remote sensing

MDPI

Article

MIRB-Net: Masked Inpainting and Residual Blending for Conditional SAR Target Generation on Backgrounds

Yang Zheng [1], Zhikai Wang [1], Ruimin Li [2,*], Junling Fan [3] and Wenbo Jia [3]

[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China; zhengy@xidian.edu.cn (Y.Z.); 24021211654@stu.xidian.edu.cn (Z.W.)
[2] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China
[3] National Key Laboratory of Strength and Structural Integrity, Aircraft Strength Research Institute of China, Xi'an 710065, China; fanjl@avic.com (J.F.); jiawb013@avic.com (W.J.)
* Correspondence: rmli@xidian.edu.cn

Highlights

What are the main findings?

- A masked inpainting and residual blending network (MIRB-Net) is proposed for instance-free, background-preserving, multi-class coupled generation directly onto authentic measured backgrounds.
- A comprehensive evaluation methodology is established to assess target-on-background generation across tri-region spatial realism (target, background, transition), distribution-level consistency, and downstream ATR utility.

What are the implications of the main findings?

- The established multidimensional criteria provide a robust benchmark for quantitatively assessing the local physical realism, background preservation, and transition naturalness of coupled target-on-background SAR imagery.
- MIRB-Net produces more structurally seamless and physically coupled SAR images than representative generation models, effectively enhancing downstream ATR performance.

Abstract

Enhancing the environmental generalization of Synthetic Aperture Radar (SAR) Automatic Target Recognition (ATR) requires diverse target–clutter combinations. Despite abundant target-free clutter, exhaustively capturing multi-background target samples is prohibitive. However, existing generative methods either distort authentic clutter statistics through full-scene reconstruction or lack autonomous generation capabilities by strictly relying on pre-existing target instances. To overcome these limitations, we propose the Masked Inpainting and Residual Blending Network (MIRB-Net) for instance-free, background-preserving, multi-class coupled generation directly onto authentic measured backgrounds. Treating the background as a measured clutter carrier, MIRB-Net synthesizes class-conditional targets in situ using three synergistic mechanisms. First, a Background-Preserving Residual Editing (BPRE) mechanism explicitly decouples spatial blending from residual prediction, effectively confining generative perturbations to the target vicinity. Second, a self-supervised Masked Inpainting Pretraining (MIP) strategy specifically designed for BPRE equips the network with crucial priors of target scattering and background statistics. Third, joint global–local adversarial supervision alongside tri-region spatial constraints ensures seamless radiometric coupling across target–clutter boundaries. Evaluated under a dedicated multidimensional assessment framework on the MSTAR dataset, MIRB-Net demonstrates

Check for updates

Academic Editors: Yingying Kong, Henry Leung, Kefeng Ji and Jinglu He

Received: 19 September 2026
Revised: 25 October 2026
Accepted: 30 October 2026
Published: 31 October 2026

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

superior local structural fidelity, faithful background preservation, and natural target–clutter coupling. Furthermore, hybrid data-augmentation experiments confirm that the autonomously generated samples consistently improve downstream ATR performance across diverse deep learning architectures.

Keywords: synthetic aperture radar (SAR); automatic target recognition (ATR); image generation; target–background coupling; residual editing

1. Introduction

Synthetic aperture radar (SAR) provides high-resolution, all-weather, and day-and-night observation capabilities [1], serving as an indispensable sensing modality for ground surveillance. Nevertheless, data-driven automatic target recognition (ATR) models remain vulnerable to shifts in training distributions [2]. Unlike optical imaging, SAR acquisition is strictly governed by platform trajectories, observation geometries, and coherent electromagnetic backscattering, rendering large-scale data collection and manual annotation across diverse environments economically and logistically prohibitive [1,3]. Furthermore, background clutter is not merely peripheral context but an intrinsic component of the radar signature [1,4]; significant domain shifts between training and operational clutter distributions frequently induce severe performance degradation in ATR classifiers [2]. Expanding target-class and clutter-environment configurations from scarce measured data is therefore paramount for enhancing the environmental generalization of SAR ATR systems [2].

Existing non-generative data expansion strategies—primarily empirical data augmentation and computational electromagnetic (CEM) simulation—fail to balance physical authenticity with scalable scene generation. Conventional geometric manipulations, such as rotation, translation, flipping, and speckle noise injection [2,4], expand the raw sample volume but cannot synthesize genuinely novel target–clutter configurations or reflect coupled backscattering interactions [1]. To achieve physical authenticity, high-frequency scattering models, such as shooting and bouncing rays (SBRs), combine geometrical optics with physical optics [5]. Representative physics-based SAR tools include RaySAR, which relates image signatures to three-dimensional geometry and multiple reflections [6], and SARCASTIC v2.0, developed for coherent phase-history simulation and ATR studies [7]. While these approaches support highly controlled studies of scattering and observation geometry, their application requires explicit geometric descriptions, appropriate material or reflection models, and specific collection parameters. Consequently, the prohibitive complexity of fine-grained computer-aided design (CAD) modeling and extensive computation severely bottlenecks their scalability across diverse macroscopic clutter environments [8]. Data-driven deep generative models have thus emerged as a complementary and promising alternative for scalable SAR data expansion by learning image-domain appearance directly from measured samples [9–11].

Prevailing deep generative frameworks have evolved from unconditional synthesis to highly controllable paradigms [2], yet they predominantly adopt a target-centric, full-chip generation paradigm that overlooks operational scene diversity. Existing generative adversarial networks (GANs) [9,10,12–16] and denoising diffusion probabilistic models (DDPMs) [11,17,18] typically treat a complete target chip (e.g., 128×128 pixels) as a unified distribution, focusing almost exclusively on the fidelity and class/aspect controllability of localized target scattering. In this holistic formulation, surrounding clutter is treated merely as passive ambient padding. Because these models are optimized to minimize global distributional distances [19,20], they inevitably distort the complex non-Gaussian statistics and coherent speckle correlation inherent to authentic measured clutter [4,21,22].

Consequently, achieving high full-chip structural similarity [23] does not guarantee the preservation of authentic background physics, confining the synthesized targets to the specific, limited clutter distributions of the training set.

In real-world operational scenarios, deploying non-cooperative targets across varied geographic terrains for exhaustive imaging is physically and logistically impossible [1,3]. Conversely, acquiring large-scale, target-free measured SAR scene clutter across diverse environments is straightforward and abundant [8]. Under this data asymmetry, re-synthesizing the entire scene from scratch is fundamentally flawed. Therefore, rather than holistically hallucinating both target and clutter, directly forming class-specific target signatures in situ on authentic measured backgrounds—while effectively preserving the underlying clutter statistics—provides a physically grounded and scalable solution to enhance ATR environmental generalization.

Recognizing the necessity of scene diversity in ATR generalization, recent studies have shifted focus from isolated target synthesis to explicitly modeling SAR target–environment interactions [2,8,24,25]. Specifically, BCNet pioneered background conversion by translating the original clutter surrounding target-of-interest (TOI) chips into diverse measured scenes while utilizing specialized mask modules and scattering-preservation losses [8]; SHFGAN addressed discordant radiometric transitions in SAR compositing by employing semantic regularization and multi-task adversarial fusion to harmoniously blend pre-extracted target patches into new backgrounds [25]; and Ship-Go leveraged an instance-to-image diffusion architecture conditioned on environmental descriptors (e.g., sea-state conditions) and explicit instance layouts to synthesize targets across varied maritime scenarios [24]. These studies establish that localized target–environment coordination is as crucial as target realism. However, their generative pipelines remain strictly instance-dependent. They fundamentally rely on pre-existing target instances—such as segmented TOI chips [8], cropped target patches [25], or explicit geometric layout masks [24]—as mandatory structural inputs. Consequently, they operate as target-to-background adaptation or fusion pipelines rather than autonomous target formation mechanisms. In contrast, decoupling the generation process from any prior target structural inputs forces an instance-free model to autonomously solve target scattering construction, strict clutter preservation, and localized physical coupling simultaneously.

Realizing this autonomous, instance-free generation necessitates overcoming two fundamental hurdles: acquiring localized contextual priors and enforcing physically coherent spatial coupling. While image inpainting architectures [26–28] inherently capture deep structural correlations between masked regions and visible boundaries, they conventionally recover pre-existing content rather than synthesize novel, class-conditional signatures. Repurposing masked inpainting as a self-supervised pretraining mechanism provides a critical pathway to establish these structural priors prior to latent-guided target formation. Furthermore, standard global–local adversarial supervision [29] typically treats all non-target pixels homogeneously, often inducing abrupt boundary artifacts or target occlusion in SAR composites [25]. Authentic SAR coupling strictly demands spatially differentiated constraints: granting generative freedom to the target centroid, enforcing gradient continuity within the near-field transition zone, and ensuring absolute radiometric invariance in the far-field clutter [1,4].

To overcome the strict dependency on pre-existing target instances while preventing background clutter distortion, we propose a Masked Inpainting and Residual Blending Network, termed MIRB-Net, which acts as an instance-free, multi-class, background-preserving framework designed specifically for autonomous SAR target–background coupled generation. In this paradigm, an authentic, target-free measured SAR image serves as a underlying clutter carrier. Driven solely by a target class label y and a stochastic latent vector z, the

network autonomously constructs class-discriminative target backscattering signatures that are coherently coupled with the surrounding clutter field directly within a designated local region. To achieve this, MIRB-Net incorporates three synergistic mechanisms: (1) a Background-Preserving Residual Editing (BPRE) mechanism that jointly predicts target-induced scattering residuals and continuous spatial blending weights, effectively confining generative perturbations to the target vicinity via an explicit background passthrough; (2) a self-supervised Masked Inpainting Pretraining (MIP) strategy for BPRE initialization that establishes robust priors for localized scattering reconstruction and spatial contextual modeling; and (3) a Tri-Region Spatial Coupling constraint across Core, Ring, and Outer functional zones, enforced alongside Global–Local Adversarial Supervision to ensure radiometric and structural continuity across target–clutter boundaries. In addition, to address the distinct evaluation requirements of this coupled generation task, we establish a task-oriented multidimensional evaluation protocol that systematically benchmarks image-level spatial realism (spatially evaluated across target, background, and transition regions), distribution-level statistical consistency, and extrinsic utility in downstream ATR tasks.

The major contributions of this paper are summarized as follows:

- We formulate an instance-free, background-preserving generative paradigm (MIRB-Net) that autonomously synthesizes class-conditional targets directly onto authentic measured backgrounds, effectively expanding diverse target–clutter configurations.
- We design a Background-Preserving Residual Editing (BPRE) mechanism that explicitly decouples where to modify (via spatial gating) from what to modify (via local residual prediction). This is empowered by a tailored self-supervised Masked Inpainting Pretraining (MIP) strategy, which equips the network with crucial priors for localized scattering reconstruction.
- We introduce a joint global–local adversarial supervision framework integrated with customized tri-region spatial constraints (Core, Ring, and Outer zones). This optimization strategy facilitates physically coherent target–clutter coupling.
- We establish a task-oriented multidimensional evaluation protocol spanning tri-region spatial realism, distribution-level consistency, and downstream ATR utility. Extensive experiments on the MSTAR benchmark demonstrate that MIRB-Net achieves superior structural fidelity and consistently enhances environmental generalization across diverse deep learning architectures.

2. Materials and Methods

2.1. Dataset

The proposed method is evaluated on the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [2,3], a widely used benchmark for synthetic aperture radar (SAR) target recognition. Our experiments focus on the specific task of generating ten representative categories of ground vehicles on target-free measured backgrounds. The model is trained and tested under the standard operating conditions (SOCs), meaning that the serial numbers and target configurations in the test set are identical to those in the training set. Specifically, measured images captured at a 17° depression angle are used to train both the generative and recognition models, while independent images captured at a 15° depression angle are reserved for testing the downstream ATR model. The dataset configuration is detailed in Table 1.

To facilitate the target-free background conditioning required by our framework, a pure-background set is additionally constructed by extracting target-free regions from large MSTAR scene images acquired under the SOCs. These background chips serve as the environmental carriers during both the initial Masked Inpainting Pretraining (MIP) pretraining and the main background-conditional target generation phases. All SAR im-

ages—including measured targets and pure backgrounds—are standardized to a unified resolution of 128 × 128 pixels as single-channel grayscale images.

For the aforementioned MIP stage, the dataset is formulated into eleven domains: ten measured target domains ($y \in \{1, \ldots, 10\}$) corresponding to the 17° MSTAR categories in Table 1, and one unified pure-background domain ($y = 0$). This background domain comprises 429 chips extracted from three large-scale 17° MSTAR scene images (1472 × 1784 pixels): Desert (HB06172), Grass (HB06241), and Field (HB06244). During the MIP phase, approximately 80% of the images from each domain are used for parameter optimization, while the remaining 20% are held out as a validation pool. This held-out set is strictly reserved for checkpoint selection without participating in parameter updates. For the subsequent main generative training, the entire pool of 2746 17° target images and all 429 background chips are utilized, including those previously held out during the MIP validation (i.e., the background pool is shared across both stages). Ultimately, MIRB-Net synthesizes one sample per target category on each background chip, yielding a total of 4290 generated samples (429 per class) to augment the downstream ATR training.

Table 1. Configuration of the MSTAR samples used for target generation and ATR evaluation under the standard operating condition (SOC).

Target Category	Serial Number	Training Set		Testing Set	
		Depression	Number	Depression	Number
2S1	b01	17°	299	15°	274
BMP2	9566	17°	232	15°	196
BRDM2	E-71	17°	298	15°	274
BTR60	k10yt7532	17°	256	15°	195
BTR70	c71	17°	233	15°	196
D7	92v13015	17°	299	15°	274
T62	A51	17°	299	15°	273
T72	132	17°	232	15°	196
ZIL131	E12	17°	299	15°	274
ZSU23-4	d08	17°	299	15°	274

2.2. Methodological Framework

To achieve instance-free SAR target generation on measured clutter, we propose the Masked Inpainting and Residual Blending Network (MIRB-Net). Built upon a conditional generative adversarial network (cGAN) paradigm [30], the task is formulated as a localized generation problem utilizing a fixed scene carrier. Let $x_b \in \mathbb{R}^{C \times H \times W}$ denote the input target-free measured background, y the target class label, and $z \sim \mathcal{N}(0, I)$ a stochastic latent vector. A mapping network F projects the class condition and latent noise into a class-specific scattering embedding:

$$s = F(y, z). \tag{1}$$

Driven by the background carrier x_b and the embedding s, the generator G synthesizes the target–clutter coupled image $\hat{x}$:

$$\hat{x} = G(x_b, s). \tag{2}$$

As illustrated in Figure 1, the overall architecture of MIRB-Net is built upon three primary components. These include a conditional generator executing the Background-Preserving Residual Editing mechanism, a set of Global–Local Discriminators, and Target–Background Spatial Coupling constraints. These spatial constraints serve as the key task-specific loss design within the broader optimization objective. During the generation phase, it executes the Background-Preserving Residual Editing (BPRE) mechanism: rather than

predicting a full-image array, the generator injects the embedding s via adaptive instance normalization (AdaIN) and explicitly predicts a target-induced scattering residual Δ alongside a continuous spatial blending weight α. The final coupled image $\hat{x}$ is formulated by fusing the weighted residual directly onto the background passthrough x_b. To empower this precise localized synthesis with robust scattering priors, the generator backbone is initialized via Masked Inpainting Pretraining (MIP). Second, to enforce adversarial realism at multiple spatial scales, the framework employs Global–Local Discriminators. The Global Discriminator evaluates the full scene to ensure macroscopic distributional consistency, while the Local Discriminator scrutinizes a center-cropped region to enforce high-frequency localized scattering fidelity. Finally, to enforce physical coherence across the target–clutter interface, Tri-Region Spatial Coupling constraints ($\mathcal{L}_{core}$, $\mathcal{L}_{ring}$, $\mathcal{L}_{outer}$) are imposed across defined functional zones. These spatial constraints operate in tandem with the adversarial supervision to facilitate seamless boundary transitions and robust far-field clutter preservation.

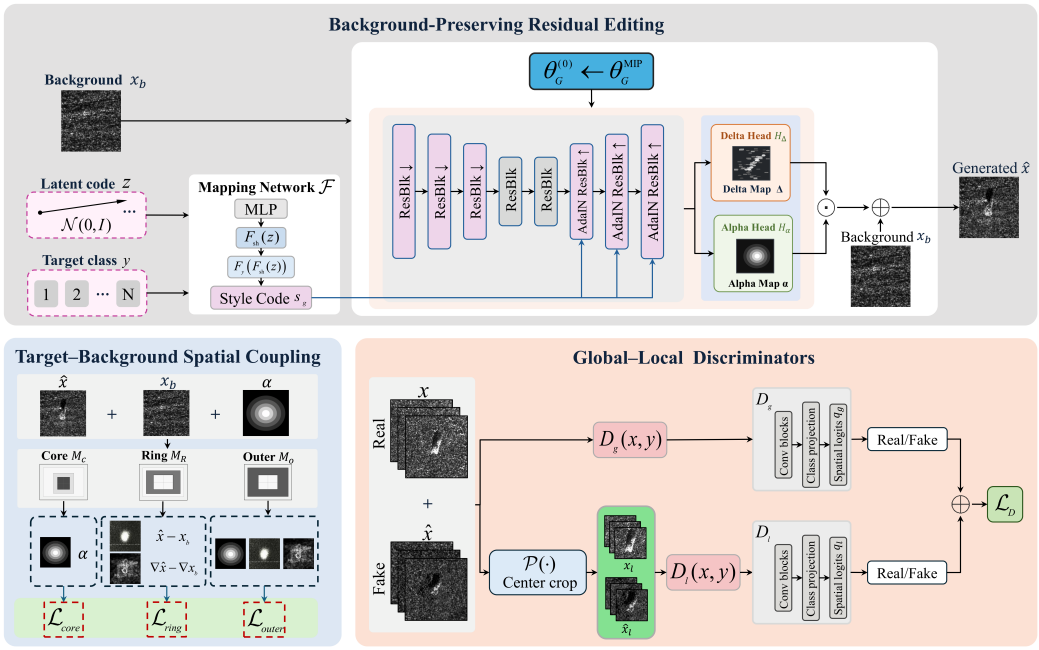


Figure 1. Framework of the proposed MIRB-Net. Built upon a conditional generative adversarial network (cGAN) paradigm, the architecture comprises a conditional generator and dual discriminators, which are systematically organized into three primary modules: (1) a Background-Preserving Residual Editing (BPRE) generator (initialized via masked inpainting pretraining) for localized in situ synthesis; (2) Global–Local Discriminators for multi-scale adversarial supervision; and (3) Target–Background Spatial Coupling constraints to facilitate seamless physical integration across the core, ring, and outer zones.

2.3. Background-Preserving Residual Editing

The core generation module of MIRB-Net is designed to synthesize user-specified, multi-class targets on a target-free measured SAR background while minimizing alterations to the measured content outside the necessary editing region. To achieve this localized generation without requiring a reference target image during inference, we propose a Background-Preserving Residual Editing (BPRE) architecture that combines latent-driven class conditioning with spatial gating.

Following the shared-mapping and class-specific output design of multi-domain conditional generation [31], a random latent variable z is first transformed by a shared Mapping Network and then projected through the branch associated with the desired target class y:

$$s_g = F_y(F_{\text{sh}}(z)), \qquad z \sim \mathcal{N}(0, I), \quad y \in \{1, \ldots, 10\}, \tag{3}$$

where F_{sh} denotes the shared mapping layers and F_y denotes the class-specific output branch. The resulting representation s_g provides the target-class condition while preserving latent-dependent variation within the same category.

The generator adopts a multiscale residual encoder–decoder architecture, the explicit structural details of which—including layer counts, channel widths, and downsampling factors—are specified in Table 2. The encoder progressively extracts contextual and structural features from the target-free measured background x_b, whereas the decoder restores the spatial resolution. The synthesized style code s_g is injected into all decoder stages through adaptive instance normalization (AdaIN) [32]:

$$\text{AdaIN}(h, s_g) = \left[1 + \gamma(s_g)\right] \odot \text{IN}(h) + \beta(s_g), \tag{4}$$

where $\text{IN}(\cdot)$ denotes instance normalization, and $\gamma(\cdot)$ and $\beta(\cdot)$ are learned affine transformations.

Table 2. Network architecture of MIRB-Net.

Module	Layers in the Module	Output Shape
Generator encoder	Conv2d(128) / k3s1	128 × 128 × 128
	ResBlk↓ ×3 / k3s1	16 × 16 × 512
	ResBlk ×2 / k3s1	16 × 16 × 512
Generator decoder	AdaIN ResBlk ×2 / k3s1	16 × 16 × 512
	AdaIN ResBlk↑ ×3 / k3s1	128 × 128 × 128
Delta head H_Δ	Conv2d / k1s1 → Tanh	128 × 128 × 3
Alpha head H_α	Conv2d / k1s1 → Sigmoid	128 × 128 × 1
Mapping network F	FC(512) ×4 (shared)	512-d
	FC(512) ×3 (class-specific)	512-d
	FC(64) (class-specific)	64-d
Style encoder E	Conv2d(128) / k3s1	128 × 128 × 128
	ResBlk↓ ×5 / k3s1	4 × 4 × 512
	Conv2d(512) / k4s1	1 × 1 × 512
	Class-specific FC(64)	64-d
Global discriminator D_g	Conv2d(128) / k3s1	128 × 128 × 128
	ResBlk↓ ×5 / k3s1	4 × 4 × 512
	Class projection (512-d)	4 × 4 × 1
Local discriminator D_l	Conv2d(256) / k3s1	64 × 64 × 256
	ResBlk↓ ×4 / k3s1	4 × 4 × 512
	Class projection (512-d)	4 × 4 × 1

Shapes are denoted as $H \times W \times C$; "d" indicates vector dimensionality; "k3s1" represents a 3 × 3 convolution with stride 1; ↓/↑ denotes 2× spatial downsampling/upsampling.

Rather than synthesizing an entirely new image and risking background destruction, the generator modifies the measured background through a dual-head design. Let h_d denote the final decoder feature. The Residual Head predicts a bounded image correction,

$$\Delta_g = \tanh(H_\Delta(h_d)), \tag{5}$$

while the Alpha Head predicts a pixel-wise editing gate,

$$\alpha_g = \sigma(H_\alpha(h_d)), \qquad 0 \le \alpha_g \le 1. \tag{6}$$

The final generated image is obtained through residual blending:

$$\hat{x} = \mathrm{clip}\big(x_b + \alpha_g \odot \Delta_g, -1, 1\big). \tag{7}$$

In this formulation, the measured background x_b enters the blend at its original pixel coordinates, providing a spatially aligned base for the output. This explicitly separates what should be modified (Δ_g) from where the modification should occur (α_g), inherently coupling the target's spatial structure with the background geometry. However, directly optimizing this highly localized editing mechanism from scratch via adversarial training is notoriously unstable [33], as the generator struggles to simultaneously learn target scattering, spatial restriction, and background preservation. Therefore, a tailored initialization strategy is required before adversarial training.

2.4. Background-Preserving Residual Editing Pretraining

To properly initialize the dual-head architecture proposed in Section 2.3 and mitigate the aforementioned optimization instability, we propose a pretraining strategy formulated as a Masked Inpainting Pretraining (MIP) task. MIP equips the generator with spatial gating and local editing priors by solving a unified style-conditioned masked reconstruction problem [26–28]. Figure 2 shows the framework of BPRE pretraining.

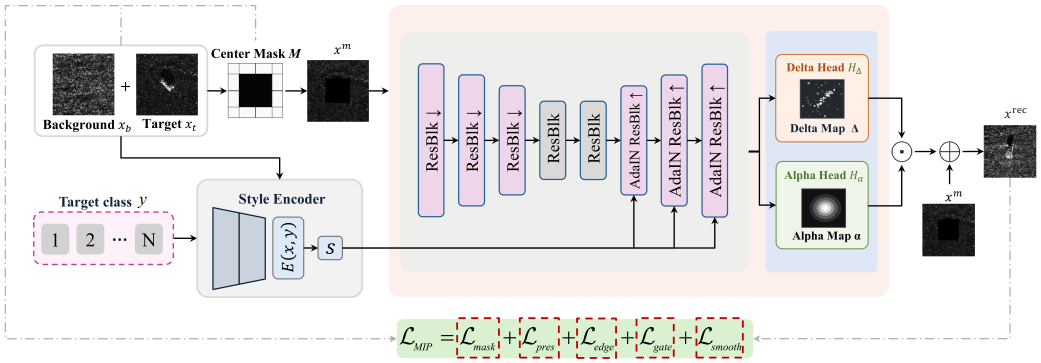


Figure 2. Framework of the Background-Preserving Residual Editing Pretraining. Formulated as a unified style-conditioned Masked Inpainting Pretraining (MIP) task, this self-supervised pipeline optimizes the dual-head parameterization (Alpha and Delta heads) under comprehensive spatial constraints. It effectively equips the generator with robust local-editing and spatial-gating priors to initialize the subsequent BPRE module.

The pretraining data contain eleven domains: one pure-background domain ($y = 0$) and ten measured target domains ($y \in \{1, \ldots, 10\}$). Target and pure-clutter images follow exactly the same reconstruction procedure, with the background treated simply as a unified additional domain. For each complete real image x, a centered square mask M covering 40% of the image dimension is applied to obtain the masked observation:

$$x^m = (1 - M) \odot x. \tag{8}$$

Unlike the main generation stage where the style is derived from latent noise, during pretraining, the complete image x is encoded by an auxiliary Style Encoder to provide a deterministic real-image style representation $s = E(x, y)$.

Conditioned on x^m and s, the generator executes the exact same forward pass parameterization defined in Equations (5) and (6) to predict Δ and α. The reconstruction image is thus formulated identically to the main stage:

$$x^{\text{rec}} = \text{clip}(x^m + \alpha \odot \Delta, -1, 1). \tag{9}$$

This network is optimized using a unified objective comprising five complementary constraints:

$$\mathcal{L}_{\text{MIP}} = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}. \tag{10}$$

Specifically, $\mathcal{L}_{\text{mask}} = \|M \odot (x^{\text{rec}} - x)\|_1/(C\|M\|_1)$ and $\mathcal{L}_{\text{edge}} = \|M \odot (\nabla x^{\text{rec}} - \nabla x)\|_1/(2C\|M\|_1)$ compare the reconstruction with the exact measured reference at corresponding spatial locations. Together, they forcefully encourage the recovery of the target's contours, scattering layout, and image-plane orientation by minimizing pixel and first-order Sobel gradient differences. Meanwhile, $\mathcal{L}_{\text{pres}} = \|(1 - M) \odot (x^{\text{rec}} - x)\|_1/(C\|1 - M\|_1)$ suppresses unintended modification of the visible periphery. The spatial gate is regularized via $\mathcal{L}_{\text{gate}} = \|(1 - \alpha) \odot M\|_1/\|M\|_1 + \eta_{\text{out}}\|\alpha \odot (1 - M)\|_1/\|1 - M\|_1$, which maintains sufficient editing response inside the mask while suppressing leakage outside. Finally, $\mathcal{L}_{\text{smooth}} = \mathbb{E}_{i,j}\left[|\alpha_{i+1,j} - \alpha_{i,j}| + |\alpha_{i,j+1} - \alpha_{i,j}|\right]$ encourages spatial continuity of the blending field and suppresses abrupt variations between neighboring pixels.

Upon completion of MIP, the pretrained generator parameters are directly transferred to initialize the BPRE module for subsequent adversarial training:

$$\theta_G^{(0)} \leftarrow \theta_G^{\text{MIP}}. \tag{11}$$

Importantly, the auxiliary Style Encoder E is discarded during generation. By maintaining complete structural isomorphism between the MIP reconstruction task and the BPRE synthesis task, the generator seamlessly transitions into the target-generation phase. The transferred generator weights therefore provide a robust initialization for structure-aware local editing; when the class-conditioned latent representation replaces the real-image-derived style code during main synthesis, the target's shape and orientation are generated coherently under adversarial and spatial supervision. Crucially, because the MIP phase utilizes a fixed, centered occlusion mask, the generator naturally acquires a spatial prior, explicitly learning to synthesize targets precisely at the image center. Building upon this centered structural prior, the spatial scale, physical dimension, and aspect angle of the generated targets are not sampled via independent explicit control variables. Instead, they are implicitly learned from the standardized MSTAR training data during the subsequent generation phase, dynamically driven by the specified target class and the stochastic latent code.

2.5. Global–Local Discriminators

While the BPRE architecture provides an explicit mechanism to restrict modifications to localized regions, it does not inherently guarantee that the synthesized target exhibits realistic SAR scattering characteristics. This realism must be injected through adversarial supervision. However, relying solely on a single full-image discriminator is suboptimal; since target signatures occupy a disproportionately small spatial footprint compared to the extensive surrounding clutter, the full-image gradient signals are easily dominated by the vast unchanged background. To resolve this, we introduce a Global–Local Discriminators module comprising two complementary branches [29]: a Global Discriminator D_g and a Local Discriminator D_l. Structurally, both discriminators adopt a fully convolutional network architecture equipped with projection-based condition injection, yielding dense spatial outputs to evaluate image fidelity at different scales.

Specifically, the Global Discriminator D_g operates on the complete image. It is designed to evaluate image-level realism and contextual consistency conditioned on the domain label y, ensuring that the overall generated image aligns macroscopically with the target domain.

Conversely, the Local Discriminator D_l operates exclusively on a target-centered patch extracted by a cropping operation $\mathcal{P}(\cdot)$. By applying $\mathcal{P}(\cdot)$ identically to both real and generated target images, D_l filters out the distant background clutter. This forces the local discriminator to concentrate entirely on class-specific scattering structures and fine-grained target appearances.

Both discriminators are optimized using the binary cross-entropy objective. For a discriminator logit q and binary label $t \in \{0, 1\}$, we define:

$$\ell_{\text{adv}}(q, t) = -t \log \sigma(q) - (1 - t) \log[1 - \sigma(q)]. \tag{12}$$

For the generator, the combined global and local adversarial objective is formulated as:

$$\mathcal{L}_{\text{adv}}^{G} = \lambda_g \, \ell_{\text{adv}}\big(D_g(\hat{x}, y), 1\big) + \lambda_l \, \ell_{\text{adv}}(D_l(\mathcal{P}(\hat{x}), y), 1). \tag{13}$$

Here, the first term encourages full-image domain consistency, while the second term provides concentrated, high-frequency adversarial supervision on the localized target region.

For the discriminators, D_g treats measured pure-background images x_b and real target images x_t as real samples in their corresponding domains, while generated images $\hat{x}$ are treated as fake:

$$\begin{aligned} \mathcal{L}_D^g = \ell_{\text{adv}}\big(D_g(x_b, 0), 1\big) + \ell_{\text{adv}}\big(D_g(x_t, y), 1\big) \\ + \ell_{\text{adv}}\big(D_g(\hat{x}, y), 0\big). \end{aligned} \tag{14}$$

Meanwhile, D_l is trained solely on target-centered regions extracted from real and generated target samples:

$$\mathcal{L}_D^l = \ell_{\text{adv}}(D_l(\mathcal{P}(x_t), y), 1) + \ell_{\text{adv}}(D_l(\mathcal{P}(\hat{x}), y), 0). \tag{15}$$

To improve adversarial stability and mitigate mode collapse, R_1 regularization [33] is applied to the Global Discriminator on real samples:

$$\mathcal{L}_{R1} = \frac{1}{2} \mathbb{E}_{(x,y) \sim p_{\text{real}}} \Big[\big\| \nabla_x D_g(x, y) \big\|_2^2 \Big]. \tag{16}$$

The complete objective for the Global–Local Discriminators is thus given by:

$$\mathcal{L}_D = \lambda_g^D \mathcal{L}_D^g + \lambda_l^D \mathcal{L}_D^l + \lambda_{\text{reg}} \mathcal{L}_{R1}. \tag{17}$$

In summary, this dual-scale adversarial design ensures that the generator learns high-fidelity, class-conditional SAR synthesis. By reliably injecting target-specific realism, the Global–Local Discriminators pave the way for the fine-grained spatial coupling constraints introduced in the overall objective. To provide a complete and reproducible specification of the entire MIRB-Net architecture, Table 2 explicitly details the layer-by-layer configurations for all the aforementioned modules.

2.6. Spatial Coupling Constraints and Overall Objective

While the Global–Local Discriminators introduced in Section 2.5 enforce target realism, relying solely on adversarial supervision can lead the generator to cheat. Since the discriminators evaluate only the absolute realism of the final output rather than the preservation of the original input, the generator may freely alter the distant measured background to ease the synthesis of target-like features. To counterbalance these unconstrained adversarial gradients and regulate the exact spatial extent of target-related modifications, we introduce a soft Core–Ring–Outer spatial coupling mechanism, complemented by conditional consistency regularizations.

2.6.1. Target–Background Spatial Coupling

Target formation and background preservation impose inherently conflicting spatial requirements. To resolve this, we construct a soft spatial decomposition centered at the prescribed target location $\mathbf{p}_0$ (the image center). For any pixel location $\mathbf{p}$, a Gaussian soft weighting function is defined as $\Phi_\rho(\mathbf{p}) = \exp\left(-\|\mathbf{p}-\mathbf{p}_0\|_2^2/(2\rho^2)\right)$. Using two scales $\rho_c < \rho_o$, three soft region masks are defined:

$$M_c = \Phi_{\rho_c}, \quad M_k = \Phi_{\rho_o}, \quad M_r = \mathrm{clip}(M_k - M_c, 0, 1), \quad M_o = 1 - M_k, \tag{18}$$

where M_c, M_r, and M_o denote the Core, Ring, and Outer regions, respectively. This soft weighting avoids artificial hard boundaries.

Distinct constraints are imposed on the three regions according to their respective functional roles.The Core region prevents the spatial gate α from collapsing, preserving sufficient editing capacity for target generation: $\mathcal{L}_{\text{core}} = \mathbb{E}[\|(1-\alpha)\odot M_c\|_1]$. The Ring region regulates the local transition between the synthesized target and the measured surroundings to avoid artificial pasting effects, constrained by both pixel and gradient similarities:

$$\mathcal{L}_{\text{ring}} = \lambda_{rp}\frac{\|M_r \odot (\hat{x}-x_b)\|_1}{C\|M_r\|_1} + \lambda_{rg}\frac{\|M_r \odot (\nabla\hat{x}-\nabla x_b)\|_1}{2C\|M_r\|_1}. \tag{19}$$

The Outer region explicitly protects the distant measured scene. It penalizes any deviation in image content, gradient structure, and non-zero gate activation:

$$\mathcal{L}_{\text{outer}} = \lambda_{op}\frac{\|M_o \odot (\hat{x}-x_b)\|_1}{C\|M_o\|_1} + \lambda_{og}\frac{\|M_o \odot (\nabla\hat{x}-\nabla x_b)\|_1}{2C\|M_o\|_1} + \lambda_{oa}\mathbb{E}[\|\alpha \odot M_o\|_1]. \tag{20}$$

Combined with the spatial smoothness regularizer $\mathcal{L}_{\text{smooth}}$ that encourages spatial coherence of the blending field [34], the complete spatial-coupling objective is $\mathcal{L}_{\text{cpl}} = \lambda_c\mathcal{L}_{\text{core}} + \mathcal{L}_{\text{ring}} + \mathcal{L}_{\text{outer}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}$.

2.6.2. Conditional Consistency and Regularization

To guarantee that the latent condition is faithfully expressed and to prevent intra-class mode collapse, we introduce a set of complementary regularizers. (1) Style Reconstruction ($\mathcal{L}_{\text{sty}}$): Ensures the generated target retains the conditional information encoded in s_g by penalizing $\|E(\mathcal{D}(\hat{x}), y) - s_g\|_1$. (2) Intra-class Diversity ($\mathcal{L}_{\text{ds}}$, $\mathcal{L}_{\text{sdiv}}$): Encourages two distinct latent variables z_1, z_2 to yield distinguishable target appearances and style representations for the same class and background. (3) Class Separation ($\mathcal{L}_{\text{sep}}$): Employs a margin-based constraint to push the recovered style away from representations of incorrect target classes. (4) Background Self-Reconstruction ($\mathcal{L}_{\text{cyc}}$): Requires the generator to perfectly reproduce x_b when conditioned on the pure-background label ($y = 0$).

2.6.3. Overall Objective

The total objective function for the BPRE generator integrates the adversarial supervision, spatial coupling, and conditional regularizations:

$$\begin{aligned}\mathcal{L}_G = \mathcal{L}_{\text{adv}}^G + \lambda_{\text{cpl}}\mathcal{L}_{\text{cpl}} + \lambda_{\text{sty}}\mathcal{L}_{\text{sty}} \\ - \lambda_{\text{ds}}\mathcal{L}_{\text{ds}} - \lambda_{\text{sdiv}}\mathcal{L}_{\text{sdiv}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}.\end{aligned} \tag{21}$$

Ultimately, the entire MIRB-Net framework is optimized through a unified minimax zero-sum game. The Global–Local Discriminators aim to maximize their capability to distinguish between real and synthesized localized targets, while the BPRE generator learns to fool the discriminators subject to explicit background-preservation constraints:

$$\min_G \max_{D_g, D_l} \mathcal{L}_{\text{MIRB}}(G, D_g, D_l) = \mathcal{L}_G - \mathcal{L}_D. \tag{22}$$

3. Results

To comprehensively evaluate the proposed MIRB-Net for class-conditional SAR target generation on measured backgrounds, we design a task-oriented evaluation protocol encompassing two primary dimensions: intrinsic generative quality and extrinsic practical utility. The intrinsic generative quality is systematically assessed across both image-level physical realism and distribution-level statistical consistency. Specifically, at the image level, the evaluation is spatially partitioned into three semantic zones: (1) the target region, to quantify the structural fidelity of the synthesized signatures; (2) the background region, to verify the preservation of the unedited clutter; and (3) the transition region, to examine the naturalness and continuity of the target–background coupling. Meanwhile, the distribution-level evaluation guarantees global statistical alignment between the generated and authentic datasets. Conversely, the extrinsic practical utility evaluates the downstream value of the synthesized samples. This is validated through automatic target recognition (ATR) experiments, specifically assessing whether the generated targets can effectively support bidirectional recognition verification and measured-data augmentation across diverse classifier architectures.

3.1. Implementation Details

All networks are implemented in PyTorch 2.7.1 and optimized on a single NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA). The comprehensive training configurations are summarized in Table 3. Both the Masked Inpainting Pretraining (MIP) and the main generative training stages are executed for 20,000 iterations using a batch size of 16. Optimization is driven by the Adam optimizer [35] with momentum parameters $\beta_1 = 0$ and $\beta_2 = 0.99$, alongside a weight decay of 1×10^{-4}.

During the MIP stage, the base learning rate is set to 1×10^{-4}. Validation is performed every 250 iterations on the held-out subset, and the weights achieving the lowest composite reconstruction loss are strictly transferred to initialize the generator for the subsequent stage. The weighting coefficients for the MIP objective in Equation (10) are empirically set to $\lambda_{\text{mask}} = 6$, $\lambda_{\text{pres}} = 1$, $\lambda_{\text{edge}} = 2$, $\lambda_{\text{gate}} = 0.5$, and $\lambda_{\text{smooth}} = 0.1$.

In the main generative training phase, the mapping network F utilizes a reduced learning rate of 1×10^{-5} to ensure stable latent space projection, while the generator and discriminators maintain the 1×10^{-4} learning rate. For the Tri-Region Spatial Coupling constraints, the Gaussian standard deviations defining the soft masks are fixed at $\rho_c = 0.09$ and $\rho_o = 0.17$. The Local Discriminator D_l processes a centered 64×64 patch to enforce localized scattering fidelity. The trade-off parameters balancing the overall minimax objective in Equation (21) are set to $\lambda_{\text{cpl}} = 1$, $\lambda_{\text{sty}} = 1$, $\lambda_{\text{ds}} = 0.3$, $\lambda_{\text{sdiv}} = 0.05$, $\lambda_{\text{sep}} = 0.05$, and $\lambda_{\text{cyc}} = 0.25$.

To validate the effectiveness of the proposed MIRB-Net, we select CycleGAN [36], BCNet [8], and SHFGAN [25] as baseline methods. CycleGAN serves as a general unpaired image-to-image translation benchmark. BCNet performs SAR background conversion by replacing the surrounding clutter of existing target instances, while SHFGAN specializes in the harmonious fusion of existing target patches into new background scenes. Notably, while BCNet and SHFGAN fundamentally rely on manipulating existing target instances, MIRB-Net addresses a distinctly more challenging task: synthesizing class-specific target signatures entirely from scratch directly onto a target-free measured background.

3.2. Evaluation of Background-Preserving Residual Editing Pretraining

To verify whether the Background-Preserving Residual Editing Pretraining effectively establishes the local content-recovery and context-modeling capabilities, we evaluate the pretrained generator on the held-out MIP validation subsets described in the Dataset subsec-

tion. Specifically, the reconstructed images are evaluated against their original unmasked counterparts. We report the Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM) [23,37] to assess spatial and structural fidelity, alongside the one-dimensional Wasserstein-1 (W_1) distance to evaluate the preservation of macroscopic pixel-intensity distributions. All metrics are computed over both the full image and the specific masked region.

Table 3. Training parameters of MIRB-Net.

Parameter	Value
Optimizer	Adam
Batch Size	16
Training Iterations	20,000
Initial Learning Rate	1×10^{-4}
Mapping Network Initial Learning Rate *	1×10^{-5}
Adam Coefficients (β_1, β_2)	$(0, 0.99)$
Weight Decay	1×10^{-4}

Note: All other parameters apply uniformly to both the MIP and main training stages. * Applicable exclusively to the main training phase.

Figure 3 illustrates representative qualitative reconstructions for measured target images. Each column corresponds to one of the ten MSTAR target classes, while the rows show the original image (REAL), the masked input (MASKED), and the reconstructed image (RECON). Under the style-conditioned MIP objective, the reconstructions successfully reproduce the main target contours and scattering layout, while maintaining spatial continuity across the masked-region boundary without producing artificial square seams.

Figure 4 presents the corresponding qualitative reconstructions for pure-background images. The three column groups correspond to the Desert (HB06172), Grass (HB06241), and Field (HB06244) scenes, respectively. The reconstructed images preserve local intensity variations and clutter textures while maintaining spatial continuity with the surrounding background. These results demonstrate that the generator learns background-context reconstruction in addition to target-structure recovery. Both figures visualize the same unified MIP procedure; the pure-background examples in Figure 4 represent the additional background domain ($y = 0$). The residual discrepancies are predominantly localized around fine edges, low-contrast structures, and isolated strong scattering points, which are inherently challenging to infer from partial contextual information.

Quantitatively, Table 4 summarizes the performance across the ten target classes and the pure-background domain. These metrics confirm that the generator has effectively learned to recover structural details and preserve macroscopic intensity distributions through the pretraining process.

3.3. Intrinsic Generative Quality

Before delving into the quantitative assessments, we first present a comprehensive visual overview of the class-conditional generation performance of MIRB-Net. Figure 5 illustrates a generation matrix, where each of the three rows represents a distinctly different measured background, and each of the ten columns corresponds to a specific MSTAR target class. As demonstrated, MIRB-Net successfully synthesizes target-specific scattering structures—such as accurate contours and dominant scattering centers—across all ten categories. Simultaneously, the model seamlessly blends these diverse targets into varying background environments without introducing artificial block boundaries or altering the global scene layout. With this qualitative foundation established, we proceed to rigorously

quantify the generative quality from three spatial perspectives: local structural fidelity, background preservation, and target–background coupling naturalness.

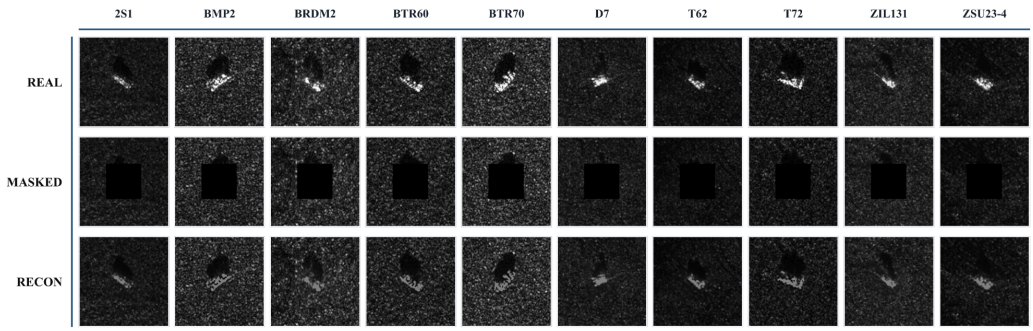


Figure 3. Qualitative masked-reconstruction results for measured target images during Background-Preserving Residual Editing pretraining, formulated as Masked Inpainting Pretraining (MIP). Each column corresponds to one of the ten MSTAR target classes, and the rows show the original image (REAL), the masked input (MASKED), and the reconstructed image (RECON).

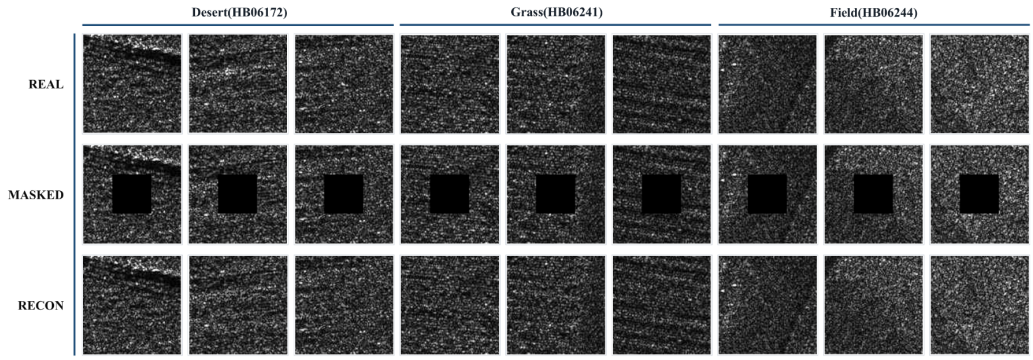


Figure 4. Qualitative masked-reconstruction results for the pure-background domain during Background-Preserving Residual Editing pretraining, formulated as Masked Inpainting Pretraining (MIP). The three column groups correspond to Desert (HB06172), Grass (HB06241), and Field (HB06244), respectively. The rows show the original background (REAL), the masked input (MASKED), and the reconstructed background (RECON).

Table 4. Quantitative reconstruction results of the Background-Preserving Residual Editing (BPRE) pretraining.

	Full Image				Masked Region			
Class	**MAE↓**	**PSNR (dB)↑**	**SSIM↑**	**W_1↓ ($\times 10^{-3}$)**	**MAE↓**	**PSNR (dB)↑**	**SSIM↑**	**W_1↓ ($\times 10^{-3}$)**
ZIL131	5.394	31.682	0.966	3.347	33.961	23.689	0.800	21.084
D7	5.805	30.283	0.964	4.039	36.543	22.290	0.790	25.447
T62	5.953	30.498	0.966	3.945	37.469	22.506	0.801	24.852
ZSU23-4	6.095	29.831	0.967	4.243	38.364	21.838	0.804	26.724
2S1	6.883	29.644	0.966	4.474	43.320	21.652	0.798	28.196
BRDM2	10.154	27.135	0.965	6.593	63.887	19.143	0.794	41.551
T72	10.449	26.168	0.965	7.369	65.775	18.175	0.795	46.434
BTR60	11.642	25.760	0.962	7.864	73.271	17.768	0.775	49.568
BMP2	12.172	25.510	0.965	8.193	76.617	17.518	0.789	51.638
BTR70	12.739	25.212	0.964	8.657	80.164	17.220	0.782	54.580
Background	1.938	41.448	0.997	1.074	12.206	33.455	0.985	6.767

Note: ↑ and ↓ indicate that higher and lower values are preferred, respectively.

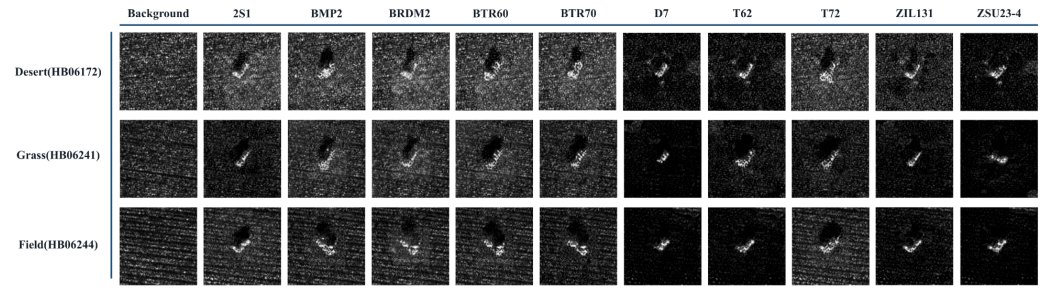


Figure 5. Qualitative results of class-conditional target generation across diverse measured backgrounds. The columns correspond to ten distinct MSTAR target categories, while the rows represent three diverse measured backgrounds.

3.3.1. Image-Level Realism

To evaluate the image-level realism in the physical pixel domain, we assess three spatial regions corresponding to the model's design: the localized target, the target–background transition, and the distant measured background. To explicitly specify the evaluation geometry, we define two concentric square boundaries centered in the image domain based on empirical observations. The inner boundary forms a 40×40 reference box that empirically encompasses the primary scattering signatures of the localized target. Expanding this inner box by 12 pixels on each side yields a 64×64 outer bounding box. Based on this concentric layout, the spatial domains are strictly partitioned: the area within the inner 40×40 box designates the localized target region, the annular area between the inner and outer boxes constitutes the target–background transition (Ring), and the entire area outside the outer 64×64 box represents the distant background. It is important to note that while the training phase employs continuous Gaussian soft masks (as formulated in Equation (18)) to ensure differentiable and seamless blending optimization, this evaluation phase utilizes discrete rectangular boundaries. This hard partitioning conceptually mirrors the functional roles of the Core, Ring, and Outer regions, providing the unweighted, binary pixel sets strictly required for standard quantitative metrics (e.g., SSIM, HOG), which are otherwise inapplicable to continuous spatial weighting fields.

For the localized target region, since the target is synthesized ab initio on a pure background via a random latent variable, no paired physical ground truth exists. To address this, we employ the unified Hungarian-matched similarity framework proposed in our previous work [17]. Specifically, the similarity is computed between the generated set and the real MSTAR training set on a per-class basis. This class-wise set-level matching [38] effectively eliminates arbitrary sequential pairing bias. Within this framework, two complementary metrics are adopted to measure local similarity: structural similarity (SSIM) [23] and directional gradient histogram similarity (HOG) [39]. The Hungarian-matched SSIM (HM-SSIM) evaluates the overall structural consistency between the real and generated target patches, while HM-HOG$_x$ and HM-HOG$_y$ capture fine-grained edge and geometric accuracy by comparing directional gradients along the horizontal and vertical axes. The scores of the ten target classes are finally averaged to obtain the overall results. As reported in Table 5, MIRB-Net achieves the highest mean HM-SSIM of 0.129, and directional gradient similarities of 0.994 and 0.993, outperforming CycleGAN, BCNet, and SHFGAN. This confirms the synthesis of highly accurate localized target contours and dominant scattering centers.

Table 5. Quantitative Evaluation of Image-Level Realism Across Three Spatial Regions. Results are reported as mean ± standard deviation over three independent runs.

Method	Localized Target			Target–Background Transition			Distant Background	
	HM-SSIM↑	HM-HOG_x↑	HM-HOG_y↑	Ring-Mean↓	Ring-Var↓	Ring-Grad↓	BG-SSIM↑	BG-HistSim↑
CycleGAN	0.051 ± 0.009	0.958 ± 0.022	0.952 ± 0.026	29.995 ± 3.312	689.392 ± 55.714	84.415 ± 8.623	0.482 ± 0.036	0.422 ± 0.043
BCNet	0.087 ± 0.011	0.949 ± 0.018	0.942 ± 0.020	22.349 ± 2.417	585.750 ± 42.806	169.404 ± 10.847	0.511 ± 0.032	**0.805 ± 0.029**
SHFGAN	0.084 ± 0.002	0.970 ± 0.008	0.975 ± 0.007	23.580 ± 2.840	519.930 ± 22.760	81.560 ± 5.620	0.492 ± 0.054	0.584 ± 0.010
MIRB-Net	**0.129 ± 0.008**	**0.994 ± 0.005**	**0.993 ± 0.006**	**20.450 ± 2.019**	**390.070 ± 34.872**	**74.804 ± 6.281**	**0.666 ± 0.027**	0.684 ± 0.045

Note: ↑ and ↓ indicate that higher and lower values are preferred, respectively. The best results are highlighted in bold.

For the target–background transition, evaluating coupling naturalness is essential because synthesizing a target locally on an original measured background can easily introduce artificial intensity shifts, texture discontinuities, or broken edges at the interface [25,40]. To quantify this, we compute three deviation metrics between the generated image and the original input background within the Ring: Ring Mean Deviation, Ring Variance Deviation, and Ring Gradient Difference. Specifically, the mean deviation measures unnatural global brightening or darkening of the local background; the variance deviation captures changes in texture fluctuation and scattering dispersion; and the gradient difference evaluates structural continuity and boundary artifacts. Since each generated image possesses an exact paired original background, these metrics are calculated for all generated samples across the ten target classes, and the overall averages are reported. As shown in Table 5, MIRB-Net obtains the lowest values across all three metrics. This indicates that the proposed residual blending mechanism and Ring-consistency constraint effectively mitigate abrupt intensity jumps and structural discontinuities, yielding a smooth and natural target–background transition.

For the distant measured background, we aim to evaluate how well the original clutter is preserved outside the intended editing area [1,4]. Since the original target-free input serves as an exact paired physical reference, we directly compute two paired metrics within the distant background region: background Structural Similarity (BG-SSIM) [23] and background Histogram Similarity (BG-HistSim). Specifically, BG-SSIM evaluates the precise preservation of local textures, spatial structures, and intensity variations at their exact pixel locations, whereas BG-HistSim measures the overall consistency of the global gray-level distribution. As shown in Table 5, MIRB-Net achieves the highest BG-SSIM of 0.666, significantly surpassing CycleGAN (0.482), BCNet (0.511), and SHFGAN (0.492). Although BCNet yields a higher BG-HistSim (0.805), its low BG-SSIM indicates that merely matching the global intensity histogram does not guarantee the structural preservation of individual background scattering centers. This comprehensive spatial evaluation verifies that MIRB-Net successfully confines modifications to the specified target neighborhood while effectively maintaining the structural and statistical integrity of the supplied measured background.

3.3.2. Distribution-Level Realism

While image-level metrics measure physical spatial quality, they do not inherently guarantee that the generated samples align with the real MSTAR data in the high-dimensional

feature space. To comprehensively assess this distribution-level realism, we evaluate the generated set using deep feature metrics and bidirectional recognition tasks.

To quantify the feature distribution discrepancy, we introduce the SAR Kernel Inception Distance (SAR-KID). The original KID [41] is a widely used metric in optical image generation that quantifies the underlying gap in visual fidelity and high-level semantics between real and generated images. It achieves this by computing the squared maximum mean discrepancy (MMD) between their deep feature sets, without making parametric distribution assumptions. Given the feature sets X and Y extracted from real and generated images, respectively, KID is calculated as:

$$\mathrm{KID}(X, Y) = \mathrm{MMD}^2(X, Y). \tag{23}$$

Considering the significant differences between SAR and optical imagery, we extend this metric to the SAR domain, following the adaptation strategy proposed in our previous work [17]. Specifically, instead of employing the conventional Inception-v3 network pretrained on natural images, we utilize an Inception-v3 model pretrained exclusively on the 17° depression-angle MSTAR dataset as the feature extractor. This adaptation ensures the extraction of SAR-specific deep representations, enabling a more reliable characterization of the feature distributions corresponding to SAR scattering properties.

Concurrently, we evaluate class separability using the SAR Inception Score (SAR-IS), an adaptation of the original Inception Score (IS) [42] for the SAR domain, as established in our previous work [17]. The IS measures both the clarity of individual samples and the overall diversity of the generated set, and it is defined as:

$$\mathrm{IS} = \exp\Big(\mathbb{E}_{x\sim p_g}[D_{\mathrm{KL}}(p(y \mid x) \,\|\, p(y))]\Big), \tag{24}$$

where $p(y \mid x)$ denotes the class distribution predicted by a pretrained classification network, and $p(y)$ represents the marginal distribution across all generated samples. A high IS requires (i) a sharp conditional distribution $p(y \mid x)$, indicating that each generated image contains a clear and recognizable structure, and (ii) a diverse marginal distribution $p(y)$, indicating broad category coverage without mode collapse. To adapt this metric to SAR imagery, we strictly used the same feature-extraction framework employed in the SAR-KID computation.

As reported in Table 6, MIRB-Net consistently achieves the lowest SAR-KID (0.091 ± 0.022) and the highest SAR-IS (1.978 ± 0.085) across independent runs. Relative to CycleGAN, BCNet, and SHFGAN, MIRB-Net reduces the SAR-KID by 85.9%, 73.3%, and 62.6%, while improving the SAR-IS by 33.6%, 31.5%, and 0.9%, respectively. The elevated SAR-KID and generally reduced SAR-IS of the baselines, along with their higher variances, suggest that although they may produce target-like visual responses, their deep features deviate significantly from real SAR distributions, suffering from indistinct class-related structures. In contrast, MIRB-Net explicitly and stably aligns the generated manifold with the real MSTAR distribution while maintaining distinct, discriminative class boundaries.

To further verify the distributional consistency of the generated images, we introduce bidirectional recognition evaluations. The fundamental premise is that if the synthetic samples genuinely capture the underlying feature manifold of the real data, the two distributions should share equivalent class-discriminative decision boundaries, allowing a classifier trained on one domain to generalize robustly to the other. To this end, we utilize VGG16 and ResNet101 backbones under two complementary protocols: Train on Real, Test on Synthetic (TRTS), and Train on Synthetic, Test on Real (TSTR).

Table 6. Quantitative Evaluation of Distribution-Level Realism. Results are reported as mean ± standard deviation over three independent runs.

Method	Statistical Metrics		TRTS (%)		TSTR (%)	
	SAR-KID↓	SAR-IS↑	VGG16	ResNet101	VGG16	ResNet101
CycleGAN	0.644 ± 0.089	1.480 ± 0.142	10.18 ± 1.31	7.27 ± 1.02	9.00 ± 5.18	7.45 ± 1.95
BCNet	0.341 ± 0.047	1.504 ± 0.098	52.72 ± 1.78	15.78 ± 1.41	59.97 ± 2.94	33.21 ± 3.69
SHFGAN	0.243 ± 0.021	1.960 ± 0.117	69.70 ± 2.11	71.50 ± 2.32	68.93 ± 1.91	77.54 ± 1.69
MIRB-Net	**0.091 ± 0.022**	**1.978 ± 0.085**	**82.24 ± 1.47**	**77.50 ± 2.23**	**85.29 ± 1.09**	**89.18 ± 1.39**

Note: ↑ and ↓ indicate that higher and lower values are preferred, respectively. The best results are highlighted in bold.

Under the TRTS protocol, classifiers trained purely on real MSTAR data accurately recognize the generated targets (Figure 6a,b). MIRB-Net achieves accuracies of 82.24% (VGG16) and 77.50% (ResNet101), indicating that the synthesized targets successfully preserve authentic, machine-recognizable class semantics rather than merely exhibiting generic bright spots.

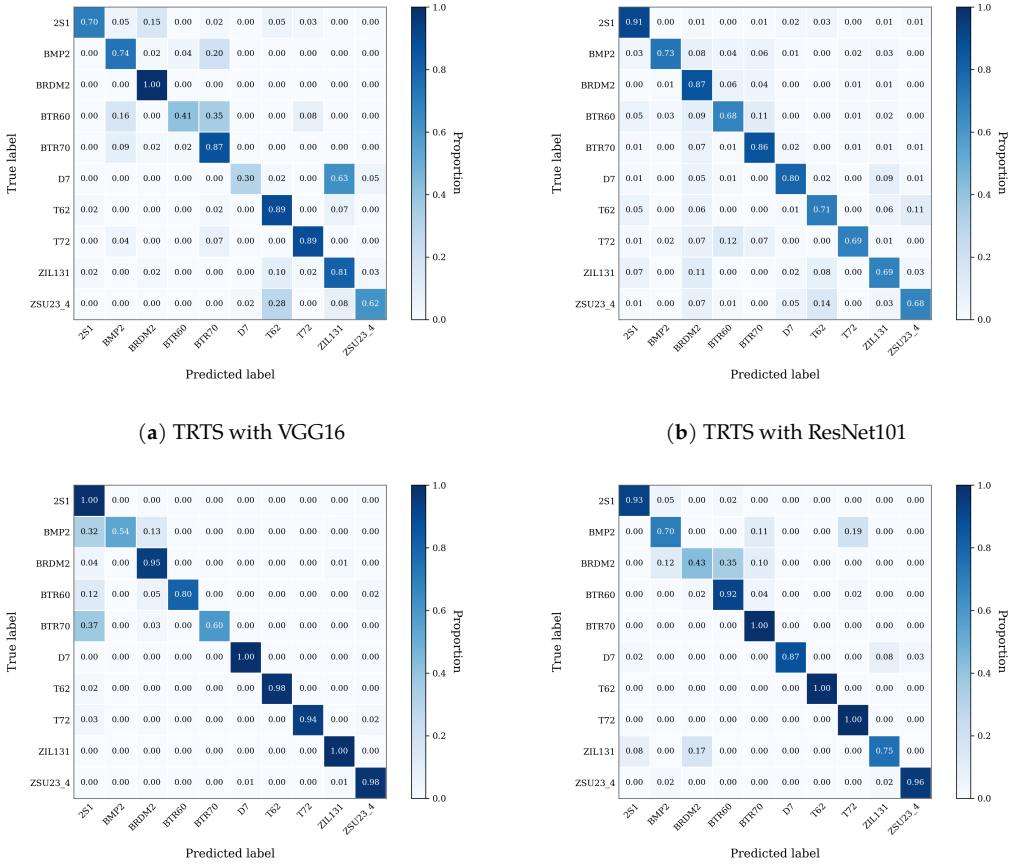
The TSTR protocol provides an even stricter evaluation, assessing whether the generated data can act as a standalone training source to teach classifiers features that transfer back to the physical world. As shown in Figure 6c,d and Table 6, classifiers trained solely on MIRB-Net samples stably achieve remarkable accuracies of 85.29 ± 1.09% and 89.18 ± 1.39% on the real independent test set. Although SHFGAN provides partial recovery (achieving up to 77.54%), the catastrophic failure of CycleGAN and BCNet under TSTR (falling below 60% with significant variance) proves that naively increasing the quantity of synthetic data is futile without strict, stable distribution-level realism. Taken together, the TRTS and TSTR protocols demonstrate that MIRB-Net samples possess robust semantic consistency and substantial, reliable downstream utility.

3.4. Extrinsic Practical Utility

Ultimately, the practical value of a generative model for SAR ATR hinges on whether its synthetic samples can provide transferable, class-discriminative information to enhance the recognition of real measured targets. To validate this downstream utility, hybrid ATR classification experiments were conducted. We employ four representative recognition backbones covering distinct inductive biases: Inception-v3 [43], VGG16 [44], ResNet18 [45], and the Transformer-based Swin-T [46]. The training set for the ATR models consisted of real MSTAR images from ten target categories captured at a 17° depression angle, augmented with the corresponding synthetic samples at a 1:1 mixing ratio. The evaluation is conducted strictly on the independent real MSTAR test set, ensuring that the reported metrics reflect true generalization to physical SAR observations rather than mere overfitting to the synthetic distribution.

As reported in Table 7, augmenting the training set with synthetic data does not uniformly guarantee performance improvements. Specifically, CycleGAN and SHFGAN cause a negative transfer on Swin-T (dropping from 84.38 ± 0.59% to 83.29 ± 0.91% and 83.10 ± 0.55%, respectively), while BCNet degrades the ResNet18 accuracy (from 97.82 ± 0.31% to 96.98 ± 0.37%). Additionally, SHFGAN degrades the accuracy and introduces severe instability on Inception-v3 (dropping to 96.17 ± 2.38%). Since all methods share the identical augmentation ratio and training protocol, these performance drops and elevated variances clearly indicate that naively expanding the dataset size is detrimental if the

generated samples suffer from unnatural target–background coupling, domain shift, or semantic ambiguity.



(**a**) TRTS with VGG16

(**b**) TRTS with ResNet101

(**c**) TSTR with VGG16

(**d**) TSTR with ResNet101

Figure 6. Confusion matrices of MIRB-Net under the TRTS and TSTR protocols using VGG16 and ResNet101.

Table 7. Recognition Accuracy (%) on the Real MSTAR Test Set Under 1:1 Real-Synthetic Hybrid Training. Results are reported as mean ± standard deviation over three independent runs.

Backbone	Real Only	Real + CycleGAN	Real + BCNet	Real + SHFGAN	Real + MIRB-Net
Inception-v3	96.69 ± 0.37	97.98 ± 0.63	97.49 ± 0.42	96.17 ± 2.38	**98.06 ± 0.48**
VGG16	82.35 ± 0.72	89.84 ± 1.07	90.38 ± 0.65	92.35 ± 0.38	**95.99 ± 0.53**
Swin-T	84.38 ± 0.59	83.29 ± 0.91	86.04 ± 0.52	83.10 ± 0.55	**87.91 ± 0.39**
ResNet18	97.82 ± 0.31	98.18 ± 0.43	96.98 ± 0.37	98.30 ± 0.72	**98.58 ± 0.24**

Note: The best results are highlighted in bold.

Conversely, MIRB-Net is the only method that consistently and stably improves the Real Only baseline across all four heterogeneous architectures, achieving the highest cross-architecture average accuracy of 95.14% (a substantial absolute gain of 4.83 percentage points). Notably, MIRB-Net adapts robustly to different operating regimes with generally narrowed standard deviations: it secures crucial complementary gains even when the classifier is near saturation (e.g., pushing ResNet18 from 97.82 ± 0.31% to

98.58 ± 0.24%), while delivering massive assistance when the real-data baseline leaves ample room for improvement (e.g., boosting VGG16 from 82.35 ± 0.72% to 95.99 ± 0.53%). This cross-architecture consistency and statistical stability provide compelling complementary evidence to the preceding structural and distributional evaluations: by restricting generation to localized target insertion on authentic backgrounds, MIRB-Net successfully produces synthetic samples that possess strict semantic alignment and genuine downstream utility.

3.5. Ablation Study

To isolate the individual contributions of the primary components in MIRB-Net, we conduct a comprehensive single-factor ablation study. The evaluated modules include the Background-Preserving Residual Editing (BPRE) pretraining strategy (formulated as the MIP task), the BPRE dual-head mechanism, and the local discriminator (Local D). For each variant, exactly one component is modified or deactivated, while the data splits, optimization configurations, and evaluation protocols remain strictly identical to those of the complete model.

Specifically, the three ablation variants are defined as follows:

1. w/o BPRE Pretraining (MIP): The BPRE pretraining stage is omitted. The generator is initialized randomly from scratch rather than leveraging the pretrained weights θ_G^{MIP}.
2. w/o BPRE Mechanism: The dual-head residual blending mechanism (Alpha Head and Residual Head) is removed. The generator is modified to directly predict the complete output image $\hat{x}$ rather than computing $x_b + \alpha \odot \Delta$.
3. w/o Local D: The patch-level local discriminator and its associated adversarial objective are removed, leaving only the global discriminator to provide adversarial supervision.

As reported in Table 8, the full MIRB-Net achieves the best performance across all metrics. Removing any single module incurs a measurable degradation in specific aspects, revealing a highly decoupled and synergistic architecture.

Table 8. Quantitative Ablation Results of the Main Components in MIRB-Net.

Variant	HM-HOG$_x$ ↑	HM-HOG$_y$ ↑	Ring-Grad↓	BG-SSIM↑	SAR-KID↓
w/o BPRE Mechanism	0.988	0.990	88.241	0.451	0.101
w/o Local D	0.972	0.974	75.601	0.632	0.147
w/o BPRE Pretrain	0.985	0.984	78.122	0.653	0.607
Full MIRB-Net	**0.994**	**0.993**	**74.804**	**0.666**	**0.091**

Note: ↑ and ↓ indicate that higher and lower values are preferred, respectively. The best results are highlighted in bold.

Effect of the BPRE Mechanism: When the BPRE mechanism is removed (*w/o BPRE Mechanism*), the explicit separation of *what* to modify (Δ) and *where* to modify (α) is lost. The generator is forced to reconstruct the entire output image directly. Consequently, this variant suffers a catastrophic 32.3% relative drop in BG-SSIM (from 0.666 to 0.451) and a significant increase in Ring-Grad (from 74.804 to 88.241), despite maintaining high target-structure scores. This asymmetric degradation indicates that the dual-head BPRE plays a key role in preventing unnecessary global modifications, restricting generation primarily to the intended local region and facilitating a natural target–background transition.

Effect of Local D: The *w/o Local D* variant omits the patch-level adversarial objective. Without this focused supervision, the global discriminator alone fails to adequately penalize localized structural artifacts. This leads to the most severe drop in directional structural metrics (HM-HOG$_x$ and HM-HOG$_y$ fall to 0.972 and 0.974, respectively), while the background metrics remain largely unaffected. This validates that Local D is essential for enforcing precise scattering-center organization and distinct target contours.

Effect of BPRE Pretraining: In the *w/o BPRE Pretraining* variant, the generator is trained from random initialization rather than inheriting the spatial gating and local editing priors established during the masked reconstruction task. Interestingly, while the physical spatial metrics (BG-SSIM, Ring-Grad, and HOG) suffer only marginal declines, the distribution-level discrepancy spikes dramatically, with SAR-KID surging by roughly 6.7 times (from 0.0911 to 0.6067). This phenomenon indicates that while pure conditional adversarial training can visually approximate basic target shapes, it fundamentally struggles to capture the authentic SAR feature manifold. MIP serves as a crucial prior, embedding both target scattering patterns and background statistics into the network initialization to guarantee high-dimensional distributional realism.

Taken together, these results demonstrate a clear division of responsibilities: BPRE restricts spatial modifications to preserve background context, Local D enhances physical structural details, and MIP governs global feature-space alignment. The complete MIRB-Net successfully unifies these three aspects to generate highly realistic, semantically reliable, and naturally coupled SAR samples.

4. Discussion

The central challenge in synthesizing SAR targets onto measured clutter is ensuring that the generated samples provide transferable, physically plausible features for downstream automatic target recognition. The following discussion comprehensively analyzes MIRB-Net from two critical perspectives. First, we compare it with existing SAR image augmentation paradigms to highlight its instance-free, background-preserving formulation. Second, we examine the inherent limitations of current SAR image evaluation protocols, discussing how our multidimensional framework addresses these issues while outlining the remaining physical limitations and future directions.

4.1. Comparison with Existing SAR Augmentation Paradigms

MIRB-Net should be evaluated according to its specific task setting—instance-free, localized target formation on measured backgrounds—rather than through a universal ranking of generative methods. Deep generative models typically approach SAR synthesis via holistic, full-chip regeneration [10,13,15,18]. Such formulations are well-suited for unconstrained data augmentation but risk distorting authentic clutter statistics when the background must act as an invariant scene carrier. Target–background fusion methods address scene integration, yet they fundamentally operate as adaptation pipelines requiring preexisting target chips or patches as structural inputs [8,24,25]. In contrast, MIRB-Net relies solely on a class label and latent noise, formulating the generation as a localized residual editing problem to achieve instance-free target synthesis while effectively maintaining the surrounding measured background.

More importantly, data-driven synthesis models like MIRB-Net provide a complementary route to physics-based SAR simulation tools. Established physics-based simulators, such as RaySAR [6] and SARCASTIC v2.0 [7], derive SAR signatures directly from three-dimensional target geometry and electromagnetic scattering models. These tools excel in physical interpretability, allowing for the rigorous simulation of complex phase histories, polarimetric responses, and multi-path scattering under precisely controlled collection parameters. However, they typically require detailed CAD models, material dielectric properties, and explicit scene definitions. MIRB-Net, conversely, learns class-conditioned macroscopic appearances and spatially localized edits directly from measured intensity images. A key advantage lies in enabling rapid, varied target synthesis directly onto authentic operational backgrounds without requiring prior geometric modeling. The intended application domain of MIRB-Net is image-level ATR dataset augmentation under the studied

imaging conditions, whereas physics-based simulators remain indispensable when precise physical accuracy or complex electromagnetic phase analysis is required.

4.2. Evaluation Limitations and Future Work

Establishing an objective and comprehensive evaluation framework is critical for validating generative models [19]. Conventional SAR synthesis studies [10,13–16,18] predominantly rely on visual inspection, feature-space distances (e.g., FID) [20], image similarity metrics (e.g., global SSIM) [23,37], and downstream ATR experiments [2,13]. While these metrics capture isolated properties, a single global metric rarely delineates structural realism, statistical fidelity, and contextual naturalness simultaneously [19,47]. For target–background coupled generation, traditional full-image similarity scores are disproportionately dominated by the large ratio of invariant background pixels, often obscuring critical anomalies such as boundary intensity discontinuities, scattering gradient errors, or localized speckle distortion [4].

To precisely align with the objectives of background-conditional local generation and address the shortcomings of global metrics, we established a task-oriented multidimensional evaluation protocol. This framework systematically decompounds evaluation into four decoupled dimensions: (1) local target realism [1], (2) background preservation [1,4], (3) target–background coupling naturalness [40], and (4) ATR utility [13]. By employing this region-wise spatial decoupling, the evaluations provide strong evidence of the physical integration within MIRB-Net: achieving high structural fidelity in the localized target region (HM-SSIM, HM-HOG), maintaining stable radiometric transitions at the boundary interface (Ring-Grad), and reliably preserving the distant measured clutter (BG-SSIM).

Despite these robust spatial constraints and comprehensive metrics, the absolute physical validation of target–environment interactions remains a fundamental limitation of the current evaluation framework. Since MIRB-Net is optimized on non-cooperative measured SAR datasets (e.g., MSTAR), paired physical ground-truth images representing the exact same target situated identically within the exact same background are inherently unavailable. Consequently, while the adopted multidimensional metrics rigorously evaluate spatial compatibility and radiometric fidelity in the image domain, they cannot fully guarantee the precise electromagnetic accuracy of complex physical interactions, such as exact layover overlap or target–ground multiple bounce scattering. Furthermore, the current framework is primarily validated on the MSTAR dataset and focuses on localized single-target generation at relatively fixed scales.

Future endeavors will extend this framework to accommodate arbitrary target placement, multi-scale generation, and multi-target complex scene construction. Finally, exploring physics-guided machine learning frameworks—integrating explicit electromagnetic constraints, 3D geometric priors, and complex operational conditions (e.g., targets obscured by multi-band camouflage or active deception equipment)—will be investigated to bridge the gap between data-driven appearance learning and rigorous physical interaction modeling for real-world reconnaissance applications.

5. Conclusions

Synthesizing realistic targets directly onto authentic, target-free clutter backgrounds has emerged as a crucial strategy for enhancing the generalization capability of SAR ATR models. However, conventional generation or target-insertion methods often struggle to maintain natural local transitions, frequently introducing abrupt artificial boundaries or violating the physical coupling inherent in SAR imagery. To address these limitations, our proposed MIRB-Net formulates this task as a localized target editing problem. By explicitly decoupling *where* to modify (via spatial gating) and *what* to modify (via local residual

prediction), MIRB-Net synthesizes class-consistent targets while facilitating a structurally seamless and physically natural target–background transition.

The success of MIRB-Net is underpinned by three synergistically designed components. First, the Background-Preserving Residual Editing (BPRE) mechanism restricts structural modifications primarily to the target region, effectively minimizing unnecessary alterations to the distant clutter. Second, the BPRE pretraining strategy (formulated via masked inpainting) equips the generator with crucial prior knowledge of target scattering, background statistics, and spatial gating before adversarial optimization. Third, the joint global–local adversarial supervision, augmented by boundary consistency constraints, promotes fine-grained target fidelity and localized coupling. To rigorously validate the quality of this target-on-background generation task, we adopted a comprehensive evaluation framework encompassing both image-level and distribution-level metrics. Extensive evaluations under this framework demonstrate the effectiveness of the proposed method in synthesizing high-fidelity targets, effectively preserving measured backgrounds, promoting natural local target–background coupling, and maintaining global distribution consistency. Furthermore, hybrid data-augmentation classification experiments demonstrate that the generated samples deliver valuable, transferable discriminative information, effectively boosting downstream ATR performance across diverse CNN and Transformer architectures.

Author Contributions: Conceptualization, Y.Z. and R.L.; methodology, Y.Z., Z.W. and R.L.; investigation, W.J. and J.F.; formal analysis, Y.Z. and R.L.; validation, Y.Z. and Z.W.; visualization, Y.Z., Z.W. and R.L.; supervision, R.L.; writing—original draft, Y.Z. and Z.W.; writing—review and editing, Y.Z., J.F. and R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research and Development Program of Shaanxi (Program No. 2024SF2-GJHX-17), the Shaanxi Provincial Innovation Capability Support Program (Grant No. 2025ZC-KJXX-105), the Xidian University Specially Funded Project for Interdisciplinary Exploration (Grant No. TZJH2024026), 2024 Cross-Innovation Special Project of 986 Hospital, Air Force Military Medical University (Grant No. 2024JC064), China Postdoctoral Science Foundation (Grant No. 2025M781519).

Data Availability Statement: The MSTAR dataset analyzed in this study is publicly available from the Air Force Research Laboratory. The generated SAR images and related experimental data are available from the corresponding author upon reasonable request.

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

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

References

1. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A Tutorial on Synthetic Aperture Radar. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–43. [CrossRef]
2. Geng, Z.; Xu, Y.; Wang, B.N.; Yu, X.; Zhu, D.Y.; Zhang, G. Target recognition in SAR images by deep learning with training data augmentation. *Sensors* **2023**, *23*, 941. [CrossRef] [PubMed]
3. Keydel, E.R.; Lee, S.W.; Moore, J.T. MSTAR extended operating conditions: A tutorial. In *Proceedings of the SPIE Algorithms for Synthetic Aperture Radar Imagery III*; SPIE: Bellingham, WA, USA, 1996; Volume 2757, pp. 228–242. [CrossRef]
4. Argenti, F.; Lapini, A.; Bianchi, T.; Alparone, L. A Tutorial on Speckle Reduction in Synthetic Aperture Radar Images. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–35. [CrossRef]
5. Ling, H.; Chou, R.-C.; Lee, S.-W. Shooting and bouncing rays: Calculating the RCS of an arbitrarily shaped cavity. *IEEE Trans. Antennas Propag.* **1989**, *37*, 194–205. [CrossRef] [PubMed]
6. Auer, S.; Bamler, R.; Reinartz, P. RaySAR—3D SAR simulator: Now open source. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 6730–6733. [CrossRef]

7. Woollard, M.; Blacknell, D.; Griffiths, H.; Ritchie, M.A. SARCASTIC v2.0—High-Performance SAR Simulation for Next-Generation ATR Systems. *Remote Sens*. **2022**, *14*, 2561. [CrossRef]
8. Luan, J.; Xu, Z.; Li, B.; Ding, J. BCNet: Background conversion network for SAR data generation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. **2024**, *17*, 3211–3225. [CrossRef]
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27, pp. 2672–2680.
10. Guo, J.; Lei, B.; Ding, C.; Zhang, Y. Synthetic aperture radar image synthesis by using generative adversarial nets. *IEEE Geosci. Remote Sens. Lett*. **2017**, *14*, 1111–1115. [CrossRef]
11. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6840–6851.
12. Zheng, Y.; Wang, R.; Guo, K.; Liang, J. ARWCGAN: A Method for High-Quality Multi-Category SAR Image Generation. *J. Xidian Univ*. **2025**, *52*, 101–112. [CrossRef]
13. Cui, Z.; Zhang, M.; Cao, Z.; Cao, C. Image data augmentation for SAR sensor via generative adversarial nets. *IEEE Access* **2019**, *7*, 42255–42268. [CrossRef]
14. Cao, C.; Cao, Z.; Cui, Z. LDGAN: A synthetic aperture radar image generation method for automatic target recognition. *IEEE Trans. Geosci. Remote Sens*. **2020**, *58*, 3495–3508. [CrossRef]
15. Song, Q.; Xu, F.; Zhu, X.X.; Jin, Y.Q. Learning to generate SAR images with adversarial autoencoder. *IEEE Trans. Geosci. Remote Sens*. **2022**, *60*, 5210015. [CrossRef]
16. Du, S.; Hong, J.; Wang, Y.; Qi, Y. A high-quality multicategory SAR images generation method with multiconstraint GAN for ATR. *IEEE Geosci. Remote Sens. Lett*. **2022**, *19*, 4011005. [CrossRef]
17. Zheng, Y.; Liu, D.; Li, R.; Wang, R.; Fan, J.; Guo, K.; Liang, J. CA-DDPM: Conditionally Embedded Attention-Aided Denoising Diffusion Probabilistic Model for High-Quality SAR Image Generation. *Remote Sens*. **2026**, *18*, 1994. [CrossRef]
18. Qosja, D.; Wagner, S.A.; O'Hagan, D.W. SAR image synthesis with diffusion models. In Proceedings of the IEEE Radar Conference (RadarConf24), Denver, CO, USA, 6–10 May 2024; pp. 1–6. [CrossRef]
19. Theis, L.; van den Oord, A.; Bethge, M. A Note on the Evaluation of Generative Models. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
20. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 6626–6637.
21. Oliver, C.; Quegan, S. *Understanding Synthetic Aperture Radar Images*; SciTech Publishing: Raleigh, NC, USA, 2004. Available online: https://books.google.com/books?id=IeGKe40S77AC (accessed on 29 October 2026).
22. Gao, G. Statistical Modeling of SAR Images: A Survey. *Sensors* **2010**, *10*, 775–795. [CrossRef] [PubMed]
23. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process*. **2004**, *13*, 600–612. [CrossRef] [PubMed]
24. Zhang, X.; Li, Y.; Li, F.; Jiang, H.; Wang, Y.; Zhang, L.; Zheng, L.; Ding, Z. Ship-Go: SAR ship images inpainting via instance-to-image generative diffusion models. *ISPRS J. Photogramm. Remote Sens*. **2024**, *207*, 203–217. [CrossRef]
25. Yang, J.; Wang, W.; Deng, C. SHFGAN: SAR image harmonization fusion with generative adversarial network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. **2026**, *19*, 17550–17565. [CrossRef]
26. Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544. [CrossRef]
27. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 89–105. [CrossRef]
28. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4470–4479. [CrossRef]
29. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph*. **2017**, *36*, 107. [CrossRef]
30. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
31. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.W. StarGAN v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8185–8194. [CrossRef]
32. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1510–1519. [CrossRef]

33. Mescheder, L.; Geiger, A.; Nowozin, M. Which training methods for GANs do actually converge? In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 3481–3490.
34. Rudin, L.I.; Osher, S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D Nonlinear Phenom*. **1992**, *60*, 259–268. [CrossRef]
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
36. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251. [CrossRef]
37. Horé, A.; Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In Proceedings of the 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369. [CrossRef]
38. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Quart*. **1955**, *2*, 83–97. [CrossRef]
39. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
40. Tsai, Y.H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; Yang, M.H. Deep Image Harmonization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2799–2807. [CrossRef]
41. Binkowski, M.; Sutherland, D.J.; Arbel, M.; Gretton, A. Demystifying MMD GANs. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
42. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29, pp. 2234–2242.
43. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
46. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
47. Sajjadi, M.S.M.; Bachem, O.; Lucic, M.; Bousquet, O.; Gelly, S. Assessing Generative Models via Precision and Recall. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 5234–5243. Available online: https://proceedings.neurips.cc/paper/2018/hash/f7696a9b362ac5a51c3dc8f098b73923-Abstract.html (accessed on 29 October 2026).

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