Noise-Tolerant Novel-View SAR Synthesis via Denoising Diffusion

Amir Rahimi, *Member, IEEE,* Stella X. Yu, *Member, IEEE*

***Abstract*—Synthetic Aperture Radar (SAR) enables robust imaging under all weather and lighting conditions, but the scarcity of labeled SAR data limits the use of modern vision models. Novel-view synthesis offers a promising way to augment training data, yet existing methods struggle with speckle noise and radiometric variability inherent to SAR imagery.**

We introduce a SAR-specific self-supervised representation learning framework based on co-domain augmentations that operate directly on pixel magnitudes. By combining multiplicative Rayleigh speckle and random monotonic intensity remapping, our method learns features that are invariant to speckle realizations while preserving structural and geometric cues. These learned representations are then used to supervise a latent-diffusion novel-view generator adapted from zero-1-to-3 through a projected feature-matching loss, replacing fragile pixel-space comparisons with noise-robust feature-space supervision.

Experiments on MSTAR and MSTAR-OOD demonstrate substantial improvements in identity preservation, pose consistency, and perceptual quality for both seen and unseen targets. Although evaluated on object-centric SAR for automatic target recognition, the proposed framework is content-agnostic and naturally extends to scene-level SAR novel-view synthesis.

I. Introduction

SYNTHETIC Aperture Radar (SAR) imaging is an essential element of remote sensing, enabling the capture of high-resolution images under any weather and lighting condition. Its utility spans a wide variety of fields, including surveillance, environmental monitoring, and disaster management. Yet, the automatic interpretation of SAR images poses significant challenges due to their unique, high imaging noise (Figure 1), limited labeled data, and complex objects and scene structures.

Despite recent advances in supervised SAR-based automatic target recognition (ATR), performance remains limited when labeled data are scarce, particularly in single-view settings. Simulated data have been used to mitigate this limitation [1], but such approaches suffer from domain gaps between synthetic and real SAR imagery and generalize poorly to novel objects. Given the persistent scarcity of labeled SAR datasets, unsupervised representation learning and single-image novel-view synthesis have become increasingly important. In scenarios where multi-view acquisition is impractical, generating additional views from a single observation enables richer target characterization and improves ATR robustness and accuracy.

We propose a latent diffusion-based framework for generating novel SAR views at varying azimuth and depression angles from a single unseen target image, inspired by the zero-1-to-3 method [2]. Unlike conventional diffusion models that rely on pixel-space losses (e.g., mean squared error), our approach is tailored to SAR imagery, whose speckle noise and nonuniform illumination render such comparisons uninformative.

A. Rahimi and S. Yu are with the department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48109 USA. Correspondence: stellayu@umich.edu

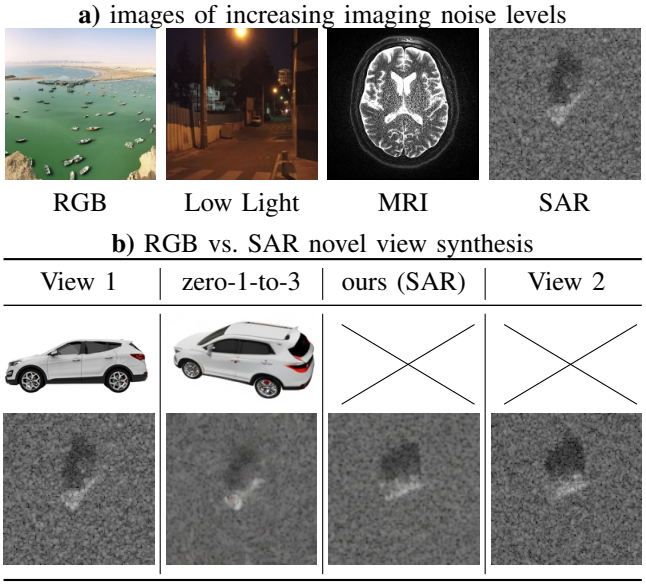

Fig. 1: **Novel-view synthesis for SAR imagery under high imaging noise**. **a)** Compared to RGB, low-light, and MRI images, SAR exhibits substantially higher noise levels, posing challenges for automatic target recognition and view synthesis. **b)** Given a single SAR image and a relative pose, the task is to synthesize a novel view. The RGB-oriented *zero-1-to-3* baseline relies on pixel-level comparisons, which perform poorly on SAR, whereas our method employs noise-tolerant, feature-level comparisons to enable effective SAR view synthesis.

We introduce a feature-space similarity loss that compares synthesized and real SAR images in a learned representation, capturing the essential structural characteristics of SAR targets while remaining invariant to particular speckle realizations. To enable efficient training within a diffusion framework, we further adopt a single-step approximation of the generated image, avoiding the need for full sampling in optimization. We term this objective the *Projected Feature Matching* (PFM) loss.

Leveraging feature-space comparisons enables our second contribution: a SAR-specific data augmentation strategy. SAR imagery exhibits complex speckle patterns and coherent backscatter responses that are often distorted by conventional image-space augmentations, such as geometric transformations. In contrast, co-domain augmentations operate in a manner that preserves the underlying spatial structural integrity and statistical properties of SAR data, producing more realistic and physically meaningful variations.

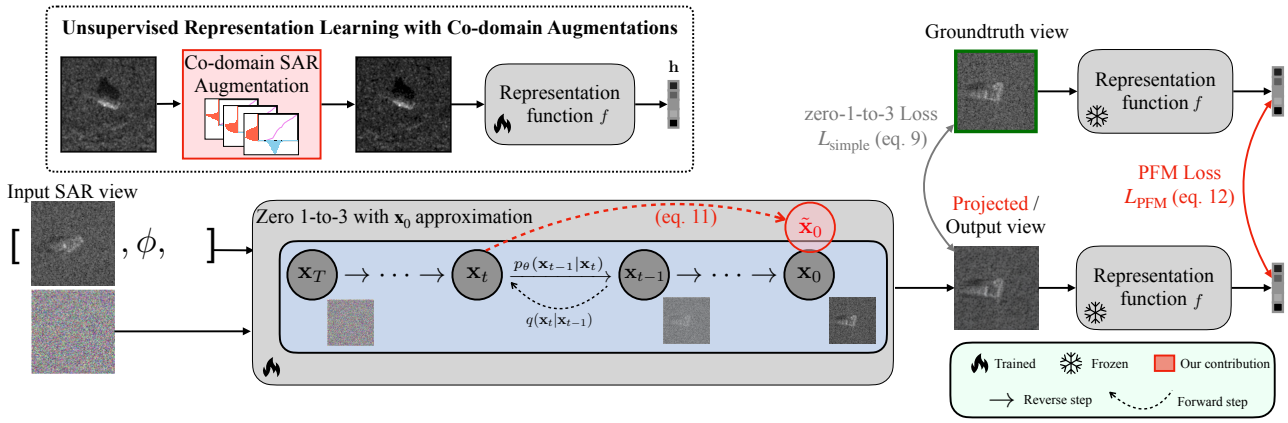

Fig. 2: **SAR novel view synthesis with Projected Feature Matching (PFM) loss**. Conventional novel-view synthesis uses image-space comparisons and is ill-suited for SAR imagery due to noise. We replace pixel-level losses in the zero-1-to-3 model with a feature matching loss on a learned representation robust to SAR-specific noise. **Top)** Unsupervised representation learning using SAR-specific co-domain augmentations, including multiplicative speckle noise and random monotonic transformations. **Bottom)** During novel-view synthesis, to avoid backpropagation through the full reverse diffusion process, we adopt a single-step approximation $\tilde{\mathbf{x}}_0$ of the generated image $\mathbf{x}_0$, used to compute the PFM loss in representation space.

We incorporate contrastive learning with custom co-domain augmentation techniques, including speckle noise modulation and a randomly sampled monotonic pixel-value transformation. These augmentations model illumination variability while preserving SAR structural consistency, yielding more robust representations than conventional image-space methods such as brightness or contrast jittering, without introducing artifacts.

Figure 2 illustrates our overall method. While these techniques can improve data efficiency for SAR ATR, their primary role in our framework is to support the synthesis of structurally coherent SAR images from novel viewpoints, an objective not fully addressed by existing augmentation-based approaches.

Extensive experiments show that our augmentation strategy not only outperforms conventional methods in azimuth angle regression and target classification, but also achieves classification performance comparable to full SimCLR [3] augmentations without relying on geometric transformations. This result is particularly important for SAR analysis, as our augmentations preserve orientation attributes that are critical for interpreting target configurations. Moreover, when combined with our proposed representation loss, the latent diffusion model produces higher-quality novel views of previously unseen objects, outperforming conventional diffusion models that do not exploit representation learning.

While our experiments focus on object-centric SAR data typical of SAR-ATR, the proposed framework is not limited to object-level imagery. The augmentation strategy, diffusion model, and feature learning are broadly applicable and extend naturally to scene-level SAR synthesis. Extending the method to complex scenes remains promising future work.

II. Background

This section reviews the background and notation used in this work, including representation learning with SimCLR and diffusion-based image generation.

SimCLR [3] is an unsupervised contrastive representation learning framework that encourages consistency between representations of different augmentations of the same sample via a contrastive loss. Given a minibatch of N samples $\{\mathbf{x}_i\}_{i=1}^{N}$, SimCLR generates two stochastic augmentations per sample, resulting in $2N$ augmented views $\{\mathbf{x}'_i\}_{i=1}^{2N}$. A base encoder $f(\cdot)$ maps each augmented view to a representation $\mathbf{h}_i = f(\mathbf{x}'_i)$, which is further projected by a small MLP $g(\cdot)$ to $\mathbf{z}_i = g(\mathbf{h}_i)$, where the contrastive loss is applied. Let $\text{sim}(\cdot,\cdot)$ denote cosine similarity, $\tau \in \mathbb{R}_+$ a temperature parameter, and $\mathbb{1}_{[\cdot]}$ the indicator function. The contrastive loss is defined as

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

for a positive pair (i, j). In Section III-A, we introduce SAR-specific augmentations in SimCLR.

Diffusion models are latent variable models whose *forward process* is a fixed Markov chain that progressively adds Gaussian noise to the data $\mathbf{x}_0$ according to a variance schedule $\beta_1, \ldots, \beta_T$. We follow [4] to index the Markov chain by time step t, and compute the approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$ using

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \ . \tag{1}$$

As the number of diffusion steps T increases, the noisy sample $\mathbf{x}_T$ converges to a standard Gaussian distribution, i.e., $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ as $T \to \infty$. The goal is to learn a parametric form of the *reverse process* $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, enabling sampling from a standard Gaussian and iteratively reversing the diffusion process to generate samples from the data distribution:

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{2}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \ . \tag{3}$$

A key property of the *forward process* is that, given $\mathbf{x}_0$, the noisy sample $\mathbf{x}_t$ can be drawn directly in closed form from a

Gaussian distribution with known parameters. Defining $\alpha_t \coloneqq 1-\beta_t$ and $\bar{\alpha}_t \coloneqq \prod_{s=1}^{t} \alpha_s$, we obtain

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{4}$$

The training objective is to minimize the variational bound on the negative log-likelihood of the data distribution:

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right]. \tag{5}$$

Assuming a fixed variance schedule β_t, and the fact that $\mathbf{x}_{t-1}$ follows a Gaussian distribution given $\mathbf{x}_0, \mathbf{x}_t$:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}), \tag{6}$$

$$\text{where } \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) \coloneqq \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t, \tag{7}$$

$$\tilde{\beta}_t \coloneqq \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \ , \tag{8}$$

minimizing the right-hand side in Equation (5) is equivalent to

$$\min \quad \mathbb{E}_q\left[L_0 + \sum_{t\geq 1} L_{t-1}\right] \tag{9}$$

$$\text{where } L_{t-1} = D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \tag{10}$$

$$L_0 = -\log p_\theta(\mathbf{x}_0|\mathbf{x}_1). \tag{11}$$

By fixing $\Sigma_\theta(\mathbf{x}_t, t) = \sigma_t\mathbf{I}$ to time-dependent constants and reparameterizing Equation (4) as

$$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{12}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, Ho et al. [4] further simplify the training objective to a denoising loss in which the model predicts the sampled Gaussian noise at time step t:

$$L_{\text{simple}}(\theta) \coloneqq \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right]. \tag{13}$$

In Section III-B, we use these equations to define the PFM loss based on the previously learned SAR representation.

III. Method

Our method has two components: SAR-specific data augmentation and diffusion-based novel-view synthesis. We introduce co-domain augmentations for SAR imagery and a zero-1-to-3-based diffusion framework enhanced with a novel feature matching loss for feature-space view consistency.

A. SAR Data Augmentation in the Co-domain

SAR imagery exhibits characteristics distinct from natural images, necessitating specialized augmentation strategies for unsupervised representation learning. We define the *co-domain* of a SAR image as the range of its pixel magnitudes, typically normalized to $[0, 1]$. Unlike geometric augmentations that modify the image domain (e.g., translations or rotations), co-domain augmentations act directly on pixel intensities. These transformations preserve spatial structure while introducing radiometric variability, making them well suited for SAR imagery dominated by noise and contrast variations. We propose the following SAR-specific co-domain augmentations.

1. Random monotonic transformation augmentation. SAR image intensities vary due to acquisition and environmental factors, including sensor calibration, speckle realization, quantization, and surface conditions [5]. In particular, environmental variables such as soil moisture strongly influence SAR backscatter, with higher moisture levels producing increased returns due to elevated dielectric constants [6].

Since these effects typically alter the global intensity distribution but not the relative ordering of pixel values, we introduce a family of monotonic intensity remapping functions to simulate radiometric shifts without disturbing spatial arrangements. This design encourages the learning of order-invariant yet structurally faithful representations. Related rank-based strategies have proven effective in SAR matching, including the RLSS descriptor of Xiong *et al.* [7].

Similar observations are reported in [1] when comparing simulated and measured SAR images. To mitigate these variations and promote invariance in the learned representations, we incorporate random monotonic pixel-value transformations. By preserving intensity ordering while altering the global radiometric scale, these augmentations encourage features that are robust to variations in SAR intensity values.

We apply random monotonic transformations to the magnitude values, i.e., the *co-domain*, of SAR image pixels. With a slight abuse of notation, we represent a SAR image as a continuous function $\mathbf{x} : \mathbb{R}^2 \to [0, 1]$, where the domain corresponds to 2D spatial coordinates and the co-domain represents the normalized pixel magnitude. For a pixel at location (u, v), the magnitude is given by $m = \mathbf{x}(u, v)$.

A co-domain transformation T is a scalar mapping applied pointwise to the image co-domain, yielding image $\tilde{\mathbf{x}}$:

$$m \mapsto \tilde{m} = T(m), \tag{14}$$

$$\tilde{\mathbf{x}}(u, v) = T(\mathbf{x}(u, v)). \tag{15}$$

Definition 1. A function $T : [0, 1] \to [0, 1]$ is *monotonic* if $T(m_0) \leq T(m_1)$ for all $0 \leq m_0 \leq m_1 \leq 1$. T is further assumed to be boundary-preserving and range-covering, such that $T(0) = 0$ and $T(1) = 1$. Such a T is denoted by $T_{\text{monotonic}}$.

We construct random instances of $T_{\text{monotonic}}$ by sampling a discrete monotonic mapping on Q uniformly quantized levels over $[0, 1]$, given by $\{0, \frac{1}{Q-1}, \frac{2}{Q-1}, \ldots, 1\}$. Specifically, we draw $Q-1$ independently and identically distributed random values, form their cumulative sum to enforce monotonicity, and normalize the result to $[0, 1]$. The resulting mapping preserves intensity ordering and introduces bounded radiometric perturbations, whose smoothness – but not necessarily amplitude – is governed by the quantization resolution Q. Larger values of Q yield smoother monotonic transformations by reducing interpolation artifacts between discrete intensity levels.

To obtain a broader family of monotonic transformations, we optionally apply a random gamma remapping of the form

$$T_\gamma(m) = m^{e^\alpha}, \text{where } \alpha \sim \mathcal{N}(0, \gamma). \tag{16}$$

The resulting discrete mapping is applied to continuous pixel magnitudes via linear interpolation. Algorithm 1 summarizes the pseudocode, and Figure 3 illustrates 50 sampled transformations for $\gamma \in \{0, 1\}$ and $Q \in \{16, 256\}$.

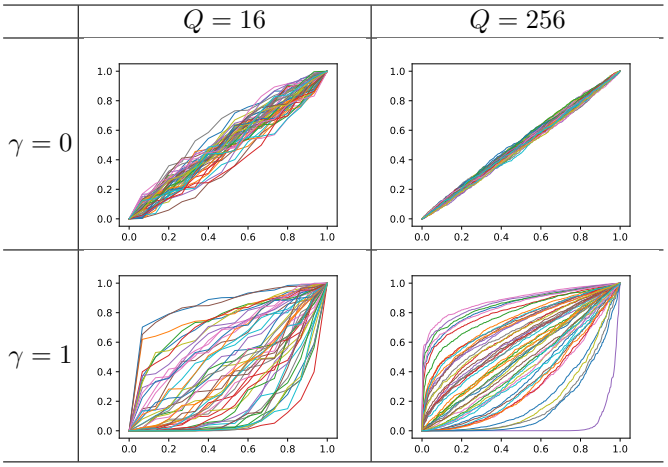

Fig. 3: **Effect of γ and Q on random monotonic transformations** $T_\gamma \cdot T_{\text{monotonic}}$. Each panel shows 50 randomly sampled monotonic mappings generated with specific (γ, Q) values. Larger values of Q yield smoother transformations while preserving intensity ordering. γ increases the expressivity of the transformations by introducing stronger nonlinear distortions.

Algorithm 1: Random monotonic co-domain transform
Input: Image $\mathbf{x}$, quantization level Q, gamma γ
Output: Transformed image $\tilde{\mathbf{x}}$

1. Construct a monotonic intensity mapping $T_{\text{monotonic}}$ by cumulatively summing 0 and $Q-1$ random values, followed by normalization to $[0, 1]$.
2. If $\gamma \neq 0$, further apply a random gamma remapping $T_\gamma(m)$ to the monotonic mapping $T_{\text{monotonic}}$.
3. Apply the resulting monotonic mapping pointwise to the magnitude image $\mathbf{x}$ via interpolation.
4. Return $\tilde{\mathbf{x}} = T_\gamma(T_{\text{monotonic}}(\mathbf{x}))$.

2. Speckle noise augmentation. The presence of speckle noise, arising from the physics of coherent radar imaging, poses a major challenge in modeling SAR images [8]. Speckle noise can significantly degrade image quality and complicate interpretation, making its accurate modeling a critical component of SAR image processing pipelines. By incorporating augmentation strategies that mimic speckle noise, our unsupervised learning framework is encouraged to learn representations that are more robust and invariant to such distortions.

Similar to the monotonic transformation augmentation, we apply speckle noise in the co-domain by perturbing pixel magnitudes. We define multiplicative speckle noise as

$$T_{\text{speckle}}(m; \sigma) \coloneqq m\,\epsilon_\sigma, \qquad \epsilon_\sigma \sim \text{Rayleigh}(\sigma) \tag{17}$$

where m denotes the original pixel magnitude and ϵ_σ is a random variable drawn from a Rayleigh distribution with scale parameter σ. We adopt the Rayleigh distribution since the magnitude of complex Gaussian noise – commonly assumed in SAR signal models – follows a Rayleigh distribution.

Although SAR intensities are not globally Rayleigh-distributed, especially in the presence of dominant scatterers, the use of Rayleigh-distributed speckle in our augmentation is motivated by the classical multiplicative model of SAR image formation [9], [10]. In that model, the observed intensity is expressed as the product of a texture term and a speckle term, with the latter commonly modeled as Rayleigh-distributed in homogeneous or local patch regions [11]. While this approximation does not capture all scattering effects in object-centric SAR imagery, it provides a practical and physically interpretable mechanism for injecting realistic stochastic variation.

In our contrastive learning framework, each training instance is augmented multiple times using independent speckle realizations. This diversity encourages the model to learn robust, geometry-aware representations by exposing it to a range of speckle patterns in local texture regions. Although more flexible speckle models, e.g., the G^0 distribution [11], could be explored in future work, our simple Rayleigh-based augmentation has proven effective in practice and aligns well with both SAR imaging physics and representation learning objectives.

We follow [12] and adopt a weighted combination of the original magnitude values and multiplicative speckle noise to simulate realistic SAR distortions:

$$T_{\text{speckle}}(m; \sigma, \lambda_1, \lambda_2) \coloneqq \lambda_1 m + \lambda_2 m \epsilon_\sigma \ , \tag{18}$$

where λ_1, λ_2 are positive hyperparameters selected via cross-validation. Compared to the conventional multiplicative form $m\epsilon_\sigma$, this generalized formulation enables controllable variability by independently adjusting the intensity scaling (λ_1) and speckle amplitude (λ_2). We find this parameterization to be more effective, as it facilitates the simulation of diverse yet physically plausible SAR radiometric conditions.

In homogeneous regions, this transformation may produce a variance slightly exceeding that predicted by the Rayleigh model. Such mild overdispersion can be beneficial, as it exposes the generative model to a broader range of noise conditions, enhancing robustness without altering the main conclusions.

To stabilize training, we normalize the constructed mappings to the range $[0, 1]$ via min-max normalization, ensuring that all transformations map $[0, 1]$ to $[0, 1]$. The mappings are defined independently of pixel values and applied to the image co-domain through linear interpolation, as used in Algorithm 1 for the monotonic transformation.

Compared to additive Gaussian noise as a standard image-space baseline, Figure 4 shows that, our two co-domain augmentations – random monotonic transform and speckle noise – better preserve object structure and fine details in SAR images while introducing realistic radiometric variability.

When evaluated on downstream classification tasks, with all hyperparameters selected via cross-validation, multiplicative speckle augmentation preserves image structure most effectively and yields strongest performance.

We adopt **SimCLR for unsupervised representation learning on SAR imagery**. Specifically, we generate multiple views of each SAR magnitude image by sampling combinations of the random monotonic transformation $T_{\text{monotonic}}$ and the speckle augmentation T_{speckle}. The resulting encoder produces a learned representation $\mathbf{h} = f(\mathbf{x})$ for a SAR image $\mathbf{x}$, which we subsequently use to define the projected feature matching (PFM) loss within the zero-1-to-3 framework.

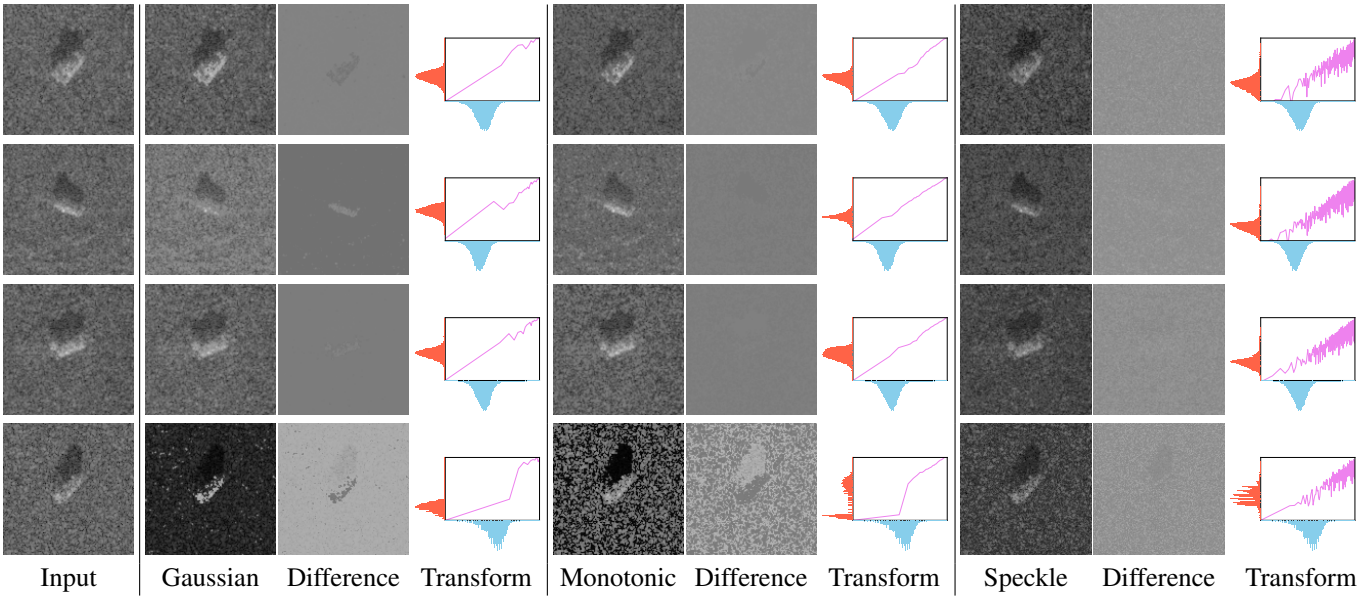

Fig. 4: **Examples of co-domain augmentations.** Each row shows a different MSTAR instance. Columns illustrate the input image (Column 1), additive Gaussian noise (Columns 2-4), monotonic transform (Columns 5-7), and multiplicative Rayleigh speckle (Columns 8-10), with corresponding difference images to their right. The "Transform" columns visualize the log-domain pixel distributions before and after augmentation, along with the mapping function. Co-domain transformations avoid introducing spatial artifacts, and Rayleigh speckle best preserves object structure, as seen in the difference images.

B. Zero-1-to-3 with Projected Feature Mapping Loss

Zero-1-to-3 [2] synthesizes novel object viewpoints from a single RGB image using a conditional latent diffusion model. It takes the input image and a relative camera transformation – defined by rotation and translation – as conditioning inputs and generates the corresponding target view of the object. It leverages large-scale diffusion models pretrained on natural images and fine-tuned on rendered images of synthetic 3D objects to support viewpoint-conditioned generation.

When both the original and transformed object views are clean RGB images, direct image-level comparison in zero-1-to-3 is effective. In contrast, for noisy SAR imagery, such pixel-space comparisons can degrade performance. To address this limitation, we introduce a *feature matching loss* based on unsupervised representation learning for SAR images.

For SAR data, the conditional inputs to the latent diffusion model consist of a SAR image $\mathbf{x}'$, a relative azimuth angle ϕ, and a relative depression (or elevation) angle ψ. The model is trained to generate a novel view $\mathbf{x}_0$ of the input $\mathbf{x}'$ given the relative angles ϕ and ψ (Figure 2).

To compute the feature matching loss, the diffusion model would require executing the full *reverse process* to generate the transformed image from noise. However, backpropagating through the full reverse process is computationally intractable during training. Recent methods based on distillation or ODE formulations of diffusion models reduce the number of reverse steps [13], [14], [15], but they still require running a solver during training and backpropagating through it, which remains expensive in our setting. Moreover, we cannot leverage a large-scale pretrained model such as zero-1-to-3 for this task.

To address these challenges, we simply use Equation (13) to replace $\boldsymbol{\epsilon}$ in Equation (12) with $\boldsymbol{\epsilon}_\theta$ to obtain an approximation $\tilde{\mathbf{x}}_0$ of $\mathbf{x}_0$ given the sample $\mathbf{x}_t$ at time step t:

$$\tilde{\mathbf{x}}_0(\boldsymbol{\epsilon}_\theta, t) = \frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1}\boldsymbol{\epsilon}_\theta \ . \tag{19}$$

This procedure is analogous to DDIM sampling [16]; however, we employ it during training rather than generation. This design enables efficient supervision without incurring the computational cost of full reverse diffusion and backpropagation.

Since the representation network f is fixed, gradients propagate only to the diffusion model parameters θ. We introduce the *Projected Feature Matching* (PFM) loss as a feature-space regularization of the diffusion denoising objective:

$$L_{\text{PFM}}(\theta) = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[\|f(\mathbf{x}_0) - f(\tilde{\mathbf{x}}_0(\boldsymbol{\epsilon}_\theta, t))\|^2\right] \ . \tag{20}$$

The total training objective is then given by

$$L_{\text{total}} = L_{\text{simple}} + \lambda\, L_{\text{PFM}} \ , \tag{21}$$

where $\lambda > 0$ controls the strength of the feature-matching regularization. See an analysis of its impact on benchmark performance in Supplementary Material.

IV. Related Work

Unsupervised representation learning aims to learn informative representations without relying on human annotations, enabling effective transfer to downstream tasks [17], [18], [19], [20], [3], [21]. Existing approaches are broadly categorized into contrastive [17], [20], [3] and non-contrastive [22], [21] methods. Both paradigms learn representations through diverse input augmentations, and our work follows the contrastive learning framework.

Diffusion models [23], [4], [16], [24], [25], [26] have emerged as state-of-the-art approaches for image generation [26], demonstrating strong performance in producing high-fidelity and diverse samples. Their success is commonly attributed to solid theoretical foundations in stochastic processes, scalability to large-scale training, and an ability to capture fine-grained details in generated images, making them well suited for high-quality generative modeling tasks.

Latent diffusion models (LDMs) [27] perform the diffusion process in the latent space of an autoencoder rather than directly in pixel space. This choice substantially reduces computational cost and inference time, since semantic and structural information is preserved after compression, enabling efficient yet effective image generation [28].

For SAR imagery, diffusion models have been applied to tasks such as despeckling [29], [30], [31], [32], [12] and SAR-to-optical translation [33], [34], which typically require access to noise-free reference images. Diffusion-based methods have also been explored for training-set expansion [35], [36], [37]. This work is the first to leverage a large-scale pretrained diffusion model for *novel-view synthesis* of SAR images.

SAR automatic target recognition (ATR) is important in environmental monitoring, disaster response, and remote sensing. Early SAR ATR systems relied on hand-crafted feature representations combined with classical classifiers such as Support Vector Machines (SVMs) [38], [39] and k-Nearest Neighbors (k-NN) [40]. With the rise of deep learning, convolutional neural networks (CNNs) achieved substantial performance gains by learning discriminative features directly from data [41], [42], [43], [44], [45].

However, the scarcity of labeled training data remains a persistent challenge for SAR ATR, motivating research into semi-supervised and unsupervised approaches. Generative Adversarial Networks (GANs), for example, have been used to improve robustness to noise [46].

Recent advances also include transformer-based architectures [45], [47] and self-supervised learning methods [48], [49], [50], which helps capture complex spatial patterns in SAR data without requiring extensive labeled datasets. Despite this progress, speckle noise and varying imaging conditions continue to challenge SAR ATR.

SAR novel view generation has become an important direction for improving target recognition in low-data regimes. Most existing SAR image generation methods rely on Generative Adversarial Networks (GANs) due to their strong synthesis capabilities [51], [52], [53], [54], [55]. GAN-based approaches have demonstrated the ability to synthesize realistic SAR images from novel viewpoints [56], [57].

However, these methods often struggle with speckle noise and image-space training limitations. They are also typically evaluated under restricted settings, e.g., sparse azimuth sampling, and do not address the challenge of generating novel views of previously *unseen* objects. In contrast, our approach leverages a large-scale pretrained diffusion backbone, SAR-specific co-domain augmentations, and feature-space matching, enabling robust and geometrically consistent novel-view synthesis even for unseen targets.

V. Experiments

We describe the datasets, evaluate our co-domain representation learning on classification and regression, assess our representation learning for synthetic-to-real SAR image classification, and evaluate our proposed PFM loss for novel-view SAR synthesis on both seen and unseen targets.

A. Datasets

MSTAR is an X-band SAR dataset containing images of 10 target classes captured across varying azimuth and elevation angles [58]. Following standard protocol, we use 6059 images collected at a 17° elevation for training and 5392 images collected at a 15° elevation for testing. Their azimuth angles span the full $0^\circ - 359^\circ$ range.

MSTAR-OOD is constructed to test generalization to unseen objects. We use a leave-one-class-out setting: 9 MSTAR classes are used for training, and the remaining class (`t72`) is held out for novel-view synthesis. Additionally, we exclude 12 randomly selected 3° azimuth intervals from the training set and reserve the corresponding images as a validation split for representation learning. The resulting training and test sets contain 5003 and 5850 images, respectively. Figure 5 shows representative images for each MSTAR class.

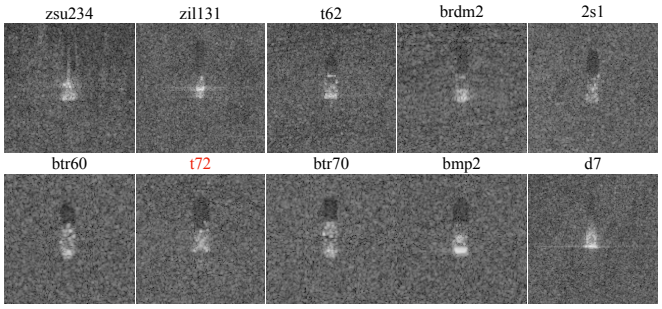

Fig. 5: Sample MSTAR-OOD images. Unseen class is `t72`.

SAMPLE (Synthetic and Measured Paired Labeled Experiment) provides paired synthetic-measured images derived from MSTAR [59]. The public release includes azimuth angles from 10° to 80°. It has 1345 synthetic images paired with their 1345 real counterparts across 10 MSTAR categories. We use the synthetic images for training and the corresponding real images for evaluation in our synthetic-to-real experiments.

B. Augmentations for Unsupervised Representation Learning

Implementation. We perform unsupervised representation learning in PyTorch [60] using a ResNet-18 backbone [61]. Training uses a batch size of 256 for 500 epochs on 128×128 images, with stochastic gradient descent, cosine annealing, an initial learning rate of 0.06, and a weight decay of 5×10^{-4}.

All augmentations are applied in the original magnitude domain, followed by a decibel transformation, $m \mapsto 20 \log_{10}(m + 10^{-3})$. This step is required to ensure compatibility with the pretrained zero-1-to-3 model. The transformed pixel values are then globally normalized to $[0, 1]$ across the dataset. Standard SimCLR augmentations, with and without geometric transformations, are applied in the decibel domain.

Augmentation	CoD	MLP		1-NN	
		Accu%↑	Error°↓	Accu%↑	Error°↓
SimCLR (full)		**98.7**	28.5	99.6	12.8
SimCLR (jitter only)		77.8	52.2	81.4	41.8
Additive Gaussian	✓	97.6	12.7	98.9	5.0
Speckle Rayleigh	✓	98.5	14.6	99.3	3.9
Monotonic	✓	89.7	25.1	93.6	10.8
Additive Gaussian+Monotonic	✓	97.9	21.3	99.4	4.9
Speckle Rayleigh+Monotonic	✓	**98.7**	**9.8**	**99.7**	**3.1**

TABLE I: **Co-domain augmentations yield the best joint performance for both classification and azimuth regression on MSTAR.** We compare different augmentation strategies using MLP and 1-NN probing of learned representations. All models are trained with SimCLR using the specified augmentations. Augmentations marked as "co-domain" (CoD) (✓) operate in the co-domain of pixel intensities. The performance is reported using classification accuracy (Accu%) and azimuth regression error (Error°). The MLP for linear probing on the learned feature is trained without any data augmentation. Remarkably, without geometric domain transformation, the combination of multiplicative Rayleigh speckle and random monotonic transformations proves most effective for both tasks.

We evaluate the learned representations $\mathbf{h} = f(\mathbf{x})$ using lightweight multi-layer perceptron (MLP) probes for both target classification and azimuth regression. A single linear layer is used for classification, while a two-layer network with a sigmoid activation is used for regression. Although the relative importance of these tasks depends on the application, both are critical for novel-view synthesis: Classification captures categorical differences, while regression ensures accurate viewpoint estimation. In addition to MLP probing, we also apply K-nearest neighbor (KNN) probing for both classification and regression in the representation space to compare different augmentation strategies. Although our experiments focus on object-centric SAR data due to dataset availability, our method itself is applicable to scene-level SAR synthesis and can be extended to more complex scenarios.

Results. We compare our augmentations with the original SimCLR augmentations (SimCLR (full)) and a variant without geometric transformations, which includes random brightness and contrast jittering and Gaussian blur; this variant is denoted by SimCLR (jitter only). Our co-domain augmentations – random monotonic remapping, multiplicative Rayleigh speckle, and additive Gaussian noise – are denoted by Monotonic, Speckle Rayleigh, and Additive Gaussian, respectively.

Table I shows the classification and regression results on MSTAR, obtained using different augmentations within the SimCLR framework. In most cases, combining our random monotonic transformation with other augmentations improves performance for both MLP and KNN probing.

When geometric transformations are excluded, additive Gaussian noise and multiplicative Rayleigh speckle consistently outperform the SimCLR (jitter only) baseline. Although the full SimCLR setting achieves the highest classification accuracy, its geometric augmentations (e.g., horizontal flipping) severely degrade azimuth regression. As shown in the t-SNE visualization in Figure 6, object categories are well separated, yet

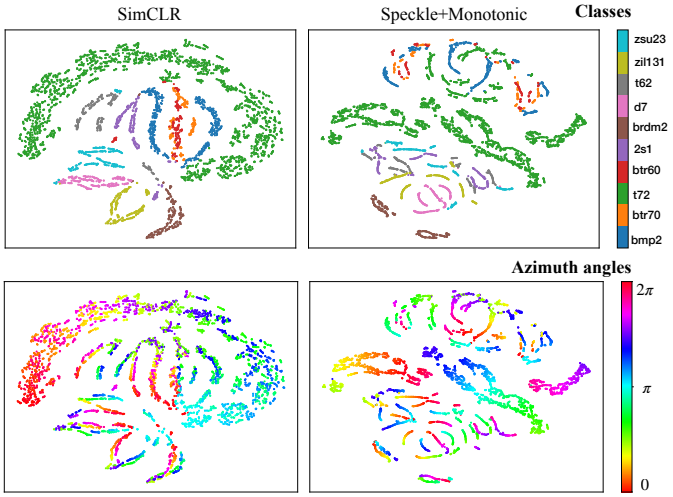

Fig. 6: **Co-domain augmentations preserve both class structure and azimuthal continuity in the learned representation.** t-SNE visualizations compare SimCLR augmentations (left) with our co-domain augmentations (Rayleigh speckle + monotonic, right). The top row is colored by object class and the bottom row by azimuth angle. While SimCLR yields strong class separation, it induces local mixing across azimuths, whereas our co-domain augmentations maintain smooth azimuthal trajectories while retaining clear class clusters.

local neighborhoods exhibit confusion across azimuth angles. In contrast, our co-domain augmentations preserve viewpoint structure while maintaining strong class separation.

Overall, the combination of multiplicative speckle noise using the Rayleigh distribution and random monotonic transformations proves most effective.

C. Synthetic-to-real

We evaluate the impact of different augmentation strategies on synthetic-to-real transfer. Specifically, we train full classification networks (i.e., without representation learning) on synthetic SAR images using different augmentations and test them on the corresponding real images in the SAMPLE dataset.

Table II shows that different augmentation strategies substantially improve synthetic-to-real transfer, with combinations involving Monotonic achieving the best performance. For example, combining additive Gaussian noise with monotonic remapping increases accuracy from 79.6% to 91.8%. The distributional gap between synthetic and real SAR images has been analyzed in [1]. Our co-domain augmentations, particularly Monotonic, effectively reduce this domain gap.

D. Novel View Synthesis

Implementation. We initialize zero-1-to-3 with the pretrained Objaverse model [62] after 105,000 iterations and fine-tune it using our PFM loss. The model employs a latent diffusion backbone with a fixed autoencoder [27]; the decoder is frozen while gradients are applied to the diffusion parameters θ. Only the conditioning and training objective are modified for SAR data; the model architecture and sampling procedure

Augmentation	Co-domain	Accuracy %	Gain %
No Augmentation [1]		66.5	
+Additive Gaussian [1]		85.2	
+Additive Gaussian*	✓	79.6	
+Monotonic	✓	78.0	
+Speckle Rayleigh	✓	80.3	
+Additive Gaussian+Monotonic	✓	**91.8**	12.2
+Speckle Rayleigh+Monotonic	✓	87.5	7.2

TABLE II: **Co-domain augmentations substantially improve synthetic-to-real SAR classification**. The model is trained on SAMPLE synthetic data under different augmentation strategies, and then tested on SAMPLE real images. Augmentations marked as "co-domain" (✓) operate in the co-domain of pixel intensities. Combinations involving our monotonic remapping consistently outperform standard image-space augmentations, with the best result achieved by Gaussian noise plus monotonic remapping. Gains are reported relative to the corresponding non-monotonic baseline. * denotes our implementation.

follow zero-1-to-3. Training is performed for 20,000 iterations with an effective batch size of 80. All models are trained on two NVIDIA A40 GPUs.

Qualitative Results. Figures 7 and 8 present representative novel-view synthesis results on MSTAR and MSTAR-OOD, respectively. With our PFM loss, the generated images preserve object identity (as verified by a trained classifier) and remain consistent with the target class, whereas fine-tuned zero-1-to-3 without PFM often alters object identity. In addition, the last two samples of MSTAR and the last sample of MSTAR-OOD show that incorporating PFM yields generated views whose orientations more closely match the target images. Novel-view synthesis for unseen categories also improves with PFM compared to zero-1-to-3 fine-tuning alone. Despite these gains, a performance gap between seen and unseen classes remains.

Quantitative Results. We evaluate novel-view synthesis by comparing each generated image to a held-out ground-truth image of the same object under the target viewpoint. For classification-based evaluation, we use a ResNet-18 classifier trained from scratch on real SAR images using supervised learning. For azimuth regression, we use a ResNet-18 backbone followed by a two-layer MLP head with 256 hidden units per layer and a sigmoid output, trained to predict the azimuth angle scaled by 2π. Note that these models are trained for the sake of evaluation only and are different from the learned representation for the PFM loss calculation. See an ablation study of the evaluation models in Supplementary Material.

We report classification accuracy and azimuth regression error to assess identity preservation and pose consistency. All results are averaged over 1,000 generated samples, evenly drawn from seen and unseen classes. Each sample is formed by selecting two views of the same object from the test set, one as input and the other as the target, and using their relative pose as conditioning for the diffusion model. The generated view is then compared with the target image to compute the reported metrics. To further assess compliance with SAR image statistics, we additionally report the following metrics.

- **SFD**: A SAR-specific FID-like metric [63], following the protocol in [64], [65], computed as a Fréchet distance in a SAR feature space. We extract features (before the regression MLP head) from a ResNet-18 trained from scratch for azimuth regression on real SAR data to compute SFD.
- **Kuan+SSIM**: To assess structural similarity between real and generated SAR images, we use the SSIM metric [66] that measures the similarity in luminance, contrast, and structure. Because speckle noise degrades the reliability of SSIM on SAR data, we follow prior work [67], [68] to first apply 7×7 Kuan filtering [69] to suppress speckle while preserving structural content, yielding more interpretable SSIM scores

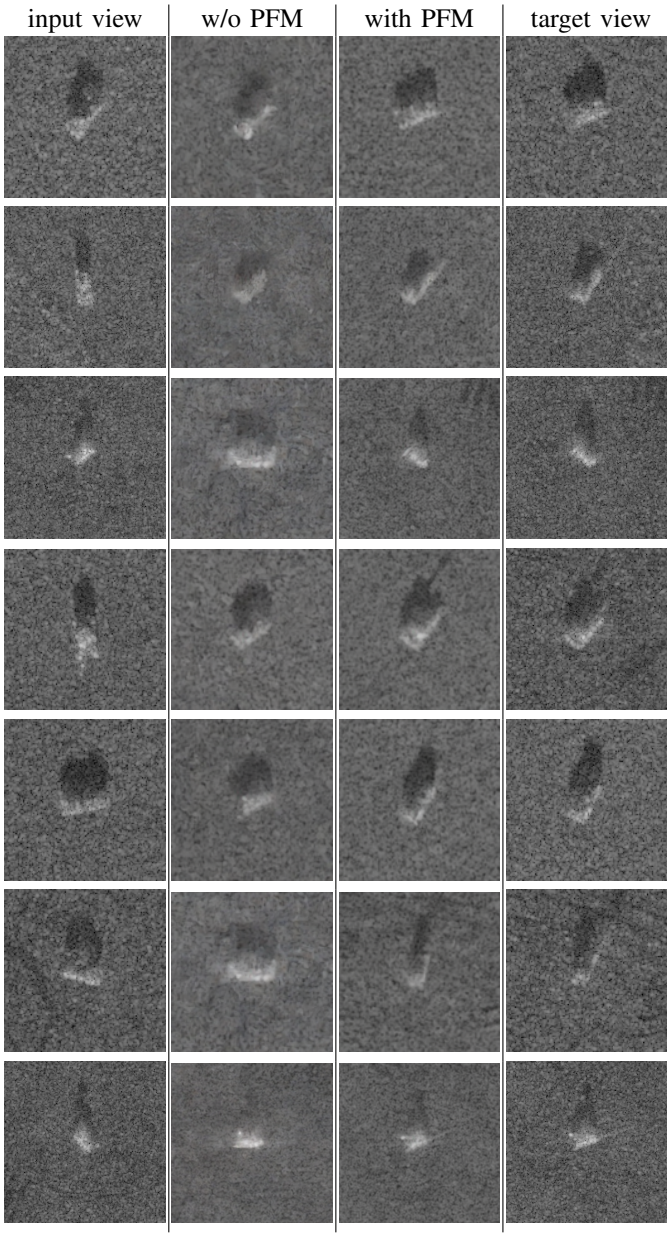

Fig. 7: **Our PFM loss improves novel-view synthesis for seen SAR targets.** Columns 1-4 show the input view from MSTAR, diffusion output without PFM ($\lambda=0$), diffusion output with PFM ($\lambda>0$), and the target view, respectively. Compared to the baseline diffusion model, adding PFM leads to sharper and more stable reconstructions on targets observed during training.

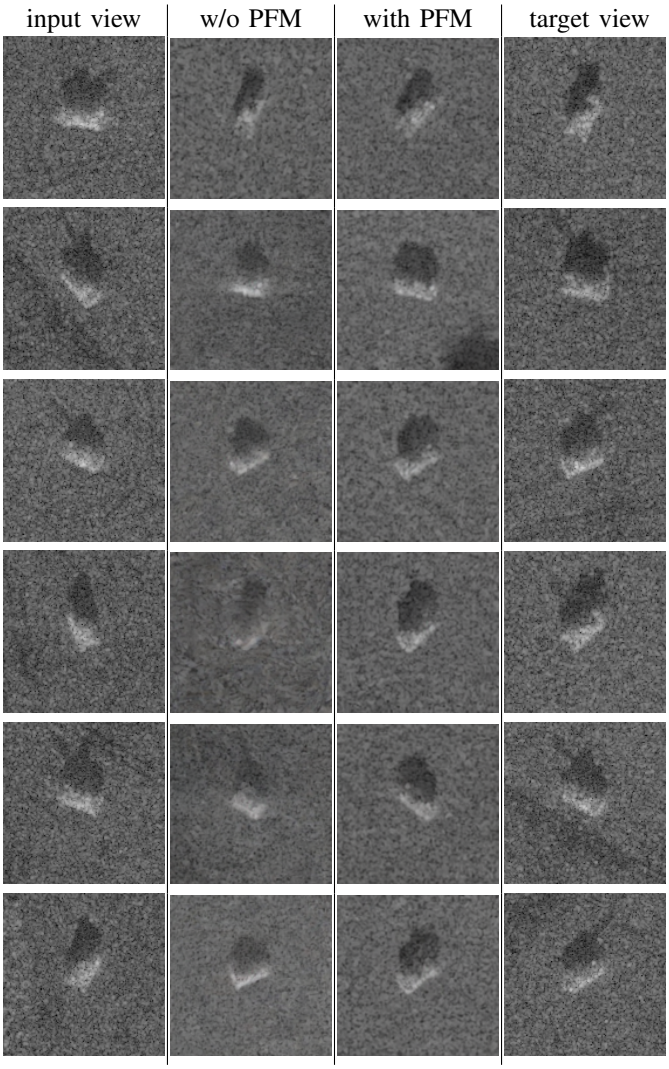

Fig. 8: **PFM improves generalization in novel-view synthesis for unseen SAR targets.** Columns 1-4 show the input view from MSTAR-OOD, diffusion output without PFM ($\lambda = 0$), diffusion output with PFM ($\lambda > 0$), and the target view, respectively. PFM substantially improves geometric consistency and structural sharpness for previously unseen targets.

for SAR data. We evaluate on the 64×64 center crop to focus on the target of interest in MSTAR images.

- **MS-SSIM**: We report Multi-Scale Structural Similarity (MS-SSIM) [70] to evaluate perceptual quality across multiple spatial scales. By capturing both fine details and global structure, MS-SSIM is well suited for SAR imagery, where target identity depends on both local texture and overall shape. As with SSIM, evaluation is performed on the 64×64 center crop. MS-SSIM is computed over three scales using weights $\beta = (0.0448, 0.3001, 0.1333)$, data range 1.0, kernel size 7, and a Gaussian kernel with standard deviation 1.0, yielding a robust measure of image fidelity under varying noise and resolution levels.

Table III compares novel-view synthesis on both seen and unseen SAR targets. For seen classes, input and target views are drawn from held-out MSTAR images at 15° elevation, while unseen-class results are evaluated on MSTAR-OOD objects not used during training. Across all metrics, adding PFM to zero-1-to-3 consistently improves performance, increasing classification accuracy, reducing azimuth error, and improving perceptual similarity, which indicates stronger structural and semantic fidelity in the generated views.

Method	**Accu%↑**	**Error°↓**	**SFD↓**	**MS-SSIM↑**	**Kuan+SSIM↑**
Seen Classes (MSTAR)					
zero-1-to-3	86.6	28.5	3.89	0.6680	0.7330
zero-1-to-3+PFM	**89.0**	**23.5**	**1.84**	**0.6720**	**0.7374**
Unseen Classes (MSTAR-OOD)					
zero-1-to-3	25.8	53.4	8.43	0.6165	0.6731
zero-1-to-3+PFM	**46.6**	**44.0**	**6.28**	**0.6195**	**0.6786**

TABLE III: **Our PFM significantly improves identity and pose fidelity for both seen and unseen targets.** We report classification accuracy (Accu%), azimuth regression error (Error°), and image-quality metrics (SFD, MS-SSIM, Kuan-SSIM) for novel-view synthesis on MSTAR (seen classes) and MSTAR-OOD (unseen classes). Some metrics (e.g., MS-SSIM, Kuan+SSIM) are computed on 64×64 center crops; SFD refers to FID with SAR Encodings; see Section V-D for metric-specific details. Adding PFM consistently improves recognition accuracy, reduces pose error, and yields images that better match the SAR statistics of the target views.

Although perceptual metrics such as Kuan+SSIM and MS-SSIM are useful statistical proxies for visual fidelity, they are limited in SAR imagery because speckle introduces stochastic pixel-level variation that is not tightly coupled to semantic content. Regression error and classification accuracy are more informative for semantics. Azimuth regression directly measures viewpoint consistency, while classification reflects preservation of target identity, both of which are critical for the downstream use of novel-view synthesis in SAR ATR.

As expected, performance on seen classes exceeds that on unseen classes across all metrics, since the model has access to similar object instances and viewpoints during training. Nevertheless, incorporating PFM consistently improves results even for unseen classes, where generalization is more challenging. Note that unseen classes exhibit higher azimuth error, often due to front–back ambiguity: Since the model has never observed these objects during training, it may confuse opposing orientations, leading to 180° errors. This also highlights the benefit of semantic-level supervision in preserving both identity and viewpoint. The remaining gap between seen and unseen classes underscores the need for future work on more robust and class-agnostic novel-view synthesis.

VI. Discussion

We introduced co-domain augmentation strategies based on random monotonic remapping and multiplicative Rayleigh speckle for self-supervised representation learning in SAR imagery. By operating directly on magnitude values, these augmentations better preserve orientation and structural cues than standard SimCLR transforms which also involves geometric transformations, making them well suited for novel-view SAR synthesis. When combined with other augmentations, random

monotonic remapping improves both classification and azimuth regression, particularly in synthetic-to-real settings.

We further leverage the learned representations to enhance diffusion-based SAR novel-view synthesis. Using a single-step approximation avoids backpropagation through the full reverse process, while feature matching in representation space provides a noise-robust alternative to pixel-space losses. Both quantitative and qualitative results demonstrate consistent gains from our feature matching loss on seen and unseen classes.

Although our approach is purely data-driven, it complements physics-based SAR simulators. Physics-based models offer interpretability and sensor control but require detailed scene and sensor modeling and are computationally expensive. Moreover, Table II and prior work [1] show that purely synthetic data can suffer from domain shift, limiting downstream performance. In contrast, learning directly from real SAR imagery enables our method to produce viewpoint-consistent and identity-preserving outputs that generalize across targets. Future work could integrate physics-based modeling with our learned augmentations to further improve realism and robustness.

In summary, carefully designed co-domain augmentations and representation-level supervision substantially improve the quality and generalization of SAR novel-view synthesis. This work encourages deeper integration of data-driven and physics-based methods across SAR applications.

Acknowledgement

This work was supported in part by the National Science Foundation under awards NSF 2215542 and NSF 2313151, with additional compute support provided by the NAIRR Pilot under CIS240421.

References

[1] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2942–2955, 2021.

[2] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[5] A. W. Doerry, "Sar image scaling dynamic range radiometric calibration and display," tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); General âĂę, 2019.

[6] Z. Lu and D. Meyer, "Study of high sar backscattering caused by an increase of soil moisture over a sparsely vegetated area: implications for characteristics of backscattering," *International Journal of Remote Sensing*, vol. 23, no. 6, pp. 1063–1074, 2002.

[7] X. Xiong, Q. Xu, G. Jin, H. Zhang, and X. Gao, "Rank-based local self-similarity descriptor for optical-to-sar image matching," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 10, pp. 1742–1746, 2019.

[8] E. E. Kuruoglu and J. Zerubia, "Modeling sar images with a generalization of the rayleigh distribution," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 527–533, 2004.

[9] J. W. Goodman, "Some fundamental properties of speckle," *JOSA*, vol. 66, no. 11, pp. 1145–1150, 1976.

[10] C. Oliver and S. Quegan, *Understanding synthetic aperture radar images*. SciTech Publishing, 2004.

[11] A. C. Frery, H.-J. Muller, C. d. C. F. Yanasse, and S. J. S. Sant'Anna, "A model for extremely heterogeneous clutter," *IEEE transactions on geoscience and remote sensing*, vol. 35, no. 3, pp. 648–659, 1997.

[12] S. Guha and S. T. Acton, "Sddpm: Speckle denoising diffusion probabilistic models," *arXiv preprint arXiv:2311.10868*, 2023.

[13] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[14] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

[15] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," *arXiv preprint arXiv:2303.01469*, 2023.

[16] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[17] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[19] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794, Springer, 2020.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[21] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[22] X. Liu, Z. Wang, Y.-L. Li, and S. Wang, "Self-supervised learning via maximum entropy coding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34091–34105, 2022.

[23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, pp. 2256–2265, PMLR, 2015.

[24] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*, pp. 8162–8171, PMLR, 2021.

[25] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[28] L. Weng, "What are diffusion models?," *lilianweng.github.io*, Jul 2021.

[29] M. V. Perera, N. G. Nair, W. G. C. Bandara, and V. M. Patel, "Sar despeckling using a denoising diffusion probabilistic model," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[30] X. Hu, Z. Xu, Z. Chen, Z. Feng, M. Zhu, and L. Stankovic, "Sar despeckling via regional denoising diffusion probabilistic model," *arXiv preprint arXiv:2401.03122*, 2024.

[31] Y. Ma, P. Ke, H. Aghababaei, L. Chang, and J. Wei, "Despeckling sar images with log-yeo-johnson transformation and conditional diffusion models," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[32] S. Xiao, L. Huang, and S. Zhang, "Unsupervised sar despeckling based on diffusion model," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 810–813, IEEE, 2023.

[33] X. Bai, X. Pu, and F. Xu, "Conditional diffusion for sar to optical image translation," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[34] M. Seo, Y. Oh, D. Kim, D. Kang, and Y. Choi, "Improved flood insights: Diffusion-based sar to eo image translation," *arXiv preprint arXiv:2307.07123*, 2023.

[35] Y. Qi, L. Wang, K. Li, H. Liu, and C. Zhao, "Latent diffusion model-based t2t-vit for sar ship classification," in *CCF Conference on Computer Supported Cooperative Work and Social Computing*, pp. 296–305, Springer, 2023.

[36] Y. Xu, C. Lin, Y. Zhong, Y. Huang, and X. Ding, "Recognizer embedding diffusion generation for few-shot sar recognization," in

Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 418–429, Springer, 2023.

[37] A. Tuel, T. Kerdreux, C. Hulbert, and B. Rouet-Leduc, "Diffusion models for interferometric satellite aperture radar," *arXiv preprint arXiv:2308.16847*, 2023.

[38] Q. Zhao and J. Principe, "Support vector machines for sar automatic target recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 2, pp. 643–654, 2001.

[39] H. Liu and S. Li, "Decision fusion of sparse representation and support vector machine for sar image target recognition," *Neurocomputing*, vol. 113, pp. 97–104, 2013.

[40] A. K. Mishra, "Validation of pca and lda for sar atr," in *TENCON 2008 - 2008 IEEE Region 10 Conference*, pp. 1–6, 2008.

[41] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4806–4817, 2016.

[42] Z. Huang, Z. Pan, and B. Lei, "Transfer learning with deep convolutional neural network for sar target classification with limited labeled data," *Remote Sensing*, vol. 9, no. 9, 2017.

[43] 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 Geoscience and Remote Sensing Letters*, vol. 14, no. 9, pp. 1484–1488, 2017.

[44] Y. Yan, "Convolutional neural networks based on augmented training samples for synthetic aperture radar target recognition," *Journal of Electronic Imaging*, vol. 27, no. 2, pp. 023024–023024, 2018.

[45] J. Fein-Ashley, T. Ye, R. Kannan, V. Prasanna, and C. Busart, "Benchmarking deep learning classifiers for sar automatic target recognition," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–6, 2023.

[46] Y. Guo, L. Du, D. Wei, and C. Li, "Robust sar automatic target recognition via adversarial learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 716–729, 2020.

[47] W. L, W. Yang, Y. Hou, L. Liu, Y. Liu, and X. Li, "Saratr-x: A foundation model for synthetic aperture radar images target recognition," 2024.

[48] N. Inkawhich, "A global model approach to robust few-shot sar automatic target recognition," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[49] H. Pei, M. Su, G. Xu, M. Xing, and W. Hong, "Self-supervised feature representation for sar image target classification using contrastive learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[50] Z. Wen, Z. Liu, S. Zhang, and Q. Pan, "Rotation awareness based self-supervised learning for sar target recognition with limited training samples," *IEEE Transactions on Image Processing*, vol. 30, pp. 7266–7279, 2021.

[51] Y. Kuang, F. Ma, F. Li, Y. Liu, and F. Zhang, "Semantic-layout-guided image synthesis for high-quality synthetic-aperature radar detection sample generation," *Remote Sensing*, vol. 15, no. 24, 2023.

[52] J. Bao, W. Yu, K. Yang, C. Liu, and T. J. Cui, "Improved few-shot sar image generation by enhancing diversity," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[53] L. Li, C. Wang, H. Zhang, and B. Zhang, "Sar image ship object generation and classification with improved residual conditional generative adversarial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[54] 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 Sensing*, vol. 13, no. 19, 2021.

[55] J. Wang, J. Li, B. Sun, and Z. Zuo, "Sar image synthesis based on conditional generative adversarial networks," *The Journal of Engineering*, vol. 2019, no. 21, pp. 8093–8097, 2019.

[56] C. Wang, J. Pei, X. Liu, Y. Huang, D. Mao, Y. Zhang, and J. Yang, "Sar target image generation method using azimuth-controllable generative adversarial network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9381–9397, 2022.

[57] Z. F. Qijun Dai, Gong Zhang and B. Xue, "Cvgan: A cross-view sar image generation method for enhancing the view diversity," *Remote Sensing Letters*, vol. 14, no. 6, pp. 631–640, 2023.

[58] E. R. Keydel, S. W. Lee, and J. T. Moore, "Mstar extended operating conditions: A tutorial," *Algorithms for Synthetic Aperture Radar Imagery III*, vol. 2757, pp. 228–242, 1996.

[59] 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 *Algorithms for Synthetic Aperture Radar Imagery XXVI*, vol. 10987, pp. 39–54, SPIE, 2019.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[62] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," *arXiv preprint arXiv:2212.08051*, 2022.

[63] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[64] Y. Kuang, F. Ma, F. Li, Y. Liu, and F. Zhang, "Semantic-layout-guided image synthesis for high-quality synthetic-aperature radar detection sample generation," *Remote Sensing*, vol. 15, no. 24, 2023.

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

[66] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[67] M. Kang and J. Baek, "Sar image change detection via multiple-window processing with structural similarity," *Sensors*, vol. 21, no. 19, p. 6645, 2021.

[68] S. Varshini R, R. Mahadevan, B. Lakshmi S, M. Periasamy, R. C. Raman, *et al.*, "Speckle noise analysis for synthetic aperture radar (sar) space data," *arXiv preprint arXiv:2408.08774*, 2024.

[69] D. T. Kuan, A. A. Sawchuk, T. C. Strand, and P. Chavel, "Adaptive noise smoothing filter for images with signal-dependent noise," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 165–177, 1985.

[70] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003.

SUPPLEMENTARY MATERIAL

A. *Augmentation Visualization*

Higher resolution images are presented for augmentations using additive Gaussian in Figure 9, random monotonic in Figure 10, and multiplicative speckle with Rayleigh distribution in Figure 11.

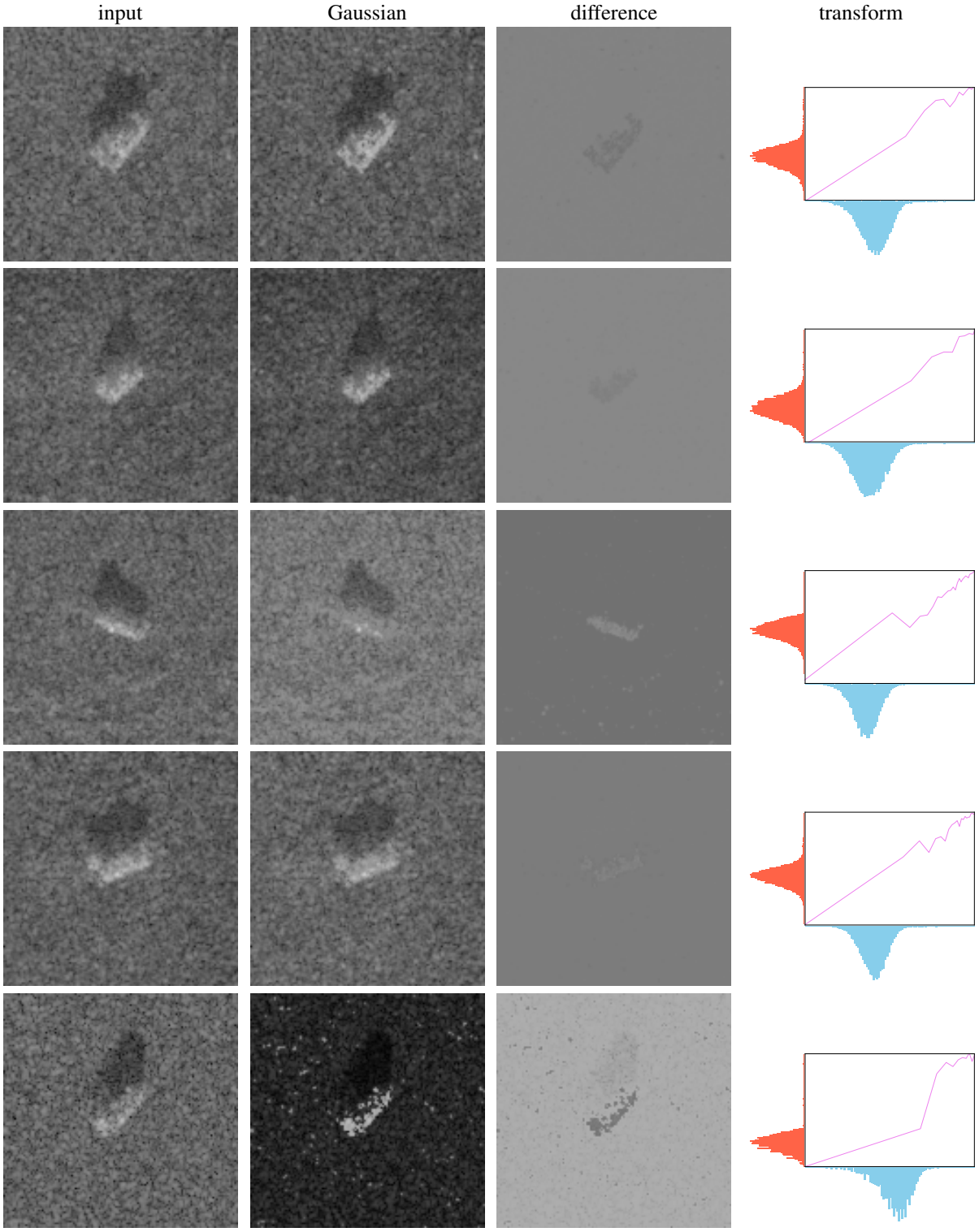

Fig. 9: Additive Gaussian augmentations.

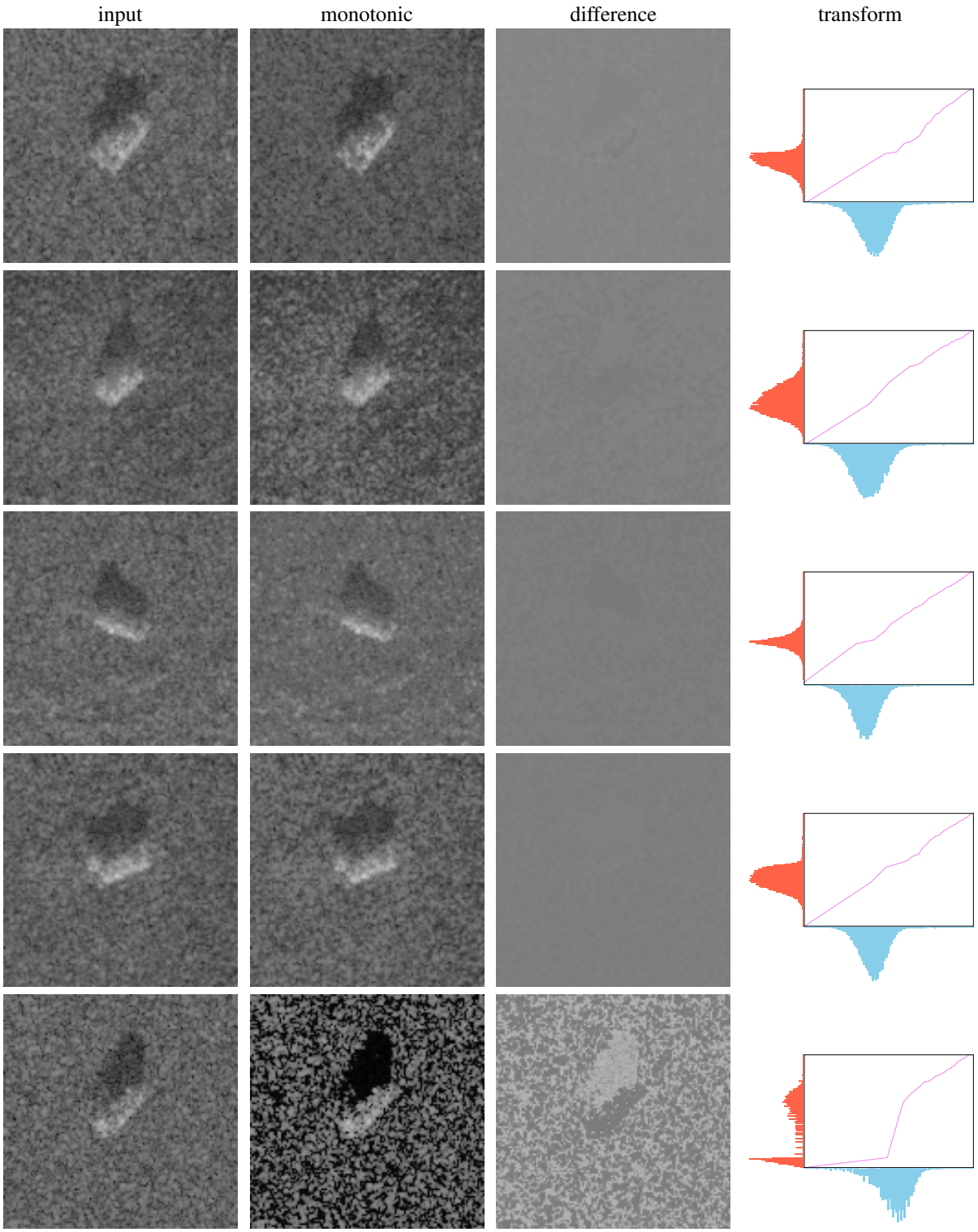

Fig. 10: Random monotonic augmentations.

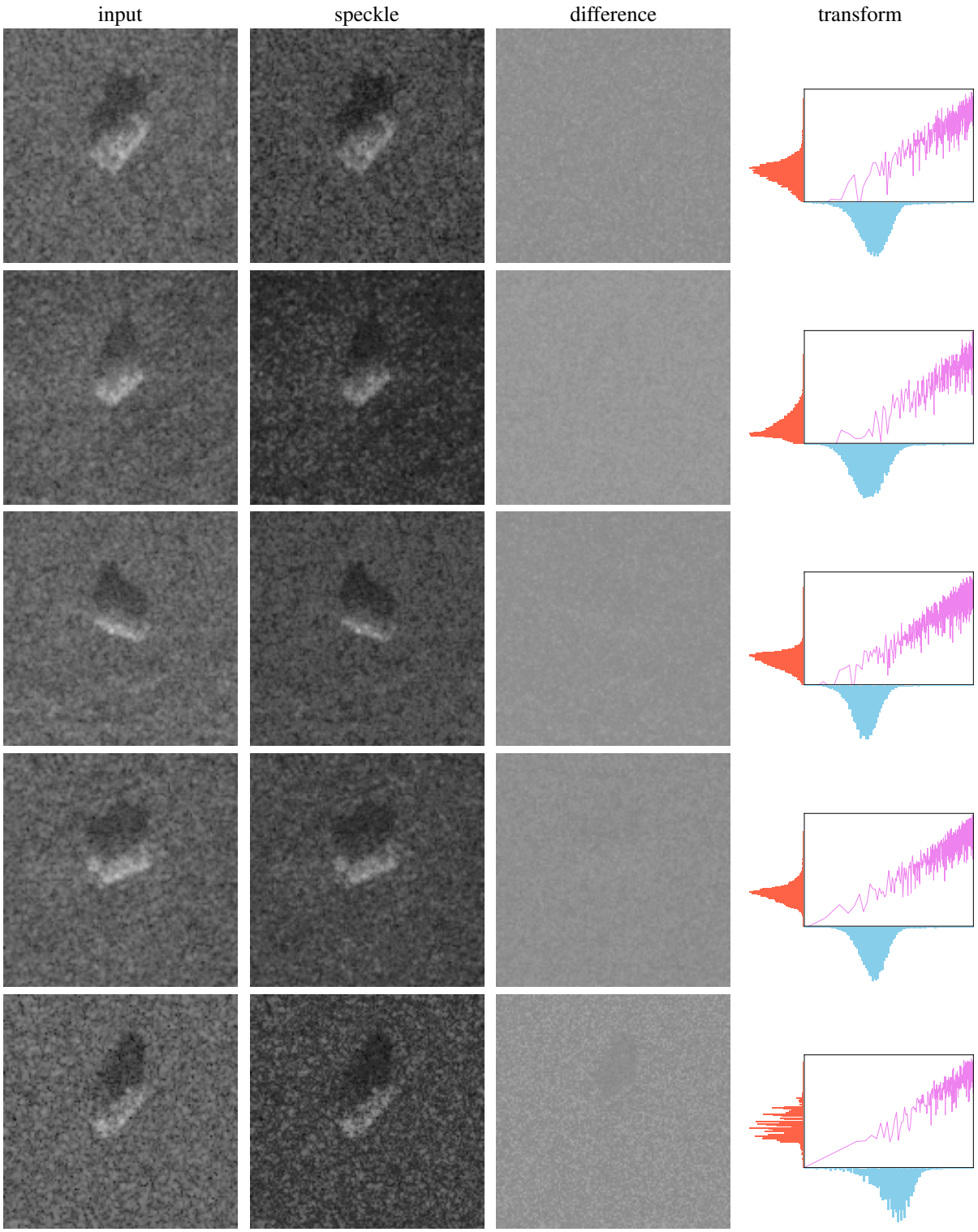

Fig. 11: Multiplicative Speckle noise augmentations.

B. Effect of Evaluation Model Choice

To assess the robustness of our evaluation protocol, we compare multiple ResNet-18 azimuth regression models applied to generated SAR images. All models share the same architecture but differ in how their representations are learned. Our primary evaluation model is trained from scratch on real SAR data (denoted *Scratch*). We additionally evaluate three SimCLR-trained models using different augmentation strategies, each followed by an MLP regression head.

We consider the following evaluation models:

- **SimCLR**: Trained using standard SimCLR augmentations.
- **Speckle Rayleigh**: SimCLR with Rayleigh speckle augmentation.
- **Speckle Rayleigh+Monotonic**: SimCLR with both Rayleigh speckle and monotonic augmentations.
- **Scratch**: A fully supervised model trained from scratch without data augmentation.

We report results for two settings: **1)** without PFM loss ($\lambda = 0$) and **2)** with PFM loss ($\lambda = 4$). Table IV shows that all models trained with SAR-specific augmentations (Speckle Rayleigh, Speckle Rayleigh+Monotonic, and Scratch) achieve comparable azimuth regression accuracy, whereas the model trained with standard SimCLR augmentations performs substantially worse. This consistency across augmented models indicates that our evaluation protocol on generated images is robust and not overly sensitive to the particular regression model, provided SAR-appropriate augmentations are used. In all cases, incorporating the PFM loss further improves performance.

Azimuth Regression Error °	**w/o PFM** ($\lambda = 0$)	**with PFM** ($\lambda = 4$)
SimCLR	44.00	40.22
Speckle Rayleigh	28.41	24.18
Speckle Rayleigh+Monotonic	29.36	24.04
Scratch	28.46	23.52

TABLE IV: **PFM improves viewpoint accuracy across all evaluation backbones.** Azimuth regression error (°) measured on generated SAR images using different ResNet-18 evaluation models. Incorporating PFM ($\lambda = 4$) consistently reduces angular error relative to standard Zero-1-to-3 ($\lambda = 0$), regardless of the representation used for evaluation.

C. Hyperparameters

Figure 12 shows the effect of the weighting parameter λ in Equation (21) across multiple evaluation metrics. Values in the range $[0.5, 5]$ consistently yield strong performance, with $\lambda = 4$ providing the best overall trade-off. When $\lambda = 0$, the model reduces to standard zero-1-to-3 fine-tuning without representation-level supervision. Introducing the PFM term (e.g., $\lambda = 1$) already improves both visual quality and downstream SAR classification, confirming the effectiveness of feature-space regularization.

All remaining hyperparameters, including $\lambda_1, \lambda_2, \sigma$, and Q, are selected by cross-validation on a held-out validation split (90% / 10% of the training set). Selection is based on linear-probe classification accuracy of the learned representation to ensure generalization. For the Speckle Rayleigh+Monotonic setting, the chosen values are $\sigma = 10$, $\lambda_1 = 1$, $\lambda_2 = 1$, and $Q = 512$.

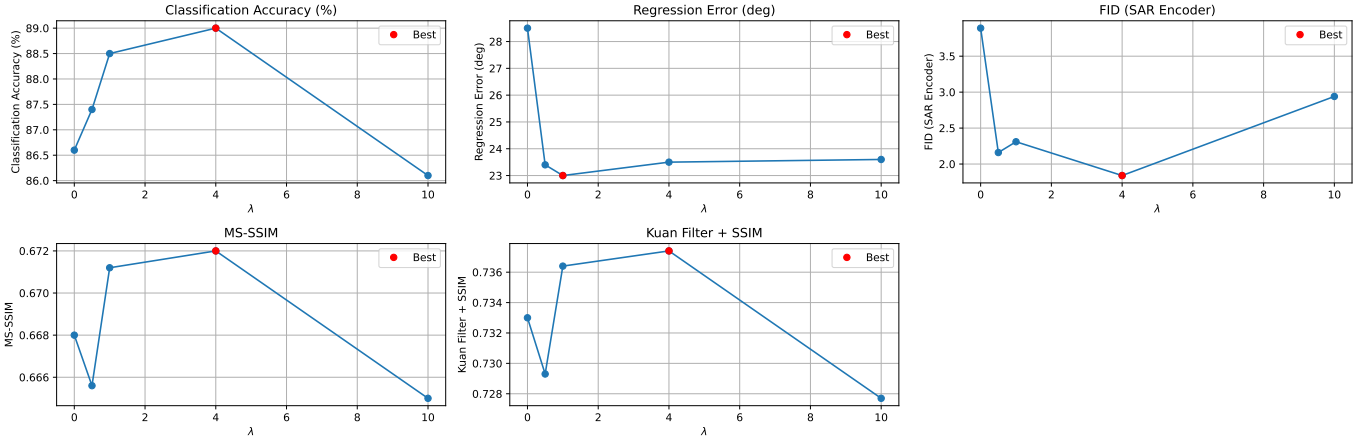

Fig. 12: **Effect of the PFM weight λ on novel-view synthesis performance.** Plots show how varying λ in the total loss (Eq. 21) affects classification accuracy (%), azimuth regression error (°), SFD (FID with SAR-encoder), MS-SSIM, and SSIM with Kuan filtering. The red marker indicates the best value for each metric. Moderate values of λ provide the best trade-off between diffusion fidelity and feature-level consistency, improving both geometric accuracy and perceptual quality.