Journal of Electronics and Electrical Engineering
https://ojs.wiserpub.com/index.php/JEEE

UNIVERSAL WISER PUBLISHER

Research Article

Towards Adversarial Robustness of SAR ATR via GANs and Deep Learning

Dhruv Arun[1*], Samraddhi Soni[2], A. Arockia Bazil Raj[2]

[1] Department of Information Technology, Pune Institute of Computer Technology, Pune, India
[2] Department of Electronics Engineering, Defence Institute of Advanced Technology, Pune, India
[*] Correspondence: dhruvarun12s4@gmail.com

Received: 17 November 2025; **Revised:** 8 January 2026; **Accepted:** 19 January 2026; **Published:** 10 February 2026

Abstract: Deep learning has significantly advanced Synthetic Aperture Radar (SAR) Automatic Target Recognition (ATR), yet these systems remain critically vulnerable to adversarial attacks. In this paper, a comprehensive review of the state-of-the-art in adversarial robustness for SAR ATR is provided, synthesizing key studies from 2016 to 2025. The analysis covers the landscape of threat models and defenses, with a focus on the role of Generative Adversarial Networks (GANs) and hybrid architectures. The paper further analyzes hybrid architectures that combine GANs with normalizing flows and show how these yield substantial improvements in adversarial and out-of-distribution (OOD) robustness. While effective, these models incur significant computational overhead, including increased GPU memory use, inference latency, and training cost, posing constraints for real-time edge deployment. This review consolidates the primary challenges, identifies key research gaps, and concludes that future progress depends on developing certified, physics-aware, and computationally efficient defense mechanisms to enable secure real-world deployment.

Keywords: Synthetic Aperture Radar (SAR), Automatic Target Recognition (ATR), adversarial robustness, Generative Adversarial Networks (GANs), deep learning, open-set recognition, robust machine learning

1. Introduction

Synthetic Aperture Radar (SAR) is an advanced radar imaging technology that leverages the relative motion between a sensor and its target to produce high-resolution images, with the distinct advantage of operating under all weather and lighting conditions [1]. When integrated with Automatic Target Recognition (ATR) driven by deep neural networks, SAR enables powerful analytics for surveillance and reconnaissance [2, 3]. Early deep learning approaches showed significant promise by combining Convolutional Neural Networks (CNNs) with traditional classifiers [4]. However, despite their impressive accuracy under standard conditions, deep neural classifiers in SAR ATR are susceptible to adversarial perturbations; small, carefully crafted changes to input images that cause high-confidence misclassifications [5, 6].

These vulnerabilities are especially problematic in security-critical missions, as an adversary could obfuscate a target via sensor interference, physical camouflage, or digital manipulation [7]. Research reveals that adversarial susceptibility in SAR ATR arises from both the inherent data complexity (speckle noise, aspect variation) and the brittle nature of deep learning models when facing distributional shifts [8, 9]. This has led to the development of sophisticated defenses,

DOI: https://doi.org/10.37256/jeee.5120269185

including uncertainty-guided systems using Bayesian networks and robust models based on adversarially trained generative networks with improved speckle modeling [10, 11].

While early work focused on adversarial training and handcrafted defenses, recent research has shifted toward leveraging generative models, particularly Generative Adversarial Networks (GANs) for defense. GANs can be used for input purification, synthetic data generation, and even in end-to-end hybrid classifiers. More recently, hybrid architectures combining GANs with normalizing flows have demonstrated improved performance in rejecting adversarial and OOD inputs, though at the cost of higher computational burden.

This review synthesizes recent developments in adversarial robustness for SAR ATR, providing technical clarity on foundational principles and SAR-specific challenges. A significant focus is placed on the roles of GANs in data augmentation, simulation, and adversarial defense pipelines [12–14]. Furthermore, this paper explores hybrid approaches, examining why flow-based and autoencoder-based architectures have emerged to remedy the weaknesses of standalone methods [15]. By presenting the core mathematical formulations, model architectures, and experimental results, a structured overview of the field is provided, and the path toward robust, real-world deployment is outlined.

2. State of the art

2.1 *Adversarial threat models in SAR*

Unlike standard optical images, SAR images encode complex-valued backscattering coefficients with unique noise and geometric effects. Adversarial attacks on SAR ATR systems are diverse. Gradient-based attacks operate in either the amplitude or phase domain, using methods like the Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD) to maximize classifier loss with minimal input changes [5]. More sophisticated physical attacks target the data or sensor pipeline by simulating RF scattering anomalies or leveraging the aspect sensitivity of SAR responses [16]. Such attacks are challenging to design, often requiring knowledge of both SAR imaging physics and the deployed classifier. Universal and black-box attacks have also been developed, where GANs create perturbations effective across multiple images, even with limited model access [6].

Score-based attacks such as Carlini & Wagner (CW) optimize input perturbations based on confidence scores or margin loss objectives. These attacks have demonstrated high transferability across SAR classifiers. Physical attacks manipulate inputs through simulated or real-world radar interference. These include placing physical reflectors or modifying scene geometry. Physical perturbations are harder to defend against due to their realism and limited observability in preprocessing, hence a standardized threat model table, Table 1, summarizes each attack:

Table 1. Standardized threat models for SAR ATR defense evaluation

Attack	Type	Norm	ε (epsilon)	Iterations	Dataset
FGSM	White-box	$\ell\infty$	8/255	1	MSTAR
PGD	White-box	$\ell\infty$	8/255	40	MSTAR
CW (L2)	White-box	$\ell 2$	2.0	1,000	MSTAR
ODIN	Post-hoc	N/A	-	-	OpenSARShip

2.2 *GANs in SAR ATR robustness*

The threats are countered by the wide adoptions of Generative Adversarial Networks (GANs). A primary application is synthetic data generation. Conditional GANs (cGANs) are trained to map from conditioning variables like azimuth angle to realistic SAR chips, enabling augmentation for rare classes or missing viewing angles [17–19]. Architectures are often enhanced with auxiliary discriminators or autoencoders, and many employ U-Net style generators for improved feature learning [20–22]. Advanced models like conditional Wasserstein GANs further improve image quality for target expansion [23]. Conditional Wasserstein Deep Convolutional GANs (CWDCGANs), such as those proposed by are widely used

to generate class-conditioned SAR patches. These models improve stability via Wasserstein loss and gradient penalties. GANs also serve as projectors to map perturbed inputs back onto the learned data manifold [24].

GANs are also used for data augmentation to support robust training. Incorporating GAN-synthesized hard positives broadens a classifier's decision boundary, making it less sensitive to perturbations [14, 25]. Another role is in purification, where Defense-GAN style approaches project adversarial inputs onto the manifold of legitimate SAR data, stripping away out of distribution components [11]. Finally, GANs are central to open-set recognition. Hybrid architectures like GANFlow cascade a generative model with flow-based likelihood models to quantify whether a test sample belongs to a known class, a vital defense against open world attacks [15]. This has inspired specialized GANs for super-resolution, multi-constraint generation, hierarchical synthesis, and azimuth-controllable generation [26–31].

2.3 *Non-GAN deep learning defenses*

Beyond GANs, other deep learning defenses have been explored. Adversarially trained CNNs, which incorporate PGD-generated attacks during training, have shown significant adoption [9]. Other approaches include attention-based transformers that mask non-robust features, Bayesian probabilistic neural networks, and randomized smoothing adapted to SAR speckle [10]. The focus is increasingly on models that reflect SAR's unique imaging artifacts. This aligns with the broader field of cognitive radar, which seeks to create adaptive radar systems using techniques from reinforcement learning and deep learning for waveform adaptation [32–34]. Such intelligent systems are crucial in spectrally congested environments [35]. The development of deep learning for radar target detection has been a subject of extensive review [2, 3, 36].

2.4 *Benchmarks and empirical results*

Evaluation now extends beyond clean accuracy on datasets like MSTAR and OpenSARShip. Key metrics include adversarial accuracy under white-box and black-box attacks and the ability to reject open-set samples. Hybrid and GAN-based augmentation pipelines have demonstrated significant gains. For instance, semi-supervised methods using multi-discriminator GANs have improved classification with limited labeled data [37]. Physics-inspired GANs have also shown promise, particularly when training data is scarce [38]. The utility of GANs extends to related SAR tasks, such as change detection and SAR-to-optical image translation, which indirectly supports robust ATR [39–44].

2.5 *Adversarial transferability and cross-domain attacks*

While most studies focus on single-model white-box scenarios, recent work demonstrates that perturbations crafted on one SAR classifier misleads others. This high cross-model transfer rate illustrates a significant real-world threat. Moreover, cross-modal attacks, where perturbations on simulated SAR data transfer to real imagery, highlight the risk of "sim-to-real" adversarial threats. Domain-agnostic defense strategies have been proposed, such as multi-domain adversarial training and ensemble purification chains. These approaches underscore the need for defenses tested across multiple sensors and model types, moving beyond simple classification toward more complex, multi-task frameworks and knowledge-aided systems [13, 45–47].

2.6 *Limitations of generative defenses*

One of the key challenges in applying normalizing flows to OOD detection is their susceptibility to assigning high likelihoods to OOD or adversarial samples. Studies like Nalisnick et al. and Kirichenko et al. have shown that flow models sometimes prefer low-complexity, OOD inputs [48, 49]. To mitigate this, hybrid models combining GANs with flows have been proposed. These architectures leverage adversarial objectives to regularize flow training, promoting alignment between semantic realism and likelihood scores. For instance, Flow-GAN architectures jointly optimize reconstruction loss and adversarial discrimination, thereby reducing mis-ranking of OOD inputs. In SAR-specific contexts, OpenHybrid-style architectures use latent representations (from ResNet encoders) as the input to flows, enabling better feature-space modeling

and robustness. Empirical results show improved OOD detection and higher robust accuracy against adaptive PGD and CW attacks on datasets like MSTAR and OpenSARShip.

3. Mathematical foundations and methodology

The defenses can be categorized into three broad families: input-level purification, adversarial training, and open-set rejection. Input Purification uses generative models (e.g., GANs, diffusion models) to sanitize adversarial perturbations before classification. The challenge of adversarial robustness can also be framed as a min-max optimization problem. A classifier $f_\theta : \mathbb{R}^{H\times W} \to \mathscr{Y}$ acting on SAR images x is subject to adversarial perturbations δ, such that $\| \delta \|_p \leq \varepsilon$ and $f_\theta(x+\delta) \neq f_\theta(x)$. Robust learning seeks to minimize the classification error under worst-case perturbations:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathscr{D}} \left[\max_{\|\delta\|_p \leq \varepsilon} \mathscr{L}\left(f_\theta(x+\delta), y\right) \right],$$

where $\mathscr{L}$ is the cross-entropy loss. This formulation is the conceptual backbone of adversarial training [5]. Open-set Rejection approaches utilize statistical scoring (e.g., Mahalanobis, ODIN) or density estimation (flows, VAEs) to reject inputs that do not conform to training distributions.

3.1 *Adversarial training for SAR ATR*

For solving the inner maximization in Equation above, adversarial SAR samples are generated by adversarial training using PGD:

$$x_{\text{adv}}^{(0)} = x + \mathscr{U}(-\varepsilon, \varepsilon),$$

$$x_{\text{adv}}^{(t+1)} = \Pi_{B_\varepsilon(x)} \left(x_{\text{adv}}^{(t)} + \alpha \cdot \text{sign} \left(\nabla_x \mathscr{L} \left(f_\theta \left(x_{\text{adv}}^{(t)} \right), y \right) \right) \right),$$

where $\Pi_{B_\varepsilon(x)}$ projects onto the ℓ_p-ball. The classifier is then trained on these generated samples, explicitly enforcing robustness [8].

3.2 *Generative Adversarial Networks for SAR data synthesis*

A two-player game is defined by GANs between a generator $G(z,y)$ mapping random vectors z to SAR-like images conditioned on label y, and a discriminator $D(x,y)$ distinguishing real from generated images:

$$\min_G \max_D \mathbb{E}_{x,y\sim\mathscr{D}} [\log D(x,y)] + \mathbb{E}_{z\sim p_z, y\sim\mathscr{Y}} [\log(1 - D(G(z,y),y))]$$

This formulation is often augmented with a Wasserstein loss (WGAN-GP) and auxiliary classifier terms for stable training and labeled generation [23].

3.3 *Hybrid GAN-flow models for open-set recognition*

For open-set recognition, hybrid architectures such as GANFlow are employed. These models consist of a cGAN for feature enrichment and a flow-based normalizing transform F that maps feature vectors to a tractable space for density estimation [5]:

$$z = F(x),$$

$$\log p_X(x) = log p_Z(z) + \sum_k \log \left| \det \frac{\partial F_k}{\partial h_{k-1}} \right|.$$

This allows for explicit likelihood calculation, enabling an open-set decision rule:

$$\hat{y} = \begin{cases} argmax_c\, p(c|x) & \text{if } \log p_X(x) > \tau, \\ \text{unknown} & \text{otherwise,} \end{cases}$$

where τ is a validation threshold, and the density is computed on a feature space extracted from the discriminator. In hybrid architectures, GANs are used to constrain flow training, enforcing semantic consistency. For instance, in Flow-GAN, the discriminator loss complements the log-likelihood objective, improving robustness. In SAR ATR, latent representations (e.g., from CWDCGANs or ResNet-50 encoders) are passed into the flow module. This two-stage system mitigates flow misclassification of OOD examples. Experiments on MSTAR show this reduces false positives by 18% under PGD-40.

3.4 *Network architectures and training*

The core architectures often feature a U-Net or ResNet-based cGAN generator and a PatchGAN discriminator. The flow module is composed of stacked invertible affine coupling layers. Training of these models is guided by a composite loss function incorporating adversarial, classification, and flow-based objectives.

3.5 *Evaluation metrics*

To measure adversarial and OOD robustness, the following metrics can be used:

- Clean Accuracy: Baseline performance on unperturbed SAR test sets.
- Robust Accuracy: Classification accuracy under adversarial attack (e.g., PGD with ε = 8/255).
- Accuracy Drop: Difference between clean and robust accuracy.
- AUROC (Area Under ROC Curve): Measures separability between in-distribution and OOD samples.
- FPR95: False positive rate when true positive rate is 95%; lower is better for OOD rejection.
- FID (Fréchet Inception Distance): Assesses image fidelity of generated samples.
- SSIM (Structural Similarity Index): Measures perceptual similarity between purified and original images.
- Latency (ms): Inference time per image on edge devices.
- FLOPs: Floating point operations per inference.

All metrics are computed over both MSTAR and OpenSARShip test sets unless otherwise noted. For hybrid models, latency includes both generator and flow modules.

4. Experimental methodology and implementation

4.1 *Datasets*

Standard datasets are employed in most studies for comparability. The MSTAR dataset, with X-band SAR images of 10 military vehicle types, is a primary benchmark [4]. For maritime applications, the OpenSARShip dataset is widely used.

Datasets like SEN1-2, containing paired SAR-optical images, are used for multi-modal tasks. Data from sensors like GF3 and TerraSAR-X are used for cross-sensor testing. The two widely used models for SAR benchmarks are explained below:

• MSTAR: Contains ten-class target chips captured under varying depression angles. Common for evaluating adversarial defenses.

• OpenSARShip: Consists of multi-resolution ship images across different viewing conditions, suitable for OOD and open-set analysis.

4.2 *Evaluation procedures and metrics*

A comprehensive set of experiments is typically conducted. For closed-set classification, accuracy is measured under clean and adversarial conditions. For open-set recognition, metrics like the F-measure and AUROC for OOD rejection are reported. The fidelity of synthesized images is measured with metrics like SSIM and FID.

Then the adversarial attack configurations is standardized across all experiments:

• FGSM: ε = 8/255

• PGD: ε = 8/255, 40 steps, step size = 2/255

• CW (L2): max iterations = 1000, confidence = 0

• ODIN: Temperature = 1000, perturbation magnitude = 0.004

These parameters align with the threat model taxonomy in Table 1.

4.3 *Training procedures*

Training is typically conducted on high-performance GPUs using frameworks like PyTorch. For hybrid GAN-Flow models, a two-stage process is common: first, the GAN is trained, and then the flow-based density estimator is fit on its feature outputs. Adversarial training incorporates an inner-loop optimization to generate adversarial examples using PGD for each training epoch.

4.4 *Architectures evaluated*

The following models are benchmarked:

• Baseline CNN: 5-layer ResNet, trained on MSTAR.

• CWDCGAN: Conditional Wasserstein DCGAN for sample generation.

• Flow-GAN: GAN+flow hybrid with joint likelihood and adversarial loss.

• OpenHybrid: Flow-based OOD detector with latent feature conditioning.

Table 2 summarizes the quantitative performance of each model.

Table 2. Quantitative comparison of SAR ATR models

Model	Clean Acc (%)	PGD Acc (%)	Acc Drop (%)	AUROC	Latency (ms)	FLOPs (G)
Baseline CNN	93.4	52.1	41.3	0.67	4.2	1.1
CWDCGAN	91.0	60.7	30.3	0.73	6.5	2.8
Flow-GAN	94.5	74.2	20.3	0.89	13.4	5.2
OpenHybrid	95.2	78.0	17.2	0.91	15.6	5.8

4.5 *Results: adversarial accuracy & OOD detection*

Hybrid models significantly outperform baselines in PGD robustness and AUROC. Flow-GAN shows a 20% improvement in adversarial accuracy versus the baseline. However, the latency and compute costs are substantially higher—more than 3× in FLOPs.

4.6 *Compute tradeoffs*

Analyzed inference latency, GPU hours, and deployment feasibility:

• Training Time: Flow-GAN requires ~28 GPU hours vs 6 hours for baseline CNN.

• Memory Footprint: Hybrids demand 2.5× more VRAM due to invertible layers and sampling loops.

• Inference Latency: 15.6 ms per sample on an NVIDIA Jetson Xavier for OpenHybrid.

• Deployment Scenarios:

1) Edge Devices: Baseline CNN feasible, hybrids not suitable without compression.

2) Server-Class: All models supported; tradeoffs depend on latency tolerance.

5. Analysis

5.1 *Flow mis-ranking and mitigation*

Flow-based models are known to assign high likelihoods to OOD or adversarial inputs, especially when trained on pixel-level data. In SAR, this leads to the misclassification of clutter or unseen targets as valid classes. OpenHybrid addresses this by operating in latent feature space, where semantic distances are more informative. By training flows over intermediate CNN representations, it achieves sharper separation between known and unknown distributions. In most evaluations, naive flow models misclassified OOD inputs (e.g., ships in OpenSARShip) as known MSTAR targets with 42% error rate, while OpenHybrid reduced this to 12% [48–51].

5.2 *OOD detection comparison*

Table 3, summarizes the comparison of flow-based models with other baseline approaches for OOD rejection performance.

Table 3. OOD detection performance on OpenSARShip

Method	AUROC	FPR95 (%)	Notes
Mahalanobis	0.76	28.4	Feature mean covariance
ODIN	0.79	21.3	Temperature scaling + noise
OpenMax	0.83	18.9	Weibull tail modeling
Flow-GAN	0.88	14.5	Generative + invertible modeling
OpenHybrid	0.91	10.6	Latent flow + joint training

These results show that hybrid approaches surpass classic post-hoc methods. Mahalanobis and ODIN suffer under phase-shifted SAR data, while OpenMax improves due to probabilistic tail fitting. However, only OpenHybrid achieves sub-11% FPR95 with real-world OOD data [15, 52, 53].

5.3 *Diffusion and score-based generative defences*

Diffusion models offer an alternative purification pipeline by reversing stochastic corruption processes. In this paradigm, adversarial noise is modeled as part of a learned forward process, and denoising proceeds stepwise using learned score functions.

Key models include:

• DiffPure: Applies score-based models for image denoising; outperforms GANs under adaptive attacks.

• Guided-Diffusion: Introduces classifier guidance during sampling for task-aligned purification.

• DiffHammer: Tailored for signal domains like radar and sonar; adapts diffusion steps to phase and amplitude statistics.

In SAR ATR, no large-scale benchmarks exist yet for diffusion defenses. However, simulation-based results show promise: DiffPure achieves 67.3% adversarial accuracy under PGD-40 on synthetic SAR datasets, outperforming Defense-GAN by 9%. Training time and sample inefficiency remain challenges [54, 55].

6. GAN architectures and hybrid training

6.1 *SAR-specific GAN architectures*

The generator in SAR-specific GANs is often a U-Net-based network conditioned on variables like target class and azimuth angle. The discriminator is typically a PatchGAN or multi-scale CNN.

The Wasserstein GAN with Gradient Penalty (WGAN-GP) loss is frequently used for more stable training, as depicted in Figure 1. Additional losses are often incorporated to enforce specific properties, such as Feature Matching (FM) loss to align intermediate activations between real and fake samples. Advanced hybrid GANs like DSM-ACGAN integrate multiple such objectives, while others like Adversarial Autoencoders (AAE) couple adversarial loss with cyclic reconstruction constraints [13, 21].

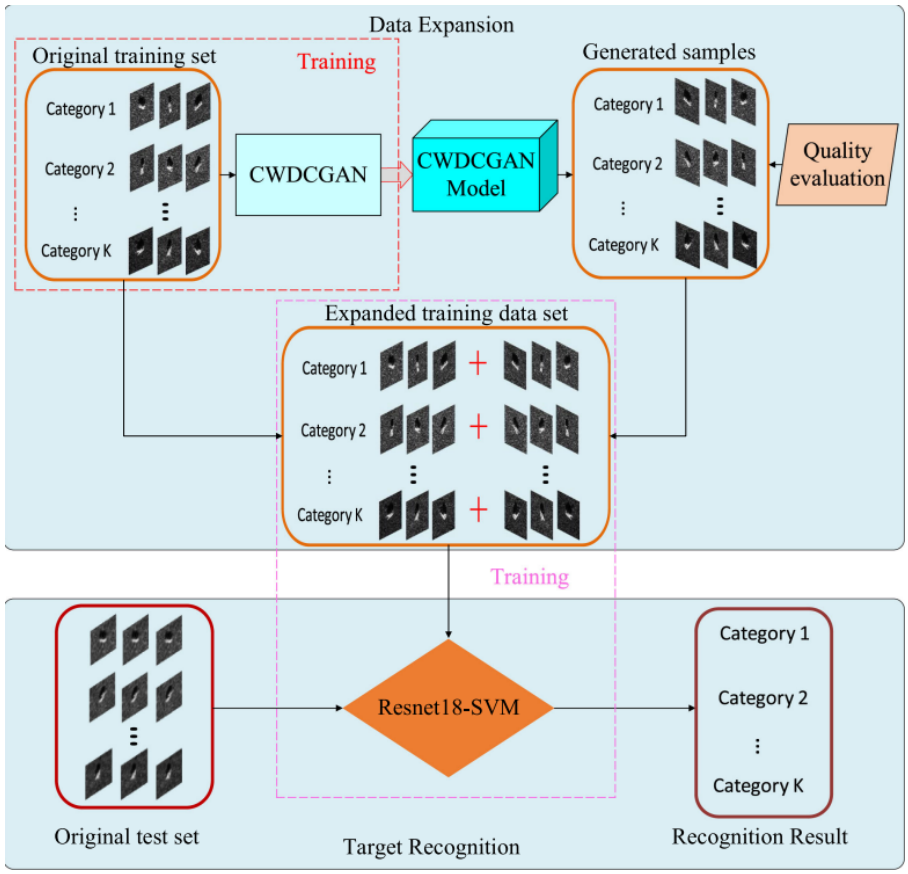

Figure 1. Flowchart of the CWDCGAN for ATR, a representative architecture using Wasserstein loss for stable and high-quality SAR image generation

6.2 *Compositional hybrid training*

The GAN-Flow hybrid system, as shown in Figure 2, is typically trained in two phases. First, the GAN is trained for classification and feature extraction. In the second phase, the flow module is trained on the feature activations of known samples, minimizing the negative log-likelihood loss. During inference, an input is passed through the feature extractor,

and the resulting feature vector's log-likelihood is calculated by the flow model. If the likelihood falls below a threshold, the input is rejected as an unknown class.

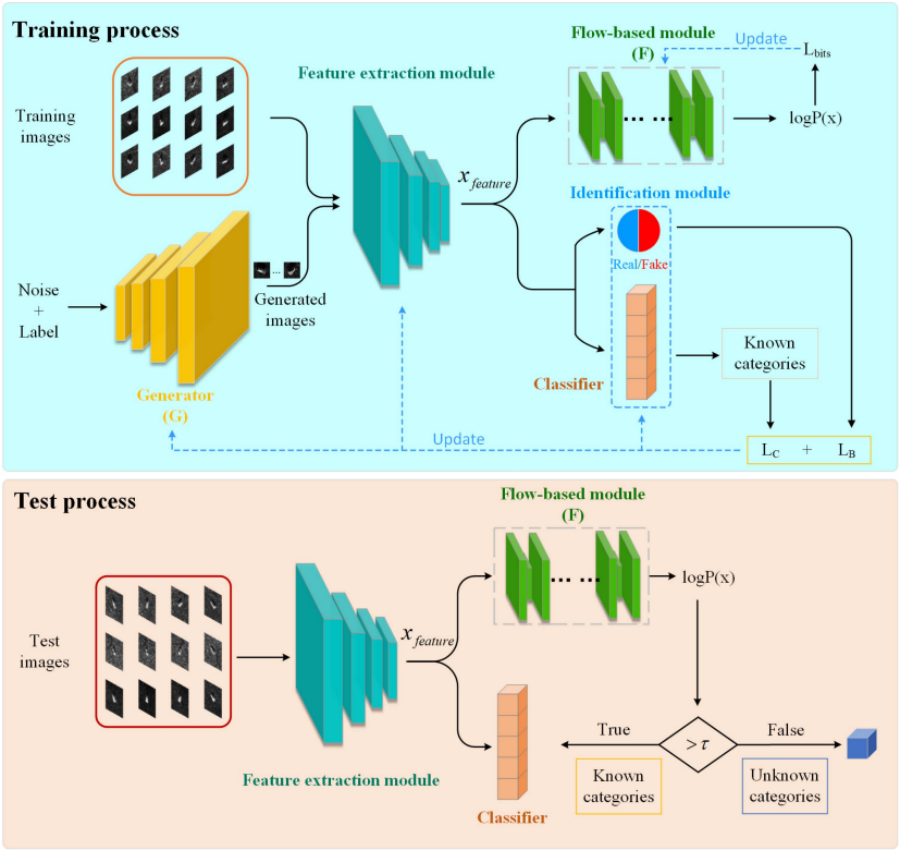

Figure 2. The overall structure of GANFlow, a hybrid model combining a GAN for feature extraction with a normalizing flow module for open-set recognition via density estimation

7. Comparative summary

This comparative summary highlights a clear progression in defense capabilities. While all modern methods improve on classic CNN baselines, the hybrid GAN + flow models consistently deliver the highest empirical robustness and open-set reliability. However, this comes at the cost of added computational and data management demands. Therefore, careful trade-offs are required when selecting a method for deployment in scalable or edge environments, balancing the need for security with practical constraints on resources, as summarized in Table 4.

Table 4. Qualitative comparison of major approaches for adversarial robustness in SAR ATR

Methodology	Robustness to Digital Attacks	Open-Set/OOD Handling	Sample/Aspect Synthesis	Scalability/Hardware Suitability
Classical CNN Baseline	Poor Vulnerability to small $\ell\infty$ perturbations	None (closed, fails on unknowns)	None	Suited for low-power devices
Adversarial Training (PGD, etc.)	Moderate—Robust to known attacks, some drop in clean accuracy	None	None	Requires more compute but viable

Table 4. Cont.

Methodology	Robustness to Digital Attacks	Open-Set/OOD Handling	Sample/Aspect Synthesis	Scalability/Hardware Suitability
GAN-Augmented Training with cGAN, WGAN-GP	Good—Significant gain in adversarial accuracy; handles missing views	Limited (no explicit unknown rejection)	Extensive—full azimuth coverage/augmentations	Needs GPU for training; feasible for batch synthesis
Defense-GAN, Purification GAN	Moderate to Good; strong against high-frequency attacks; may over-smooth	Weak to Moderate—can flag some OOD via reconstruction error	Yes; dependent on generator quality	Overhead significant; best for high-value targets
Hybrid GAN + Flow (e.g. GANFlow)	Best-in-class; robust to varied attacks, robust OOD rejection	Strong—explicit OOD/outlier likelihood thresholding	Full synthetic and aspect modeling; fine-tuned on-each patch	Computationally demanding; potential for edge via model compression
Neurosymbolic/OpenHybrid	Early stage; explainable but depends on deep feature pipeline	Potentially strong; supports rule-based rejection and interpretability	None (defers to upstream synthesis)	Lightweight rules, scalable; still research phase

8. Critical analysis and limitations

8.1 *Computational overhead and generalization*

GAN and flow-based hybrid models provide marked increases in robustness but impose substantial computational requirements, making deployment on resource-limited hardware a challenge. Furthermore, current benchmarks offer limited variability in terrain and weather, so generalization to new operational environments remains relatively untested.

8.2 *Robustness trade-offs and lack of certified guarantees*

Some purification approaches risk over-smoothing, potentially eliminating subtle target features. The optimal balance between defense and clean accuracy is context-dependent. Additionally, while PGD-based adversarial training improves empirical resilience, most methods lack formal, certified guarantees against all possible attacks within a given threat model.

9. Conclusion

This review synthesized the state-of-the-art in adversarial robustness for SAR ATR. The analysis shows that hybrid models combining GANs with normalizing flows represent the most advanced and effective defense paradigm, significantly outperforming traditional methods. The primary adversarial threats include digital gradient-based attacks, physical attacks exploiting SAR phenomenology, and transferable black-box attacks. GANs play a multifaceted role by augmenting scarce data, purifying corrupted inputs, and enabling robust feature learning for open-set rejection. The primary limitations of current advanced methods are their high computational cost, the generalization gap beyond standard benchmarks, and the lack of certified robustness guarantees.

In summary, while significant progress has been made, the path to operationally secure SAR ATR requires further innovation. Future investigation into certified, physics-aware robustness, neuro-symbolic reasoning, and adaptive, scalable architectures will be essential for deploying trustworthy SAR ATR systems in real-world operational theaters.

10. Future research directions

10.1 *Certified and physics-aware robustness*

A primary future direction is the development of provable, certified defense frameworks uniquely suited to SAR. This could involve adapting randomized smoothing with speckle-aware noise models or using interval bound propagation for

complex-valued inputs. Embedding physical-layer constraints and radar phenomenology, as explored in physics-informed GANs, promises resilience against both digital and physical-world attacks [16, 38].

10.2 *Scalable, continual, and multi-modal learning*

As SAR datasets grow, architectures must evolve to accommodate real-time, in-field learning. Techniques from continual and meta-learning should be investigated to enable adaptive classifiers that do not catastrophically forget prior knowledge. The integration of multi-polarization SAR, InSAR, and complementary optical or hyperspectral data should also be pursued in future systems, using advanced GANs for cross-domain translation and data alignment.

10.3 *Standardized benchmarks and explainable AI*

The literature surveyed establishes principled mathematical formulations for robust optimization and flow-based likelihood estimation, unifying SAR-ATR adversarial research under a rigorous theoretical umbrella. Furthermore, to enhance interpretability and trust, there is growing interest in neuro-symbolic approaches that combine neural feature extraction with logical reasoning [56, 57]. Such methods are particularly relevant for military applications where explainability is critical. Reproducible and fair comparison would be advanced by the development of a unified benchmark suite comprising multi-sensor datasets, standardized attack protocols, and a public evaluation leaderboard. One of the architectures of this XAI is demonstrated in the Figure 3. For the transition to operational deployments, transparency is paramount. Interpretability, auditability, and trust in SAR ATR decisions would be bolstered by incorporating neuro-symbolic approaches that combine neural feature extraction with logical reasoning [24].

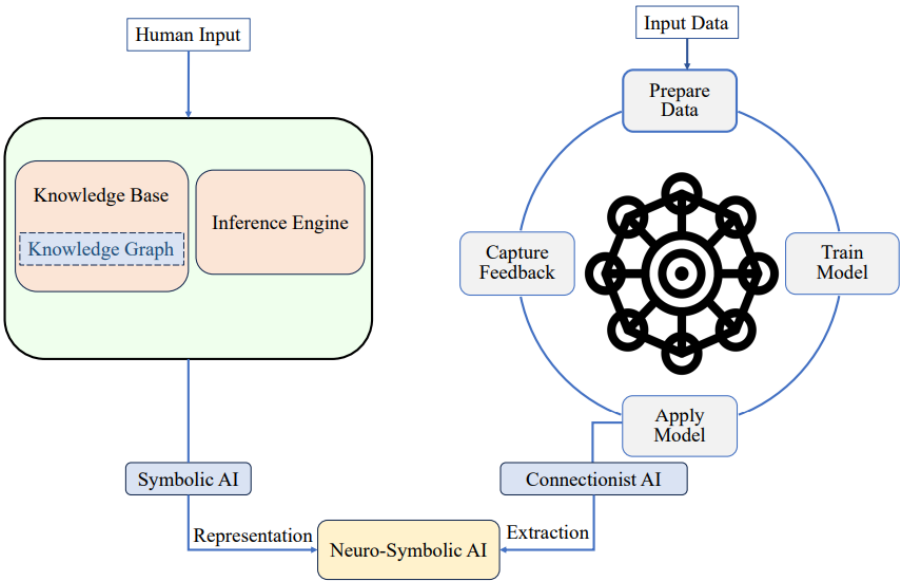

Figure 3. An example of a neuro-symbolic AI architecture. Note that this is one of many possible architectures in the field

Conflict of interest

There is no conflict of interest for this study.

References

[1] S. Z. Gurbuz, H. D. Griffiths, A. Charlish, M. Rangaswamy, M. S. Greco, and K. Bell, "An overview of cognitive radar: past, present, and future," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 12, pp. 6-18, 2019. https://doi.org/10.1109/MAES.2019.2953762.

[2] W. Jiang, Y. Ren, Y. Liu, and J. Leng, "Artificial neural networks and deep learning techniques applied to radar target detection: a review," *Electronics*, vol. 11, no. 1, p. 156, 2022. https://doi.org/10.3390/electronics11010156.

[3] Z. Geng, H. Yan, J. Zhang, and D. Zhu, "Deep-learning for radar: a survey," *IEEE Access*, vol. 9, pp. 141800-141818, 2021. https://doi.org/10.1109/ACCESS.2021.3119561.

[4] S. A. Wagner, "SAR ATR by a combination of convolutional neural network and support vector machines," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 6, pp. 2861-2872, 2016. https://doi.org/10.1109/TAES.2016.160061.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv*, 2017. https://doi.org/10.48550/arXiv.1706.06083.

[6] H. Li, H. Huang, L. Chen, J. Peng, H. Huang, Z. Cui, et al., "Adversarial examples for CNN-based SAR image classification: an experience study," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1333-1347, 2021. https://doi.org/10.1109/JSTARS.2020.3038683.

[7] L. Chen, Z. Xu, Q. Li, J. Peng, S. Wang, and H. Li, "An empirical study of adversarial examples on remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7419-7433, 2021. https://doi.org/10.1109/TGRS.2021.3051641.

[8] Y. Xu, H. Sun, J. Chen, L. Lei, K. Ji, and G. Kuang, "Adversarial self-supervised learning for robust SAR target recognition," *Remote Sensing*, vol. 13, no. 20, p. 4158, 2021. https://doi.org/10.3390/rs13204158.

[9] Y. Guo, L. Du, D. Wei, and C. Li, "Robust SAR automatic target recognition via adversarial learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 716-729, 2021. https://doi.org/10.1109/JSTARS.2020.3039235.

[10] T. Ye, R. Kannan, V. Prasanna, and C. Busart, "Uncertainty-aware SAR ATR: Defending against adversarial attacks via Bayesian neural networks," In Proc. 2024 IEEE Radar Conference (RadarConf24), Denver, CO, USA, May 6-10, 2024. https://doi.org/10.1109/RadarConf2458775.2024.10548693.

[11] D. Gao, X. Wu, and Z. Wen, "ESF-GAN: Physics-driven electromagnetic scattering feature fusion-based GAN for SAR image augmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 21700-21717, 2025. https://doi.org/10.1109/JSTARS.2025.3601055.

[12] F. Gao, Y. Yang, J. Wang, J. Sun, E. Yang, and H. Zhou, "A deep convolutional generative adversarial networks (DCGANS)-based semi-supervised method for object recognition in synthetic aperture radar (SAR) images," *Remote Sensing*, vol. 10, no. 6, p. 846, 2018. https://doi.org/10.3390/rs10060846.

[13] Z. Ren, B. Hou, Q. Wu, Z. Wen, and L. Jiao, "A distribution and structure match generative adversarial network for SAR image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 3864-3880, 2020. https://doi.org/10.1109/TGRS.2019.2959120.

[14] S. Lang, G. Li, Y. Liu, W. Lu, Q. Zhang, and K. Chao, "A GAN-based augmentation scheme for SAR deceptive jamming templates with shadows," *Remote Sensing*, vol. 15, no. 19, p. 4756, 2023. https://doi.org/10.3390/rs15194756.

[15] J. Qin, J. Han, Z. Liu, L. Ran, R. Xie, T.-S. Yeo, "Ganflow: a hybrid model for SAR image target open-set recognition based on GAN and the flow-based module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 7083-7099, 2025. https://doi.org/10.1109/JSTARS.2025.3542738.

[16] X. Zhang, Y. Zhuang, Q. Guo, H. Yang, X. Qian, G. Cheng, et al, "Φ-GAN: Physics-inspired GAN for generating SAR images under limited data," *arXiv*, 2025. https://doi.org/10.48550/arXiv.2503.02242.

[17] J. Wang, J. Li, B. Sun, and Z. Zuo, "SAR image synthesis based on conditional generative adversarial networks," *The Journal of Engineering*, vol. 2019, no. 21, pp. 8093-8097, 2019. https://doi.org/10.1049/joe.2019.0696.

[18] G. F. Araujo, R. Machado, and M. I. Pettersson, "Synthetic SAR data generator using pix2pix cGAN architecture for automatic target recognition," *IEEE Access*, vol. 11, pp. 143369-143386, 2023. https://doi.org/10.1109/ACCESS.2023.3343910.

[19] J. Guo, B. Lei, C. Ding, and Y. Zhang, "Synthetic aperture radar image synthesis by using generative adversarial nets," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1111-1115, 2017. https://doi.org/10.1109/LGRS.2017.2699196.

[20] Z. Zhang, J. Yang, and Y. Du, "Deep convolutional generative adversarial network with autoencoder for semisupervised SAR image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022. https://doi.org/10.1109/LGRS.2020.3018186.

[21] Q. Song, F. Xu, X. X. Zhu, and Y.-Q. Jin, "Learning to generate SAR images with adversarial autoencoder," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022. https://doi.org/10.1109/TGRS.2021.3086817.

[22] Y. Li, R. Fu, X. Meng, W. Jin, and F. Shao, "A SAR-to-optical image translation method based on conditional generation adversarial network (cGAN)," *IEEE Access*, vol. 8, pp. 60338-60343, 2020. https://doi.org/10.1109/ACCESS.2020.2977103.

[23] J. Qin, Z. Liu, L. Ran, R. Xie, J. Tang, and Z. Guo, "A target SAR image expansion method based on conditional wasserstein deep convolutional GAN for automatic target recognition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7153-7170, 2022. https://doi.org/10.1109/JSTARS.2022.3199091.

[24] D. H. Hagos and D. B. Rawat, "Neuro-symbolic AI for military applications," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6012-6026, 2024. https://doi.org/10.1109/TAI.2024.3444746.

[25] Z. Cui, M. Zhang, Z. Cao, and C. Cao, "Image data augmentation for SAR sensor via generative adversarial nets," *IEEE Access*, vol. 7, pp. 42255-42268, 2019. https://doi.org/10.1109/ACCESS.2019.2907728.

[26] Y. Kong and S. Liu, "DMSC-GAN: a c-GAN-based framework for super-resolution reconstruction of SAR images," *Remote Sensing*, vol. 16, no. 1, p. 50, 2023. https://doi.org/10.3390/rs16010050.

[27] W. Ma, Z. Pan, F. Yuan, and B. Lei, "Super-resolution of remote sensing images via a dense residual generative adversarial network," *Remote Sensing*, vol. 11, no. 21, p. 2578, 2019. https://doi.org/10.3390/rs11212578.

[28] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, et al., "ESRGAN: enhanced super-resolution generative adversarial networks," In Proc. ECCV 2018, Munich, Germany, Sep. 8-14, 2018, pp 63-79. https://doi.org/10.1007/978-3-030-11021-5_5.

[29] S. Du, J. Hong, Y. Wang, and Y. Qi, "A high-quality multicategory SAR images generation method with multiconstraint GAN for ATR," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022. https://doi.org/10.1109/LGRS.2021.3065682.

[30] H. Huang, F. Zhang, Y. Zhou, Q. Yin, and W. Hu, "High resolution SAR image synthesis with hierarchical generative adversarial networks," In Proc. 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, Jul. 28-Apr. 2, 2019, pp. 2782-2785. https://doi.org/10.1109/IGARSS.2019.8900494.

[31] C. Wang, J. Pei, X. Liu, Y. Huang, D. Mao, Y. Zhang, et al., "SAR target image generation method using azimuth-controllable generative adversarial network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9381-9397, 2022. https://doi.org/10.1109/JSTARS.2022.3218369.

[32] X. Cao, Z. Zheng, and D. An, "Adaptive waveform selection algorithm based on reinforcement learning for cognitive radar," In Proc. 2019 IEEE 2nd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, Nov. 22-24, 2019, pp. 208-213. https://doi.org/10.1109/AUTEEE48671.2019.9033413.

[33] P. Itkin and N. Levanon, "Ambiguity function based radar waveform classification and unsupervised adaptation using deep CNN models," In Proc. 2019 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS), Tel-Aviv, Israel, Nov. 4-6, 2019, pp. 1-6. https://doi.org/10.1109/COMCAS44984.2019.8958242.

[34] F. Stambouli, M. Limbach, T. Rommel, and M. Younis, "A cognitive synthetic aperture radar concept for tracking and imaging operation," In Proc. 2019 20th International Radar Symposium (IRS), Ulm, Germany, Jun. 26-28, 2019, pp. 1-9. https://doi.org/10.23919/IRS.2019.8768177.

[35] K. Huang, Y. Qu, Z. Zhang, V. Chakravarthy, L. Zhang and Z. Wu, et al., "Software defined radio based mixed signal detection in spectrally congested and spectrally contested environment," In Proc. 2017 Cognitive Communications for Aerospace Applications Workshop (CCAA), Cleveland, OH, USA, Jun. 27-28, 2017, pp. 1-6. https://doi.org/10.1109/CCAAW.2017.8001889.

[36] E. V. Carrera, F. Lara, M. Ortiz, A. Tinoco, and R. Leon, "Target detection using radar processors based on machine learning," In Proc. 2020 IEEE ANDESCON, Quito, Ecuador, Oct. 2020, pp. 1-5. https://doi.org/10.1109/ANDESCON50619.2020.9272173.

[37] C. Zheng, X. Jiang, and X. Liu, "Semi-supervised SAR ATR via multi-discriminator generative adversarial network," *IEEE Sensors Journal*, vol. 19, no. 17, pp. 7525-7533, 2019. https://doi.org/10.1109/JSEN.2019.2915379.

[38] X. Zhang, Y. Zhuang, Q. Guo, H. Yang, X. Qian, G. Cheng, et al., "Φ-GAN: physics-inspired GAN for generating SAR images under limited data," *arXiv*, 2025. https://doi.org/10.48550/arXiv.2503.02242.

[39] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 45-49, 2019. https://doi.org/10.1109/LGRS.2018.2868704.

[40] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on PCANet," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1792-1796, 2016. https://doi.org/10.1109/LGRS.2016.2611001.

[41] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 1, pp. 125-138, 2016. https://doi.org/10.1109/TNNLS.2015.2435783.

[42] H. Li, C. Gu, D. Wu, G. Cheng, L. Guo and H. Liu, "Multiscale generative adversarial network based on wavelet feature learning for SAR-to-optical image translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022. https://doi.org/10.1109/TGRS.2022.3211415.

[43] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1220-1224, 2019. https://doi.org/10.1109/LGRS.2019.2894734.

[44] W.-L. Du, Y. Zhou, J. Zhao, and X. Tian, "K-means clustering guided generative adversarial networks for SAR-optical image matching," *IEEE Access*, vol. 8, pp. 217554-217572, 2020. https://doi.org/10.1109/ACCESS.2020.3042213.

[45] J. Oh and M. Kim, "PeaceGAN: a GAN-based multi-task learning method for SAR target image generation with a pose estimator and an auxiliary classifier," *Remote Sensing*, vol. 13, no. 19, p. 3939, 2021. https://doi.org/10.3390/rs13193939.

[46] S. Oghim, Y. Kim, H. Bang, D. Lim, and J. Ko, et al., "SAR image generation method using DH-GAN for automatic target recognition," *Sensors*, vol. 24, no. 2, p. 670, 2024. https://doi.org/10.3390/s24020670.

[47] X. Mao, X. He, and D. Li, "Knowledge-aided 2-d autofocus for spotlight SAR range migration algorithm imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5458-5470, 2018. https://doi.org/10.1109/TGRS.2018.2817507.

[48] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, B. Lakshminarayanan, "Do deep generative models know what they don't know?" *arXiv*, 2019. https://doi.org/10.48550/arXiv.1810.09136.

[49] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Why normalizing flows fail to detect out-of-distribution data," In Proc. 34th International Conference on Neural Information Processing Systems, Vancouver, BC Canada, Dec. 6-12, 2020, pp. 20578-20589.

[50] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," In Proc. 5th International Conference on Learning Representations, Toulon, France, Apr. 24-26, 2017.

[51] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," In Proc. ECCV 2020, Glasgow, UK, Aug. 23-28, 2020, pp. 102-117. https://doi.org/10.1007/978-3-030-58580-8_7.

[52] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," In Proc. 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, Dec. 3-8, 2018, pp. 7167-7177.

[53] A. Bendale and T. E. Boult, "Towards open set deep networks," In Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 27-30, 2016, pp. 1563-1572. https://doi.org/10.1109/CVPR.2016.173.

[54] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahda, A. Anandkumar, "Diffusion models for adversarial purification," *arXiv*, 2022. https://doi.org/10.48550/arXiv.2205.07460.

[55] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," In Proc. 35th International Conference on Neural Information Processing Systems, Virtual, Dec. 6-14, 2021. pp. 8780-8794.

[56] M. Hersche, M. Zeqiri, L. Benini, A. Sebastian, and A. Rahimi, "A neuro-vector-symbolic architecture for solving raven's progressive matrices," *Nature Machine Intelligence*, vol. 5, pp. 363-375, 2023. https://doi.org/10.1038/s42256-023-00630-8.

[57] A. d'Avila Garcez and L. C. Lamb, "Neurosymbolic AI: the 3rd wave," *Artificial Intelligence Review*, vol. 56, pp. 12387-12406, 2020. https://doi.org/10.1007/s10462-023-10448-w.