 | 98.86 | 92.61 | 96.70 | 98.08 | 92.25 | 96.45 | 98.38 | 91.82 |
| Synt-Bound | Synthetic | 43.58 | 47.89 | 58.30 | 42.83 | 45.38 | 52.47 | 42.56 | 43.89 | 46.70 |
| CycleGAN | Translated | 45.10 | 53.92 | 68.80 | 45.10 | 53.65 | 68.73 | 44.53 | 51.25 | 58.93 |
| Pix2Pix | Translated | 49.52 | 46.85 | 63.84 | 49.52 | 46.79 | 62.80 | 46.82 | 55.60 | 57.84 |
| CRN | Translated | 51.32 | 49.55 | 63.92 | 51.32 | 49.21 | 63.25 | 45.79 | 45.81 | 59.79 |
| AdaIN | Translated | 51.20 | 56.53 | 65.81 | 51.20 | 55.53 | 65.77 | 50.71 | 52.92 | 62.50 |
| SPADE | Translated | 65.85 | 48.79 | 73.62 | 65.85 | 48.60 | 71.94 | 63.38 | 46.46 | 65.11 |
| ASAPNet | Translated | 52.61 | 68.83 | 65.33 | 52.61 | 68.80 | 65.30 | 48.97 | 59.85 | 58.39 |
| cond.DDPM | Translated | 73.28 | 80.95 | 78.76 | 70.09 | 78.55 | 76.28 | 68.33 | 73.52 | 71.80 |
| MIDMs | Translated | 83.66 | 86.05 | 86.36 | 85.20 | 82.83 | 82.06 | 76.92 | 81.51 | 81.95 |
| BBDM | Translated | 91.58 | 93.45 | 87.33 | 88.45 | 86.52 | 85.58 | 83.06 | 84.36 | 82.50 |
| LDM | Translated | 79.83 | 81.63 | 80.82 | 82.06 | 82.65 | 82.30 | 78.85 | 78.88 | 78.55 |
| LFCS2M (Ours) | Translated | 96.56 | 98.12 | 90.15 | 93.35 | 95.72 | 92.08 | 92.33 | 94.92 | 88.82 |




(a) (b) (c) (d) (e)

Fig. 11. Confusion matrix plot of recognition results for five scenarios.

TABLE V

PERFORMANCE COMPARISON OF VARIOUS METHODS IN SCENARIOS 4 AND 5. FOR REFERENCE, MEASURED AND SYNTHETIC BOUNDARIES ARE USED DIRECTLY FOR TARGET RECOGNITION WITH BOTH MEASURED AND SYNTHETIC DATA, FACILITATING A COMPARISON OF THE EFFECTIVENESS OF USING TRANSLATED IMAGES IN ATR

| Methods | Training Data Types | Scenario 4 | | | Scenario 5 | | |
|---|---|---|---|---|---|---|---|
| | | $\mathcal{R}_1$ | $\mathcal{R}_2$ | $\mathcal{R}_3$ | $\mathcal{R}_1$ | $\mathcal{R}_2$ | $\mathcal{R}_3$ |
| Mea-Bound | Measured | 97.86 | 98.73 | 92.14 | 96.91 | 97.38 | 92.14 |
| Synt-Bound | Synthetic | 65.41 | 68.89 | 70.32 | 40.62 | 43.85 | 56.37 |
| CycleGAN | Translated | 69.53 | 72.86 | 75.81 | 48.81 | 50.93 | 57.15 |
| Pix2Pix | Translated | 70.30 | 75.52 | 72.95 | 45.92 | 48.33 | 52.80 |
| CRN | Translated | 70.56 | 69.11 | 72.27 | 47.73 | 47.27 | 53.56 |
| AdaIN | Translated | 71.26 | 76.10 | 73.45 | 48.61 | 52.57 | 65.20 |
| SPADE | Translated | 75.25 | 77.24 | 70.51 | 58.43 | 45.59 | 53.86 |
| ASAPNet | Translated | 71.88 | 75.90 | 73.23 | 49.12 | 61.85 | 60.93 |
| cond.DDPM | Translated | 83.81 | 85.38 | 84.62 | 78.59 | 80.56 | 80.89 |
| MIDMs | Translated | 84.92 | 86.86 | 85.97 | 83.45 | 84.78 | 76.94 |
| BBDM | Translated | 87.49 | 88.71 | 83.02 | 83.84 | 85.95 | 78.32 |
| LDM | Translated | 81.73 | 80.68 | 82.85 | 76.25 | 78.61 | 72.36 |
| LFCS2M (Ours) | Translated | 92.28 | 93.92 | 85.13 | 86.85 | 88.79 | 83.55 |

TABLE VI

COMPARISON WITH THE BBDM BASELINE AND ABLATION OF FDFRM AND MIGCA ON PARAMETERS AND SCENARIO 1 PERFORMANCE

| Variant | FDFRM | MIGCA | Params | Trainable Params | FID | DISTS | LPIPS | $\mathcal{R}_1$ | $\mathcal{R}_2$ | $\mathcal{R}_3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| BBDM baseline | × | × | 292.42M | 228.52M | 25.95 | 0.884 | 0.620 | 91.58 | 93.45 | 87.33 |
| BBDM + MIGCA | × | ✓ | 294.58M | 237.09M | 24.69 | 0.436 | 0.620 | 92.79 | 95.28 | 88.95 |
| BBDM + FDFRM | ✓ | × | 294.83M | 229.36M | 21.06 | 0.861 | 0.594 | 92.27 | 94.50 | 87.41 |
| LFCS2M | ✓ | ✓ | 300.80M | 238.06M | 19.57 | 0.115 | 0.477 | 96.56 | 98.12 | 90.15 |

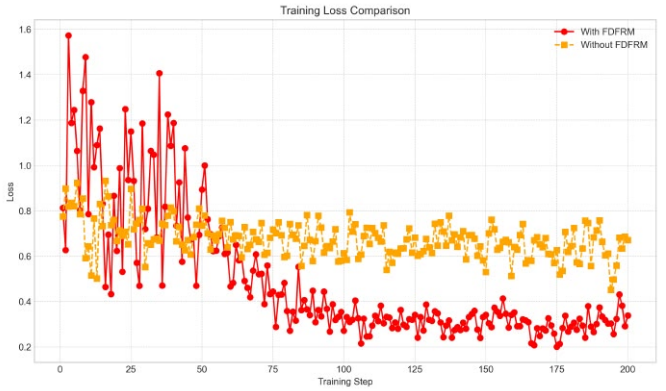



Fig. 12. Training loss values for LFCS2M with and without the FDFRM.

FDFRM (approximately 0.6). This comparison shows that the learned spectral reweighting changes the optimization trajectory and improves the final objective.

Figure 13 further compares the early translated outputs. Without FDFRM, the outputs retain pronounced color artifacts and approach the measured-SAR appearance more slowly . With FDFRM, the outputs acquire measured-domain intensity and speckle characteristics earlier and are visually closer to the measured samples by the final displayed stage. Together with the spectral analysis in Figs. 8–10, this behavior supports interpreting FDFRM as adaptive full-spectrum refinement.

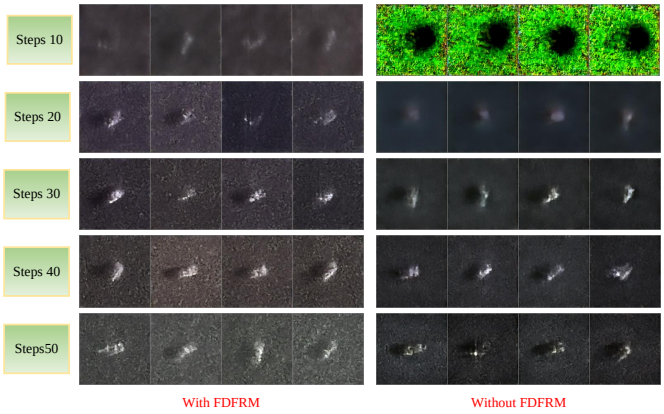



Fig. 13. Visualization of LFCS2M-generated images with and without FDFRM for the first 50 epochs of training.

*2) Impact of the MIGCA Mechanism:* MIGCA enhances the learned reverse process by using the synthetic terminal latent $s$ as a fixed source condition for the evolving bridge state $r_t$. The same condition is used during training and inference; the paired measured latent contributes only to the training bridge state and regression target . Figure 14 presents the FID, DISTS, and LPIPS values for translated images with and without MIGCA. The results show that images translated with MIGCA consistently exhibit superior fidelity throughout the training process.

Figure 15 compares the translated images, showing that MIGCA improves fidelity by retaining background clutter, cross-scatter spots, and other subtle features of the measured SAR images. For instance, when translating M35 without MIGCA, the resulting image exhibits distorted cross-scatter spots, which are absent in the measured and synthetic images. With MIGCA, these features are preserved, indicating the

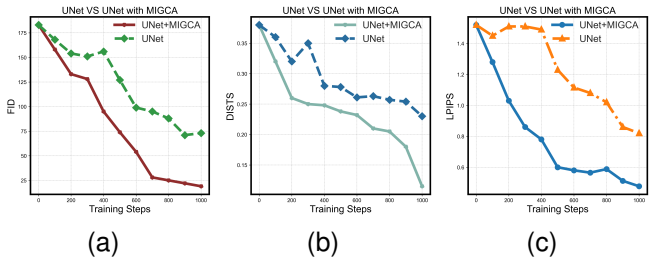



Fig. 14. Analysis of the FID, DISTS, and LPIPS Values of Translated Measured SAR Images Using the UNet Network with and without the MIGCA Mechanism.

mechanism's role in improving translation quality.

## V. Discussion

### A. Complementary Roles of FDFRM and MIGCA

The ablation results reveal that FDFRM and MIGCA address different aspects of the S2M domain discrepancy. Relative to the bridge-only BBDM baseline, FDFRM alone reduces FID from 25.95 to 21.06 and LPIPS from 0.620 to 0.594, whereas MIGCA alone produces a larger reduction in DISTS, from 0.884 to 0.436. This difference is consistent with their respective functions: FDFRM adaptively reweights the two-dimensional latent spectrum, while MIGCA conditions the evolving bridge state on the fixed synthetic terminal latent to retain source-dependent target structure during translation. Combining both modules yields the lowest FID, DISTS, and LPIPS values of 19.57, 0.115, and 0.477, respectively, and increases ResNet18 accuracy from 93.45% for the bridge-only baseline to 98.12%. The joint improvement across distributional, structural, perceptual, and recognition measures indicates that the two modules are complementary rather than redundant.

The optimization and visualization results provide further insight into this complementarity. FDFRM changes the early optimization trajectory and converges to a lower training objective, while the translated outputs acquire measured-domain intensity and speckle characteristics earlier. This behavior should be interpreted as adaptive full-spectrum refinement: the dominant correction at low and lower intermediate frequencies observed on SAMPLE is learned from the data. MIGCA, in contrast, reduces distortions in localized patterns such as clutter and cross-scatter spots. Together, these observations explain why combining spectral reweighting with source-conditioned bridge denoising produces a larger gain than either component alone.

### B. Relation Between Translation Fidelity and Recognition

The three evaluation levels, comprising frequency statistics, perceptual similarity, and downstream recognition, provide mutually supporting evidence. Spectral analysis shows that LFCS2M reduces the mean gap over the structural band from 1.71 dB to 0.19 dB and the energy-weighted gap from 0.308 dB to 0.005 dB. Consistent with this reduction, LFCS2M achieves competitive similarity performance, with

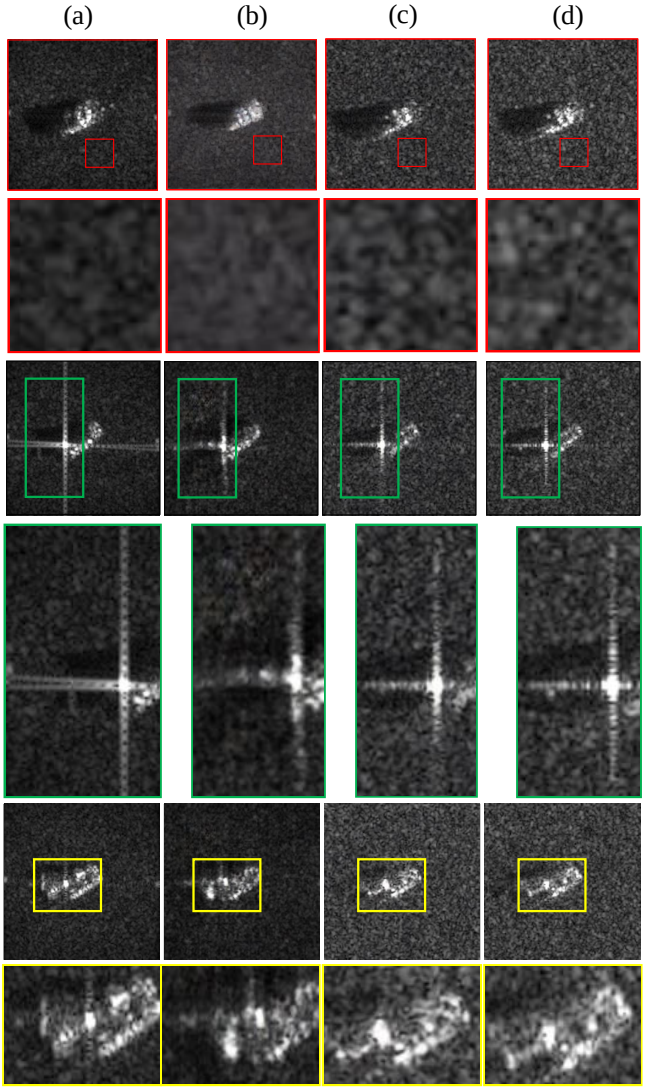



Fig. 15. Comparison of the impact of the MIGCA mechanism. (a) Synthetic SAR target; (b) Translated image of LFCS2M without the MIGCA mechanism; (c) Translated image of LFCS2M with the MIGCA mechanism; (d) Measured SAR target.

FID, DISTS, and LPIPS values of 19.57, 0.115, and 0.477, respectively. More importantly for SAR ATR, the translated images remain useful across three classifiers with different architectures, indicating that the benefit is not tied to a single recognition backbone.

The five recognition scenarios further clarify the scope of this generalization. In Scenario 1, ResNet18 trained on LFCS2M translations reaches 98.12%, close to the 98.86% measured bound. Accuracy decreases gradually as unseen azimuths and categories are introduced, but LFCS2M retains 93.92% when only the classes are unseen and 88.79% when both the classes and azimuth angles are unseen. In each case it remains above the synthetic bound and the competing translation methods.

### C. Limitations

The current results are bounded by two considerations. First, the most difficult Scenario 5 retains a gap relative to the

measured bound, showing that extrapolation to simultaneously unseen categories and azimuth angles remains more challenging than interpolation within observed conditions. Second, LFCS2M currently requires paired synthetic and measured samples to construct the diffusion bridge and its training target. Future work will therefore investigate translation under unpaired data and more parameter-efficient frequency-coupled designs while retaining the observed recognition benefits.

## VI. Conclusion

This article presents LFCS2M to reduce the domain discrepancy between synthetic and measured SAR data through latent diffusion in a frequency-coupled latent space. FDFRM applies a learnable two-dimensional spectral mask that can amplify or attenuate individual frequency coordinates and directional components without imposing a fixed band-selection prior. The resulting synthetic latent representation is transported toward its measured counterpart through a diffusion bridge optimized with MIGCA. Similarity analysis, target recognition experiments , and frequency-domain evaluation show that LFCS2M reduces the domain discrepancy while preserving target structure. In particular, the dominant spectral gap observed at low and lower intermediate frequencies in the SAMPLE dataset is substantially reduced after translation, consistent with how FDFRM learns an adaptive response over the full spectrum from the data . The translated images can therefore augment measured-domain training data for SAR ATR while reducing the need for additional measured acquisitions.

## References


[1] P. Lang, X. Fu, J. Dong, H. Yang, J. Yin, J. Yang, and M. Martorella, "Recent advances in deep-learning-based sar image target detection and recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6884–6915, 2025.

[2] X. X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler, "Deep learning meets sar: Concepts, models, pitfalls, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 143–172, 2021.

[3] K. Ni, W. Fang, N. Huang, Q. Liu, and Z. Zheng, "Complex scattering-aware and globally enhanced hybrid network for sar target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, pp. 1–15, 2026.

[4] E. A. Underwood, T. McKechnie, and L. M. Collins, "Data-efficient deep learning for sar automatic target recognition: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, pp. 1–29, 2026.

[5] F. Gao, F. Zhong, R. Lang, J. Wang, J. Sun, and A. Hussain, "Frequency characteristics guided network for few-shot sar target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 25 821–25 832, 2025.

[6] Y. Wen, X. Wang, L. Peng, and Y. Qiao, "A coarse-to-fine hierarchical feature learning for sar automatic target recognition with limited data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 13 646–13 656, 2024.

[7] Y. Zhao, L. Zhao, S. Zhang, L. Liu, K. Ji, and G. Kuang, "Decoupled self-supervised subspace classifier for few-shot class-incremental sar target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 15 845–15 861, 2024.

[8] S. Wang, Y. Wang, H. Liu, Y. Sun, and C. Zhang, "A few-shot sar target recognition method by unifying local classification with feature generation and calibration," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–19, 2024.

[9] J. Slesinski and D. Wierzbicki, "Review of synthetic aperture radar automatic target recognition: A dual perspective on classical and deep learning techniques," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, pp. 1–56, 2025.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[11] Z. Huang, X. Zhang, Z. Tang, F. Xu, M. Datcu, and J. Han, "Generative artificial intelligence meets synthetic aperture radar: A survey," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–44, 2024.

[12] J. Oh and M. Kim, "Peacegan: A gan-based multi-task learning method for sar target image generation with a pose estimator and an auxiliary classifier," *Remote Sens.*, vol. 13, no. 19, p. 3939, 2021.

[13] D. Xiang, Y. Liu, J. Cheng, X. Lu, Y. Xie, and D. Guan, "Sar target recognition with image generation and azimuth angle feature constraints," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 18 561–18 580, 2025.

[14] D. Guan, R. Feng, Y. Xie, H. Ding, Y. Cui, and D. Xiang, "Sar vehicle data generation with scattering features for target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 5520–5538, 2026.

[15] B. Jones, A. Ahmadibeni, and A. Shirkhodaie, "Physics-based simulated sar imagery generation of vehicles for deep learning applications," in *Proc. SPIE 11511, Applications of Machine Learning 2020*, vol. 11511. SPIE, 2020, p. 115110T.

[16] N. Inkawhich, M. J. Inkawhich, E. K. Davis, U. K. Majumder, E. Tripp, C. Capraro, and Y. Chen, "Bridging a gap in sar-atr: Training on fully synthetic and testing on measured data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2942–2955, 2021.

[17] A. Ahmadibeni, L. Borooshak, B. Jones, and A. Shirkhodaie, "Aerial and ground vehicles synthetic sar dataset generation for automatic target recognition," in *Proc. SPIE 11393, Algorithms for Synthetic Aperture Radar Imagery XXVII*, International Society for Optics and Photonics. SPIE, 2020, p. 113930N. [Online]. Available: https://doi.org/10.1117/12.2558258

[18] B. Lewis, T. Scarnati, E. Sudkamp, J. Nehrbass, S. Rosencrantz, and E. Zelnio, "A sar dataset for atr development: the synthetic and measured paired labeled experiment (sample)," in *Proc. SPIE 10987, Algorithms for Synthetic Aperture Radar Imagery XXVI*, International Society for Optics and Photonics. SPIE, 2019, p. 109870H. [Online]. Available: https://doi.org/10.1117/12.2523460

[19] Y. Sun, L. Lei, D. Guan, X. Li, and G. Kuang, "SAR image speckle reduction based on nonconvex hybrid total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1231–1249, 2021.

[20] Z. Ying, W. Ke, Y. Zhai, Z. Long, J. Zhou, H. Zhu, and C. L. P. Chen, "Diffusar: Frequency domain-aware diffusion model for sar image generation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 11 851–11 866, 2025.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33. Vancouver, Canada: Curran Associates, Inc., 2020, pp. 6840–6851.

[22] Z. Zhang, Z. Zhao, J. Yu, and Q. Tian, "Shiftddpms: exploring conditional diffusion models by shifting diffusion trajectories," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 3, Washington, USA, 2023, pp. 3552–3560.

[23] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," in *in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34. Montreal, Canada: Curran Associates, Inc., 2021, pp. 1415–1428.

[24] Y. Li and T. Balz, "A multiple-level unsupervised domain adaptation network for sar automatic target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 27 857–27 871, 2025.

[25] M. Han, B. Sun, Z. Men, Y. Zhou, and Y. Zhi, "Bistatic sar target recognition using transferred monostatic knowledge and cross-domain feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, pp. 1–18, 2026.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 10 684–10 695.

[27] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Bbdm: Image-to-image translation with brownian bridge diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (CVPR)*, June 2023, pp. 1952–1961.

[28] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Trans. Multimedia*, vol. 24, pp. 3859–3881, 2021.

[29] Y. Sun, L. Lei, and G. Kuang, "Change-prior-guided unsupervised change detection of heterogeneous remote sensing images," *IEEE Trans. Image Process.*, vol. 35, pp. 7119–7133, 2026.

[30] T. R. Shaham, M. Gharbi, R. Zhang, E. Shechtman, and T. Michaeli, "Spatially-adaptive pixelwise networks for fast image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Nashville, TN, USA, June 2021, pp. 14 882–14 891.

[31] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct 2017.
[32] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Beach, CA, USA, June 2019.
[33] 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. (CVPR)*, Honolulu, HI, USA, July 2017, pp. 1125–1134.
[34] 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. (ICCV)*, Venice, Italy, Oct 2017.
[35] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM Trans. Graph.*, 2022, pp. 1–10.
[36] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Vqbb: Image-to-image translation with vector quantized brownian bridge," *arXiv preprint arXiv:2205.07680*, vol. 2, 2022.
[37] M. Zhao, F. Bao, C. LI, and J. Zhu, "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, vol. 35. Curran Associates, Inc., 2022, pp. 3609–3623.
[38] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Trans. Multimedia*, vol. 24, pp. 3859–3881, 2022.
[39] D. Malmgren-Hansen, A. Kusk, J. Dall, A. A. Nielsen, R. Engholm, and H. Skriver, "Improving sar automatic target recognition models with transfer learning from simulated data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1484–1488, 2017.
[40] J. M. Arnold, L. J. Moore, and E. G. Zelnio, "Blending synthetic and measured data using transfer learning for synthetic aperture radar (sar) target classification," in *Proceedings of SPIE*, vol. 10647. SPIE, Apr 2018.
[41] S. R. Sellers, P. J. Collins, and J. A. Jackson, "Augmenting simulations for sar atr neural network training," in *Proc. IEEE Int. Radar Conf. (RADAR)*. IEEE, 2020, pp. 309–314.
[42] B. Lewis, K. Cai, and C. Bullard, "Adversarial training on sar images," in *Proc. SPIE*, vol. 11394. SPIE, 2020, pp. 83–90.
[43] S. Du, J. Hong, Y. Wang, K. Xing, and T. Qiu, "Physical-related feature extraction from simulated sar image based on the adversarial encoding network for data augmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[44] G. Youk and M. Kim, "Transformer-based synthetic-to-measured sar image translation via learning of representational features," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.
[45] L. Liao, G. Zhang, Z. Liu, and Y. Shi, "Complex-valued conditional diffusion model with physical scattering information for sar image target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 16 580–16 595, 2026.
[46] P. Wang, Y. Liu, X. Liang, D. Zhu, X. Gong, Y. Ye, H. F. Lee, and B. Huang, "CIRSM-Net: A cyclic registration network for SAR and optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–19, 2025.
[47] Z. He, P. Wang, Y. Zou, B. Huang, D. Zhu, H. F. Lee, and H. Leung, "DADIGAN: A dual attention blocks-based disentangled iterative generative adversarial network for cloud and shadow removal on SAR and optical images," *Information Fusion*, vol. 125, p. 103487, 2026.
[48] D. Du, Y. Gu, and T. Liu, "A diffusion-guided task decomposition framework for sar-to-optical image translation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 27 597–27 614, 2025.
[49] H. Wang, S. Chen, F. Xu, and Y.-Q. Jin, "Application of deep-learning algorithms to mstar data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Milan, Italy, 2015, pp. 3743–3745.
[50] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, Seattle, WA, USA, 2020, pp. 6070–6079.
[51] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, 2022.
[52] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, Utah, 2018, pp. 586–595.
[53] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proc. IAPR Asian Conf. Pattern Recognit. (ACPR)*, 2015, pp. 730–734.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 770–778.
[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.(ICLR)*, Vienna, Austria, June 2021, pp. 1–21.
[56] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 1501–1510.
[57] 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.
[58] J. Seo, G. Lee, S. Cho, J. Lee, and S. Kim, "Midms: Matching interleaved diffusion models for exemplar-based image translation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 2, 2023, pp. 2191–2199.
[59] T. T. Cai and R. Ma, "Theoretical foundations of t-sne for visualizing high-dimensional clustered data," *J. Mach. Learn. Res.*, vol. 23, no. 301, pp. 1–54, 2022.

**Jinrui Liao** received the B.S. and M.S. degrees from Wuyi University, Jiangmen, China, in 2019 and 2023, respectively. He is currently pursuing the Ph.D. degree with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China.

His research interests include computer vision and remote sensing image processing, especially in synthetic aperture radar (SAR) target detection and recognition.

**Qingsong Wang** received the B.S. and Ph.D. degrees from National University of Defense Technology, Changsha, China, in 2006 and 2011, respectively.

Since 2019, he has worked in the School of Electronics and Communication Engineering of Sun Yat-sen University, where he is mainly engaged in basic theory and engineering application research in the field of precision processing of microwave mapping. His research interests include multi-system radar 3-D positioning, image matching and UAV vision navigation.

**Yikui Zhai (Senior Member, IEEE)** received the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013. Since October 2007, he has been working with Wuyi University, Jiangmen, China, where he is currently a Full Professor. He has also been the Associate Dean of the School of Electronics and Information Engineering, Wuyi University, since 2021. He has been a Visiting Scholar with the Department of Computer Science, Universita degli Studi di Milano, Milan, Italy, from June 2016 to June 2017, in August 2023, and in January 2024. His research interests include image processing, deep learning, optical character recognition, object detection, UAV change detection, and self-supervised learning.

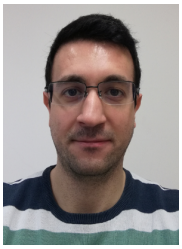
**Xingwang Hu** received the B.S. degree in software engineering and the M.S. degree in computer science and technology from Henan University. He is currently pursuing the Ph.D. degree with the School of Electronics and Communication Engineering, Sun Yat-sen University. His research interests include synthetic aperture radar imaging, signal processing and radio frequency interference mitigation.

**Nengcai Li** received the Ph.D. degree from Beijing University of Chemical Technology, Beijing, China, in 2026. He is currently a postdoctoral researcher with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China. His research interests include PolSAR image interpretation and polarimetric scattering mechanism analysis, region structure modeling, and intelligent remote sensing image analysis.

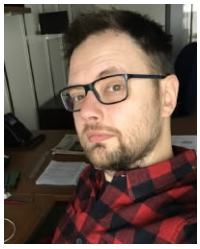
**Haifeng Huang (Member, IEEE)** received the B.S. degree in mathematics and the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 1997 and 2005, respectively. He is a Professor and the Associate Dean of the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China.

His research interests include basic theory and key technology in the field of space electronics and intelligent sensing and are mainly oriented to intelligent remote sensing, surveying, oceanography, surveillance, geological hazards, and other applications.

**C. L. Philip Chen (Life Fellow, IEEE)** received the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988. He is the Chair Professor and the Dean of the College of Computer Science and Engineering, South China University of Technology. His current research interests include cybernetics, systems, and computational intelligence. He is a Fellow of AAAS, IAPR, CAA, and HKIE; and a member of the Academia Europaea (AE) and European Academy of Sciences and Arts (EASA). He received the IEEE Norbert Wiener Award in 2018 for his contributions in systems and cybernetics and machine learning, received two times Best Transactions Paper Award from IEEE Transactions on Neural Networks and Learning Systems for his papers in 2014 and 2018, respectively, and he is a Highly Cited Researcher by Clarivate Analytics from 2018 to 2022. He was a recipient of the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University, in 1988, after he graduated from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA, in 1985. He was the Editor-in-Chief of IEEE Transactions on Cybernetics, the Editor-in-Chief of IEEE Transactions on Systems, Man, and Cybernetics: Systems, and the President of the IEEE Systems, Man, and Cybernetics Society.



**Pasquale Coscia (Senior Member, IEEE)** received the Ph.D. degree in industrial and information engineering from the Università degli Studi di Milano, Milan, Italy, in 2019. From 2019 to 2022, he was a Postdoctoral Researcher with the Università degli Studi di Padova, Padua, Italy. His research interests include the areas of applied aspects of computational intelligence for signal and image processing, computer vision, machine learning and probabilistic models applied to multiagents systems. Dr. Coscia was the recipient of the 2021 Best Paper Award of the Elsevier's Image and Vision Computing Journal for his work on human motion prediction. He is a member of the program committees for numerous IEEE international conferences and workshops. Since 2023, he has held the position of Co-chair for the Intelligent Measurement Systems Technical Committee (TC-22) within the IEEE Instrumentation and Measurement Society.



**Angelo Genovese (Senior Member, IEEE)** received the Ph.D. degree in computer science from the Università degli Studi di Milano, Crema, Italy, in 2014., He has been a Postdoctoral Research Fellow in computer science with the Università degli Studi di Milano, since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. He has authored or coauthored over 30 papers in international journals, proceedings of international conferences, books, and book chapters. His research interests include signal and image processing, 3-D reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems. Dr. Genovese is an Associate Editor of the Journal of Ambient Intelligence and Humanized Computing (Springer).