SAR-DiSCO: Differentiable Scattering Coefficient Optimization for 3D Adversarial Attacks on SAR Target Recognition

Anonymous authors
Paper under double-blind review

Abstract

Synthetic aperture radar (SAR) automatic target recognition (ATR) systems are highly vulnerable to adversarial attacks, which can cause classification errors and threaten system reliability. Many existing attacks on SAR ATR inject perturbations directly into the 2D image domain without explicitly modeling target-surface scattering or the SAR image formation process. Such perturbations do not necessarily correspond to consistent target-level modifications across 3D viewing conditions. To address this gap, we propose SAR-DiSCO, a simulation-based adversarial attack framework that optimizes target-surface scattering coefficients through differentiable SAR rendering. The proposed approach employs a custom differentiable SAR renderer to perform end-to-end gradient optimization of vertex-wise scattering coefficients, producing adversarial perturbations while preserving target geometry. Specifically, we (1) build an end-to-end differentiable SAR imaging and recognition framework that links 3D surface coefficients to classification predictions, and (2) design a loss function that balances misclassification and scattering-coefficient deviation. We then (3) iteratively update the coefficients via backpropagation for view-specific attacks or shared multi-view optimization, without requiring a separately fitted surrogate mapping. Experiments on a simulated five-class vehicle dataset demonstrate an average view-specific fooling rate of 88.96% across five CNN- and Transformer-based classifiers. Further experiments demonstrate strong cross-architecture transferability and partial cross-view effectiveness when a single coefficient field is shared across viewing conditions.

1 Introduction

Deep neural networks (DNNs) have substantially advanced synthetic aperture radar (SAR) automatic target recognition (ATR) by learning discriminative representations directly from SAR imagery (Zhu et al., 2021; Kechagias-Stamatis & Aouf, 2021; Huang et al., 2019). Despite their strong recognition performance, DNN-based SAR ATR models can be vulnerable to adversarial examples (Szegedy et al., 2013). Small, deliberately optimized changes to the input representation may alter a model's prediction without changing the underlying target class.

Existing adversarial attacks on SAR ATR can be broadly divided into image-domain and signal-domain approaches. Image-domain attacks perturb SAR pixels or image representations and include global (Huang et al., 2020a;b; Wan et al., 2024), local (Peng et al., 2021; Zhang et al., 2022; Du et al., 2022), and sparse perturbations (Wan et al., 2025; Gao et al., 2025). Although these methods provide direct control over the digital input, they generally do not model how target-level changes propagate through SAR imaging (Chen et al., 2021). Signal-domain attacks instead manipulate radar echoes (Xia et al., 2022; Zhang et al., 2025; Liu et al., 2025) or formulate perturbations using attribute scattering-center models (Peng et al., 2022b; Zhang et al., 2024; Zeng et al., 2025). By operating closer to the SAR acquisition process, these attacks incorporate more imaging-domain structure than purely pixel-level methods. However, existing methods generally do not directly optimize spatially varying surface scattering coefficients of a three-dimensional target through an end-to-end SAR imaging and recognition pipeline. Consequently, it remains unclear whether a single target-level parameterization can retain its adversarial effectiveness across different three-dimensional

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

viewing conditions. Figure 1 illustrates the different perturbation locations considered by these attack paradigms.

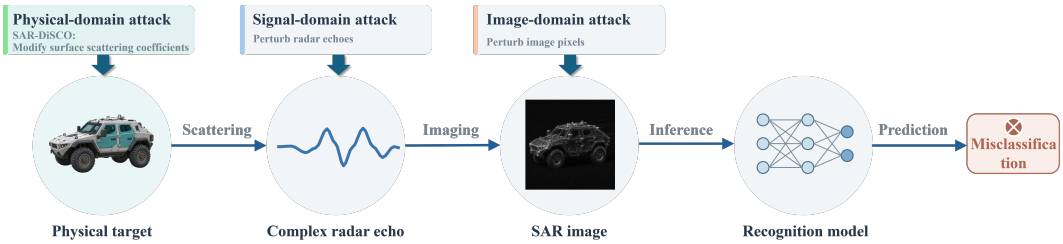

Figure 1: Attack locations in the SAR recognition pipeline. SAR-DiSCO modifies target-surface scattering coefficients before imaging; signal- and image-domain attacks perturb radar echoes and formed SAR images, respectively.

These limitations motivate moving adversarial optimization upstream, from formed SAR images or radar-echo representations to the physical properties of the target or its surroundings. This perspective is embodied by physical-domain adversarial attacks, which modify such properties so that the resulting adversarial effect is propagated through the sensing process. In optical imaging, adversarial patches, textures, and camouflage patterns have been used to induce errors in recognition and detection systems under physical transformations (Wen et al., 2023; Cui et al., 2024; Labarbarie et al., 2024). Differentiable rendering further enables gradients to be propagated from image-space objectives to scene or material parameters (Li et al., 2018; Nimier-David et al., 2020; Yong et al., 2025), and has supported several optical physical-attack frameworks (Zhang et al., 2018; Wang et al., 2021; Duan et al., 2021; Zheng et al., 2024; Li et al., 2025). However, optical rendering models do not capture the coherent electromagnetic scattering and imaging mechanisms that govern SAR observations. An end-to-end differentiable connection between target-surface scattering parameters, simulated SAR images, and recognition loss therefore remains underexplored.

To address this gap, we introduce SAR-DiSCO (SAR Differentiable Scattering Coefficient Optimization), a simulation-based physical-domain attack framework for SAR target recognition. SAR-DiSCO couples a custom differentiable SAR renderer with a recognition model and optimizes spatially varying surface scattering coefficients parameterized at mesh vertices while keeping the target geometry fixed. At each optimization step, a dynamically guided untargeted objective selects the highest-scoring incorrect class as the current rival and promotes misclassification toward that class. A sigmoid parameterization bounds the coefficients to the admissible range, while a squared ℓ_2 penalty constrains their deviation from the clean coefficients. For view-specific attacks, the coefficient tensor is optimized independently for each observation. In the multi-view setting, observations from different sensor and azimuth configurations sequentially update the same coefficient tensor without reinitialization, producing a target-specific surface parameterization shared across views. On a simulated five-class vehicle dataset, SAR-DiSCO achieves an average fooling rate of 88.96% across five CNN- and Transformer-based classifiers in the primary evaluation and demonstrates cross-architecture transferability.

Main contributions.

- **Target-level attack formulation.** We introduce SAR-DiSCO, which optimizes vertex-wise surface scattering coefficients on a fixed target mesh through simulated SAR image formation.
- **Controlled coefficient optimization.** A dynamically guided untargeted objective, sigmoid-bounded coefficients, and a squared ℓ_2 penalty support both view-specific optimization and a coefficient field shared across views.
- **Evaluation across models and views.** Experiments on five recognition architectures measure view-specific effectiveness, cross-model transfer, and shared multi-view performance, separating vulnerability at one view from persistence across viewing conditions.

2 Related Work

2.1 SAR Adversarial Attacks

Existing studies on SAR adversarial attacks broadly encompass image-domain attacks and approaches based on radar signals or scattering models. Image-domain attacks construct perturbations at the pixel or image-representation level to mislead recognition models. General-purpose methods developed for image classification, such as FGSM (Goodfellow et al., 2014), BIM (Kurakin et al., 2016), DeepFool (Moosavi-Dezfooli et al., 2016), and C&W (Carlini & Wagner, 2017), provide foundational approaches to pixel-level adversarial optimization. SAR-specific studies further incorporate target structure and image characteristics, including target-region perturbations in TRPG (Meng et al., 2022), speckle variation in SVA (Peng et al., 2022a), and partial target occlusion in TPO (Ma et al., 2023). RaLP (Duan et al., 2024) extends region-adaptive local perturbation design to SAR object detection. These studies explore attack effectiveness, transferability, and perturbation localization, without necessarily establishing physical realizability.

Signal-domain and scattering-based attacks incorporate the radar acquisition process or electromagnetic scattering mechanisms into perturbation design. Representative approaches include phase-modulated jamming in SAR-GPA (Liu et al., 2021) and the conversion of image-domain universal perturbations into interference signals in TUAA (Liu et al., 2025). Other methods introduce explicit scattering representations: SMGAA (Peng et al., 2022b) combines a scattering-center model with gradient-based optimization, the full-aspect attack of Wei et al. (2026) optimizes scattering-center amplitudes across azimuth angles, and SAR-PAA (Ma et al., 2024) considers the deployment of additional scatterers near the target. SAR-PATT (Luo et al., 2024) modifies the surface parameters of 3D targets through a two-stage procedure: it first optimizes intermediate scattering data and then maps the resulting perturbations to target-surface parameters. These approaches connect adversarial perturbations more closely to SAR image formation and target-level modifications.

Although these studies have advanced SAR adversarial research from different perspectives, the perturbation representation remains an important distinction. Image-domain approaches optimize digital representations, which do not by themselves ensure correspondence to a consistent physical modification across viewing conditions. Signal-domain and scattering-based approaches provide a more explicit connection to radar imaging, but differ in their parameterization and optimization mechanisms. For example, the full-aspect attack uses particle swarm optimization, while SAR-PAA employs differential evolution; SMGAA, in contrast, already supports gradient-based optimization of scattering-model parameters. SAR-PATT maps optimized intermediate scattering data to target-surface parameters through a separate mapping stage. SAR-DiSCO instead directly optimizes vertex-wise surface scattering coefficients on a fixed three-dimensional target mesh, propagating classification gradients through a differentiable SAR renderer. Its distinction therefore lies in mesh-based surface parameterization and direct optimization through simulated image formation, rather than in introducing scattering-aware or gradient-based attacks in general.

2.2 Differentiable Rendering

Differentiable rendering enables the optimization of scene parameters through rendered images. Rasterization-based methods are widely used for efficient image formation and geometric optimization. Representative works such as SoftRas (Liu et al., 2019) and DIB-R (Chen et al., 2019) introduce soft rasterization or gradient approximations to support differentiable geometric processing. Mesh Splatting (Zhang et al., 2026) further combines a softened mesh representation with splatting-based rendering for end-to-end multiview surface reconstruction. Differentiable light-transport methods, including Redner (Li et al., 2018) and Radiative Backpropagation (Nimier-David et al., 2020), enable gradient computation through physically based rendering for inverse-rendering applications. Complementing these methods, Neural Light Transport (Zhang et al., 2021) combines physically based diffuse rendering with learned transport effects for relighting and view synthesis. These approaches offer different trade-offs in computational efficiency, scene representation, and transport modeling. In this work, differentiable rendering connects target-surface scattering coefficients to a classification objective through simulated SAR image formation, enabling direct optimization of adversarial surface perturbations.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

SAR-DiSCO seeks a bounded change to the surface scattering representation of a three-dimensional target such that the corresponding simulated SAR image is misclassified, while the target geometry remains fixed. The framework couples a differentiable SAR renderer to a pretrained recognition model, allowing the classification loss to be differentiated with respect to the surface parameters. As shown in Figure 2, it contains two components. The differentiable rendering module maps the target geometry, its surface coefficients, and the imaging configuration to a SAR image. The adversarial optimization module then updates the surface coefficients using gradients propagated through both the classifier and the renderer.

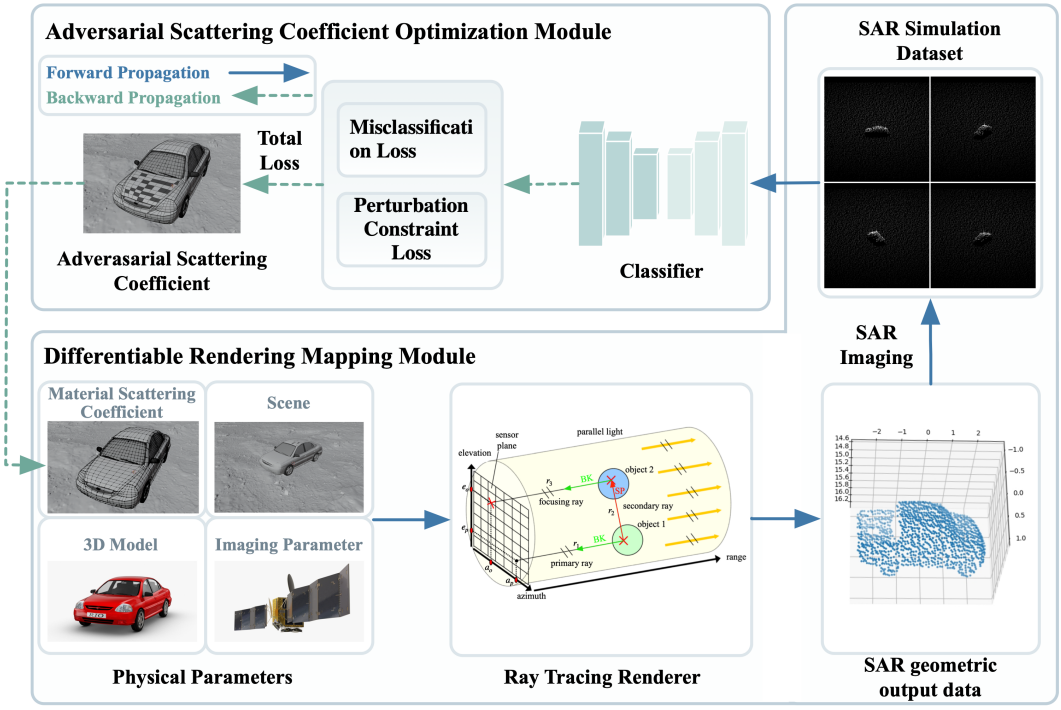

Figure 2: Overview of SAR-DiSCO. A fixed target mesh and imaging configuration are passed through the differentiable SAR renderer and the recognition model. The resulting classification gradient updates only the vertex-wise surface scattering coefficients; the target geometry and classifier parameters remain fixed.

Let $M = (\mathbf{V}, \mathbf{F})$ denote a triangular target mesh with q vertices and fixed geometry. We represent its normalized surface scattering coefficients by $\mathbf{T} \in [0,1]^{q\times 3}$, following the three-channel vertex representation used by the renderer. At a ray–triangle intersection, the coefficient is obtained by barycentric interpolation of the coefficients at the three incident vertices. This produces a continuous coefficient field over the mesh without modifying $\mathbf{V}$ or $\mathbf{F}$. Given an imaging configuration $\mathbf{\Theta} = (\boldsymbol{\theta}_c, \boldsymbol{\theta}_s)$, the differentiable renderer $\mathcal{R}$ produces

$$\mathbf{I} = \mathcal{R}(M, \mathbf{T}; \mathbf{\Theta}), \qquad \mathbf{I} \in [0,1]^{H\times W}, \tag{1}$$

where H and W are the range and azimuth image dimensions. The vector $\boldsymbol{\theta}_c$ specifies the observation geometry, including sensor position, aspect angle, and depression angle, whereas $\boldsymbol{\theta}_s$ contains the sampling, resolution, and dynamic-range parameters. The coefficients in $\mathbf{T}$ parameterize the diffuse return used by the adopted simulator; they should not be interpreted as a complete electromagnetic material model.

Definition 1 (Differentiable SAR recognition map)**.** For fixed M and $\mathbf{\Theta}$, the composite map from surface coefficients to classifier logits is $\mathbf{\Phi}_{\mathbf{\Theta}}(\mathbf{T}) = \mathcal{C}(\mathcal{R}(M, \mathbf{T}; \mathbf{\Theta})) \in \mathbb{R}^n$. Where differentiable, its Jacobian with respect to $\mathrm{vec}(\mathbf{T})$ has shape $n \times 3q$.

The classifier $\mathcal{C}$ is fixed during an attack. SAR-DiSCO therefore updates only $\mathbf{T}$, using gradients of a guided untargeted loss through the composite map in Definition 1. Attack success is defined by $\arg\max_i [\mathbf{\Phi}_{\mathbf{\Theta}}(\mathbf{T})]_i \neq y_{\mathrm{gt}}$, where y_{gt} is the ground-truth class. This formulation changes the simulated surface response while preserving the target geometry.

3.2 Differentiable SAR image formation

For each valid ray–surface intersection, the renderer evaluates a coefficient- and incidence-dependent return and associates it with range–azimuth coordinates. The image-formation stage adds the two-way propagation phase, coherently accumulates complex responses on the range–azimuth grid, approximates point-spread-function broadening, and applies logarithmic dynamic-range compression. With the mesh geometry fixed, ray intersections and grid assignments for each prescribed view remain unchanged as the surface coefficients are optimized. Gradients are therefore required only through the coefficient-dependent rendering operations. This gives a piecewise-differentiable path from the image loss to the vertex coefficients, despite the discrete geometric assignments. Appendix A specifies the forward equations and implementation details.

3.3 Guided and constrained coefficient optimization

For an image $\mathbf{I}$, the fixed classifier produces logits $\mathbf{z} = \mathcal{C}(\mathbf{I}) \in \mathbb{R}^n$ and class probabilities $\mathbf{p}$,

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)}, \qquad i \in \{1, \ldots, n\}, \tag{2}$$

where n is the number of classes. We guide each update towards the currently most competitive incorrect class while retaining an untargeted success criterion.

Definition 2 (Guided rival class)**.** Given logits $\mathbf{z} = \mathbf{\Phi}_{\mathbf{\Theta}}(\mathbf{T})$ and ground-truth label y_{gt}, the iteration-wise guided rival is $c_{\mathrm{g}}(\mathbf{T}; \mathbf{\Theta}) = \arg\max_{j \in \{1,\ldots,n\} \setminus \{y_{\mathrm{gt}}\}} z_j$. The rival class guides the current update, but attack success remains untargeted and is defined by $\arg\max_i z_i \neq y_{\mathrm{gt}}$.

Ties are resolved by a fixed class-index ordering. At each iteration, the ground-truth logit is masked when selecting c_{g}. The class index is then treated as fixed for that update, and the classification term is

$$\mathcal{L}_{\mathrm{cls}}(\mathbf{T}; \mathbf{\Theta}) = -\log p_{c_{\mathrm{g}}(\mathbf{T};\mathbf{\Theta})}(\mathbf{T}; \mathbf{\Theta}). \tag{3}$$

This term favors the current rival class; minimizing the regularized objective does not guarantee an incorrect prediction.

We constrain the coefficient field without applying hard clipping. An unconstrained latent tensor $\boldsymbol{\omega} \in \mathbb{R}^{q \times 3}$ is mapped to the admissible range using an element-wise sigmoid,

$$\mathbf{T}_{\mathrm{adv}} = \sigma(\boldsymbol{\omega}), \qquad [\sigma(\boldsymbol{\omega})]_{v,h} = \frac{1}{1 + \exp(-\omega_{v,h})}. \tag{4}$$

Here $v \in \{1, \ldots, q\}$ indexes vertices and $h \in \{1, 2, 3\}$ indexes channels. Finite latent values give $\mathbf{T}_{\mathrm{adv}} \in (0, 1)^{q \times 3}$. We initialize $\boldsymbol{\omega}$ by the element-wise logit of $\mathbf{T}_{\mathrm{orig}} = 0.5$, so $\boldsymbol{\omega} = \mathbf{0}$. A squared Frobenius penalty limits the aggregate coefficient change, equivalent to $\sum_{v=1}^{q} \sum_{h=1}^{3} (T_{v,h} - T_{v,h}^{\mathrm{orig}})^2$. For a single imaging configuration, the complete objective is

$$\mathcal{J}(\boldsymbol{\omega}; \mathbf{\Theta}) = \mathcal{L}_{\mathrm{cls}}(\sigma(\boldsymbol{\omega}); \mathbf{\Theta}) + c \, \|\sigma(\boldsymbol{\omega}) - \mathbf{T}_{\mathrm{orig}}\|_F^2, \tag{5}$$

We obtain an approximate solution $\widehat{\boldsymbol{\omega}}$ by minimizing $\mathcal{J}$ and set $\mathbf{T}_{\mathrm{adv}} = \sigma(\widehat{\boldsymbol{\omega}})$. A finite Adam run does not guarantee a minimizer. The weight $c \geq 0$ controls the trade-off between the guided classification objective and deviation from the clean coefficients. We minimize Eq. (5) with Adam at learning rate α, while keeping the classifier, geometry, and imaging parameters fixed. PyTorch automatic differentiation (Paszke et al., 2017) propagates the loss gradient through the classifier and the coefficient-dependent operations of the renderer to $\boldsymbol{\omega}$.

For view-consistent optimization, let $\mathcal{V} = \{\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_K\}$ contain the selected azimuth and observation configurations. We share one latent tensor across all views and optimize

$$\min_{\boldsymbol{\omega}} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\mathrm{cls}}(\sigma(\boldsymbol{\omega}); \mathbf{\Theta}_k) + c \, \|\sigma(\boldsymbol{\omega}) - \mathbf{T}_{\mathrm{orig}}\|_F^2. \tag{6}$$

The implementation processes the views sequentially and applies an Adam update after each view while retaining the same $\boldsymbol{\omega}$ and optimizer state. This is an incremental optimization of the view-averaged objective, not a full-batch gradient step. Each view updates the same target-level coefficient field. Setting $K = 1$ recovers the single-view objective. Algorithm 1 summarizes the optimization, and Appendix C reports the hyperparameters and imaging configurations.

4 EXPERIMENTS

4.1 SETUP

We construct a simulated SAR dataset from five vehicle meshes (bus, car, jeep, truck, and van), rendering 1,800 images per class over elevation angles of 15°, 17°, 25°, 35°, and 45° and 360 azimuths at 1° intervals over $[0^\circ, 360^\circ)$. Each mesh is scaled to realistic dimensions, centered on the horizontal plane, and rendered above a common ground background. The renderer traces a 150×150 ray grid and produces 224×224 intensity images. The resulting 9,000 images are divided into training and test subsets with an 8:2 split; mesh statistics and representative renders are provided in Appendix C.

We evaluate ResNet-50 He et al. (2016), VGG-16 Simonyan & Zisserman (2014), ViT-B/16 Dosovitskiy et al. (2021), Swin-V2-S Liu et al. (2022), and EfficientNetV2-S Tan & Le (2021) initialized from ImageNet-pretrained weights. Each classifier is fine-tuned for at most 50 epochs using SGD with learning rate 0.001, momentum 0.9, and batch size 32. For SAR-DiSCO, all clean vertex coefficients are initialized to 0.5. The point-spread approximation uses diffusion range $D = 100$ and decay factor 0.3, followed by dynamic-range clipping to $[-40, 30]$ dB. In the view-specific experiments, the latent coefficients are optimized for 50 steps with Adam, learning rate 0.05, and regularization weight $c = 0.25$. In the multi-view experiment, one shared coefficient tensor is sequentially updated over the five sensor configurations, each with azimuths sampled over $[0^\circ, 360^\circ)$ at 1° intervals, for 30 epochs using the same optimizer settings. Experiments run under PyTorch 2.4.1 on an NVIDIA RTX 4070 Ti; Appendix C contains the complete configuration.

Our primary evaluation assumes white-box access to the classifier and renderer while keeping the mesh geometry fixed and treating each view's imaging configuration as prescribed; only vertex-wise scattering coefficients are optimized. In the view-specific setting, coefficients are optimized independently for each observation. In the multi-view setting, all views update one shared coefficient tensor. Black-box transfer uses coefficients generated on a source classifier without accessing the target classifier's parameters or gradients. These settings separate optimization effectiveness, view consistency, and cross-architecture transferability.

We report fooling rate (FR), the proportion of attacked samples classified incorrectly:

$$\mathrm{FR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ \underset{j \in \{1,\ldots,n\}}{\arg\max} \left[\mathcal{C}(\mathbf{I}_i^{\mathrm{adv}}) \right]_j \neq y_i^{\mathrm{gt}} \right\}. \tag{7}$$

Here N denotes the number of evaluated samples and $\mathbf{1}\{\cdot\}$ is the indicator function. We use the complete test set, with $N = 360$ per category and $N = 1{,}800$ per source–target transfer pair. Tables report $100\,\mathrm{FR}$ in percent, with the Original rows representing clean classification error rates. Means are computed from unrounded rates and rounded to two decimal places for reporting.

4.2 ATTACK PERFORMANCE

We evaluated all 360 test images per category, giving 1,800 test images across the five categories. In this view-specific evaluation, each attack used the sample's original imaging configuration and ran for the prescribed iteration budget.

We compared SAR-DiSCO with BIM (Kurakin et al., 2016), adversarial patches (Patch) (Brown et al., 2017), scattering-center perturbations (SMGAA) (Peng et al., 2022b), and randomized surface coefficients (Random). BIM, Patch, and SMGAA used the same iteration count as SAR-DiSCO; SMGAA used three adversarial scatterers. Table 1 reports the results. Equal iteration counts do not imply matched perturbation budgets across these different parameterizations.

Table 1: **View-specific attack performance.** Fooling rate (%, ↑) across five target categories and five recognition architectures. Shaded rows identify SAR-DiSCO; bold values mark the best result within each model block.

Model	Attack	Target category					Mean
		Bus	Car	Jeep	Truck	Van	
ResNet-50	Original	1.11	1.39	1.67	1.11	1.39	1.33
	Random	8.33	6.67	3.61	9.17	12.78	8.11
	BIM	**66.39**	74.17	58.06	82.50	70.00	70.22
	Patch	32.50	38.61	54.17	47.22	29.72	40.44
	SMGAA	37.22	46.94	52.50	54.17	48.33	47.83
	SAR-DiSCO	42.78	**83.33**	**60.00**	**100.00**	**100.00**	**77.22**
VGG-16	Original	0.83	1.94	2.22	3.33	2.78	2.22
	Random	5.56	12.50	8.33	6.39	10.00	8.56
	BIM	72.22	65.56	71.39	87.78	68.33	73.06
	Patch	47.22	22.78	39.44	58.33	25.83	38.72
	SMGAA	52.22	17.22	29.44	73.06	65.83	47.56
	SAR-DiSCO	**95.83**	**78.89**	**75.83**	**100.00**	**99.17**	**89.94**
ViT-B/16	Original	0.83	1.39	1.67	2.22	2.78	1.78
	Random	14.72	3.89	7.22	19.44	18.06	12.67
	BIM	74.72	79.72	68.33	93.06	59.72	75.11
	Patch	48.33	63.61	59.17	70.00	66.39	61.50
	SMGAA	35.28	42.78	57.78	63.06	79.44	55.67
	SAR-DiSCO	**95.00**	**92.22**	**83.89**	**100.00**	**100.00**	**94.22**
Swin-V2-S	Original	0.56	0.83	1.67	1.67	1.11	1.17
	Random	7.22	2.50	11.11	15.83	18.89	11.11
	BIM	78.33	61.67	74.72	89.17	69.44	74.67
	Patch	45.56	68.33	53.89	59.44	67.50	58.94
	SMGAA	53.33	45.83	58.06	79.72	71.94	61.78
	SAR-DiSCO	**90.56**	**86.39**	**93.06**	**96.39**	**100.00**	**93.28**
EfficientNetV2-S	Original	0.83	0.56	1.39	1.39	1.11	1.06
	Random	4.44	7.22	1.94	9.44	13.33	7.28
	BIM	63.89	77.78	**87.50**	95.28	55.83	76.06
	Patch	32.50	48.89	36.11	63.61	58.33	47.89
	SMGAA	15.83	29.72	42.22	48.06	63.61	39.89
	SAR-DiSCO	**90.83**	**98.61**	61.11	**100.00**	**100.00**	**90.11**

Under bounded coefficient constraints within the simulator, SAR-DiSCO deceives all five recognition architectures evaluated. Across the five target categories and five recognition models, the unperturbed condition (Original) and randomized coefficient perturbation (Random) yield mean fooling rates of 1.51% and 9.54%, respectively. BIM, Patch, and SMGAA achieve mean fooling rates of 73.82%, 49.50%, and 50.54%, respectively, whereas SAR-DiSCO reaches 88.96%. These results show that optimizing vertex-wise surface scattering coefficients can substantially alter the simulated SAR responses and mislead the evaluated recognition models.

We additionally compare SAR-DiSCO with SAR-PATT (Luo et al., 2024), an existing attack that modifies the surface parameters of 3D targets to induce misclassification in simulated SAR images. Appendix D presents this comparison and describes the evaluation settings used to account for differences between the imaging pipelines.

4.3 Black-box transferability

We generated attacks on each source classifier and evaluated them on the other four classifiers without access to their parameters or gradients. FR is evaluated on all 1,800 test samples for each source–target pair. Figure 3 reports all source–target pairs, with sources in rows and targets in columns.

Table 2: **Scope of the surface perturbation.** Fooling rate (%, ↑) on ResNet-50 under three optimization settings.

Setting	Bus	Car	Jeep	Truck	Van	Mean
View-specific	42.78	83.33	60.00	100.00	100.00	**77.22**
Multi-view	54.17	56.39	48.61	61.39	64.17	56.94
External	26.39	36.39	31.67	65.28	75.83	47.11

Transfer FR ranged from 36.44% for ResNet-50 to Swin-V2-S to 67.39% for VGG-16 to ViT-B/16. VGG-16 had the highest mean outgoing transfer FR, while ResNet-50 had the lowest mean incoming transfer FR among the evaluated models.

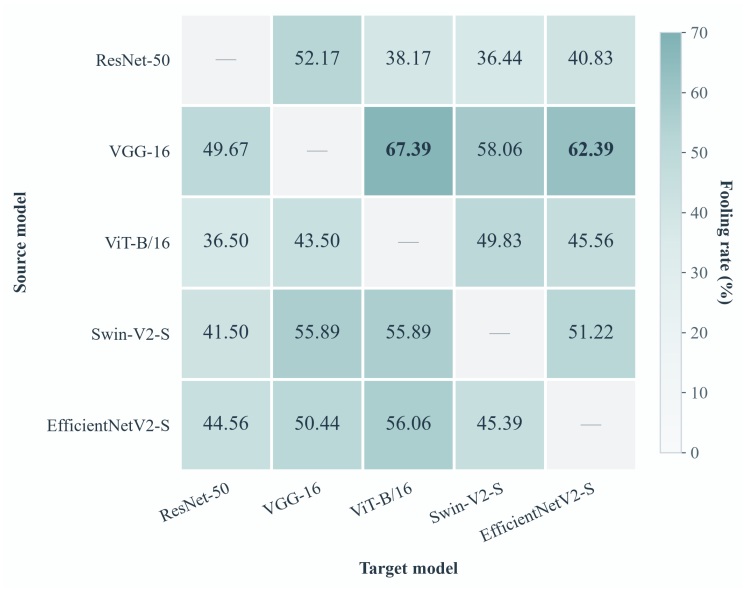

Figure 3: Cross-model attack transferability matrix.

4.4 View-consistent and external-component attacks

Qualitative comparisons show that the optimized coefficient changes are spatially concentrated while the target geometry remains fixed (Appendix E). Across all five classifiers, the t-SNE projections show greater overlap among adversarial representations than among their clean counterparts (Appendix F). Because the attack is untargeted, examples from one class can cross different decision boundaries rather than converging toward a single incorrect class. These projections provide a descriptive view of the representation shift induced by the attack, rather than evidence for a unique causal mechanism.

We further evaluate whether one coefficient tensor can remain effective across viewing conditions. During multi-view optimization, observations from the five sensor configurations and azimuths sampled over $[0°, 360°)$ at $1°$ intervals sequentially update the same tensor without reinitialization. This view-shared setting reaches a mean FR of 56.94% on ResNet-50 (Table 2). Although it is weaker than view-specific optimization, its performance remains substantially above the unperturbed condition, indicating that one surface parameterization can retain partial cross-view effectiveness within the simulator.

Finally, we attach a small rooftop-like component to each target, then hold the assembled geometry fixed and optimize only the component's scattering coefficients (Appendix G). Despite modifying only this external region, the attack obtains a mean FR of 47.11%. The truck and van remain especially vulnerable, with fooling rates of 65.28% and 75.83%, respectively. These results indicate that the attack does not require unrestricted modification of the complete target surface, although validation with a higher-fidelity electromagnetic solver is still needed before claiming real-world realizability.

Table 3: **Optimization budget.** Fooling rate (%) on ResNet-50. The shaded column is the main setting (50 iterations).

Target	10	20	30	40	50	60	70	80
Bus	7.78	22.50	30.83	37.50	42.78	44.72	46.94	47.22
Car	22.78	65.56	75.28	81.67	83.33	87.22	88.89	90.56
Jeep	8.61	29.44	45.00	60.56	60.00	70.56	73.89	76.11
Truck	64.44	95.83	98.61	98.61	100.00	100.00	100.00	100.00
Van	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Mean	40.72	62.67	69.94	75.67	**77.22**	80.50	81.94	82.78

Table 4: **Regularization strength.** Fooling rate (%) on ResNet-50. The shaded column is the main setting ($c = 0.25$).

Target	0	0.1	0.2	0.25	0.3	0.4	0.5
Bus	63.61	56.11	43.61	42.78	25.28	12.50	5.28
Car	93.33	87.22	83.33	83.33	73.61	65.28	54.17
Jeep	73.61	67.22	59.72	60.00	48.33	39.72	26.67
Truck	100.00	100.00	100.00	100.00	95.56	86.39	77.22
Van	100.00	100.00	100.00	100.00	98.61	93.33	86.39
Mean	86.11	82.11	77.33	**77.22**	68.28	59.44	49.94

4.5 Optimization sensitivity

We evaluated the optimization budget on ResNet-50 by recording fooling rates every 10 iterations up to 80 iterations. As shown in Table 3, the mean fooling rate increased from 40.72% at 10 iterations to 75.67% at 40 iterations, after which improvement became more gradual. At 50 iterations, it reached 77.22%; we therefore use this budget in the main experiments.

Convergence nevertheless varied substantially across categories: the van reached 100% FR after 10 iterations, whereas the bus increased from 7.78% to 47.22% over 10–80 iterations. The aggregate trend therefore masks target-dependent differences in optimization difficulty.

We also varied the regularization weight c from 0 to 0.5. Table 4 shows that the mean fooling rate decreased from 86.11% without regularization to 49.94% at $c = 0.5$, showing lower attack effectiveness at larger regularization weights. We use $c = 0.25$ in the main experiments, where the mean fooling rate is 77.22%.

5 Discussion and conclusion

SAR-DiSCO evaluates the sensitivity of SAR recognition models to surface-scattering changes on a fixed target mesh. The experiments show strong view-specific attacks, cross-model transfer, and partial effectiveness when one coefficient field is shared across views. The perturbations remain attached to the target and affect recognition through simulated image formation. This provides a controlled setting for studying scattering-aware robustness, subject to the renderer's approximations.

The principal limitation is the fidelity of the adopted differentiable renderer. Its diffuse-return approximation does not comprehensively represent multiple scattering, electromagnetic coupling, background clutter, or system noise. The optimized coefficients should therefore not yet be interpreted as directly realizable electromagnetic material modifications, and the central conclusion is limited to vulnerability within the adopted simulator. Higher-fidelity electromagnetic simulation and measurements of physically modified targets are needed to determine whether the observed effects persist in real SAR systems.

Shared multi-view optimization was weaker than view-specific optimization, which limits claims of viewpoint invariance. A shared coefficient field must accommodate different aspect-dependent responses, and sequential updates may favor some views over others. The transfer results also motivate robustness evaluation across both architectures and viewing conditions. Whether scattering-aware augmentation improves defense requires a separate evaluation.

ETHICS STATEMENT

Research on adversarial attacks against SAR recognition systems has potential dual-use implications. This work is intended to support controlled robustness evaluation and the development of more reliable SAR recognition systems. All experiments are conducted using simulated vehicle models and involve neither human subjects nor private or sensitive data. The proposed method should only be evaluated and deployed in authorized research environments.

REPRODUCIBILITY STATEMENT

The attack formulation and optimization procedure are described in Section 3 and the appendix. Dataset construction, rendering parameters, classifier training settings, hardware, attack hyperparameters, and threat-model assumptions are reported in Section 4.1 and Appendix C. Hyperparameter sensitivity analyses are reported in Section 4.5; the appendices provide qualitative results, feature-space visualizations, and external-component geometry. An anonymized implementation of SAR-DiSCO, together with the configuration files and scripts required to reproduce the experiments, is included in the supplementary material.

AI USE STATEMENT

Generative AI tools were used solely to assist with English translation and language polishing. The authors reviewed the resulting text and take full responsibility for the final content of this paper.

REFERENCES

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Li Chen, Zewei Xu, Qi Li, Jian Peng, Shaowen Wang, and Haifeng Li. An empirical study of adversarial examples on remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7419–7433, 2021.

Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3D objects with an interpolation-based differentiable renderer. *Advances in neural information processing systems*, 32, 2019.

Jiahao Cui, Wang Guo, Haikuo Huang, Xun Lv, Hang Cao, and Haifeng Li. Adversarial examples for vehicle detection with projection transformation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Meng Du, Daping Bi, Mingyang Du, Xinsong Xu, and Zilong Wu. Ulan: A universal local adversarial network for SAR target recognition based on layer-wise relevance propagation. *Remote Sensing*, 15(1):21, 2022.

Jiale Duan, Linyao Qiu, Guangjun He, Ling Zhao, Zhenshi Zhang, and Haifeng Li. A region-adaptive local perturbation-based method for generating adversarial examples in synthetic aperture radar object detection. *Remote Sensing*, 16(6):997, 2024.

Yexin Duan, Jialin Chen, Xingyu Zhou, Junhua Zou, Zhengyun He, Jin Zhang, Wu Zhang, and Zhisong Pan. Learning coated adversarial camouflages for object detectors. *arXiv preprint arXiv:2109.00124*, 2021.

Fei Gao, Mingyang Li, Jun Wang, Jinping Sun, Amir Hussain, and Huiyu Zhou. General sparse adversarial attack method for SAR images based on key points. *IEEE Transactions on Aerospace and Electronic Systems*, 2025.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Teng Huang, Yongfeng Chen, Bingjian Yao, Bifen Yang, Xianmin Wang, and Ya Li. Adversarial attacks on deep-learning-based radar range profile target recognition. *Information Sciences*, 531: 159–176, 2020a.

Teng Huang, Qixiang Zhang, Jiabao Liu, Ruitao Hou, Xianmin Wang, and Ya Li. Adversarial attacks on deep-learning-based SAR image target recognition. *Journal of Network and Computer Applications*, 162:102632, 2020b.

Zhongling Huang, Zongxu Pan, and Bin Lei. What, where, and how to transfer in SAR target recognition based on deep CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4): 2324–2336, 2019.

Odysseas Kechagias-Stamatis and Nabil Aouf. Automatic target recognition on synthetic aperture radar imagery: A survey. *IEEE Aerospace and Electronic Systems Magazine*, 36(3):56–81, 2021.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Pol Labarbarie, Adrien CHAN-HON-TONG, Stéphane Herbin, and Milad Leyli-abadi. Optimal transport based adversarial patch to leverage large scale attack transferability. In *International Conference on Learning Representations*, pp. 32485–32512, 2024.

Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018.

Yanjie Li, Kaisheng Liang, and Bin Xiao. UV-Attack: Physical-world adversarial attacks on person detection via Dynamic-NeRF-based UV mapping. In *International Conference on Learning Representations*, pp. 95147–95169, 2025.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3D reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7708–7717, 2019.

Wei Liu, Xuanshen Wan, Chaoyang Niu, Wanjie Lu, and Yuanli Li. Template-based universal adversarial attack for synthetic aperture radar automatic target recognition network. *IET Radar, Sonar & Navigation*, 19(1):e12691, 2025.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, June 2022.

Zhe Liu, Weijie Xia, and Yongzhen Lei. SAR-GPA: SAR generation perturbation algorithm. In *Proceedings of the 3rd International Conference on Advanced Information Science and System*, pp. 1–6, 2021.

Binyan Luo, Hang Cao, Jiahao Cui, Xun Lv, Jinqiang He, Haifeng Li, and Chengli Peng. SAR-PATT: A physical adversarial attack for SAR image automatic target recognition. *Remote Sensing*, 17(1): 21, 2024.

Yanjing Ma, Langjun Xu, Jifang Pei, Weibo Huo, Yin Zhang, Yulin Huang, and Jianyu Yang. Target partial-occlusion: An adversarial examples generation approach against SAR target recognition networks. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7054–7057. IEEE, 2023.

Yanjing Ma, Jifang Pei, Weibo Huo, Yin Zhang, Yulin Huang, Keyang Chen, and Jianyu Yang. SAR-PAA: A physically adversarial attack approach against SAR intelligent target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

Tianying Meng, Fan Zhang, and Fei Ma. A target-region-based SAR ATR adversarial deception method. In *2022 7th International Conference on Signal and Image Processing (ICSIP)*, pp. 142–146. IEEE, 2022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Merlin Nimier-David, Sébastien Speierer, Benoît Ruiz, and Wenzel Jakob. Radiative backpropagation: An adjoint method for lightning-fast differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(4):146–1, 2020.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS 2017 Autodiff Workshop*, 2017.

Bowen Peng, Jingyuan Xia, Bo Peng, Jie Zhou, Shuaifeng Zhi, and Yongxiang Liu. Target segmentation based adversarial attack for SAR images. In *2021 CIE International Conference on Radar (Radar)*, pp. 2146–2150. IEEE, 2021.

Bowen Peng, Bo Peng, Jie Zhou, Jingyuan Xia, and Li Liu. Speckle-variant attack: Toward transferable adversarial attack to SAR target recognition. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022a.

Bowen Peng, Bo Peng, Jie Zhou, Jianyue Xie, and Li Liu. Scattering model guided adversarial examples for SAR target recognition: Attack and defense. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022b.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller models and faster training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10096–10106. PMLR, 2021.

Xuanshen Wan, Wei Liu, Chaoyang Niu, Wanjie Lu, Meng Du, and Yuanli Li. Black-box universal adversarial attack for DNN-based models of SAR automatic target recognition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:8673–8696, 2024.

Xuanshen Wan, Wei Liu, Chaoyang Niu, Wanjie Lu, and Yuanli Li. Fast sparse adversarial attack for synthetic aperture radar target recognition. *Journal of Applied Remote Sensing*, 19(1):016502–016502, 2025.

Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8565–8574, 2021.

Bo Wei, Huagang Xiong, and Teng Huang. Full aspect angular domain-robust adversarial attack method for SAR ATR systems. *EURASIP Journal on Advances in Signal Processing*, 2026:49, 2026.

Huixiang Wen, Shan Chang, and Luo Zhou. Light projection-based physical-world vanishing attack against car detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Weijie Xia, Zhe Liu, and Yi Li. SAR-PeGA: A generation method of adversarial examples for SAR image target recognition network. *IEEE Transactions on Aerospace and Electronic Systems*, 59(2): 1910–1920, 2022.

Silong Yong, Venkata Nagarjun Pudureddiyur Manivannan, Bernhard Kerbl, Zifu Wan, Simon Stepputtis, Katia Sycara, and Yaqi Xie. OMG: Opacity matters in material modeling with gaussian splatting. In *International Conference on Learning Representations*, pp. 35832–35850, 2025.

Fei Zeng, Yuanjia Chen, Yulai Cong, Lei Zhang, Sijia Li, Jianqiang Xu, and Jia Duan. Stealthy backdoor attack in SAR target recognition with ASCM-Based physically realizable triggers. *IEEE Transactions on Radar Systems*, 2025.

Fan Zhang, Tianying Meng, Deliang Xiang, Fei Ma, Xiaokun Sun, and Yongsheng Zhou. Adversarial deception against SAR target recognition network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4507–4520, 2022.

Fan Zhang, Yameng Yu, Fei Ma, and Yongsheng Zhou. A physically realizable adversarial attack method against SAR target recognition model. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 17:11943–11957, 2024.

Guangbin Zhang, Songyi Gao, Xing Chang, and Yan Wang. Physical-domain SAR universal adversarial perturbation generation. *IEEE Transactions on Aerospace and Electronic Systems*, 2025.

Ruiqi Zhang, Jiacheng Wu, and Jie Chen. Mesh splatting for end-to-end multiview surface reconstruction. In *International Conference on Learning Representations*, 2026.

Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021.

Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2018.

Junhao Zheng, Chenhao Lin, Jiahao Sun, Zhengyu Zhao, Qian Li, and Chao Shen. Physical 3D adversarial attacks against monocular depth estimation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24452–24461, 2024.

Xiao Xiang Zhu, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, and Richard Bamler. Deep learning meets SAR: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):143–172, 2021.

A DIFFERENTIABLE SAR RENDERING DETAILS

Conventional SAR simulators contain discrete sampling, visibility, and thresholding operations that are not differentiable with respect to every scene variable. SAR-DiSCO does not attempt to differentiate these operations with respect to geometry. Instead, it fixes the mesh and, for each prescribed view, holds the corresponding ray intersections and grid assignments fixed while differentiating the coefficient-dependent image-formation operations. The forward pipeline comprises four stages: geometric projection and ray intersection, complex-response calculation, grid accumulation and point-spread approximation, and dynamic-range normalization.

The renderer uses orthographic projection to approximate far-field illumination. Let $\mathbf{p}_k = (x_k, y_k, z_k)$ be the intersection of the k-th valid ray with the target surface. The sensor geometry maps this point to azimuth and slant-range coordinates (a_k, r_k). For each prescribed view, these coordinates determine its image-grid assignment and remain fixed during coefficient optimization.

Second, the renderer relates the surface coefficients to the simulated return. At intersection $\mathbf{p}_k$, let $b_{k,j}$ be the barycentric weights of its three incident vertices v_j. The scalar coefficient in the following notation is $T_k = \sum_{j=1}^{3} b_{k,j} T_{v_j,3}$, where $b_{k,j} \geq 0$ and $\sum_{j=1}^{3} b_{k,j} = 1$. The implementation retains the third component of the weighted path response (zero-based channel index 2) in the scalar SAR contribution buffer. Let θ_k be the angle between the incident direction and the surface normal. Under the diffuse-return approximation used in our simulator, the amplitude is $A_k = T_k \cos(\theta_k)$. We introduce a propagation-dependent phase to model coherent accumulation. The resulting complex response is

$$S_k = A_k \exp\left(-\mathrm{i}\frac{4\pi R_k}{\lambda}\right), \tag{8}$$

where $\mathrm{i}^2 = -1$ and R_k is the one-way absolute slant range from the radar to the target. The phase calculation uses $\lambda = 0.031$ in scene length units; both azimuth and range grid spacings are 0.1 in the same units.

The projected coordinates determine fixed pixel assignments on the range–azimuth grid. The complex echo at pixel (u, v) is the sum of the responses assigned to that pixel:

$$E(u, v) = \sum_{k \in \Omega(u,v)} S_k, \tag{9}$$

where $\Omega(u, v)$ contains the valid intersections assigned to pixel (u, v). With assignments and phases fixed, this ideal accumulation is linear in the scalar surface coefficients. An implementation must preserve all accumulation terms in its backward pass to realize this derivative.

The simulator approximates point-spread-function (PSF) broadening with an inverse-square axial kernel applied to the envelope $A(u, v) = |E(u, v)|$. The unthresholded, additive form of this approximation is

$$A'(u, v) = A(u, v) + \sum_{d=1}^{D} \sum_{s \in \{-1,1\}} \frac{A(u + sd, v) + A(u, v + sd)}{(1 + d/\gamma)^2}. \tag{10}$$

Here D is the maximum offset in pixels, $\gamma > 0$ controls decay, and terms outside the grid are omitted. The implemented thresholded variant selects pixels satisfying $A(u, v) > 0.8 \max A$, with decay factor $\gamma = 0.3$ and spreading slices clipped to the image boundaries. During optimization, thresholding determines the active response support for each forward pass, while gradients propagate through the selected coefficient-dependent response amplitudes and the subsequent spreading operations. For fixed geometry and grid assignments, the implemented renderer therefore provides a differentiable mapping from the surface coefficients to the broadened SAR response.

The simulator then compresses the broadened envelope using the following logarithmic convention:

$$I_{\mathrm{dB}}(u, v) = 10 \log_{10}(A'(u, v)), \qquad A'(u, v) > 0. \tag{11}$$

This $10 \log_{10}$ envelope transform is the simulator's convention; calibrated amplitude ratios use $20 \log_{10}$, while power ratios use $10 \log_{10}$. Changing that factor without adjusting the clipping

thresholds would change the generated inputs. The logarithmic values are clipped to the configured bounds:

$$\hat{I}_{\mathrm{dB}}(u,v) = \max\{b_{\min}, \min\{b_{\max}, I_{\mathrm{dB}}(u,v)\}\}. \tag{12}$$

Here $b_{\min} = -40$ and $b_{\max} = 30$. For a nonconstant clipped image, instance-wise min–max normalization gives

$$I(u,v) = \frac{\hat{I}_{\mathrm{dB}}(u,v) - \min(\hat{I}_{\mathrm{dB}})}{\max(\hat{I}_{\mathrm{dB}}) - \min(\hat{I}_{\mathrm{dB}})}. \tag{13}$$

The extrema in Eq. (13) are taken over all pixels. The implementation applies $10\log_{10}$ before clipping; zero amplitudes yield $-\infty$ and are then mapped to $b_{\min}$. Min–max normalization uses the clipped extrema directly, so Eq. (13) requires a nonconstant clipped image. The normalized image is then resized to 224×224 pixels. Together, coherent accumulation, response spreading, logarithmic compression, clipping, and normalization form the implemented differentiable coefficient-to-image mapping. Automatic differentiation propagates the classification gradient through this mapping to the vertex-wise surface coefficients; the fixed geometric assignments remain outside the optimized variable set.

B Optimization algorithm

Algorithm 1 SAR-DiSCO coefficient optimization

Require: Mesh M, clean coefficients $\mathbf{T}_{\mathrm{orig}}$, view set $\mathcal{V} = \{\mathbf{\Theta}_k\}_{k=1}^{K}$, true label y_{gt}, regularization weight c, epochs N_{epoch}, learning rate α

Ensure: Adversarial scattering coefficients $\mathbf{T}_{\mathrm{adv}}$

$\boldsymbol{\omega} \leftarrow \log(\mathbf{T}_{\mathrm{orig}}/(1-\mathbf{T}_{\mathrm{orig}}))$ (element-wise; $0 < T_{v,h}^{\mathrm{orig}} < 1$)
for $e = 1$ **to** N_{epoch} **do**
  for $\mathbf{\Theta}_k \in \mathcal{V}$ **do**
    $\mathbf{T}_{\mathrm{adv}} \leftarrow \sigma(\boldsymbol{\omega})$
    $\mathbf{I}_{\mathrm{adv}} \leftarrow \mathcal{R}(M, \mathbf{T}_{\mathrm{adv}}; \mathbf{\Theta}_k)$
    $\mathbf{z} \leftarrow \mathcal{C}(\mathbf{I}_{\mathrm{adv}})$; $\mathbf{p} \leftarrow \mathrm{softmax}(\mathbf{z})$
    $c_{\mathrm{g}} \leftarrow \arg\max_{j \neq y_{\mathrm{gt}}} z_j$
    $\mathcal{L}_{\mathrm{total}} \leftarrow -\log(p_{c_{\mathrm{g}}}) + c \cdot \|\mathbf{T}_{\mathrm{adv}} - \mathbf{T}_{\mathrm{orig}}\|_F^2$
    $\boldsymbol{\omega} \leftarrow \mathrm{Adam}(\boldsymbol{\omega}, \nabla_{\boldsymbol{\omega}}\mathcal{L}_{\mathrm{total}}, \alpha)$
  end for
end for
return $\mathbf{T}_{\mathrm{adv}} \leftarrow \sigma(\boldsymbol{\omega})$

C Dataset and implementation details

Databases: We constructed a SAR simulation dataset using five categories of 3D vehicle models: bus, car, jeep, truck, and van. All models were obtained from the TurboSquid[1] website as free resources and used in compliance with the platform's and authors' licensing terms. The basic specifications of the models are listed in Table 5, and their rendered examples are shown in Figure 4. To ensure consistent imaging parameters, each model was uniformly scaled to realistic physical dimensions and translated so that its center lies at the origin in the x–y plane and its lowest point aligns with the ground level (z = 0), simulating vehicles placed on the ground. Using the differentiable SAR renderer, we generated multi-view SAR images of all targets under the imaging and simulation settings summarized in Table 6. During the simulation process, a ground background was added to all model renderings to ensure consistency of the imaging scenes. For each target, one image is generated at each azimuth in $\{0^\circ, 1^\circ, \ldots, 359^\circ\}$ for each of the five elevation angles, yielding $5 \times 360 = 1{,}800$ images per category. The resulting 9,000 images are divided into training and testing subsets with an 8:2 split, with 360 test images per category and 1,800 test images in total.

[1] https://www.turbosquid.com/

Table 5: **Target meshes.** Vertex, triangle, and component counts for the five simulated vehicles.

Statistic	Bus	Car	Jeep	Truck	Van
Vertices	55,808	117,075	50,651	9,218	76,214
Triangles	94,924	208,086	93,927	18,428	139,097
Parts	3	8	6	7	16

Table 6: **Imaging configuration.** Parameters used for dataset construction.

Parameter	Value
Elevation angles	15°, 17°, 25°, 35°, 45°
Rotation range	$[0°, 360°)$ at 1° intervals
Number of rays	22,500 (150×150)
Azimuth/Range sampling interval	0.1 (scene length units)
Phase wavelength	0.031 (scene length units)
PSF response threshold	$0.8 \max A$
PSF decay factor	0.3
Output image size	224×224

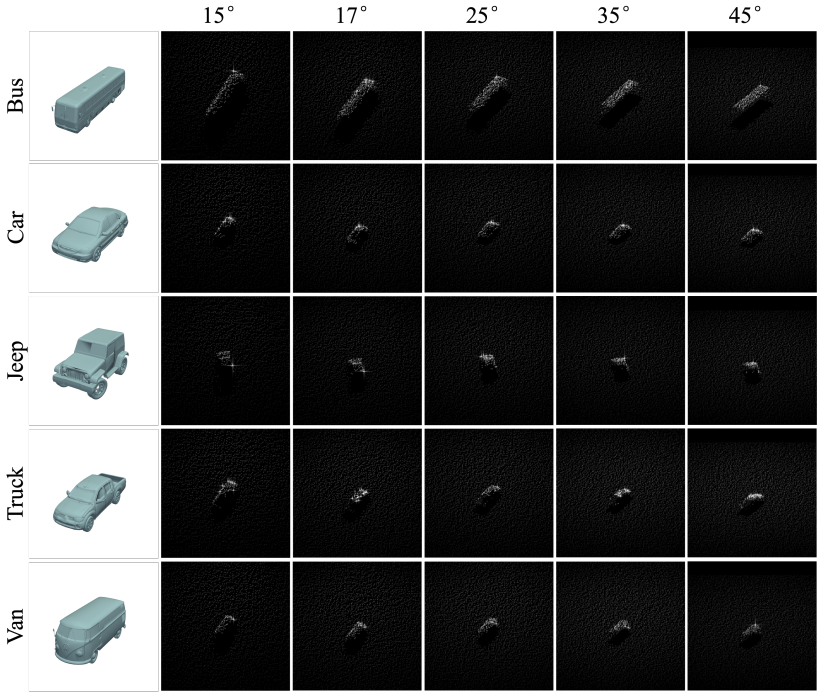

Figure 4: 3D models and SAR rendered images used in the dataset.

Experimental setup: The main paper reports the classifier training protocol, hardware, threat models, and attack hyperparameters. This appendix complements those details with mesh-level statistics (Table 5), the complete imaging configuration (Table 6), and clean-model recognition results (Figure 5).

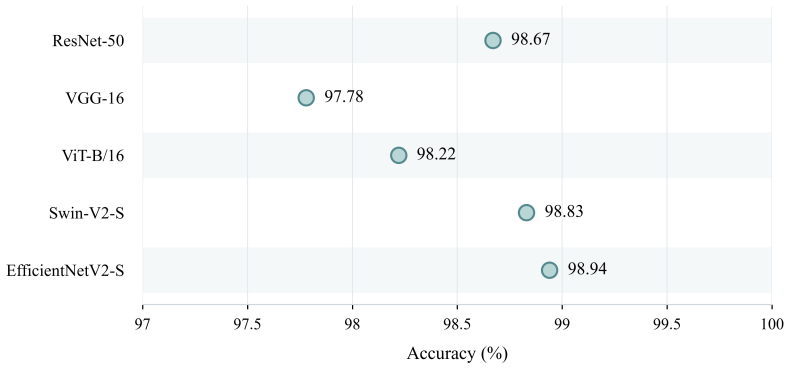

Figure 5: Classifier recognition accuracy.

The latent variable is initialized by applying the inverse-logit transform to $\mathbf{T}_{\text{orig}} = 0.5$. Logarithmic compression precedes clipping to the stated dynamic range; zero-valued responses are mapped to the lower bound after taking the logarithm. Tables 3 and 4 provide the sensitivity analyses that motivate the selected iteration count and regularization weight.

Evaluation metric: We use FR as defined in Eq. (7). It measures the fraction of all evaluated examples assigned an incorrect class. Percentages equal $100\,\mathrm{FR}$. Each category-level FR uses all 360 test samples, and each transfer FR uses all 1,800 test samples.

D Comparison with Existing 3D Surface Attack Methods

To further evaluate different attack pipelines targeting 3D object surfaces, we include SAR-PATT (Luo et al., 2024) as an additional baseline. SAR-PATT first optimizes intermediate scattering data through gradient-based updates and then fits target-surface parameters in a separate mapping stage. In contrast, SAR-DiSCO establishes an end-to-end differentiable mapping from vertex-wise surface scattering coefficients to SAR images and classifier outputs under fixed geometry and observation configurations. Classification gradients propagate directly to the surface coefficients through the coefficient-dependent imaging operations, eliminating the separate perturbation-to-surface mapping stage.

For both methods, the gradient-based optimization stage uses Adam with 50 iterations and a learning rate of 0.05. SAR-PATT uses sigmoid-bounded scattering intensities, a squared ℓ_2 penalty with weight 1, and an iterative sparsity mask with a perturbation threshold of 10^{-4}. Its mapping stage selects the scattering point with the largest intensity change among those associated with each triangular face and fits that face's scattering coefficient using Brent's method over $[0, 1]$. SAR-DiSCO uses the coefficient regularization weight $c = 0.25$ from the main experiments. The matched iteration count excludes SAR-PATT's subsequent mapping stage and does not imply equal computational or perturbation budgets.

Because the two imaging pipelines produce different image distributions, we train a separate ResNet-50 on clean images from each pipeline, using the same target meshes, training and test views, and classifier training settings. This comparison evaluates each attack in its respective imaging and classification pipeline. Table 7 reports clean and attacked FR using all 360 test images per category. Both pipelines have low clean FR across all five categories. SAR-DiSCO achieves higher attacked FR on four categories, exceeding SAR-PATT by 13.05, 28.89, and 10.56 percentage points on Car, Truck, and Van, respectively. The results on Jeep are nearly identical (60.00% versus 59.72%), whereas SAR-PATT performs better on Bus, achieving 51.39% compared with 42.78% for SAR-DiSCO.

E Qualitative attack results

Figure 6 shows where the optimized surface coefficients change, with affected regions highlighted in red. Figure 7 compares clean and adversarial renders using ResNet-50. Changes concentrate around

Table 7: **Evaluation of 3D surface attack pipelines.** Clean and attacked FR (%) on ResNet-50 classifiers trained separately in their respective imaging domains. Each category contains 360 test images.

Method	Metric	Bus	Car	Jeep	Truck	Van
SAR-PATT	Clean FR	1.94	3.06	3.89	2.22	2.78
SAR-PATT	Attacked FR	51.39	70.28	59.72	71.11	89.44
SAR-DiSCO	Clean FR	1.11	1.39	1.67	1.11	1.39
SAR-DiSCO	Attacked FR	42.78	83.33	60.00	100.00	100.00

prominent scattering regions while the target geometry remains fixed. These visualizations describe the simulated perturbations; they do not establish material realizability or identify a unique cause of classifier failure.

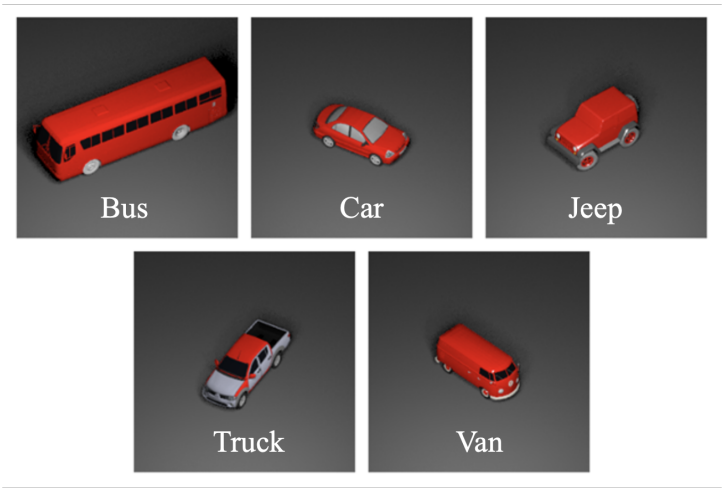

Figure 6: Visualization of applied perturbation locations on the target models.

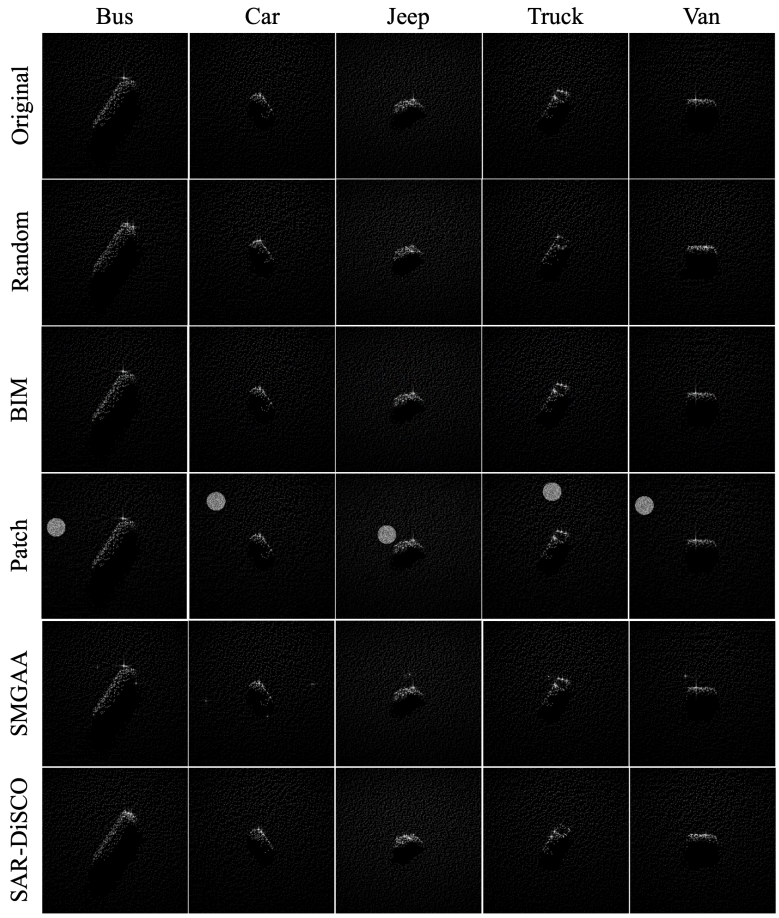

Figure 7: Clean images and adversarial images generated by the compared methods using ResNet-50.

F FEATURE-SPACE ANALYSIS

We extracted intermediate representations of clean and adversarial examples from each classifier and projected them into two dimensions using t-SNE. This visualization describes feature-space organization rather than measuring distances to the classifier's decision boundaries.

In Figure 8, colors identify the five target categories, and marker shapes distinguish clean from adversarial examples. Adversarial projections show more overlap between categories than their clean counterparts. Because t-SNE can distort distances and global structure, cluster positions alone do not establish proximity to decision boundaries.

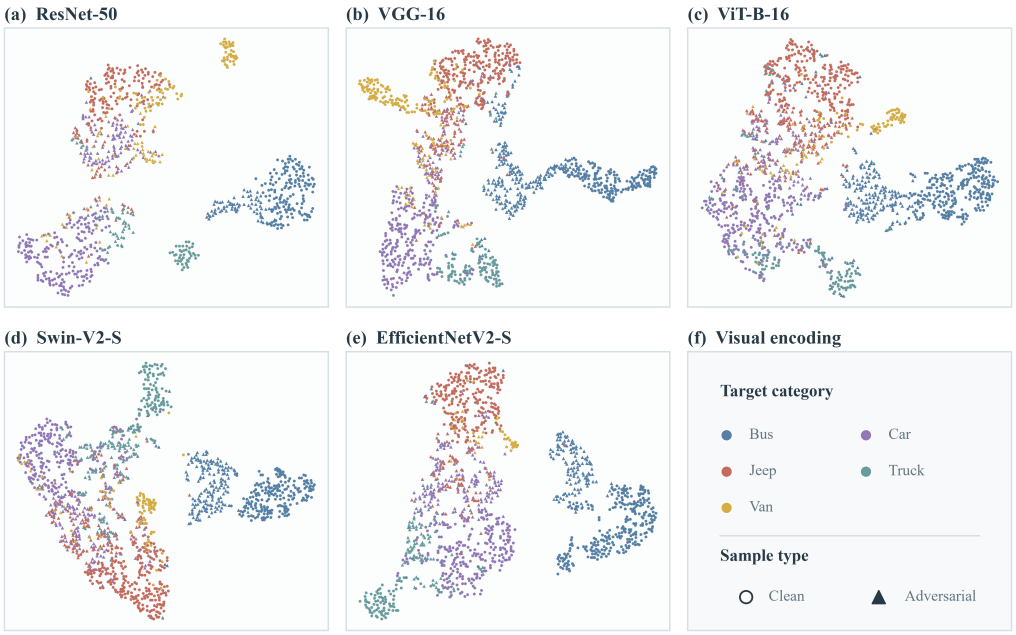

Figure 8: t-SNE projections for the five recognition architectures. Circles denote clean examples and triangles denote adversarial examples; colors identify the five target categories (bus, car, jeep, truck, and van).

The untargeted objective permits different incorrect predictions for examples from the same class, which is compatible with dispersed adversarial projections. The bus remains comparatively isolated in these plots, but this observation alone does not explain category-dependent attack success rates.

G EXTERNAL-COMPONENT GEOMETRY

For this experiment, a small rooftop-like component is attached to each target. The assembled geometry is then held fixed while only the component's scattering coefficients are optimized. The quantitative results are reported in Table 2.



Figure 9: External component used to localize the optimized scattering coefficients.