ProSR: Semantic-Prototype-Guided Discrete Modeling for Physically Consistent SAR Super-Resolution

Byoungwoo Kim and Munchurl Kim

Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{quddn826, mkimee}@kaist.ac.kr

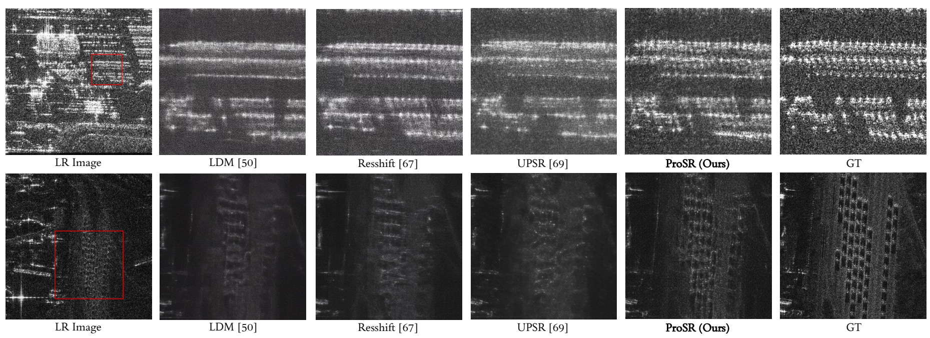
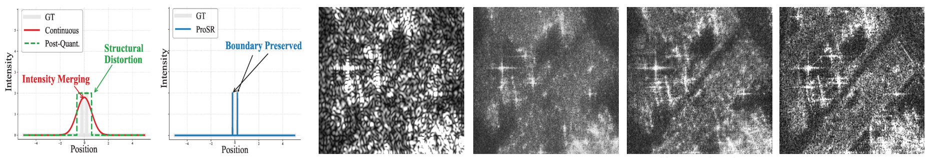


Fig. 1: Qualitative comparison on SAR ISR ($\times 4$) results on small objects. Our ProSR restores a structural integrity of tiny targets (red boxes), whereas existing diffusion methods suffer from distortion or blurring. (Zoom in for better view.)

Abstract. High-resolution Synthetic Aperture Radar (SAR) imagery is critical for precision analysis such as automatic target recognition, yet its acquisition is costly. Although generative image super-resolution (ISR) models offer a promising alternative, current smooth-approximation-based diffusion frameworks often struggle to preserve the coherent scattering statistics, causing stochastic structural distortions that are less consistent with real SAR physics. To address this, we propose Semantic-Prototype-Guided Super-Resolution (ProSR), reformulating SAR ISR as *a semantically-guided discrete token prediction task* within a quantized latent space. By mapping signal features to discrete scattering primitives, ProSR preserves the impulsive nature of SAR without over-smoothing. Furthermore, we integrate a Self-Supervised Learning backbone into SAR ISR to extract label-free semantic priors, overcoming label scarcity. Guided by these priors, we introduce Semantic-Aligned Detail Encoding to decouple high-frequency signals into discrete scattering primitives. In parallel, the Semantic Prototype Map Generator explicitly constructs semantic prototype maps, allowing Prototype-Map-Guided Attention to route the information flows within identical categories and mitigate inter-class interference. To validate our approach, we present a large-scale 0.25 m resolution benchmark from the Umbra Open Dataset. Experimental results show ProSR achieves superior visual quality while preserving essential scattering characteristics required for practical SAR applications.

Keywords: High-Resolution SAR Image Super-Resolution · Discrete Generative Modeling · Self-Supervised Semantic Guidance



(a) Theory: Vanilla Diffusion (b) Theory: ProSR (Ours) (c) LR Input (d) Vanilla Diffusion (Merged) **(e) ProSR (Ours)** (f) Ground Truth

Fig. 2: Smooth-approximation-based vs. semantic-category-based distribution learning. (*a, d*) vanilla diffusion models can exhibit stochastic structure distortion (e.g., intensity merging) by averaging neighboring signals. (*b, e*) ProSR (Ours) preserves distinct scattering peaks via predicting semantic-category-based representation.

1 Introduction

High-resolution (HR) Synthetic Aperture Radar (SAR) imagery is vital for high-precision analysis—such as automatic target recognition (ATR) [72] and infrastructure monitoring [33]—offering rich scattering information derived from unique interactions between electromagnetic wave and terrain geometry [16]. However, acquiring such high-fidelity data is costly, due to sensor costs and orbital constraints. Consequently, image super-resolution (ISR) has emerged as a powerful framework to bridge the gap between accessible low-resolution (LR) data and the high-fidelity requirements of practical applications [3, 4, 61].

Generative models, led by smooth-approximation-based diffusion processes, have demonstrated exceptional capabilities in synthesizing sharp, high-resolution natural images [26, 50, 53, 54, 67, 69]. However, when applied to the SAR domain, these models often struggle to preserve SAR physical integrity, suffering from what we term *'stochastic structural distortions'*. This degradation can be understood from mode interpolation [1]—a phenomenon where smooth transitions between disjoint data modes generate samples entirely outside the true data distribution. Fig. 2 compares a smooth-approximation-based versus a semantic-category-based distribution learning and their effects on generated SAR ISR results. The vanilla diffusion models [50, 67, 69] frequently struggle to separate edges that are very close together as shown in Fig. 2-(a). The two nearby edges (gray color) are difficult to be modeled by the vanilla diffusion models that can yield an overlapped and blurred edge (red color), creating one single fake intensity between the true scattering peaks after post quantization (green color). This structural degradation can be related to the learned score function being a smooth approximation of the data manifold [52, 55], which often makes it difficult to maintain the sharp discontinuities essential to SAR physics (Fig. 2-(d)). A detailed analysis of a toy example regarding unresolved conditional scatterer modes is provided in the *suppl. C* to offer intuition for this behavior. To address this problem, we propose Semantic-Prototype-Guided Super-Resolution (ProSR). ProSR reformulates SAR ISR as *a semantically-guided discrete token prediction task*, implemented via Prototype-Map-Guided Masked Generative Modeling (PMG). By leveraging a discrete latent space to map features into a finite set of physically valid codebook entries [21], PMG treats reconstruction as an iterative selection of categorical tokens rather than smooth density approximation. Unlike vanilla diffusion models, as illustrated in Fig. 2-(a), that tend to average out complex signal distributions, two nearby distinct edges can be faithfully generated, as shown in Fig. 2-(b), by quantized representation of our ProSR that enforces a strict reconstruction of SAR-specific scattering patterns [16]. By doing so, the generation process is inherently prevented from generating ambiguous and smeared scattering signals, which is shown in Fig. 2-(e) unlike the result shown in Fig. 2-(d).

Furthermore, while recent trends in SAR ISR [3,4,61] aim to improve structural fidelity by incorporating task-oriented objectives (e.g., detection losses), these approaches are limited by the scarcity of labeled SAR data. To overcome this label dependency, we introduce a self-supervised representation framework to the SAR ISR task, leveraging foundational models [5, 44] for label-free semantic guidance. This approach allows us (i) to establish Semantic-Aligned Detail Encoding (SADE), which maps discrete codes to specific scattering classes via a semantic-aligned detail codebook, and (ii) to define a semantic-aware perceptual loss. While these components provide a robust semantic foundation, they cannot inherently prevent contextual leakage or semantic drift during the cross-attention process. To address this, we introduce Semantic Prototype Map Generator (SPMG) and Prototype-Map-Guided Attention (PMGA). Specifically, within PMG, the SPMG constructs spatial-semantic maps that guide PMGA to restrict attention flows to semantically consistent regions. By acting as a spatial-semantic guidance, PMGA eliminates inter-class signal leakage, ensuring that the reconstruction is both physically accurate and aligned with the actual structure. Finally, we provide a 0.25 m slant range resolution SAR ISR benchmark from the Umbra Open Dataset [57] as a public resource to mitigate data scarcity. In summary, the primary contributions of our work are as follows:

- To the best of our knowledge, ProSR is the first discrete generative framework tailored for SAR ISR. It utilizes PMG to suppresses *stochastic structural distortions* by mapping signal features into a finite set of physically valid codebook entries, ensuring high-frequency fidelity.
- We establish a label-free semantic guidance strategy by integrating a self-supervised representation model. This enables SADE, SPMG, and PMGA to achieve precise structural anchoring and suppress inter-class interference without requiring scarce manual annotations.
- We establish a large-scale 0.25 m resolution SAR ISR benchmark from the Umbra Open Dataset [57] enabling robust evaluation and future research.
- Our ProSR establishes a new state-of-the-art in SAR ISR, demonstrating superior visual realism while better preserving the essential scattering characteristics required for practical SAR applications.

2 Related Works

2.1 Single Image Super-Resolution (SISR)

SISR has transitioned from traditional optimization to deep generative models [56]. While early pixel-wise architectures [19, 32, 38, 71] often produced over-smoothed results, GAN-based models [34, 37, 42, 60] significantly improved sharpness but frequently introduced unrealistic hallucinations due to adversarial instability. Consequently, recent research has shifted toward diffusion models [9, 35, 39, 51, 59, 64, 65, 67, 69] for higher perceptual fidelity and training stability; however, these diffusion frameworks *inherently rely on a smooth approximation of the data distribution* [1, 52, 55]. This induces *mode interpolation* [1], where the model generates samples that smoothly bridge disjoint data modes, resulting in structural hallucinations. Although such artifacts may appear plausible in natural images, they degrade the physical integrity required for high-precision SAR image analysis.

2.2 Discrete Latent Representations

Discrete representation learning, popularized by VQ-VAE [58] and VQ-GAN [21], maps high-dimensional data into a finite set of quantized codebook entries. Unlike continuous models, these formulations replace latent values with a categorical selection from a fixed vocabulary. By reformulating reconstruction as a discrete token prediction paradigm [2, 6, 7, 10, 47], researchers utilize quantization as a formative bottleneck. While lossy for natural images, this bottleneck uniquely favors SAR, precluding blurred intermediate values and forcing the model to select from distinct entries. This mechanism suppresses the averaging effects common in continuous spaces. Leveraging these advantages, we propose ProSR, which incorporates a scattering primitive guidance within the discrete bottleneck. This enables our ProSR to achieve realistic reconstruction while preserving physical scattering distributions, overcoming the blurring artifacts of continuous spaces.

2.3 SAR Image Super-Resolution (SAR ISR)

SAR ISR is uniquely challenging due to the presence of speckle noise and complex electromagnetic interactions with geometric structures of the ground. Recent works have moved from mathematical priors [29, 30] to generative paradigms [12, 13, 28, 68] to synthesize high-frequency signatures. These realistic reconstructions improve downstream performance (e.g., ATR), proving that realistic scattering patterns are vital for operational analysis. To further enhance fidelity, task-driven SAR ISR incorporates auxiliary structural priors such as edge maps or Electro-Optical imagery [66, 73], or downstream losses such as a detection loss [3, 4, 61]. Despite their success, these methods are heavily bottlenecked by the scarcity of labeled SAR datasets. This label dependency significantly limits the generalizability and scalability of task-driven models. Unlike these label-dependent methods, our ProSR overcomes the scarcity of labeled data by leveraging a self-supervised representation model as a semantic anchor.

2.4 Self-supervised Representation Learning in SAR

To overcome label scarcity, Self-Supervised Learning (SSL) has become a prevailing approach for learning robust representations from unlabeled data [11, 23, 49].

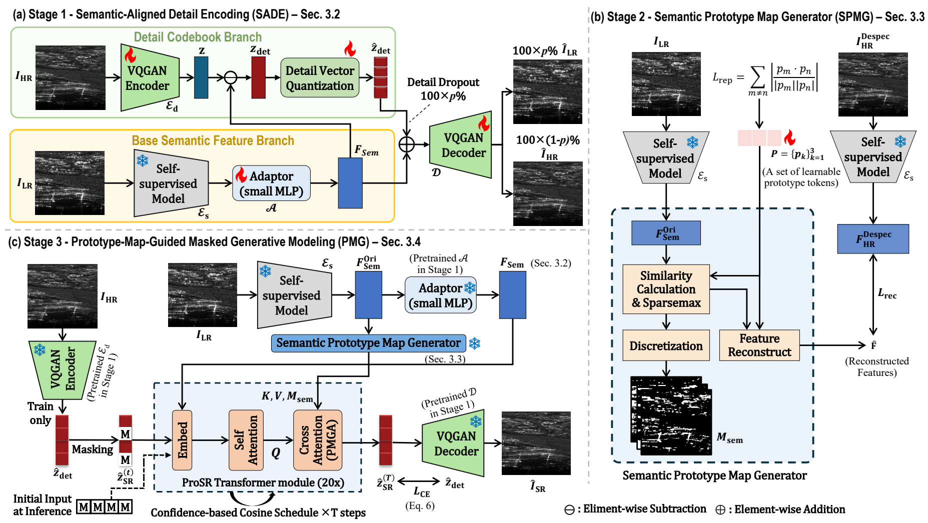


Fig. 3: Overall architecture of the proposed ProSR framework.

Although frameworks like MAE [22] and DINO [5] have shown that rich semantic features can be extracted from vast unlabeled imagery, their application in SAR has remained largely focused on high-level tasks such as classification or semantic segmentation [36, 44]. The potential of SSL-derived semantic priors to serve as a guidance mechanism for low-level reconstruction remains largely untapped. Our ProSR bridges this gap by integrating SSL-derived semantic priors into the reconstruction pipeline. Specifically, we construct a detail codebook aligned with the semantic features of LR inputs and employ a semantic prototype map to guide cross-attention. This enables input-consistent detail recovery while ensuring physical consistency without the need for manual labels.

3 Methodology

3.1 Overview of ProSR

Fig. 3 illustrates the overall architecture of our ProSR framework, which bridges continuous semantic priors with discrete high-frequency details using a spatial-semantic map to address SAR stochastic structural distortions. The framework consists of three stages: (i) Stage 1 - Semantic-Aligned Detail Encoding (SADE), (ii) Stage 2 - Semantic Prototype Map Generation (SPMG), and (iii) Stage 3 - Prototype-Guided Masked Generative Modeling (PMG). It should be noted that Stage 1 - SADE is first *pretrained* with HR SAR ground-truth (GT) images and their LR versions to learn separated representations of details (high-frequency) via the VQGAN Encoder (denoted as $\mathcal{E}_\text{d}$) [21] and the Detail Vector Quantization module, and structural information (low-frequency) via a self-supervised model [44](denoted as $\mathcal{E}_\text{s}$) and an Adaptor (denoted as $\mathcal{A}$) [8], where both representations are jointly optimized in conjunction with a shared VQGAN Decoder (denoted as $\mathcal{D}$) [21]. Then, the pretrained $\mathcal{E}_s$ is incorporated in Stage 2 for training and inference. In Stage 3, the pretrained $\mathcal{E}_{\mathrm{d}}$, $\mathcal{E}_{\mathrm{s}}$ and $\mathcal{A}$ are used for training while the pretrained $\mathcal{E}_{\mathrm{s}}$, $\mathcal{A}$ and $\mathcal{D}$ are for inference.

3.2 Stage 1 - Semantic-Aligned Detail Encoding (SADE)

SADE isolates the high-frequency SAR details of HR (GT) $I_{\mathrm{HR}} \in \mathbb{R}^{H\times W}$ into codebook representations where H and W are the height and width for I_{HR}, respectively, while maintaining strict alignment with LR input $I_{\mathrm{LR}} \in \mathbb{R}^{H/4\times W/4}$ which is upscaled to the same size of I_{HR}, as shown in Fig. 3-(a). To isolate such high-frequency SAR image details, we employ a dual-encoder architecture. We extract raw semantic features $\mathbf{F}_{\mathrm{sem}}^{\mathrm{Ori}} = \mathcal{E}_{\mathrm{s}}(I_{\mathrm{LR}}) \in \mathbb{R}^{192\times(H/8\cdot W/8)}$ for I_{LR}. For this, we used the SSL backbone [44] as $\mathcal{E}_{\mathrm{s}}$ that was pretrained with our SAR training data. $\mathbf{F}_{\mathrm{sem}}^{\mathrm{Ori}}$ is then projected into a lower-dimensional latent space as $\mathbf{F}_{\mathrm{sem}} = \mathcal{A}(\mathbf{F}_{\mathrm{sem}}^{\mathrm{Ori}}) \in \mathbb{R}^{32\times(H/8\cdot W/8)}$ via a lightweight adapter $\mathcal{A}$ [8]. While $\mathbf{F}_{\mathrm{sem}}$ serves as a compact semantic anchor for detail decomposition, $\mathbf{F}_{\mathrm{sem}}^{\mathrm{Ori}}$ can be used as high-capacity keys and values for the cross-attention mechanism in Stage 3.

Simultaneously, $\mathcal{E}_{\mathrm{d}}$ as a detail encoder maps I_{HR} into a latent representation $\boldsymbol{z} = \mathcal{E}_{\mathrm{d}}(I_{\mathrm{HR}}) \in \mathbb{R}^{32\times(H/8\cdot W/8)}$. We define the latent detail feature $\boldsymbol{z}_{\mathrm{det}}$ as the residual between $\boldsymbol{z}$ and $\mathbf{F}_{\mathrm{sem}}$, which is denoted as $\boldsymbol{z}_{\mathrm{det}} = \boldsymbol{z} - \mathbf{F}_{\mathrm{sem}}$. This residual is then quantized into $\hat{\boldsymbol{z}}_{\mathrm{det}}$ which then becomes a codebook entry. By doing so, we restrict the codebook to high-frequency details, effectively decoupling them from spatial semantic context (low-frequency components).

To ensure the isolation of high-frequency details of I_{HR}, we employ a stochastic training strategy. With a probability of p, $I_{\mathrm{LR}} = \mathcal{D}(\mathbf{F}_{\mathrm{sem}})$ is reconstructed by compelling the decoder $\mathcal{D}$ to distill the semantic structure from $\mathbf{F}_{\mathrm{sem}}$ independently. With the remaining $1-p$ probability, $\mathcal{D}$ integrates $\hat{\boldsymbol{z}}_{\mathrm{det}}$ with $\mathbf{F}_{\mathrm{sem}}$ to synthesize I_{HR}. The reconstructed outputs for the two branches are obtained as:

$$\hat{I} = \begin{cases} \hat{I}_{\mathrm{LR}} = \mathcal{D}(\mathbf{F}_{\mathrm{sem}}) & \text{with } p \\ \hat{I}_{\mathrm{HR}} = \mathcal{D}(\hat{\boldsymbol{z}}_{\mathrm{det}} + \mathbf{F}_{\mathrm{sem}}) & \text{with } 1-p \end{cases} \tag{1}$$

where p is empirically set to 0.25. This stochastic strategy prevents the detail codebook from encoding redundant structures, isolating pure details, by forcing $\mathbf{F}_{\mathrm{sem}}$ to maintain a self-sufficient structural representation. Consequently, $\hat{\boldsymbol{z}}_{\mathrm{det}}$ focuses on semantic-aligned scattering signatures (high-frequency) of I_{HR} while $\mathbf{F}_{\mathrm{sem}}$ acts as an independent structural anchor that dictates semantic structures (low-frequency) of I_{LR}, making $\hat{\boldsymbol{z}}_{\mathrm{det}}$ inherently predictable from $\mathbf{F}_{\mathrm{sem}}$ in Stage 3.

Gumbel Vector Quantization. We employ Gumbel Vector Quantization (Gumbel-VQ) [21,27] to map $\boldsymbol{z}_{\mathrm{det}}$ into a discrete codebook entry, ensuring stable and differentiable training. To prevent codebook collapse and ensure its uniform utilization, we adopt the KL regularization $\mathcal{L}_{\mathrm{KL}} = \mathbb{E}\left[\sum_{k=1}^{K} \pi_k \log(\pi_k \cdot K)\right]$, where $\pi_k = P(q(\boldsymbol{z}_{\mathrm{det}}) = \boldsymbol{e}_k)$ denotes the probability of $\boldsymbol{z}_{\mathrm{det}}$ being assigned to the

k-th entry (codebook assignment) $\boldsymbol{e}_k$ via the quantizer $q(\cdot)$. This objective maximizes codebook entropy, forcing tokens to represent a diverse and comprehensive library of SAR scattering primitives.

Training Objectives. Stage 1 is optimized using a stochastic objective function that adaptively selects loss terms from the reconstruction branches as:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{rec}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} & \text{with } p \\ \mathcal{L}_{\text{rec}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} & \text{with } 1-p \end{cases} \tag{2}$$

where $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{adv}}$ [21] are an $\mathcal{L}_1$-reconstruction and adversarial losses, respectively. To capture domain-specific signatures without labeled SAR data, we define the perceptual loss $\mathcal{L}_{\text{per}}$ as $\mathcal{L}_{\text{SAFE}}$ using pre-trained a SAR SSL model $\mathcal{E}_{\text{s}}$ [44], enabling physically-grounded supervision from unlabeled imagery. This approach ensures the preservation of essential scattering patterns that pixel-wise losses often overlook. Restricting $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{KL}}$ to the $1-p$ branch forces the codebook toward realistic high-frequency signatures, while $\mathbf{F}_{\text{sem}}$ preserves the fundamental semantic structural layout.

3.3 Stage 2 - Semantic Prototype Map Generation (SPMG)

The SPMG converts continuous semantic features into a discrete spatial-semantic map (Fig. 3-b), bridging the encoding-generation gap stemming from ignoring semantic information during cross-attention.

Step 1: Mapping Spatial Features to Semantic Prototype Tokens. We define a set of $K = 3$ learnable prototype tokens, $P = \{\boldsymbol{p}_{\text{T}}, \boldsymbol{p}_{\text{S}}, \boldsymbol{p}_{\text{C}}\} \in \mathbb{R}^{3\times C}$, representing the fundamental scattering primitives in the SAR domain: target (T), shadow (S), and clutter (C). The choice of a token ensures a corresponding semantic interpretability by aligning it with a physical scattering primitive. Empirically, we found that $K > 3$ often leads to redundant or inactive prototype tokens, whereas $K = 3$ maintains an effective representation throughout all our experiments. For the semantic feature $\mathbf{F}_{\text{Sem}}^{\text{Ori}}$, we denote $\boldsymbol{f}_i \in \mathbb{R}^C$ as a spatial feature vector at spatial position i of $\mathbf{F}_{\text{Sem}}^{\text{Ori}}$. For $\boldsymbol{f}_i$, we compute a sparse similarity assignment $S_{i,k}$ for a prototype token $\boldsymbol{p}_k \in P$ using the Sparsemax [43] as:

$$S_{i,k} = \left[\text{Sparsemax}(-\|\boldsymbol{f}_i - \boldsymbol{p}_k\|_2^2)\right]_k \tag{3}$$

Step 2: Semantic Map generation via Dynamic Discretization. Instead of a naive argmax, which often blurs the physical distinction between target and clutter, we employ dynamic discretization to isolate high-intensity regions as structural anchors, ensuring a physical reliability (see *Suppl. B* for sensitivity analysis). At each spatial position i and for each class $k \in \{\text{T, S}\}$, we define a binary indicator mask $\mathcal{I}_{i,k} \in \{0, 1\}$ as:

$$\mathcal{I}_{i,k} = \mathbb{1}(S_{i,k} \geq \tau_k \text{ and } S_{i,k} \geq S_{i,m}), \quad k, m \in \{\text{T, S}\}, \quad m \neq k \tag{4}$$

where $\tau_k = \gamma \cdot \max_j(S_{j,k})$ is a relative threshold with j spanning all spatial indices. We empirically set $\gamma = 0.7$ to prioritize high-confidence scattering centers while filtering stochastic artifacts (see *Suppl. B* for sensitivity analysis). To ensure semantic exclusivity, clutter regions (C) are defined as the regions that do not belong to target and shadow regions, satisfying $\sum_{k\in\{\mathrm{T,S}\}} \mathcal{I}_{i,k} = 0$. To isolate stable scattering signatures, we intersect the top $c\%$ global $S_{i,\mathrm{C}}$ pixels with the clutter region where c is empirically set to 30. This sparse filtering, prioritizing the classes with relatively higher similarity scores, yields a non-overlapping Semantic Prototype Map (M_{sem}) that robustly guides Stage 3 to enable semantic-guided cross-attention while eliminating boundary ambiguity.

Prototype Optimization Objectives. To anchor the learnable prototype tokens to physical reality, we reconstruct feature map $\hat{\mathbf{F}} \in \mathbb{R}^{192\times(H/8\cdot W/8)}$ using prototype tokens P. The reconstructed feature at position i is formulated as $\hat{\boldsymbol{f}}_i = \sum_{k\in\{\mathrm{T,S,C}\}} S_{i,k}\cdot\boldsymbol{p}_k$, where $S_{i,k}$ is the sparse similarity score from Eq. (3) and $\boldsymbol{p}_k \in P$ is the corresponding learnable prototype token. We optimize $\hat{\mathbf{F}}$ against despeckled HR features $\mathbf{F}_{\mathrm{HR}}^{\mathrm{despec}} = \mathcal{E}_{\mathrm{s}}(I_{\mathrm{HR}}^{\mathrm{despec}})$ using a feature reconstruction loss $\mathcal{L}_{\mathrm{rec}} = \|\mathbf{F}_{\mathrm{HR}}^{\mathrm{despec}} - \hat{\mathbf{F}}\|_1$, where $I_{\mathrm{HR}}^{\mathrm{despec}}$ is obtained following [15]. To ensure distinctness between target, shadow, and clutter, we use a Repulsion Loss $\mathcal{L}_{\mathrm{rep}} = \sum_{m\neq n}\left|\frac{\boldsymbol{p}_m\cdot\boldsymbol{p}_n}{\|\boldsymbol{p}_m\|\|\boldsymbol{p}_n\|}\right|$ that enforces the orthogonality between prototype token pairs. The final objective is $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{rep}}\mathcal{L}_{\mathrm{rep}}$, ensuring that the learnable tokens P are both physically representative and semantically discriminative.

3.4 Stage 3 - Prototype-Map-Guided Masked Generative Modeling

Prototype-Map-Guided Masked Generative Modeling (PMG) (Fig. 3-c) synthesizes HR tokens using a ProSR Transformer based on the MaskGIT framework [7]. By reformulating reconstruction *as an iterative selection of categorical tokens* rather than smooth density approximation, our ProSR suppresses *mode interpolation* [1] and ensures that generated signals remain within physically valid scattering primitives. Input embeddings for the ProSR transformers are formed by concatenating masked detail tokens with $\mathbf{F}_{\mathrm{sem}}$ from Stage 1, providing a strong semantic prior of I_{LR} throughout the iterative process.

Prototype-Map-Guided Attention (PMGA). While self-attention captures long-range spatial dependencies, unconstrained cross-attention between latent tokens and semantic priors ($\mathbf{F}_{\mathrm{sem}}^{\mathrm{Ori}}$) can lead to contextual leakage due to the absence of semantic guidance, suffering from semantic drift, where scattering signatures from disparate terrain types may interfere during the cross-attention process. To address this, we introduce PMGA as a spatial-semantic guidance. Using M_{sem} (defined in Sec. 3.3), PMGA restricts cross-attention weights $A_{ij} = \mathrm{Softmax}(Q_i K_j^T/\sqrt{d} + M_{ij})$ such that HR tokens only aggregate information from semantically consistent regions. Here, Q_i is the latent query derived from self-attention at position i, and K_j is latent key with $\mathbf{F}_{\mathrm{sem}}^{\mathrm{Ori}}$ at position j. The Semantic Prototype Mask M_{ij} is defined as:

$$M_{ij} = \begin{cases} 0, & \text{if } \ell(\mathrm{pos}_i) = \ell(\mathrm{pos}_j) \\ -\infty, & \text{otherwise} \end{cases} \tag{5}$$

where $\ell(\text{pos}_i) \in \{\text{T, S, C}\}$ denotes a semantic label at position i obtained from M_{sem}. For unassigned regions, cross-attention is bypassed to block background interference. Enforcing intra-class attention suppresses inter-class feature leakage while reinforcing the distinct signatures of individual scattering centers, preserving the sharp intensity distributions and speckle statistics inherent to SAR signals.

Training Objectives. PMG is optimized by minimizing the cross-entropy (CE) loss between predicted and HR (GT) tokens, conditioned on the visible context and semantic guides:

$$\mathcal{L}_{\text{stage3}} = \mathbb{E}_{\hat{\boldsymbol{z}}_{\text{det}},M}\left[\sum\nolimits_{i\in M}\mathcal{L}_{\text{CE}}\left(\hat{\boldsymbol{z}}_{\text{det}}^{(i)}, p_\theta(\hat{\boldsymbol{z}}_{\text{SR}}^{(i)} \mid \hat{\boldsymbol{z}}_{\text{det}}^{\bar{M}}, \mathbf{F}_{\text{sem}}^{\text{ori}}, \mathbf{F}_{\text{sem}}, M_{\text{sem}})\right)\right] \tag{6}$$

where $\hat{\boldsymbol{z}}_{\text{det}} = \{\hat{\boldsymbol{z}}_{\text{det}}^{(i)}\}_{i=1}^{N}$ is the collection of HR tokens, $\hat{\boldsymbol{z}}_{\text{det}}^{(i)}$ denotes the token at index i, and p_θ is the likelihood predicted by PMG Transformer θ. The indices are partitioned into a masked set M and a visible set $\bar{M}$, where $\hat{\boldsymbol{z}}_{\text{det}}^{\bar{M}}$ serves as the context for predicting the masked tokens. By minimizing $\mathcal{L}_{\text{stage3}}$ in the discrete latent space, PMG learns to predict high-frequency scattering primitives strictly aligned by the spatial-semantic guidance map M_{sem}. This formulation effectively circumvents the mode interpolation problem, helping the reconstructed intensities remain within the support of physically plausible SAR characteristics.

Inference: Iterative Sampling with PMGA. At inference, our ProSR synthesizes the HR detail tokens via an iterative decoding process following the MaskGIT framework [7]. Starting from a fully masked grid, the number of tokens to be unmasked at each step t is determined by a decreasing masking schedule $\gamma(t)$ [7]. At each iteration, the highest-confidence tokens are fixed while the remainder are re-masked for the next step. Throughout this process, PMGA remains active to ensure that the ProSR Transformer only attends to semantically consistent regions in $\mathbf{F}_{\text{sem}}^{\text{Ori}}$. This iterative refinement, guided by M_{sem}, produces high-fidelity SAR image details aligned with the physical scattering layout.

4 Experimental Results

4.1 Dataset Construction

To evaluate our ProSR under realistic conditions, we curated a high-fidelity SAR dataset from the Umbra Open Dataset [57]. We utilized X-band Single Look Complex (SLC) data to serve as the source for HR reference imagery.

HR-LR Pairs. The source SLC data possesses a high-precision native resolution ($\sim$0.25 m azimuth, $\sim$0.25 m range). To establish a uniform benchmark, we standardized these to 0.25 m for SAR HR references and 1.0 m for SAR LR counterparts. Following the sub-aperture decomposition methodology [44], the SAR LR references were achieved via spectral domain cropping (sub-sampling the Doppler and range spectra) rather than naive image-space interpolation. This ensures physically grounded resolution degradation while preserving authentic speckle statistics and phase integrity in slant-range geometry. Subsequently, the SAR LR images were oversampled by zero-padding GT spectra in the frequency domain prior to the inverse transform.

Amplitude Extraction and Normalization. Following [17], we adopted a normalization strategy optimized for generative SAR tasks. To manage the high dynamic range, we scaled the amplitude (A) using the scene-level mean (μ) and standard deviation (σ): $A_{\text{norm}} = A/(\mu + 3\sigma)$. The normalized values were clipped to $[0, 1]$ to preserve predominant scattering while mitigating extreme outliers, then quantized into 8-bit integers for storage.
Data Filtering and Diversity. Physical consistency is paramount for model convergence. Since over 97% of our Umbra samples [57] are in VV polarization, we constructed a VV SAR dataset to prevent training biases from heterogeneous scattering signatures. We further restricted incidence angles from 10° to 50°, excluding extreme angles due to significant physical divergence and data scarcity.
Preprocessing and Statistics. The SAR images were tiled into non-overlapping 1024×1024 patches. To ensure a balanced representation of terrestrial features, we applied a selective quality control filter that prioritized patches with significant scattering structures while retaining a controlled portion of low-signal regions. The curated dataset comprises $132,452$ VV-polarized patches from 502 unique SAR images, with heights and widths ranging from approximately 4.5K to 66K and 10K to 94K pixels, respectively. These images span a diverse global geographic distribution (See Fig. 6 in *Suppl. A*) to ensure broad environmental representation, with their detailed statistics summarized in Table 6 of *Suppl. A*. More details for the characteristics of SAR images are described in *Suppl. A*.
Data Availability. Our curated dataset is available at `https://github.com/KAIST-VICLab/ProSR`.

4.2 Implementation Details

ProSR was implemented in PyTorch [46] and trained on dual NVIDIA A6000 GPUs using 256×256 random crops. Data augmentation included horizontal flips and random rotations (90°, 180°, 270°). For a fair comparison, all baseline SR networks (Table 1) were trained from scratch on our data without metric-specific early stopping, following their official implementations with SAR-adapted inputs. For AE-based baselines, pretrained AEs remained frozen while only the diffusion ISR networks were trained from scratch. For full-image evaluation, all methods use 256×256 sliding-window inference (stride 128) and weighted averaging.
Stage 1 - SADE. Stage 1 follows the ResShift [67] architecture ($f = 8$) with the encoder ($\mathcal{E}_{\text{d}}$)'s final projection modified to 32 dimensions. We employed Gumbel-VQ [21, 27] with 1,024 codebook entries and 32 embedding dimensions, stabilized by $\lambda_{\text{KL}} = 10^{-6}$. SADE was trained for 500K iterations using AdamW (batch size 12 on a single GPU, learning rate 1×10^{-4}) [41] with a composite loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{SAFE}}\mathcal{L}_{\text{SAFE}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}$, where λ_{SAFE} is set to 0.5 and, λ_{adv} is adaptively balanced [21]. Simultaneously, semantic features were extracted via a SAFE ViT-tiny model $\mathcal{E}_{\text{s}}$ [44], pre-trained on our SAR dataset, and projected to 32 embedding dimensions via a lightweight adapter [8].
Stage 2 - SPMG. The embedding space of P was aligned with the ViT-Tiny feature space [20]. The SPMG was trained for 15k iterations (AdamW, LR 1×10^{-4}, batch size 64, $\lambda_{rep} = 2$), with other hyperparameters following Sec. 3.3.

Stage 3 - PMG. PMG Transformer [20] comprises 20 layers with 8 attention heads and a 512-dimensional hidden space. We utilized a Positional Encoding Generator [14], facilitating stable sliding-window inference. Training was conducted for 600K iterations using AdamW (LR 1×10^{-4}, weight decay 0.045, $\beta_1 = 0.9, \beta_2 = 0.96$) with a total batch size 32. After a 10K-iteration linear warm-up, the learning rate was governed by a Cosine Annealing schedule [40], decaying from 1×10^{-4} to 1×10^{-7}. Additionally, a label smoothing factor of 0.1 was applied to the CE loss to ensure robust token synthesis.
Inference and Sampling. For token synthesis, we followed the MaskGIT [7] sampling process, employing a cosine schedule $\gamma(t)$ over 8 steps with a temperature of 4.5 for logit sampling.

4.3 Performance Comparison

Evaluation Metrics. To comprehensively evaluate our ProSR with other methods, we used three categories of metrics calculated in the SAR amplitude domain:

- **Standard Fidelity:** Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure(SSIM) [62] evaluate pixel-level fidelity and local structural agreement, respectively. However, these metrics can favor spatially conservative or over-smoothed reconstructions and therefore do not fully characterize SAR-specific scattering fidelity. Furthermore, SSIM depends strongly on the local covariance (σ_{xy}) between spatially aligned structures; small scatterer displacements, or merging can reduce this covariance and lower SSIM, in some cases even below that of oversampled LR images. We therefore interpret PSNR and SSIM jointly with complementary SAR-oriented structural, perceptual, and distributional metrics.
- **Target and Structural Integrity:** To evaluate physical and semantic consistency, we prioritize metrics focusing on fine-grained structural fidelity and scattering-sensitive signal integrity. This includes Information-content Weighted SSIM (IW-SSIM) [63] to assess fidelity in target-dense areas and Haar Perceptual Similarity Index (HaarPSI) [48] for local structural coherence and edge integrity. Additionally, we employ Target-to-Clutter Ratio (TCR)—calculated using prototype-derived target masks—to verify the radiometric fidelity. Rather than simply seeking higher TCR, we evaluate the absolute deviation from the GT ($|\Delta\mathrm{TCR}|$) to ensure the reconstructed scattering intensity remains physically consistent with the original observation.
- **Statistical and Perceptual Realism:** Fréchet Inception Distance (FID) [25], Density (Dens), and Coverage (Cov) [45] are used to evaluate statistical distributions, quantifying the alignment and overlap between the synthesized and GT data manifolds. Perceptual Realism is measured via Learned Perceptual Image Patch Similarity (LPIPS) [70], and Deep Image Structure and Texture Similarity (DISTS) [18] to evaluate structural and textural similarity.

Overall Quantitative Comparison. Table 1 compares ProSR with SOTA methods. Although the evaluated diffusion baselines (UPSR [69], LDM [50], and ResShift [67]) obtain higher pixel-aligned PSNR/SSIM, these gains do not

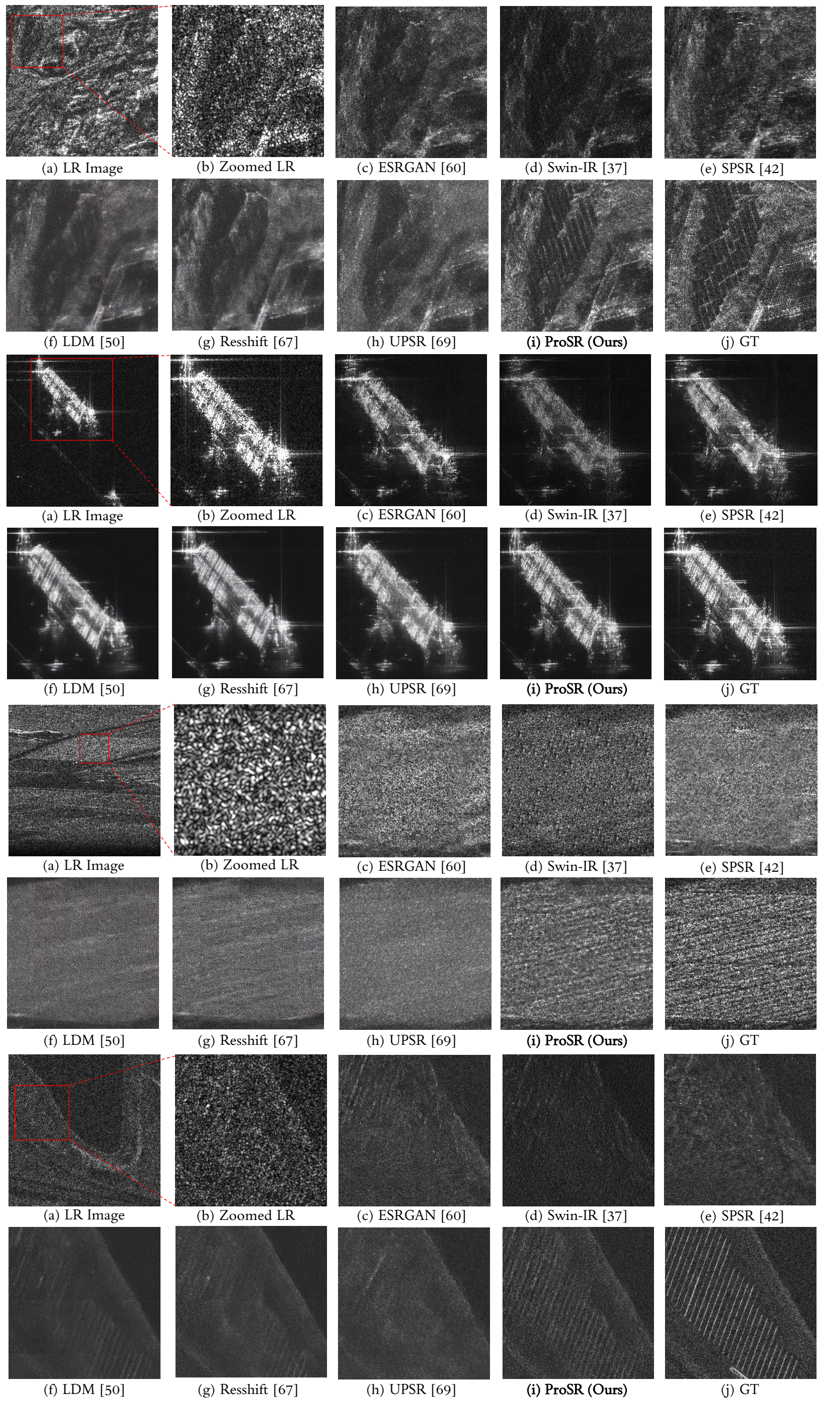


Fig. 4: Qualitative comparison of SAR ISR ($\times 4$) results (Zoom in for better view)

Table 1: Quantitative comparison on our SAR ISR ($\times 4$) benchmark. The best and second-best results are highlighted in **bold red** and blue, respectively.

Methods	Standard Fidelity PSNR ↑	SSIM ↑	Target & Structural Integrity $\lvert\Delta\text{TCR}\rvert$ ↓ (TCR)	IW-SSIM ↑	HaarPSI ↑	Statistical & Perceptual Realism FID ↓	Dens ↑	Cov ↑	LPIPS ↓	DISTS ↓
GT (Reference)	-	-	0.0 (3.3168)	-	-	-	-	-	-	-
Oversampled	16.2056	**0.1545**	0.6646 (3.9814)	0.3478	0.4211	70.86	0.2887	0.4379	0.8062	0.4370
ESRGAN [60]	16.1841	0.1012	0.2223 (3.0945)	0.3480	0.4655	50.84	0.7249	0.6912	**0.2977**	**0.1372**
SwinIR-GAN [37]	15.9363	0.0959	0.5142 (2.8026)	0.3306	0.4407	64.73	0.5156	0.4362	0.3071	0.2237
SPSR [42]	16.2756	0.1028	0.2831 (3.0337)	0.3451	0.4578	48.38	0.5886	0.5936	0.3184	0.1470
LDM-15 [50]	17.6427	0.1277	0.0892 (3.2276)	0.4065	0.4017	46.91	0.4946	0.6057	0.3888	0.2708
ResShift [67]	17.4784	0.1273	0.0657 (3.3825)	0.4066	0.4359	34.19	0.6156	0.7022	0.3575	0.2199
UPSR [69]	**17.8545**	0.1323	0.2815 (3.0353)	0.4201	0.4262	45.73	0.6429	0.6665	0.3801	0.2210
ProSR (Ours)	16.9293	0.1088	**0.0329** (3.2839)	**0.4206**	**0.4766**	**23.70**	**0.9741**	**0.8592**	0.3010	0.1532

Table 2: Stage 1 reconstruction comparison. LDM and ResShift use the same pre-trained VQGAN autoencoder [21].

Models	PSNR ↑	SSIM ↑	LPIPS ↓
LDM/ResShift AE [50, 67]	**18.3558**	**0.4072**	**0.1681**
Ours (1024)	18.1018	0.3799	0.2044

Table 3: Ablation on quantization types.

Quantization Types	PSNR ↑	SSIM ↑	LPIPS ↓
HR-VQ	16.5177	0.1141	0.2765
Dual-VQ	17.6195	0.3296	0.2165
Detail-VQ (Ours)	**18.1018**	**0.3799**	**0.2044**

Table 4: Ablations on codebook size K.

Codebook Size K	PSNR ↑	SSIM ↑	LPIPS ↓
512	17.9659	0.3641	0.2124
1024 (Selected)	**18.1018**	0.3799	0.2044
2048	18.0894	**0.3892**	**0.2020**

Table 5: Ablation study on the PMGA.

Attention Types	IW-SSIM ↑	FID ↓	Dens ↑	Cov ↑
Vanilla Cross-Attn.	0.4186	25.39	0.9404	0.8499
PMGA (Ours)	**0.4206**	**23.70**	**0.9741**	**0.8592**

guarantee better SAR scattering realism. ESRGAN [60] achieves competitive LPIPS/DISTS but struggle to preserve physically consistent speckle patterns. In contrast, our ProSR dominates in Target and Structural Integrity ($\lvert\Delta\text{TCR}\rvert$, IW-SSIM, HaarPSI), capturing both structural and radiometric fidelity. Notably, ProSR excels in FID, Density, and Coverage, demonstrating its superior statistical realism. These results indicate that our ProSR better captures the underlying distribution of SAR imagery while preserving essential scattering characteristics.

Quantitative Comparison of Autoencoder Reconstruction in Stage 1. Table 2 compares the reconstruction performances of several autoencoders (AE) in Stage 1. Since the baseline models built upon LDM [50] and ResShift [67] benefit from massive pre-training of AE at a much larger scale, the Stage 1 metrics of our ProSR are lower. Nevertheless, leveraging disentangled representations, our ProSR effectively captures discrete scattering anchors through Stages 2 and 3, providing a stronger semantic prior for the SR task, as shown in Table 1.

Qualitative Comparison. Fig. 4 visually compares the SAR ISR results. Since GAN-based models [37, 42, 60] frequently introduce gritty artifacts and vanilla diffusion baselines [50, 67, 69] yield oversmoothed textures, both paradigms are prone to structural hallucinations, resulting in a practical issue for reliable SAR-based ground observation. In contrast, our ProSR preserves sharp point-scatterer intensities and structural clarity with significantly reduced hallucination. By leveraging discrete scattering anchors, our ProSR provides categorical guidance that prevents from blurred averages or ungrounded features, ensuring high-

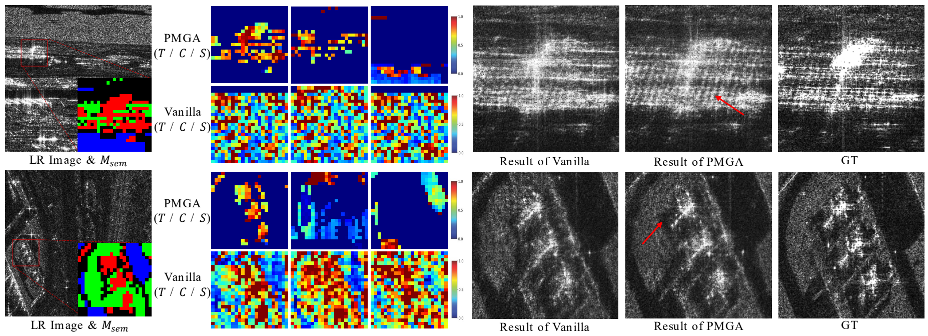


Fig. 5: Impact of M_{sem} (Red: Target, Green: Clutter, Blue: Shadow) on attention. The heatmaps represent pixel-wise averaged attention weights, normalized to the 90th percentile for improved visual clarity.

fidelity restoration of physically consistent scattering patterns. More qualitative comparisons are provided in Figs. 11, 12 and 13 of *Suppl. B*.

Downstream Utility Validation (ATR). To validate practical utility, we evaluate downstream ATR on MSTAR [31]. To fairly evaluate the recovery of HR-domain scattering characteristics, all methods use a fixed classifier [24] trained solely on original HR images without SR-specific adaptation. By accurately restoring structural primitives, our ProSR achieves 88.39% accuracy—outperforming the oversampled LR (23.87%) and the strongest baseline (84.18%, a 4.21%p gain). This confirms that preserving scattering physics is more critical for SAR imagery than merely optimizing pixel-level metrics. (See *Suppl. B* for detailed setups)

4.4 Ablation Studies

We conduct a series of ablation experiments to validate the effectiveness of the key components in our ProSR.

Effect of Detail-only Quantization. We compared three types of quantization (Table 3): standard HR-VQ, Dual-VQ (quantizing LR semantic features as well), and our Detail-VQ. (Note that across all comparative ablations, the HR codebook size is fixed to $N = 1024$, while Dual-VQ utilizes an extra $N = 256$ codebook for LR quantization.) Standard HR-VQ performs poorest across all metrics, producing *over-vivid* artifacts by forcedly discretizing stochastic speckles into fixed tokens. While Dual-VQ achieves competitive LPIPS, its pixel-level fidelity (PSNR/SSIM) drops significantly as quantizing the LR components introduces structural errors. In contrast, Detail-VQ optimizes the codebook exclusively for high-frequency details, effectively decoupling the representation from LR distortions. By doing so, Detail-VQ contributes to a cleaner latent space, preventing over-sharpening while preserving superior statistical and structural fidelity.

Impact of Codebook Size. We evaluated performance across $N \in \{512, 1024, 2048\}$, as summarized in Table 4. While $N = 512$ is insufficient for complex textures, $N = 1024$ and 2048 yield comparable LPIPS and SSIM. We selected

$N = 1024$ because it achieves highest PSNR and, more importantly, reduces the complexity of training Stage 3 classification task compared to a larger-sized codebook. This selected size ensures stable training and efficient convergence while maintaining a superior representational power for SAR primitives.

Effectiveness of PMGA Semantic-Guidance. Table 5 shows semantic guidance of PMGA reduces FID by 1.69. Furthermore, Fig. 5 illustrates that PMGA discriminates individual scattering centers, preventing indiscriminate blurred merging in vanilla cross-attention models. PMGA-driven generation, guided by M_{sem}, suppresses structural hallucinations while preserving intrinsic high-frequency integrity, leading to higher Density and IW-SSIM. Ultimately, our mechanism improves both generative fidelity (Dens to 0.9741) and diversity (Cov to 0.8592), ensuring physically reliable SAR reconstruction.

Discrete vs. Continuous (DiT-style) Modeling. To validate the advantages of discrete token modeling over score-based generation, we replaced our Stage 3 transformer with a continuous DiT-style variant—by modifying only the in/output linear projections—under the identical framework. As detailed in *Suppl. B*, the DiT-style approach struggled with dense, impulse-like scatterers of SAR imagery, often resulting in stochastic structural distortions and degraded LPIPS/FID (0.3211/40.43 for DiT vs. 0.3010/23.70 for ProSR). These results suggest discrete modeling effectively mitigates the structural merging of scatterers.

5 Conclusion

In this paper, we proposed ProSR, a generative framework to overcome the limitations of smooth-approximation-based diffusion models in SAR ISR, while addressing labeled SAR data scarcity. Unlike standard diffusion models that are often prone to *stochastic structural distortions*—misaligning complex scattering distributions and yielding over-smoothed textures—our ProSR performs *semantically-guided discrete token prediction* within a discrete detail latent space defined by SADE. To enable semantically-guided prediction, the SPMG utilizes an SSL backbone to extract label-free semantic priors, constructing explicit M_{sem} that guide PMGA to route information flows strictly within consistent categories. By mitigating inter-class confusion, ProSR restores sharp, physically consistent scattering without requiring manually labeled data. Extensive experiments on our 0.25 m resolution benchmark validate that ProSR achieves superior physical realism and structural accuracy while suppressing structural hallucinations. Future work will extend this paradigm to downstream tasks such as object detection, and develop physics-informed no-reference metrics to quantify signal authenticity.

Acknowledgements. This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government [Ministry of Science and ICT (Information and Communications Technology)] (Project Number: RS-2024-00338513, Project Title: AI-based Computer Vision Study for Satellite Image Processing and Analysis).

References

1. Aithal, S.K., Maini, P., Lipton, Z.C., Zico Kolter, J.: Understanding hallucinations in diffusion models through mode interpolation. arXiv e-prints pp. arXiv–2406 (2024)
2. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems **34**, 17981–17993 (2021)
3. Awais, C.M., Reggiannini, M., Moroni, D., Karakus, O.: A classification-aware super-resolution framework for ship targets in sar imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2026)
4. Bhattacharjee, S., Shanmugam, P., Das, S.: A hybrid algorithm for construction of super-resolution sar imagery for ship detection applications. IETE Journal of Research pp. 1–22 (2025)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
7. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
8. Chen, B., Bi, S., Tan, H., Zhang, H., Zhang, T., Li, Z., Xiong, Y., Zhang, J., Zhang, K.: Aligning visual foundation encoders to tokenizers for diffusion models. arXiv preprint arXiv:2509.25162 (2025)
9. Chen, C., Abdolshah, M., Shevchenko, V., Li, H., Xu, C., Purkait, P.: Srsr: Enhancing semantic accuracy in real-world image super-resolution with spatially re-focused text-conditioning. arXiv preprint arXiv:2510.22534 (2025)
10. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International conference on machine learning. pp. 1691–1703. PMLR (2020)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
12. Chen, Z., Zhang, C., Wan, C., Zhang, S., Xiong, B.: Dadsr: Degradation-aware diffusion super-resolution model for object-level sar image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025)
13. Chen, Z., Zhang, S., Xiong, B.: Super-resolving sar images with diffusion models: A dual evaluation of metrics and applications. In: 2024 IEEE 17th International Conference on Signal Processing (ICSP). pp. 340–346. IEEE (2024)
14. Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. In: ICLR 2023 (2023), `https://openreview.net/forum?id=3KWnuT-R1bh`, accessed 2026-06-30
15. Dalsasso, E., Denis, L., Tupin, F.: As if by magic: Self-supervised training of deep despeckling networks with merlin. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021)
16. Datcu, M., Huang, Z., Anghel, A., Zhao, J., Cacoveanu, R.: Explainable, physics-aware, trustworthy artificial intelligence: A paradigm shift for synthetic aperture radar. IEEE Geoscience and Remote Sensing Magazine **11**(1), 8–25 (2023)

17. Debuysère, S., Trouvé, N., Letheule, N., Lévêque, O., Colin, E.: Quantitative comparison of fine-tuning techniques for pretrained latent diffusion models in the generation of unseen sar images. arXiv preprint arXiv:2506.13307 (2025)
18. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
19. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
21. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
24. 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. pp. 770–778 (2016)
25. 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. Advances in neural information processing systems **30** (2017)
26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
27. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
28. Jiang, N., Zhao, W., Wang, H., Luo, H., Chen, Z., Zhu, J.: Lightweight super-resolution generative adversarial network for sar images. Remote Sensing **16**(10), 1788 (2024)
29. Karakuş, O., Achim, A.: On solving sar imaging inverse problems using nonconvex regularization with a cauchy-based penalty. IEEE Transactions on Geoscience and Remote Sensing **59**(7), 5828–5840 (2020)
30. Karimi, N., Taban, M.R.: A convex variational method for super resolution of sar image with speckle noise. Signal processing: Image communication **90**, 116061 (2021)
31. Keydel, E.R., Lee, S.W., Moore, J.T.: Mstar extended operating conditions: A tutorial. Algorithms for synthetic aperture radar imagery III **2757**, 228–242 (1996)
32. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
33. Kyriou, A., Mpelogianni, V., Nikolakopoulos, K., Groumpos, P.P.: Review of remote sensing approaches and soft computing for infrastructure monitoring. Geomatics **3**(3), 367–392 (2023)

34. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
35. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
36. Li, W., Yang, W., Hou, Y., Liu, L., Liu, Y., Li, X.: Saratr-x: Toward building a foundation model for sar target recognition. IEEE Transactions on Image Processing **34**, 869–884 (2025)
37. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
38. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
39. Liu, Z., Zhang, Z., Tang, H.: Semantic-guided diffusion model for single-step image super-resolution. arXiv preprint arXiv:2505.07071 (2025)
40. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
41. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
42. Ma, C., Rao, Y., Cheng, Y., Chen, C., Lu, J., Zhou, J.: Structure-preserving super resolution with gradient guidance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7769–7778 (2020)
43. Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: International conference on machine learning. pp. 1614–1623. PMLR (2016)
44. Muzeau, M., Frontera-Pons, J., Ren, C., Ovarlez, J.P.: Safe: a sar feature extractor based on self-supervised learning and masked siamese vits. arXiv preprint arXiv:2407.00851 (2024)
45. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International conference on machine learning. pp. 7176–7185. PMLR (2020)
46. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
47. Peng, J., Luo, X., Fu, J., Liu, D.: Confidence-based iterative generation for real-world image super-resolution. In: European Conference on Computer Vision. pp. 323–341. Springer (2024)
48. Reisenhofer, R., Bosse, S., Kutyniok, G., Wiegand, T.: A haar wavelet-based perceptual similarity index for image quality assessment. Signal Processing: Image Communication **61**, 33–43 (2018)
49. Ren, X., Wei, W., Xia, L., Huang, C.: A comprehensive survey on self-supervised learning for recommendation. ACM Computing Surveys **58**(1), 1–38 (2025)
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)

51. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2022)
52. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
53. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
54. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
55. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
56. Su, H., Li, Y., Xu, Y., Fu, X., Liu, S.: A review of deep-learning-based super-resolution: From methods to applications. Pattern Recognition **157**, 110935 (2025)
57. Umbra Space: Umbra open dataset: Very high-resolution SAR imagery. `https://umbra.space/open-data` (2023), accessed: 2026-02-17
58. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
59. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
60. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: The European Conference on Computer Vision Workshops (ECCVW) (September 2018)
61. Wang, Y., Li, S., Dong, G., Liu, H.: Metric or task: A new perspective of sar super-resolution imaging. IEEE Transactions on Geoscience and Remote Sensing (2025)
62. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
63. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Transactions on image processing **20**(5), 1185–1198 (2010)
64. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25456–25467 (2024)
65. Xiao, J., Zhang, J., Zou, D., Zhang, X., Ren, J., Wei, X.: Semantic segmentation prior for diffusion-based real-world super-resolution. arXiv preprint arXiv:2412.02960 (2024)
66. Yanshan, L., Li, Z., Fan, X., Shifu, C.: Ogsrn: Optical-guided super-resolution network for sar image. Chinese Journal of Aeronautics **35**(5), 204–219 (2022)
67. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. Advances in neural information processing systems **36**, 13294–13307 (2023)
68. Zhang, C., Zhang, Z., Deng, Y., Zhang, Y., Chong, M., Tan, Y., Liu, P.: Blind super-resolution for sar images with speckle noise based on deep learning probabilistic degradation model and sar priors. Remote Sensing **15**(2), 330 (2023)
69. Zhang, L., You, W., Shi, K., Gu, S.: Uncertainty-guided perturbation for image super-resolution diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17980–17989 (2025)

70. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
71. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
72. Zhou, J., Liu, Y., Liu, L., Li, W., Peng, B., Song, Y., Kuang, G., Li, X.: Fifty years of sar automatic target recognition: The road forward. arXiv preprint arXiv:2509.22159 (2025)
73. Zhu, Y., Huang, Y., Yang, M., Mao, D., Zhang, Y., Jiao, L., Zhang, Y., Yang, J.: Sar image super-resolution based on multi-scale edge texture-oriented gan approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025)