

Large Model-Driven Aerospace Microwave Imaging

GeoDiff-SAR: A Geometric Prior Guided Diffusion Model for SAR Image Generation

Fan ZHANG¹, Xuanting WU¹, Fei MA^{1*}, Qiang YIN¹ & Yuxin HU²

¹College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

²Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

Abstract Synthetic Aperture Radar (SAR) imaging results are highly sensitive to observation geometries and the geometric parameters of targets. However, existing generative methods primarily operate within the image domain, neglecting explicit geometric information. This limitation often leads to unsatisfactory generation quality and the inability to precisely control critical parameters such as azimuth angles. To address these challenges, we propose GeoDiff-SAR, a geometric prior guided diffusion model for high-fidelity SAR image generation. Specifically, GeoDiff-SAR first efficiently simulates the geometric structures and scattering relationships inherent in real SAR imaging by calculating SAR point clouds at specific azimuths, which serves as a robust physical guidance. Secondly, to effectively fuse multi-modal information, we employ a feature fusion gating network based on Feature-wise Linear Modulation (FiLM) to dynamically regulate the weight distribution of 3D physical information, image control parameters, and textual description parameters. Thirdly, we utilize the Low-Rank Adaptation (LoRA) architecture to perform lightweight fine-tuning on the advanced Stable Diffusion 3.5 (SD3.5) model, enabling it to rapidly adapt to the distribution characteristics of the SAR domain. To validate the effectiveness of GeoDiff-SAR, extensive comparative experiments were conducted on real-world SAR datasets. The results demonstrate that data generated by GeoDiff-SAR exhibits high fidelity and effectively enhances the accuracy of downstream classification tasks. In particular, it significantly improves recognition performance across different azimuth angles, thereby underscoring the superiority of physics-guided generation.

Keywords Synthetic Aperture Radar (SAR); Image Generation; Diffusion Models; Physics-Guided; Data Augmentation

Citation GeoDiff-SAR. Sci China Inf Sci, for review

1 Introduction

Synthetic Aperture Radar (SAR), as an active microwave imaging modality, offers unparalleled capabilities for all-time and all-weather Earth observation, playing an indispensable role in critical domains such as military reconnaissance, geological exploration, and disaster monitoring [1,2]. In recent years, the paradigm shift brought by Deep Learning (DL) has revolutionized SAR image interpretation. Automatic Target Recognition (ATR) algorithms, particularly those based on Convolutional Neural Networks (CNNs) [3] and Transformers [4], have achieved remarkable success. Unlike traditional methods relying on hand-crafted features, deep learning models automatically extract high-dimensional hierarchical representations from massive datasets, significantly elevating the intelligence and accuracy of SAR processing [5]. Beyond ATR, deep learning has also been successfully applied to various SAR image analysis tasks, including image segmentation [6, 7], target detection [8], and multi-object tracking [9].

However, the efficacy of these data-driven algorithms is heavily contingent upon the availability of massive, high-quality annotated datasets to drive parameter optimization. In practical applications, modern SAR systems face a severe “data hunger” dilemma [10]. On one hand, SAR imaging is highly sensitive to observation geometry (e.g., azimuth and depression angles), frequency bands, and polarization modes, making valid data for specific operational conditions extremely scarce. On the other hand, interpreting SAR imagery requires specialized domain expertise, rendering manual annotation prohibitively expensive and time-consuming. This data scarcity often leads to overfitting during the training of deep neural networks, causing a sharp degradation in generalization capability, especially in Few-shot or Zero-shot scenarios [11]. Additionally, SAR images are vulnerable to adversarial attacks [12], further complicating the development of robust recognition systems.

To mitigate this data shortage, researchers have historically explored various data augmentation strategies. As illustrated in the comprehensive capability comparison in Figure 1, these methods generally evolve from physics-based simulation to data-driven generation, each with distinct trade-offs in speed, accuracy, and interpretability.

* Corresponding author (email: mafei@mail.buct.edu.cn)

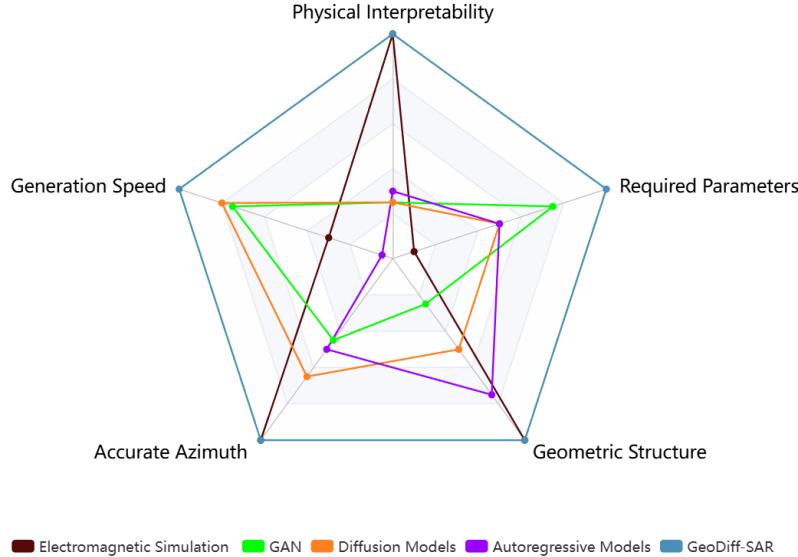


Figure 1 Radar chart comparison of five mainstream SAR image generation paradigms. The axes represent key performance metrics: Generation Speed, Accurate Azimuth Control, Physical Interpretability, Simplicity of Required Parameters, and Geometric Structure Accuracy. Compared to Electromagnetic Simulation, GANs, Autoregressive Models, and standard Diffusion Models, our proposed **GeoDiff-SAR** (represented by the green area) achieves the most balanced performance, effectively combining the high physical fidelity of simulations with the efficient generative capability of deep learning.

Electromagnetic Simulation. Traditional electromagnetic simulation methods, notably Shooting and Bouncing Rays (SBR) and Physical Optics (PO), model radar echoes by solving Maxwell's equations or their approximations [14, 15]. Balz et al. [16] demonstrated that GPU-accelerated real-time SAR simulation can achieve high geometric precision. Although these methods yield accurate geometric contours and physical interpretability, their applicability in deep learning pipelines is constrained by computationally intensive derivation processes and a deficiency in textural diversity. They often produce overly smooth images that fail to reproduce the complex speckle noise and background clutter inherent in real SAR imagery, resulting in a substantial “Sim-to-Real” gap [17].

Generative Adversarial Networks (GANs). To bridge this gap, GANs [18] emerged as a prevalent approach. Seminal frameworks such as Pix2Pix [19] and CycleGAN [20] have been utilized to translate optical images or semantic maps into the SAR domain. Guo et al. [21] leveraged GANs to generate multi-view SAR targets to facilitate ATR. Despite strides in texture synthesis, GANs suffer from inherent instability during adversarial training and are prone to “Mode Collapse” [21], yielding limited diversity. Crucially, vanilla GANs lack explicit 3D geometric constraints, frequently producing non-physical deformations (e.g., distorted wings) detrimental to interpretation tasks.

Autoregressive Models. Parallel to GANs, Autoregressive (AR) models have also made significant strides in image synthesis. Methods like VQ-VAE [22] tokenize images into discrete codes and model the distribution sequentially, demonstrating strong semantic understanding. DALL-E [23] further showcased the potential of AR models in large-scale text-to-image generation. However, these models typically model images as 1D sequences, which incurs high computational costs during inference and may suffer from error accumulation. Furthermore, their discrete latent spaces often lack the fine-grained continuous control required for the precise geometric structures characteristic of SAR targets.

Diffusion Models. Recently, Denoising Diffusion Probabilistic Models (DDPM) [24] and their variants (e.g., DDIM [25]) have established prominence due to their superior probability density modeling. By learning data distributions via a progressive denoising process, diffusion models synthesize samples with high fidelity and diversity [26]. In remote sensing, they have shown breakthroughs in optical generation [27], SAR despeckling [28], and super-resolution [29]. However, directly applying Text-to-Image (T2I) models (e.g., Stable Diffusion [30]) to SAR encounters substantial impediments. The unique SAR imaging mechanisms (layover, foreshortening) create a domain gap that natural image models lack the inductive bias to accommodate. Exclusive reliance on text prompts often results in geometric hallucinations, where targets exhibit radar-like textures but violate physical projection laws [31]. Recent advances in polarimetric SAR systems [32] and scattering characteristics analysis [33] have highlighted the importance of incorporating physical priors into generative models. Figure 2 visually summarizes these limitations and illustrates how our proposed method bridges the gap between physical simulation and data-driven

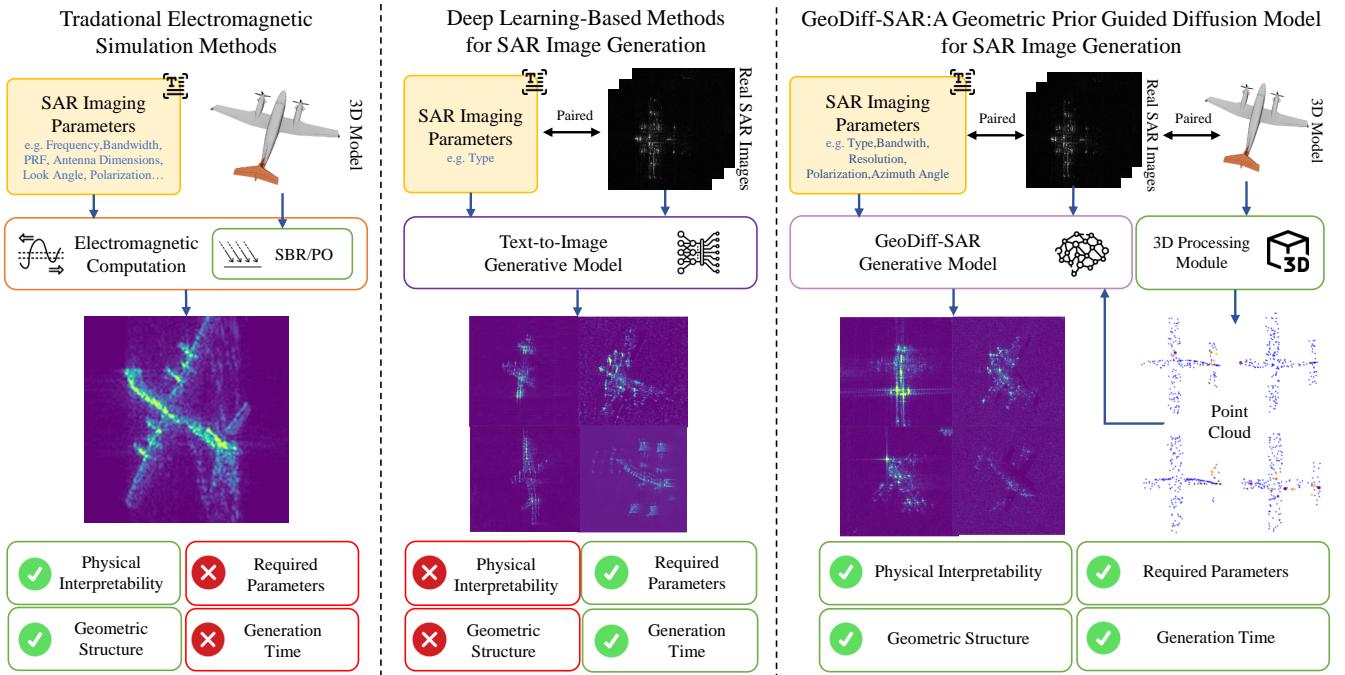


Figure 2 Illustration of the research motivation and paradigm comparison. (**Left**) Traditional electromagnetic simulations ensure physical correctness but suffer from slow generation speeds and require complex parameter configurations (e.g., antenna dimensions, PRF). (**Middle**) Generic deep learning-based text-to-image models offer fast generation but often lack physical constraints, leading to geometric hallucinations and structural instability. (**Right**) The proposed **GeoDiff-SAR** framework bridges this gap by integrating a 3D Processing Module to extract physical point clouds as geometric priors. This hybrid paradigm achieves a synergy of high generation efficiency, precise geometric structure, and azimuth controllability.

generation.

While methods like ControlNet [34] and T2I-Adapter [35] introduce spatial conditions (Canny edges, depth maps), acquiring pixel-aligned semantic edge maps for SAR is impractical. Some pioneering works utilized 3D model projections [36], but naive 2D projections forfeit depth and scattering intensity info.

To address these challenges, we propose a **Geometric Prior Guided Diffusion Model for SAR Image Generation (GeoDiff-SAR)**. This framework achieves high-quality, physically consistent SAR data augmentation by integrating explicit geometric priors with a powerful generative backbone. Specifically, we employ ray-casting on 3D models to calculate point clouds aligned with target azimuths [37], simulating geometric scattering centers and occlusion relationships. To fuse this multi-modal information, we introduce a multi-modal fusion network that dynamically regulates 3D physical priors, image controls, and textual descriptions [38]. Furthermore, we utilize Low-Rank Adaptation (LoRA) [39] to efficiently fine-tune the Stable Diffusion 3.5 (SD3.5) model [40], adapting it to the SAR domain while preserving generalization. As shown in Figure 3, GeoDiff-SAR is capable of generating high-fidelity images that strictly adhere to the physical geometry across continuous azimuth angles, demonstrating its robustness for SAR data augmentation.

The main contributions of our research are as follows.

- We propose GeoDiff-SAR, a novel physics-aware generative framework that bridges the gap between deterministic physical simulation and probabilistic diffusion models. By explicitly incorporating geometric priors, it addresses the fundamental challenge of SAR imaging's high sensitivity to observation geometries, enabling high-fidelity generation that is both physically consistent and texturally realistic.
- We design a target geometric information extraction mechanism and a geometry-text-image fusion strategy. The former efficiently transforms 3D models into azimuth-aligned point cloud representations through a physical projection process. This representation explicitly simulates the distribution of strong scattering centers and geometric occlusion relationships specific to the radar line-of-sight, thereby providing a robust structural reference that effectively suppresses geometric hallucinations. The latter employs a multi-modal fusion network based on Feature-wise Linear Modulation (FiLM) to bridge the gap between abstract conditions and visual generation. By dynamically integrating these discrete physical priors with continuous textual conditions (such as polarization and resolution) and SAR visual features, this strategy ensures precise control over both global semantic attributes and local structural details, yielding high-fidelity SAR images.

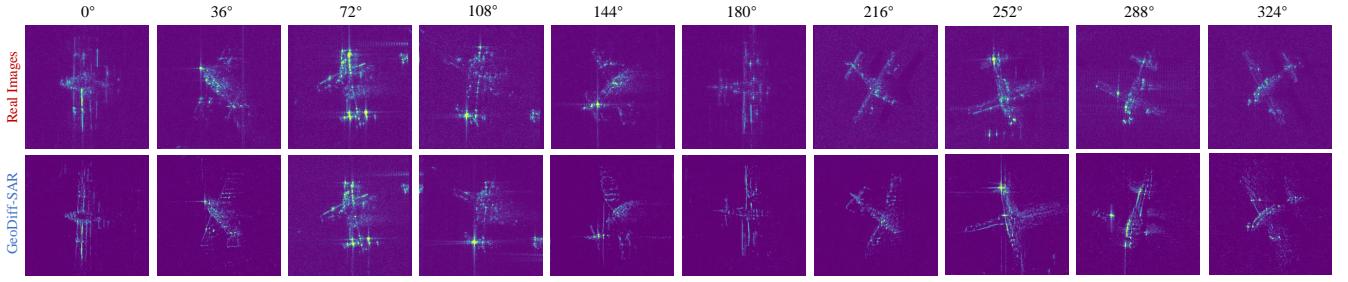


Figure 3 Visual comparison between real SAR samples (top row) and images generated by GeoDiff-SAR (bottom row) across ten distinct azimuth angles. The generated images accurately reproduce key SAR characteristics, including view-dependent geometric distortions (e.g., layover) and the distribution of strong scattering centers. The high degree of visual alignment with ground truth samples demonstrates the model's robustness in capturing both the physical structure and the speckle texture of SAR targets under varying observation geometries.

- We construct a comprehensive evaluation pipeline for downstream SAR target recognition tasks, validating the practical utility of physics-guided data augmentation. Extensive experiments demonstrate that GeoDiff-SAR significantly improves classification performance, particularly in data-scarce scenarios and sparse azimuth conditions, outperforming existing state-of-the-art generative baselines.

2 Method

2.1 Overall Framework

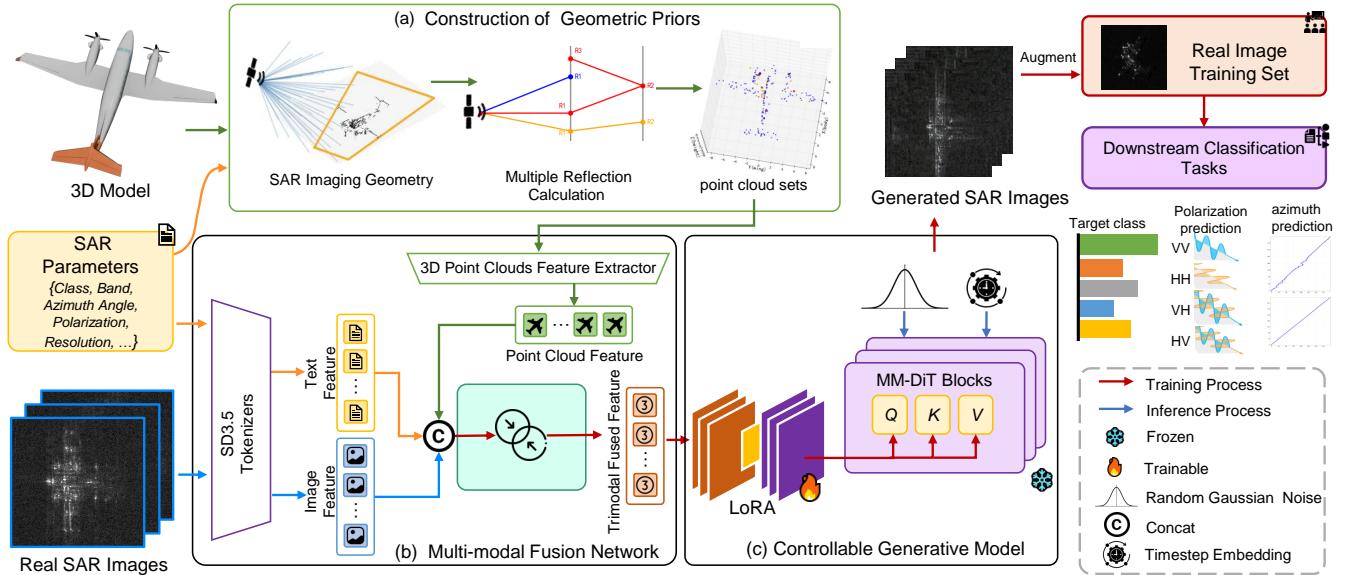


Figure 4 The overall architecture of the proposed GeoDiff-SAR framework. (a) **Construction of Geometric Priors**: 3D models are processed via a point cloud processing module to extract explicit geometric scattering characteristics, such as multi-bounce reflections, serving as a robust physical prior. (b) **Multi-modal Fusion Network**: These physical features are synthesized with textual and visual conditions through a multi-modal fusion network to condition the Stable Diffusion 3.5 backbone, which is efficiently adapted to the SAR domain using LoRA (trainable) while keeping the pre-trained weights frozen. (c) **Controllable Generative Model**: As depicted in the top-right, the generated high-fidelity SAR images are utilized to augment scarce real training data for downstream PyTorch Image Models Multi-Label Classification tasks, thereby validating the practical utility and effectiveness of the proposed data augmentation strategy.

As illustrated in Figure 4, GeoDiff-SAR bridges the gap between the physical interpretability of traditional simulations and the distribution fitting power of deep generative models. By decoupling deterministic physical simulation from probabilistic latent diffusion, the framework enforces explicit physical inductive biases to address the texture-geometry misalignment inherent in SAR generation. To enforce precise geometric control, we construct an independent physical extraction branch. By employing Ray Casting on 3D models under specified radar geometries (azimuth θ , elevation ϕ), we simulate SAR scanning to capture SAR-relevant spatial features, specifically occlusion relationships and main scattering structures. These raw point clouds are then encoded into high-dimensional

physics-informed representations via a dedicated point cloud encoder. We design a parameter-efficient fine-tuning strategy to mitigate overfitting on limited SAR datasets. A multi-modal fusion network is introduced to dynamically integrate geometric priors with textual and visual conditions. Furthermore, we employ LoRA to efficiently adapt the pre-trained SD3.5m to the SAR domain, allowing the model to capture SAR-specific texture styles while preserving its original generalization capabilities. We leverage Stable Diffusion 3.5 Medium as the generative foundation. Distinct from traditional U-Net architectures, SD3.5m employs a Multimodal Diffusion Transformer (MMDiT). This architecture enables efficient processing of long-sequence dependencies and supports independent weighting for multi-modal inputs, making it ideal for handling complex physical-semantic interactions.

2.2 Construction of Geometric Priors

To address the issue of geometric "hallucinations" common in generative models under weak constraints, we construct an explicit physical prior branch. The core logic of this branch is not to directly generate the final SAR image, but rather to simulate the target's geometric scattering structures and occlusion relationships inherent in the SAR imaging process. SAR imaging, functioning as a typical side-looking ranging system, induces unique geometric distortions—namely foreshortening, layover, and shadow—which differ fundamentally from the central projection mechanism of natural optical images. To accurately capture these distinct geometric features, we employ 3D models to perform virtual SAR scanning simulations using an object-centric observation geometry, as illustrated in Figure 5.

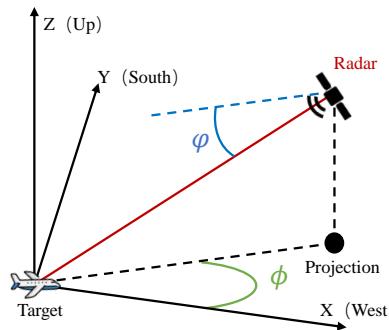


Figure 5 Schematic diagram of the object-centric radar observation geometry employed in the point cloud simulation. The coordinate system is defined with the X -axis pointing West (0°) and the Y -axis pointing South (90°). The spatial position of the radar sensor is determined by the azimuth angle ϕ and the depression angle ψ , establishing the Line-of-Sight (LOS) vector pointing towards the target center O . This geometric configuration serves as the basis for constructing the transformation matrix $\mathbf{T}_{\text{radar}}$ and executing ray casting to capture view-dependent scattering characteristics.

In this geometric configuration, we define a specific coordinate system aligned with the target's orientation, where the X -axis is designated as West (0°) and the Y -axis as South (90°), with the azimuth angle increasing in a counter-clockwise direction. Based on this setup, for the 3D mesh model $\mathcal{M} = \{\mathcal{V}, \mathcal{F}\}$ of a given target (where \mathcal{V} denotes the vertex set and \mathcal{F} the face set), we construct the transformation matrix $\mathbf{T}_{\text{radar}}$ to map coordinates from the world system to the radar Line-of-Sight (LOS) coordinate system:

$$\mathbf{T}_{\text{radar}} = \begin{bmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi \cos \psi & \cos \phi \cos \psi & -\sin \psi \\ \sin \phi \sin \psi & \cos \phi \sin \psi & \cos \psi \end{bmatrix} \quad (1)$$

where ϕ represents the azimuth angle and ψ denotes the depression angle. To cover the target and simulate the SAR imaging area, the method employs a two-level ray generation strategy: regular grid scanning and dense random scanning. Regular grid scanning generates the main ray beam with a fixed resolution $\Delta\theta$ in both azimuth and elevation directions to capture the overall contour of the target. For each azimuth step α and elevation step β , the ray direction vector \vec{D} is given by:

$$\vec{D} = \begin{bmatrix} \cos(\beta) \cos(\alpha) \\ \cos(\beta) \sin(\alpha) \\ \sin(\beta) \end{bmatrix} \quad (2)$$

Monte Carlo dense sampling is introduced to fill the gaps left by regular sampling. To enhance detail capture for complex structures (such as air intakes and landing gears), we introduce random perturbation rays directed toward the model center. For each ray \mathbf{r} , its direction vector \mathbf{d} is defined as:

$$\mathbf{d} = \text{Normalize}(\mathbf{d}_{target} + \boldsymbol{\xi}), \quad \text{where } \boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \quad (3)$$

where \mathbf{d}_{target} is the principal direction pointing to the geometric center of the model, $\boldsymbol{\xi}$ represents the Gaussian random perturbation term, and σ controls the beam divergence.

High-intensity regions in real SAR images typically originate from secondary or tertiary reflections caused by the corner reflector effect. To emulate this characteristic, we design a recursive ray tracing and energy attenuation model. Recent work on scattering characteristics analysis [33] has demonstrated the importance of capturing multi-bounce reflections for accurate SAR target representation. The calculation results are presented in Figure 6.

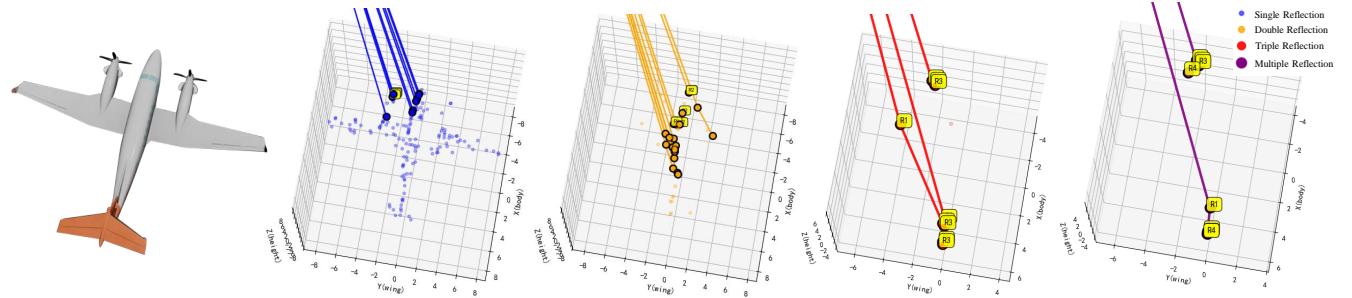


Figure 6 Illustration of multiple bounce ray tracing paths at an azimuth angle of 300°

We define the ray state as $S_k = (\mathbf{p}_k, \mathbf{d}_k, E_k)$, representing the origin, direction, and energy intensity of the k -th bounce, respectively. For the k -th iteration, we first calculate the nearest intersection point \mathbf{p}_{hit} between the ray and the mesh \mathcal{M} , as well as the surface normal vector \mathbf{n} at that location. The pixel brightness in SAR images depends on the backscattering coefficient of the target. Based on the geometric properties at the intersection point, we construct a Pseudo-RCS Response Function Ψ to modulate the current energy I_k :

$$I_{scatter} = I_k \cdot e^{-\mu L_{path}} \cdot \Psi(\mathbf{p}_{hit}, \mathbf{n}, \mathcal{F}) \quad (4)$$

where L_{path} denotes the cumulative path length, and μ represents the medium absorption coefficient. The response function Ψ is composed of the weighted product of four key geometric factors, precisely simulating the scattering logic implemented:

$$\Psi = \mathcal{W}_{base} \cdot \mathcal{H}_{edge}(\mathcal{F}) \cdot \mathcal{H}_{orient}(\mathbf{n}) \cdot \mathcal{H}_{struct}(\mathbf{p}_{hit}) \quad (5)$$

Here, \mathcal{H}_{edge} is the edge effect factor. When the area of the face containing the intersection point, $A_{\mathcal{F}}$, is smaller than the threshold τ_{area} (typically corresponding to wing edges or antennas), $\mathcal{H}_{edge} = \alpha_{edge} > 1.0$, simulating edge diffraction enhancement. \mathcal{H}_{orient} is the side-looking orientation factor, simulating the strong response of side-looking radar to vertical surfaces. When the vertical component of the normal vector $|\mathbf{n}_z| < \tau_{vert}$ (e.g., the side of the fuselage), a high gain α_{vert} is assigned; conversely, a low gain is assigned for horizontal surfaces (e.g., the top surface of wings). \mathcal{H}_{struct} is the structural scattering factor, which detects specific spatial regions (such as the junction between the wing and the fuselage) to simulate the dihedral corner reflector effect, applying a gain α_{struct} . The reflection direction of the secondary ray, \mathbf{d}_{k+1} , not only follows the law of specular reflection but also incorporates a diffuse scattering component caused by surface roughness:

$$\mathbf{d}_{spec} = \mathbf{d}_k - 2(\mathbf{d}_k \cdot \mathbf{n})\mathbf{n} \quad (6)$$

$$\mathbf{d}_{k+1} = \text{Normalize}(\mathbf{d}_{spec} + \zeta \cdot \mathbf{u}), \quad \mathbf{u} \sim \mathcal{U}(-1, 1)^3 \quad (7)$$

where ζ is the roughness coefficient. This recursive process continues until the maximum number of reflections K_{max} is reached or the energy E falls below a threshold.

As shown in Figure 7, through the aforementioned recursive calculations, we ultimately obtain a point cloud set $\mathcal{P} = \{(\mathbf{p}_{hit}, I_{final}, k) \mid I_{final} > \tau_{min}\}$ containing coordinates, intensity, and reflection type. This point cloud accurately records the distribution of scattering centers for single, double, and triple reflections, providing a physics-compliant geometric skeleton for the subsequent generative model.

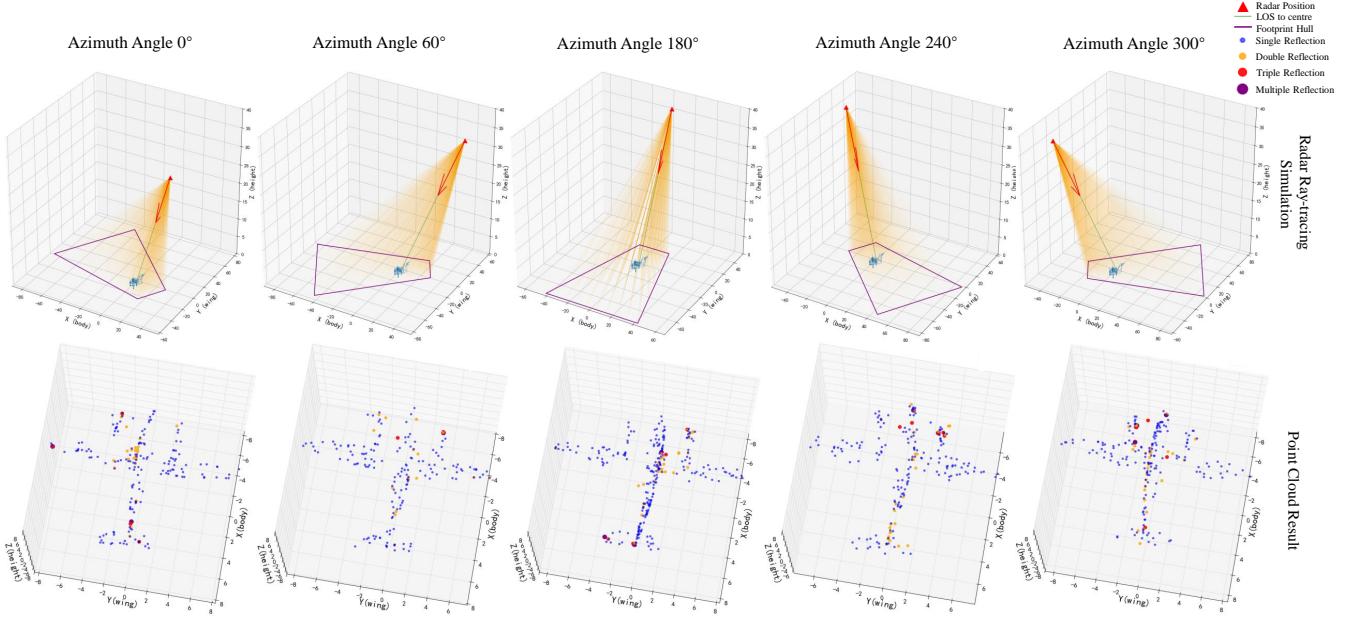


Figure 7 Schematic diagram of radar illumination at different azimuth angles and the corresponding calculated point cloud models. The point clouds are color-coded by reflection type (blue for single, yellow for double, red for triple, and purple for multiple bounces), effectively capturing the view-dependent spatial distribution of strong scattering centers and occlusion relationships.

To extract high-dimensional semantic geometric features from the simulated SAR point cloud \mathcal{P} , we adopt the Point Transformer V3 (PTv3) [41] as the point cloud encoder. Specifically, given the input point cloud $\mathcal{P} \in \mathbb{R}^{N \times (3+D_{in})}$ containing physical properties (intensity, number of reflections), we first preprocess it via voxelization using Grid Sampling to mitigate the impact of uneven local density. Subsequently, the PTv3 backbone utilizes a serialization strategy based on Space-filling Curves to map unordered 3D points into an ordered 1D sequence. During the feature extraction stage, the Multi-scale Attention Block aggregates local geometric details and global shape contexts across different resolution levels. The output feature \mathbf{F}_{geo} not only encodes the fine-grained 3D structure of the aircraft but also integrates the simulated physical scattering intensity information. To adapt to the subsequent 2D generation network, we employ Adaptive Projection Mapping to project the 3D features \mathbf{F}_{geo} onto the radar slant-range plane. This process yields a spatially aligned physical condition feature map $\mathbf{C}_{geo} \in \mathbb{R}^{H \times W \times C}$, serving as strong physical guidance for the GeoDiff-SAR generation process.

2.3 Multi-modal Fusion Network

To effectively inject 3D physical priors (point cloud features) and visual conditions (image features) into the generative space of text-guided diffusion models, we propose a multi-modal fusion network. Rather than relying on simple feature concatenation, this network is designed as a five-stage cascaded architecture. It aims to progressively reinforce the physical consistency and semantic alignment of the generated results, transitioning from coarse-grained global weighting to fine-grained feature modulation. The overall workflow is illustrated in Figure 8.

Given that the text features $F_t \in \mathbb{R}^{B \times D}$ (derived from CLIP [42] or T5 [43], where $D = 2048$), the point cloud physical features $F_p \in \mathbb{R}^{B \times 64}$, and the image features $F_i \in \mathbb{R}^{B \times 16}$ reside in distinct feature spaces with varying dimensions, it is essential to first map them into a unified high-dimensional manifold space. We employ a Linear Projection layer in conjunction with Layer Normalization to mitigate the distributional discrepancies across modalities:

$$\tilde{F}_m = \text{LayerNorm}(\mathbf{W}_m F_m + \mathbf{b}_m), \quad m \in p, i \quad (8)$$

$$\tilde{F}_t = \text{LayerNorm}(F_t) \quad (9)$$

where $\tilde{F}_p, \tilde{F}_i, \tilde{F}_t \in \mathbb{R}^{B \times 2048}$ denote the feature vectors after dimension unification.

To account for the variations in the signal-to-noise ratio between physical priors and textual descriptions across different training samples, we design an Adaptive Gating Module to dynamically allocate modal weights. This step achieves global information complementarity among modalities. After concatenating the tri-modal features, this module predicts the importance score α for each modality via a Multi-Layer Perceptron (MLP):

$$\alpha = \text{Softmax}(\mathbf{W}_{gate}(\text{Dropout}(\sigma(\mathbf{W}_{in}[\tilde{F}_t, \tilde{F}_p, \tilde{F}_i])))) \quad (10)$$

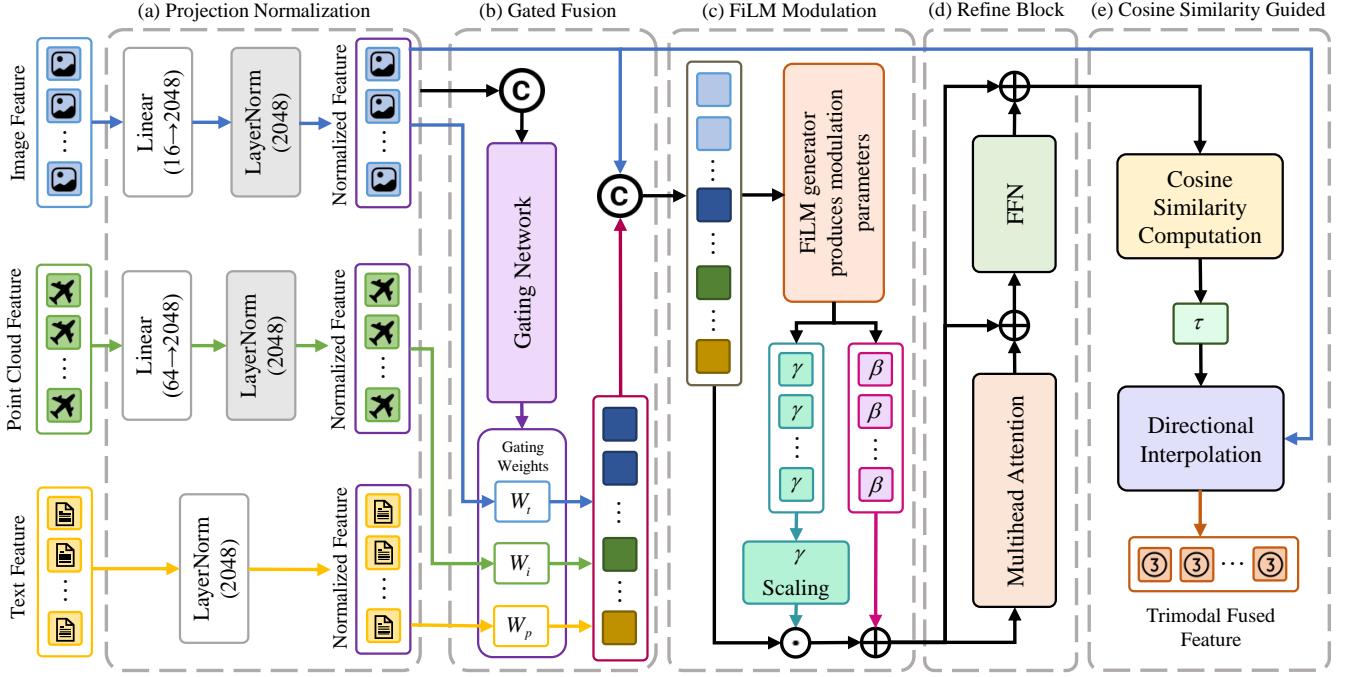


Figure 8 The overall architecture of the proposed multi-modal fusion network. It comprises five cascaded modules: (1) **Projection Normalization** aligns multi-modal features into a unified high-dimensional space; (2) **Gated Fusion** dynamically assigns weights to text, image, and point cloud features based on information density; (3) **FiLM Modulation** injects fine-grained texture details through affine transformations; (4) **Refine Block** enhances feature interaction via multi-head attention; and (5) **Cosine Similarity Guided** mechanism enforces semantic consistency by interpolating features based on visual alignment.

where $\alpha = [\alpha_t, \alpha_p, \alpha_i]$ and σ denotes the GELU activation function. Based on these weights, we obtain the preliminarily fused feature F_{pre} :

$$F_{pre} = \alpha_t \tilde{F}_t + \alpha_p \tilde{F}_p + \alpha_i \tilde{F}_i \quad (11)$$

To incorporate the texture features of SAR images more finely into the generation process, we introduce the Feature-wise Linear Modulation (FiLM) mechanism. In contrast to gated additive fusion, FiLM performs multiplicative modulation on feature channels via affine transformation. We take the concatenation of the preliminarily fused feature F_{pre} and the image feature \tilde{F}_i as the conditional input to generate the scaling coefficients γ and the shifting coefficients β :

$$[\gamma, \beta] = \text{MLP}_{film}([F_{pre}, \tilde{F}_i]) \quad (12)$$

To ensure training stability, we apply a Tanh constraint and residual scaling to the scaling coefficients:

$$\hat{\gamma} = 1 + \lambda \cdot \tanh(\gamma) \quad (13)$$

$$F_{film} = \hat{\gamma} \odot F_{pre} + \beta \quad (14)$$

where \odot denotes element-wise multiplication, and λ is a scaling factor (set to 0.5). This mechanism enables the model to dynamically enhance or suppress the activation response of specific channels based on the texture intensity of the image.

To prevent the generated results from deviating from the authentic visual semantics after multi-step transformations, we introduce a geometric constraint mechanism guided by Cosine Similarity. We calculate the cosine similarity, s_{cos} , between the refined feature $F_{refined}$ and the image feature \tilde{F}_i . When the similarity falls below a preset threshold τ_{target} , we enforce an interpolation of the feature vector towards the direction of the image semantics:

$$\tau = \text{Clamp}\left(\frac{\max(0, \tau_{target} - s_{cos})}{\max(\epsilon, 1 - s_{cos})}, 0, \tau_{max}\right) \quad (15)$$

$$F_{final} = (1 - \tau) \frac{F_{refined}}{|F_{refined}|} + \tau \frac{\tilde{F}_i}{|\tilde{F}_i|} \quad (16)$$

Finally, feature energy conservation is maintained through amplitude restoration. This strategy effectively guarantees that the generated SAR images strictly adhere to the input physical and visual conditions while possessing textual semantics.

2.4 Controllable Generative Model

Considering the massive parameter count of SD3.5m and the scarcity of labeled data in the SAR domain, Full Fine-tuning is not only computationally expensive but also prone to overfitting on small samples, thereby leading to a loss of generative diversity. Therefore, we adopt a Parameter-Efficient Fine-Tuning strategy. During the training process, we freeze the pre-trained VAE encoder \mathcal{E} , decoder \mathcal{D} , and the CLIP/T5 text encoder to preserve their constructed high-quality latent space and semantic understanding capabilities. The set of parameters Θ requiring optimization consists of only two parts:

$$\Theta = \{\theta_{LoRA}, \theta_{Fusion}\} \quad (17)$$

Here, θ_{LoRA} represents the Low-Rank Adaptation matrix parameters inserted into the attention layers of the MMDiT backbone, used for transferring the texture style of the SAR domain; θ_{Fusion} denotes the parameters of our proposed multi-modal fusion network, responsible for learning the nonlinear mapping relationships among multi-modal conditions (text, point cloud, image). The schematic diagram of the LoRA working principle is shown in Figure 9.

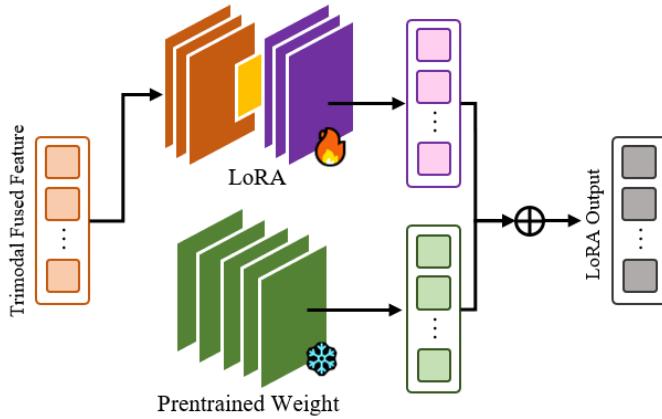


Figure 9 The schematic diagram of the LoRA working principle applied in the SD3.5 backbone.

The training objective of the model is based on the standard noise prediction paradigm of the Latent Diffusion Model. Given a real SAR image $x \in \mathcal{D}_{train}$, we first utilize the VAE encoder to map it into the latent space $z_0 = \mathcal{E}(x)$. Subsequently, Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is added to z_0 at time step $t \sim [1, T]$ to obtain the noisy latent variable z_t . The inputs to the network ϵ_θ include the noisy latent variable z_t , the time step t , and the multi-modal condition embedding C_{fused} generated by the fusion network. Our optimization objective is to minimize the Mean Squared Error (MSE) between the predicted noise and the real noise:

$$\mathcal{L}_{simple} = \mathbb{E}_{z_0, \epsilon, t, C_{fused}} [\|\epsilon - \epsilon_\theta(z_t, t, C_{fused}(\theta_{Fusion}))\|_2^2] \quad (18)$$

where C_{fused} is the output resulting from the fusion of text features F_t , physical geometric features F_p generated by SAR simulation, and the corresponding SAR image features F_i . Through end-to-end gradient backpropagation, θ_{Fusion} learns how to dynamically adjust the injection intensity of physical priors, while θ_{LoRA} learns how to translate these conditions into SAR-specific speckle textures and background clutter.

In the inference phase, GeoDiff-SAR demonstrates its "Text + Physics" dual-driven generation capability. The inference workflow is designed as an automated generation pipeline that does not require paired image inputs. Users only need to provide a text instruction T_{in} containing the target category and imaging parameters. The system first retrieves the corresponding 3D model based on the category information in T_{in} and parses the azimuth and depression angle parameters. Subsequently, it invokes the SAR simulator described in Section 3.2 to generate the point cloud P under the specific viewpoint in a virtual environment and extracts the physical geometric features F_p . Since there is no corresponding ground truth SAR image during the inference phase, we set the image feature input F_i to a zero vector or use a preset average SAR style vector. At this point, the Adaptive Gating Network described in Section 3.3 automatically adjusts the weight distribution, increasing the weights of the text F_t and physical features F_p to ensure that the generation process is primarily guided by geometric structures and semantic descriptions. To balance generation diversity and fidelity, and specifically to ensure that the generated images strictly adhere to the geometric contours of the 3D model, we extend the Classifier-Free Guidance (CFG) strategy. During the sampling process, we simultaneously compute the conditional predicted noise $\epsilon_\theta(z_t, C_{fused})$ and the unconditional predicted

noise $\epsilon_\theta(z_t, \emptyset)$ (i.e., with empty text and empty point cloud inputs). The final denoising step is determined by the following equation:

$$\tilde{\epsilon}_t = \epsilon_\theta(z_t, \emptyset) + w \cdot (\epsilon_\theta(z_t, C_{fused}) - \epsilon_\theta(z_t, \emptyset)) \quad (19)$$

where w is the Guidance Scale. By adjusting w , we can control the intensity of the physical prior: a larger w forces the model to generate SAR images that are highly aligned with the 3D point cloud contours, thereby effectively suppressing geometric hallucinations; whereas a smaller w allows for more textural variations. Finally, the denoised latent variable z_0 is mapped back to the pixel space via the VAE decoder \mathcal{D} to obtain the final high-quality SAR image. As shown in Figure 10, leveraging geometric physical constraints and the powerful generative capabilities of SD3.5m, GeoDiff-SAR is capable of generating images that are highly consistent with the azimuth of real images.

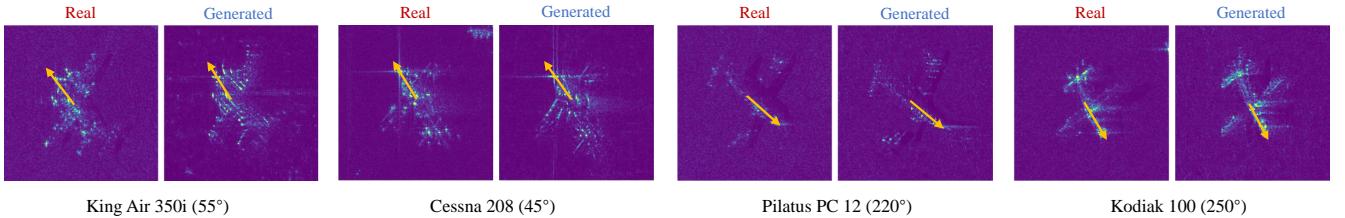


Figure 10 Qualitative generation results of GeoDiff-SAR across different aircraft categories and specific azimuth angles. The figure presents side-by-side comparisons of real SAR images and generated samples for: King Air 350i (55°), Cessna 208 (45°), and Pilatus PC 12 (220°), Kodiak 100 (250°). Guided by 3D physical priors, the generated targets exhibit consistent orientation with the specified viewpoints, mitigating the azimuth ambiguity issues often observed in purely data-driven approaches.

3 Experiments

3.1 Experimental setups

To fully characterize the generative performance of GeoDiff-SAR, this section presents a comprehensive evaluation from two complementary perspectives. First, we employ a suite of visual metrics to quantify the fidelity and diversity of the synthesized images. Second, to substantiate the model’s effectiveness in data augmentation, we conduct downstream classification experiments, thereby verifying the practical value of GeoDiff-SAR in boosting the performance of recognition algorithms.

Dataset. To strictly evaluate the efficacy of physics-guided data augmentation, we constructed a high-resolution SAR aircraft dataset based on field collection campaigns at Shanxi Yaocheng Airport. The dataset includes four distinct aircraft categories (King Air 350i, Cessna 208, Pilatus PC 12, Kodiak 100) and covers four polarization modes (HH, HV, VH, VV). Polarimetric SAR systems [32] and sub-look decomposition techniques [8] have shown great potential in enhancing target discrimination capabilities, which motivates our comprehensive evaluation across multiple polarization modes.

As illustrated in Figure 11, the real data is strategically partitioned into two groups to simulate a data-scarce scenario. For the training of generative models, we utilized only a sparse subset of real images sampled at 10° azimuth intervals (e.g., $0^\circ, 10^\circ, 20^\circ, \dots$). To bridge the angular gaps, GeoDiff-SAR was tasked with generating synthetic samples specifically at the intermediate 5° azimuth offsets (e.g., $5^\circ, 15^\circ, 25^\circ, \dots$). These generated images are then combined with the sparse real data to form a densely sampled dataset for downstream classification tasks, effectively doubling the angular resolution.

Implementation Details. Our framework was implemented in PyTorch and trained on a high-performance cluster equipped with three NVIDIA GeForce RTX 4090 GPUs. We initialized the generative backbone with SD3.5m weights. To align with the computational constraints and ensure training stability, all SAR images were resized to 256×256 pixels. We adopted the LoRA fine-tuning strategy, setting the network dimension to 32 and alpha to 16, while keeping the VAE and pre-trained text encoders frozen. The training was conducted in BF16 mixed precision to optimize memory usage. We employed the Prodigy optimizer, an adaptive algorithm with a D-coefficient of 0.5 and an initial learning rate (d_0) of $1e-4$. Specifically, the learning rates for the U-Net and Text Encoder were decoupled and set to $1e-3$ and $2e-5$, respectively, governed by a cosine annealing with minimum learning rate scheduler. The total batch size was set to 12.

Evaluation Metrics. To comprehensively assess the quality of the generated SAR imagery, we employ four complementary metrics spanning pixel-level fidelity to perceptual realism. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are utilized to evaluate the low-level reconstruction accuracy and the

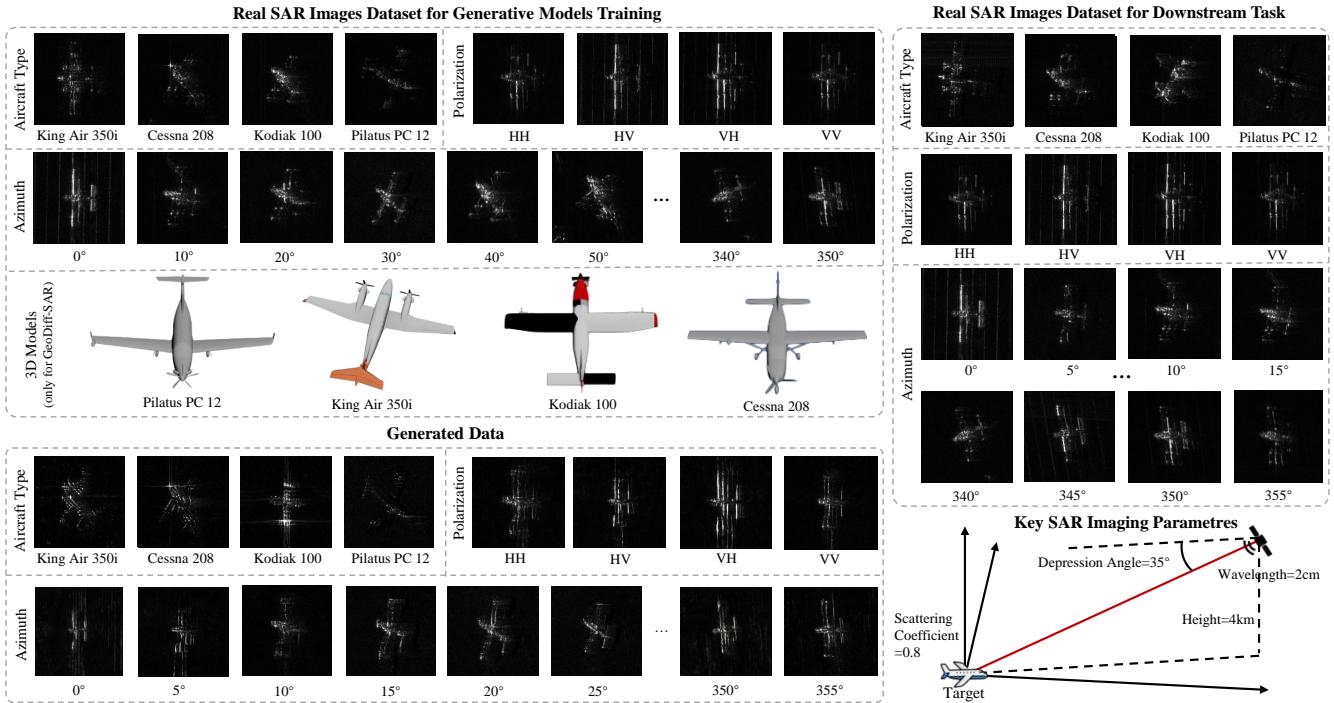


Figure 11 Data structure and composition strategy. The figure illustrates the three key components of our experimental data: (1) **Real SAR Images Dataset for Generative Models Training**: A sparse subset of real data sampled at coarse 10° azimuth intervals (e.g., $0^\circ, 10^\circ$), used to supervise the generative baseline models. (2) **Generated Data**: Synthetic samples produced by GeoDiff-SAR (guided by **3D Models**) specifically at the missing intermediate 5° azimuths (e.g., $5^\circ, 15^\circ$), designed to fill the angular gaps. (3) **Real SAR Images Dataset for Downstream Task**: A reserved split of the real SAR dataset (saved from the initial partitioning) which is subsequently merged with the generated data to construct the final augmented training set for the downstream classification task.

preservation of structural information (e.g., scattering edges), respectively. To capture high-level semantic consistency consistent with human perception, we calculate the Learned Perceptual Image Patch Similarity (LPIPS). Finally, we adopt the Fréchet Inception Distance (FID) to measure the distributional distance between the synthesized and real datasets; a lower FID score indicates that the generated samples possess higher realism and diversity, closely approximating the statistical manifold of real SAR data.

Downstream Classification Task Model. We employed the PyTorch Image Models Multi Label Classification [44] framework as our experimental testbed. Built upon the robust and widely-adopted timm ecosystem, this framework provides a unified interface to deploy deep classification architectures, including Convolutional Neural Networks and Vision Transformers. Leveraging this codebase ensures that our evaluation is performed on industry-standard backbones with reproducible training strategies, allowing for a fair comparison of data augmentation effectiveness. In this experiment, we verify whether the synthetic samples generated by GeoDiff-SAR can serve as effective training data to boost the performance of classifiers on real SAR datasets. Specifically, we design data augmentation scenarios where the generated images are mixed with limited real training samples. The classifiers are trained on these augmented sets and subsequently evaluated on a held-out set of real SAR images.

3.2 Visual Quality Evaluation

As shown in Figure 12, to comprehensively evaluate the generation performance, we selected representative baselines covering the entire spectrum of mainstream generative paradigms. These include GAN-based approaches represented by VQGAN-CLIP, Autoregressive models represented by DALL-E, and a diverse set of state-of-the-art Diffusion models ranging from SD1.5 and SDXL to PixArt and Flux. Crucially, the "SD3.5m" entry in our comparison denotes the vanilla Stable Diffusion 3.5 Medium backbone fine-tuned solely on the SAR dataset without the injection of 3D point cloud features. This setting serves as a direct ablation study, establishing a baseline to quantify the specific performance gains attributed to our proposed geometric prior module.

To further demonstrate the interpretability and precise controllability of our framework, we visualize the multi-modal generation process in Figure 13. This figure explicitly maps the one-to-one correspondence between the input **Textual Prompts** (specifying attributes such as Azimuth Angle and Polarization Mode), the intermediate **3D Point Cloud Priors** (providing geometric guidance), and the final **Generated SAR Images**. As observed,

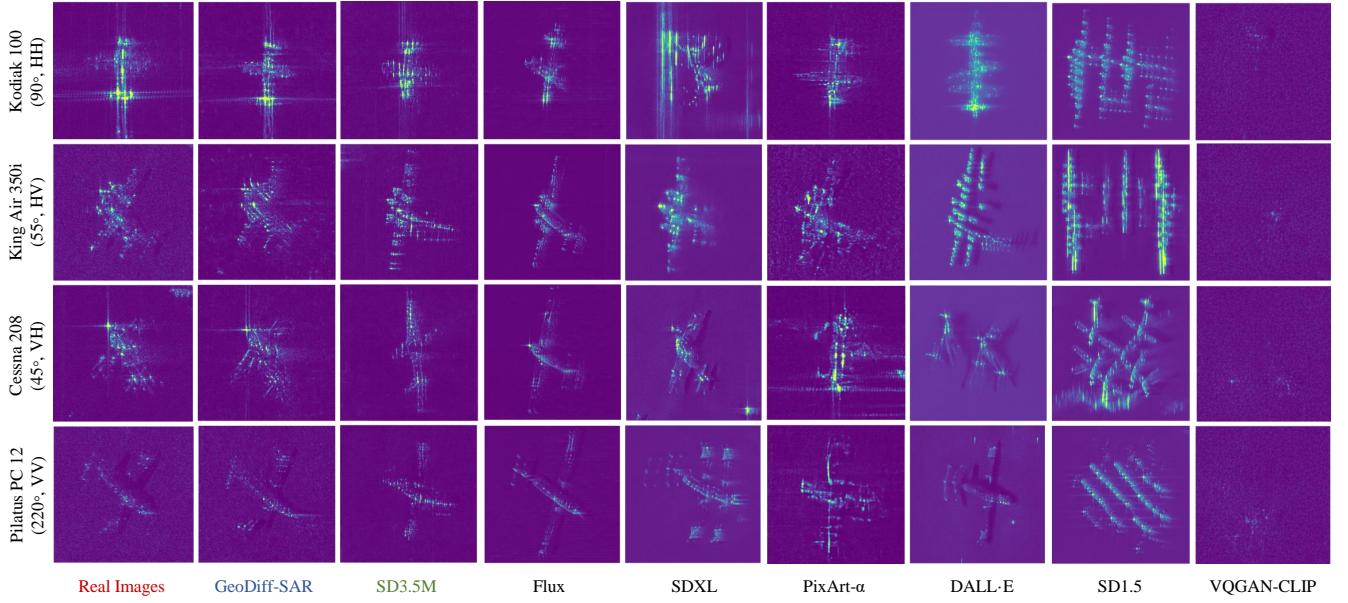


Figure 12 Visual quality comparison between GeoDiff-SAR and representative baseline models. Our method outperforms baselines in preserving geometric structures and texture details. Each column corresponds to a distinct configuration: column 1 shows Kodiak 100 (90° , HH); column 2 shows King Air 350i (55° , HV); column 3 shows Cessna 208 (45° , VH); and column 4 shows Pilatus PC 12 (220° , VV). Benefiting from the explicit constraints of physical priors, the generated images strictly adhere to the designated observation viewpoints, effectively eliminating the azimuth ambiguity and semantic confusion issues prevalent in baseline models.

Table 1 Quantitative comparison of visual quality metrics against generative models on the real SAR dataset. The best results are in bold. \uparrow means the higher the better and \downarrow means the opposite. The second best results are underlined.

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow
VQGAN-CLIP [42, 45]	8.79	0.188	0.561	35.2
SD1.5 [30]	11.15	0.378	0.472	18.7
DALL-E [23]	23.13	0.505	0.310	6.1
PixArt [46]	20.36	0.428	0.278	7.5
SDXL [47]	15.04	0.592	0.352	10.8
Flux [48]	<u>26.18</u>	0.715	<u>0.251</u>	<u>4.1</u>
SD3.5m [40]	25.23	<u>0.738</u>	0.265	5.5
GeoDiff-SAR	31.36	0.812	0.232	3.4

the generated images not only strictly adhere to the geometric structures defined by the point clouds but also accurately reflect the scattering variations induced by different polarization modes, validating the effectiveness of our physics-guided mechanism.

Evaluation Protocol. To ensure a rigorous and fair comparison of visual quality, we established a standardized "Label-to-Image" evaluation protocol. All competing models were first trained or fine-tuned on the same partition of the SAR aircraft dataset to adapt to the domain-specific scattering characteristics. During the evaluation phase, we did not rely on random generation; instead, we strictly utilized the semantic labels from the held-out test set as input conditions. For each real SAR image in the test set, we guided the trained models to generate a corresponding synthetic sample using the identical textual description. This paired generation strategy ensures that the generated distribution is compared against the real data distribution under strictly aligned semantic conditions. By calculating the discrepancy between the synthesized images and the ground-truth real images sharing the same labels, we can accurately assess each model's capability to reproduce authentic SAR textures and precise geometric structures.

Evaluation Result. Table 1 presents the quantitative comparison results on the real SAR dataset. As indicated in the table, GeoDiff-SAR achieves the FID score of 3.4, significantly outperforming the baseline SD3.5m (5.5) and the recent strong competitor Flux (4.1). This low FID indicates that our physics-guided strategy effectively aligns the generated distribution with real SAR data. Furthermore, compared to the vanilla SD3.5m, our method achieves balanced improvements in PSNR and SSIM, demonstrating that the introduction of 3D geometric priors enhances structural fidelity without compromising generative diversity.

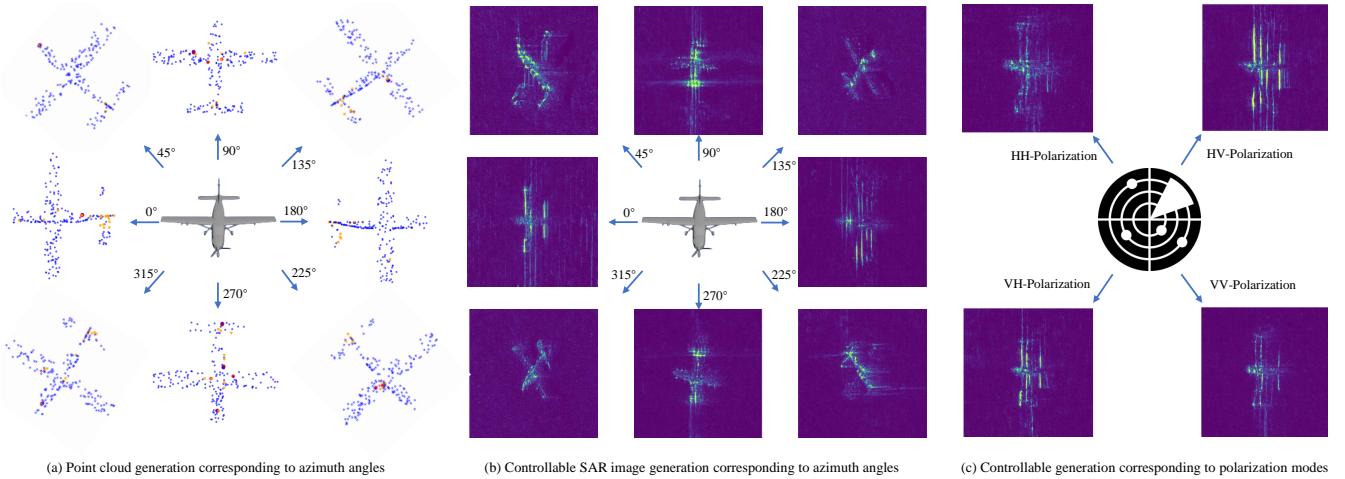


Figure 13 Visualization of the multi-modal correspondence between text, point clouds, and generated images. (a) **Point cloud generation corresponding to azimuth angles:** The intermediate 3D point cloud priors generated under varying observation geometries (e.g., 0° to 315°), serving as explicit geometric guidance. (b) **Controllable SAR image generation corresponding to azimuth angles:** The synthesized SAR images that strictly adhere to the specific azimuths defined in (a), demonstrating precise geometric alignment with the provided prompts. (c) **Controllable generation corresponding to polarization modes:** The generation results under different polarization conditions (e.g., HH, HV, VH, VV), illustrating the model’s capability to simulate distinct electromagnetic scattering characteristics.

3.3 Experiments on Downstream Classification Tasks

Experimental Pipeline and Data Partitioning. To rigorously evaluate the capability of GeoDiff-SAR in mitigating data scarcity through dense angular sampling, we designed a specific downstream classification experiment as illustrated in Figure 14. We simulated a sparse data scenario by downsampling the available real training data to coarse 10° azimuth intervals (e.g., $0^\circ, 10^\circ, 20^\circ, \dots$). This sparse subset served as the sole training supervision for the generative models.

During the data augmentation phase, GeoDiff-SAR was tasked with an "azimuth interpolation" mission: explicitly generating synthetic SAR images at the missing intermediate 5° azimuth offsets (e.g., $5^\circ, 15^\circ, 25^\circ, \dots$). These physics-guided synthetic samples were then interleaved with the sparse real data to reconstruct a dense dataset with a 5° angular resolution. Finally, this augmented mixed dataset was utilized to train the *PyTorch Image Models Multi-Label Classifier*, and its performance was evaluated on a held-out test set to verify the gain brought by this fine-grained angular completion.

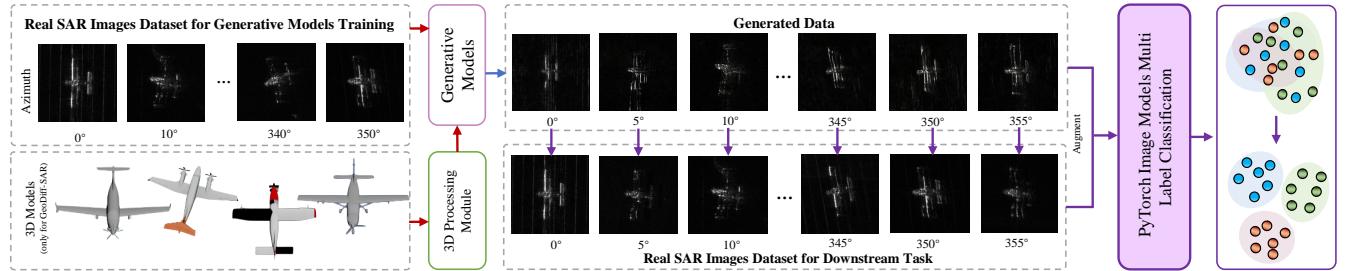


Figure 14 Experimental workflow for the downstream task focusing on azimuth completion. The pipeline consists of three phases corresponding to the labels in the figure: (1) **Generative Model Training:** The generative models are trained using the Real SAR Images Dataset for Generative Models Training, which is downsampled to sparse 10° azimuth intervals. For our method, 3D Models (processed by the 3D Processing Module) provide explicit geometric priors. (2) **Data Augment:** The models synthesize Generated Data, specifically targeting the missing intermediate azimuths (e.g., 5° offset) to fill the angular gaps. (3) **Downstream Classification:** This synthetic data is augmented with the Real SAR Images Dataset for Downstream Task to construct a Mixed Dataset. Finally, this densely sampled dataset is used to train the PyTorch Image Models Multi Label Classification to evaluate recognition performance enhancement.

In the data augmentation phase, synthetic SAR images generated by GeoDiff-SAR—guided by paired text and physical priors—were added to the scarce 20% real training set to construct an augmented dataset. This combined dataset was then utilized to train the PyTorch Image Models Multi-Label Classifier. During the training process, the best model weights were selected based on the performance on the validation set, and the final classification metrics (Precision, Recall, F1-Score) were reported on the independent 70% test set. This strict protocol ensures that the performance gains are genuinely attributed to the generalization capability learned from our synthetic data.

rather than overfitting to the training distribution.

While visual fidelity metrics provide a quantitative measure of image quality, they do not necessarily reflect the semantic correctness or the discriminatory power of the generated features. To move beyond superficial visual quality assessments and validate the effectiveness of GeoDiff-SAR at the application level, we established a downstream classification task as a critical evaluation metric. We adopted a multi-label classification framework based on PyTorch Image Models as our standardized testing platform. As shown in Table 2, GeoDiff-SAR achieves state-of-the-art performance across all three classification tasks, demonstrating superior semantic alignment compared to both baseline generative models and real data benchmarks. As shown in Table 2, GeoDiff-SAR achieves state-of-the-art performance across all three classification tasks, demonstrating superior semantic alignment compared to both baseline generative models and real data benchmarks.

To provide a more granular analysis of these results, we visualize the detailed classification performance in Figure 15. The azimuth confusion curves in Figure 15(a) reveal that while baseline methods (e.g., SD3.5m, Flux) exhibit significant angular ambiguity—manifested as jagged fluctuations deviation from the diagonal—GeoDiff-SAR maintains a smooth, highly consistent prediction trajectory that closely aligns with the ground truth. Furthermore, the comparative heatmaps in the lower panels demonstrate the model's robust discriminative power across diverse aircraft categories (e.g., Cessna 208, King Air 350i) and polarization modes (e.g., HH, HV). The distinct performance gains in these specific sub-tasks confirm that our physics-guided approach successfully preserves critical semantic features that are often lost in purely data-driven generation.

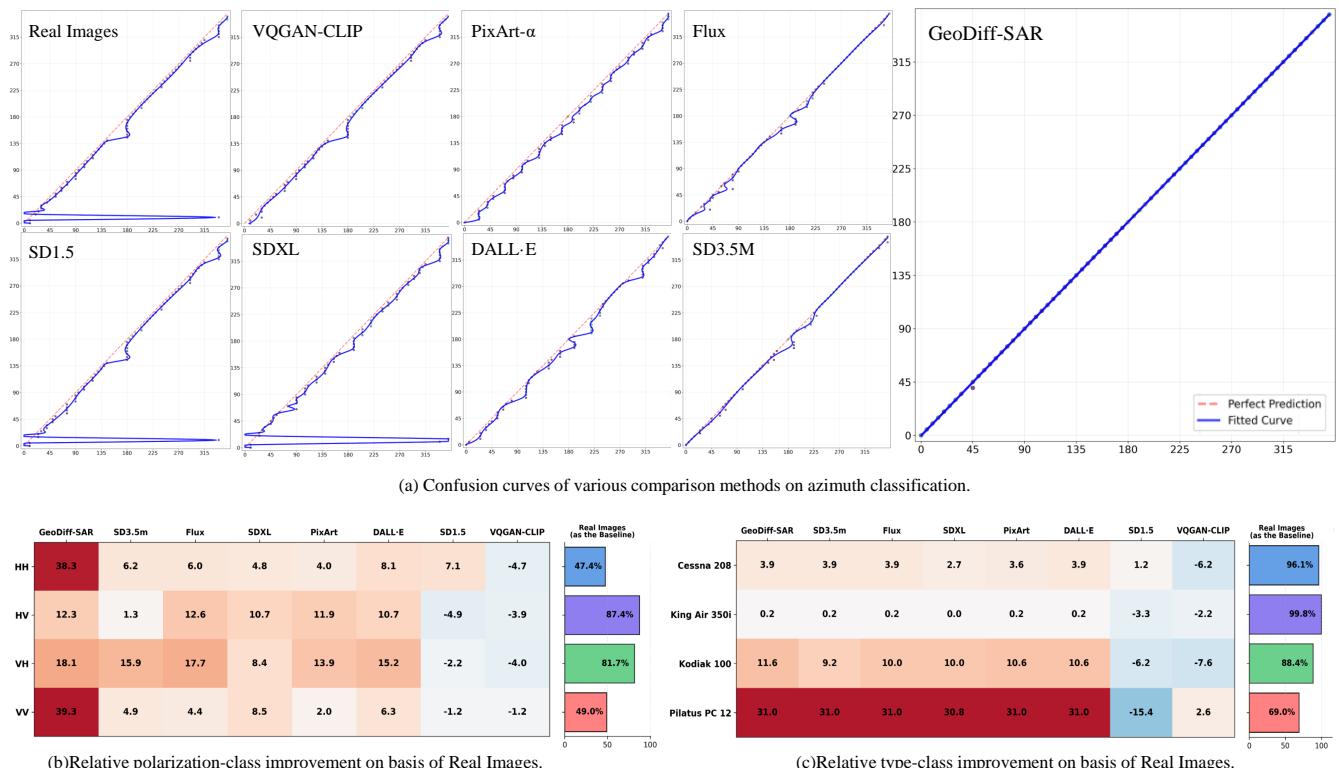


Figure 15 Visualization of downstream classification results. (a) **Confusion curves of various comparison methods on azimuth classification.** The curves depict the correlation between predicted and ground-truth azimuths; GeoDiff-SAR demonstrates a smooth trajectory that closely follows the perfect prediction diagonal, indicating superior angular consistency compared to the jagged fluctuations of baseline models. (b) **Relative polarization-class improvement on basis of Real Images.** The heatmap quantifies the classification performance gains over the real data baseline across different polarization modes (e.g., HH, HV), highlighting the model's ability to synthesize discriminative polarimetric features that enhance downstream recognition. (c) **Relative type-class improvement on basis of Real Images.** This panel illustrates the accuracy improvements for specific aircraft categories (e.g., Cessna 208, King Air 350i), confirming that the physics-guided generation effectively boosts the semantic separability of diverse targets.

Aircraft type classification task. Specifically, in the coarse-grained aircraft type classification task, while recent advanced models like Flux and SD3.5m achieve high F1-scores (over 0.99), GeoDiff-SAR reaches a perfect score of 1.000. This indicates that our method accurately captures the morphological structure of different aircraft, eliminating ambiguity in category generation.

Polarization classification task. In the polarization task, GeoDiff-SAR demonstrates exceptional capability

in simulating scattering mechanisms, achieving an F1-score of 0.933, which is significantly higher than the Real Image baseline (0.665) and Flux (0.763). This suggests that our model successfully disentangles the complex electromagnetic scattering properties associated with different polarization modes.

Azimuth classification task. The most significant advantage of our method is observed in the fine-grained azimuth estimation task, which poses a severe challenge for generic text-to-image models. As evidenced by the table, models relying solely on text prompts (e.g., SDXL, PixArt) fail to precisely control the geometric orientation, resulting in F1-scores below 0.45. Even the powerful Flux model only achieves 0.769. In contrast, GeoDiff-SAR achieves an F1-score of 0.939, outperforming the second-best method by a substantial margin. This validates that the integration of 3D physical priors provides explicit geometric guidance, ensuring the generated targets strictly follow the intended viewpoints.

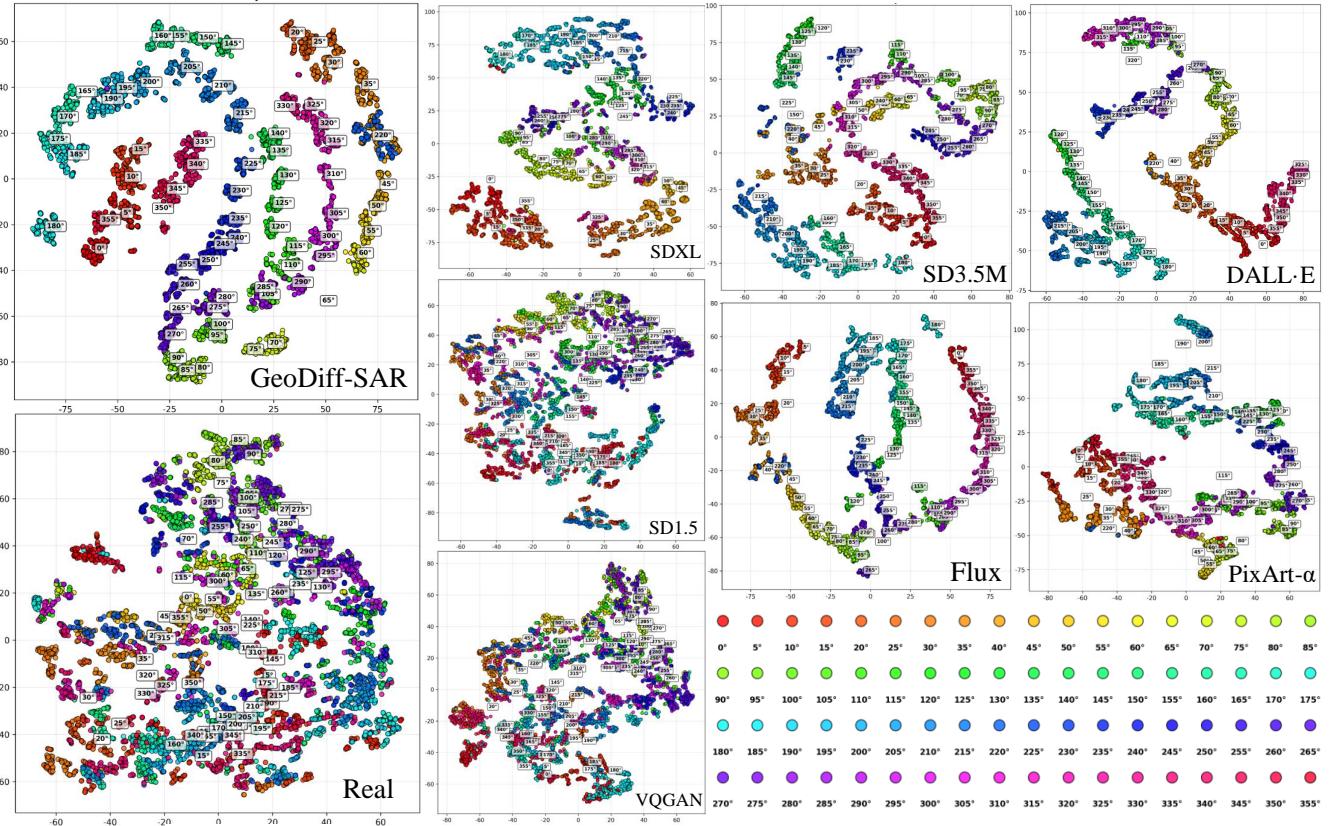


Figure 16 T-SNE visualization of feature distributions comparing GeoDiff-SAR against baseline models. The clear separation of clusters demonstrates the discriminative capability of our method in azimuth estimation.

Given the substantial performance gap observed in the Azimuth estimation task, we further employ t-SNE to visualize the feature distributions of the generated images. Figure 16 compares the feature embeddings of GeoDiff-SAR against representative baseline models. As illustrated, the feature manifold of baseline models exhibits severe entanglement across different azimuth angles, indicating that these text-driven models struggle to distinguish precise geometric orientations. In contrast, GeoDiff-SAR demonstrates clear inter-class separability and high intra-class compactness. Distinct clusters corresponding to different azimuths are formed, confirming that the injected 3D physical priors effectively guide the model to generate orientation-discriminative features, rather than merely fitting the text prompts loosely.

To conduct a fine-grained ablation study on the efficacy of the proposed 3D geometric prior, we quantitatively evaluated the azimuth estimation performance across the entire 360° spectrum. Figure 17 presents the polar plots detailing the Precision, Recall, and F1-score for three experimental settings.

As illustrated, the Real Image baseline collapses into a constricted inner cluster near the origin, indicating that classifiers trained solely on limited real data fail to generalize to the full continuous azimuth space. The vanilla SD3.5m, while significantly expanding the performance envelope, exhibits a distinct "sawtooth" pattern. This fluctuation reveals the model's sensitivity to specific viewing angles, struggling to maintain consistency at oblique orientations where text prompts alone are insufficient to resolve geometric ambiguity. In sharp contrast, GeoDiff-

Table 2 Performance comparison on downstream multi-label classification tasks (Aircraft Type, Azimuth, and Polarization) to validate data augmentation effectiveness. The best results are in bold. The second best results are underlined.

Methods	Aircraft Type			Azimuth			Polarization		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Real Image	0.883	0.903	0.887	0.078	0.225	0.109	0.664	0.668	0.665
VQGAN-CLIP [42, 45]	0.650	0.623	0.621	0.038	0.120	0.050	0.429	0.438	0.425
SD1.5 [30]	0.824	0.867	0.823	0.056	0.177	0.077	0.661	0.659	0.628
DALL-E [23]	<u>0.998</u>	<u>0.998</u>	<u>0.998</u>	0.671	0.657	0.641	0.765	0.769	0.766
PixArt [46]	0.997	0.995	0.996	0.491	0.469	0.437	0.743	0.745	0.729
SDXL [47]	0.992	0.986	0.989	0.373	0.390	0.327	0.745	0.738	0.715
Flux [48]	0.996	0.994	0.995	0.786	0.773	0.769	<u>0.766</u>	<u>0.765</u>	<u>0.763</u>
SD3.5m [40]	0.994	0.994	0.994	<u>0.797</u>	<u>0.787</u>	<u>0.782</u>	0.735	0.735	0.731
GeoDiff-SAR	1.000	1.000	1.000	0.940	0.940	0.939	0.933	0.934	0.933

SAR forms a robust, nearly perfect outer ring enveloping the other methods. It maintains high scores consistently approaching 1.0 across all metrics and angles. This visualization strongly validates our ablation hypothesis: the injection of explicit 3D physical priors effectively fills the "blind spots" of the generative backbone, ensuring omnidirectional geometric stability and physical consistency.

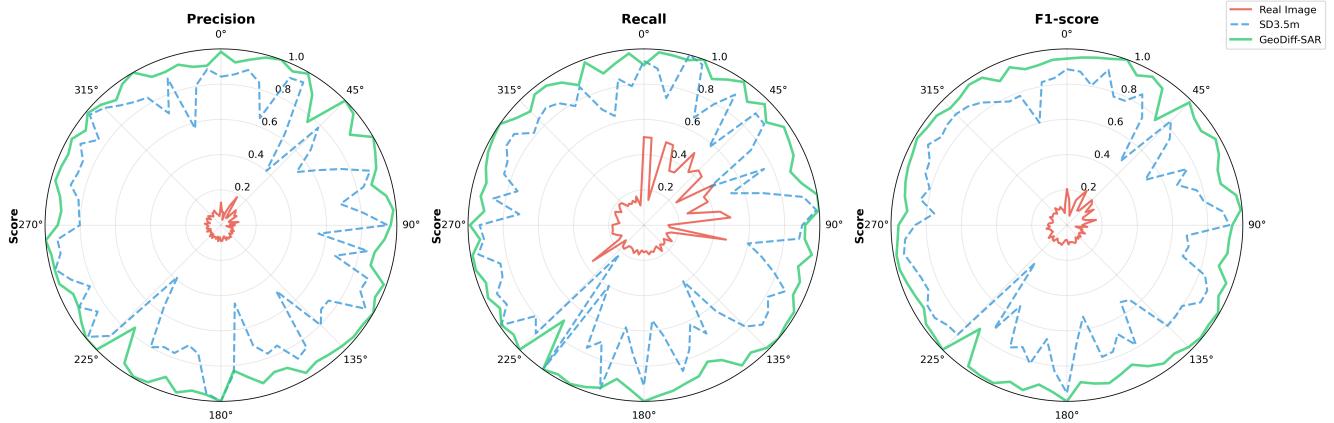


Figure 17 Fine-grained ablation study on azimuth estimation performance across the full 360° range. The polar plots compare Precision, Recall, and F1-scores among the Real Image baseline (red solid line), the vanilla SD3.5m (blue dashed line), and our GeoDiff-SAR (green solid line). Note that the Real Image baseline shows poor generalization (inner cluster), and SD3.5m exhibits angular instability (jagged fluctuations). In contrast, GeoDiff-SAR achieves a smooth, high-performance outer envelope, confirming the critical role of 3D physical priors in ensuring consistent geometric generation.

4 Conclusion

In this paper, we presented GeoDiff-SAR, a novel physics-aware generative framework designed to alleviate the severe data scarcity bottleneck and the "Sim-to-Real" gap in SAR automatic target recognition. Addressing the limitations of existing generative models—where GANs suffer from mode collapse and generic diffusion models are prone to geometric hallucinations—we proposed a paradigm shift that integrates explicit 3D geometric priors with powerful pre-trained diffusion backbones.

Our method innovatively employs a ray-casting-based point cloud simulator to capture the unique side-looking scattering characteristics of SAR targets. By injecting these physical priors through a multi-modal fusion network and utilizing LoRA for efficient fine-tuning, GeoDiff-SAR successfully adapts the Stable Diffusion 3.5 model to the SAR domain. Extensive experiments on real-world datasets demonstrate the superiority of our approach. Quantitatively, GeoDiff-SAR achieved a state-of-the-art FID score of 3.4, significantly outperforming mainstream baselines. More importantly, in downstream classification tasks, our model demonstrated exceptional performance, particularly achieving an F1-score of 0.939 in fine-grained azimuth estimation. This validates that our framework not only generates visually realistic textures but also strictly preserves physical geometric consistency, ensuring the practical utility of synthetic data for training robust ATR algorithms.

In future work, we aim to extend this physics-guided framework to more complex scenarios, such as simulating large-scale background clutter and multi-target interactions, to further broaden its applicability in diverse remote sensing missions.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 62201027 and 62271034).

References

- 1 Moreira A, Prats-Iriloa P, Younis M, et al. A tutorial on synthetic aperture radar. *IEEE Geosci Remote Sens Mag*, 2013, 1: 6–43
- 2 Oliver C, Quegan S. Understanding synthetic aperture radar images. SciTech Publishing, 2004
- 3 Morgan D A, et al. Deep learning for ATR from SAR imagery. In: Proceedings of SPIE Algorithms for Synthetic Aperture Radar Imagery, 2015
- 4 Pei J, Huang Y, Huo W, et al. SAR automatic target recognition based on multiview deep learning framework. *IEEE Trans Geosci Remote Sens*, 2018, 56: 2196–2210
- 5 Chen S, Wang H, Xu F, et al. Target classification in synthetic aperture radar images using a convolutional neural network. *IEEE Geosci Remote Sens Lett*, 2016, 13: 156–160
- 6 Ma F, et al. Fast task-specific region merging for SAR image segmentation. *IEEE Trans Geosci Remote Sens*, 2022
- 7 Ma F, et al. Fast SAR image segmentation with deep task-specific superpixel sampling and soft graph convolution. *IEEE Trans Geosci Remote Sens*, 2021
- 8 Zhang T, Xie N, Quan S, Wang W, Wei F, Yu W. Polarimetric SAR Ship Detection based on the Sub-look Decomposition Technology. *IEEE Trans Radar Syst*, 2026, 4: 35–49
- 9 Chen H, Zhao W, Zhang R, Li N, Li D. Multiple Object Tracking in Video SAR: A Benchmark and Tracking Baseline. *IEEE Geosci Remote Sens Lett*, 2025, 22: 1–5, Art no. 4010905, doi: 10.1109/LGRS.2025.3592711
- 10 Zhang F, Zhou Y, Mao S, et al. Data augmentation for SAR target recognition using generative adversarial network. *IEEE Geosci Remote Sens Lett*, 2019, 16: 1467–1471
- 11 Rostami M, Kolouri S, Eaton E, et al. Transfer learning for few-shot SAR image classification. *IEEE Trans Aerosp Electron Syst*, 2019
- 12 Gao F, Li M, Wang J, et al. General sparse adversarial attack method for SAR images based on key points. *IEEE Trans Aerosp Electron Syst*, 2025
- 13 Auer S, Hinz S, Bamler R. Ray tracing simulation techniques for understanding high-resolution SAR images. *IEEE Trans Geosci Remote Sens*, 2009, 48: 1445–1456
- 14 Franceschetti G, Migliaccio M, Riccio D, et al. SAR raw signal simulation of complex scenes. *IEEE Trans Aerosp Electron Syst*, 2003, 39: 323–331
- 15 Balz T. Real-time SAR simulation of complex scenes using programmable graphics processing units. In: Proceedings of EUSAR, 2006. 1–4
- 16 Lewis B, et al. Generative adversarial networks for SAR image realism. In: Proceedings of SPIE, 2019, 10987: 109870V
- 17 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2014
- 18 Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks. In: Proc IEEE CVPR, 2017. 1125–1134
- 19 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc IEEE ICCV, 2017. 2223–2232
- 20 Guo J, Xu B, Ju S. Synthetic aperture radar image synthesis by using generative adversarial nets. *IEEE Geosci Remote Sens Lett*, 2017, 14: 1111–1115
- 21 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. In: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017. 214–223
- 22 Van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2017
- 23 Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents (DALL-E 2). 2022. ArXiv:2204.06125
- 24 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2020, 33: 6840–6851
- 25 Song J, Meng C, Ermon S. Denoising diffusion implicit models. In: Proceedings of International Conference on Learning Representations (ICLR), 2021
- 26 Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2021
- 27 Vinker Y, Pajouheshgar E, Boissard J Y, et al. Clipasso: Semantically-aware object sketching. *ACM Trans Graph*, 2022, 41: 1–11
- 28 Perera M, Noshadi A, Fookees C. SAR despeckling using a denoising diffusion probabilistic model. *IEEE Geosci Remote Sens Lett*, 2023, 20: 1–5
- 29 Batzolos G, Stanczuk J, Schonlieb C B, et al. Conditional image generation with score-based diffusion models. 2021. ArXiv:2111.13606
- 30 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models (Stable Diffusion). In: Proc IEEE CVPR, 2022. 10684–10695
- 31 Tang X, Zhang X, Ma F, et al. Remote Sensing Image Generation With Diffusion Models. *IEEE Trans Geosci Remote Sens*, 2024
- 32 Shen B, Liu T, Gao G, Chen H, Yang J. A Low-Cost Polarimetric Radar System Based on Mechanical Rotation and Its Signal Processing. *IEEE Trans Aerosp Electron Syst*, 2025, 61: 4744–4765, doi: 10.1109/TAES.2024.3507776
- 33 Zhong F, Gao F, Liu T, et al. Scattering Characteristics Guided Network for ISAR Space Target Component Segmentation. *IEEE Geosci Remote Sens Lett*, 2025, 22: 1–5, Art no. 4009505, doi: 10.1109/LGRS.2025.3576662
- 34 Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proc IEEE ICCV, 2023. 3836–3847
- 35 Mou C, Wang X, Xie L, et al. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proc IEEE ICCV, 2023. 4296–4305
- 36 Malmgren-Hansen D, Kusk A, Dall J, et al. Improving SAR automatic target recognition using simulated data. In: Proceedings of IEEE Radar Conference, 2017. 1150–1153
- 37 Schreiber E, Gupta I J. Scattering center analysis of man-made targets. *IEEE Trans Antennas Propag*, 2005, 53: 178–184
- 38 Perez E, Strub F, De Vries H, et al. FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018
- 39 Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models. In: Proceedings of International Conference on Learning Representations (ICLR), 2022
- 40 Esser P, Kulal S, Lang A, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis (SD3). In: Proceedings of International Conference on Machine Learning (ICML), 2024
- 41 Wu X, Jiang L, Wang P S, et al. Point Transformer V3: Simpler, Faster, Stronger. In: Proc IEEE CVPR, 2024. 4840–4851
- 42 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision (CLIP). In: Proceedings of International Conference on Machine Learning (ICML), 2021. 8748–8763
- 43 Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer (T5). *J Mach Learn Res*, 2020, 21: 1–67

- 44 Yang R. PyTorch Image Models Multi-Label Classification. 2020. GitHub repository. <https://github.com/yang-ruixin/PyTorch-Image-Models-Multi-Label-Classification>
- 45 Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis (VQGAN). In: Proc IEEE CVPR, 2021. 12873–12883
- 46 Chen J, Yu J, Ge C, et al. PixArt- α : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In: Proceedings of International Conference on Learning Representations (ICLR), 2024
- 47 Podell D, English Z, Lacey K, et al. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In: Proceedings of International Conference on Learning Representations (ICLR), 2024
- 48 Black Forest Labs. FLUX.1. <https://blackforestlabs.ai>, 2024