MVAE-ACGAN: A Generative Framework for Imbalanced Few-shot Multi-band SAR Target Detection

Journal:	*IEEE Transactions on Aerospace and Electronic Systems*
Manuscript ID	TAES-2026-0048
Manuscript Type:	Regular Paper
Date Submitted by the Author:	05-Jan-2026
Complete List of Authors:	Zhou, Zihan; Beihang University School of Electronic and Information Engineering, School of Electronic and Information Engineering Huaping, Xu; Beihang University School of Electronic and Information Engineering Yao, Xianxun; Beihang University School of Electronic and Information Engineering Liu, Wei; The Hong Kong Polytechnic University Department of Electrical and Electronic Engineering Liu, Xianghua; Houtang Technology Development Co Ltd
Keywords:	Radar target recognition, Synthetic Aperture Radar Data

SCHOLARONE™
Manuscripts

MVAE-ACGAN: A Generative Framework for Imbalanced Few-shot Multi-band SAR Target Detection

Zihan Zhou
School of Electronic and Information Engineering, Beihang University, Beijing 100191, China

Huaping Xu, Member, IEEE
School of Electronic and Information Engineering, Beihang University, Beijing 100191, China

Xianxun Yao, Member, IEEE
School of Electronic and Information Engineering, Beihang University, Beijing 100191, China

Wei Liu, Senior, IEEE
Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Xianghua Liu
Houtang Technology Development Co., Ltd., Macao, China

***Abstract*— Multi-band Synthetic Aperture Radar (SAR) provides complementary scattering information of targets across different frequencies, which is of great value for target detection in a complex background. However, the multi-band SAR dataset characterized by scarce targets and abundant backgrounds constitutes an imbalanced few-shot problem for target detection, which significantly limits the effectiveness of deep learning based detectors. To address this imbalanced few-shot learning challenge, a Modified Variational Autoencoder-Auxiliary Classifier Generative Adversarial Network (MVAE-ACGAN) framework is proposed for generating high-quality multi-band SAR target imagery in this paper. The proposed framework introduces three improvements to VAE-ACGAN: a geometric transformation based pre-augmentation module, which preliminarily expands the training target dataset to prevent mode collapse in the generative network; a modified reconstruction loss including the band-weighted reconstruction pixel loss and the reconstruction statistical loss, which ensures the cross-band variability and regularity of generated samples; and a feature perceptual mechanism with a staged training strategy in the classification branch of an ACGAN-based discriminator, which enforces similarity between generated samples and real samples in a discriminative feature space. Collectively, these improvements enable the proposed framework to generate samples that not only maintain inter- and intra-band authenticity but also remain separable from the background, thereby supplying high-quality training samples for subsequent target detection networks. Experimental results on a self-built multi-band SAR dataset show that the proposed MVAE-ACGAN framework outperforms existing generative approaches in terms of reconstruction and generation quality, and significantly improves the detection performance of the deep learning detector, demonstrating excellent results across multiple evaluation metrics.**

***Index Terms*— ACGAN, VAE, multi-band SAR, target detection, few-shot learning**

I. Introduction

TARGET detection, as one of the core tasks of synthetic aperture radar (SAR), serves as the technical foundation for various SAR applications, including military reconnaissance, disaster assessment, and environmental monitoring. However, the conventional SAR system faces a bottleneck in target detection under a complex background, as manifested by its difficulty in achieving both high detection probability and low false alarm probability [1], due to the limitation of the single-band SAR system in capturing comprehensive information of target and background. Multi-band SAR has emerged as an advanced alternative, which acquires multiple SAR images through radar signals of distinct frequencies, and expands the target information acquired in the frequency dimension, enabling a more comprehensive characterization of targets [2].

Existing studies on multi-band SAR target detection mainly focus on extracting discriminative handcrafted features from each band and integrating them into a joint detector. Typical examples include band-transferable texture representations for foliage-penetrating target [3], statistical feature fusion for target covered by camouflage netting [4], scattering-mechanism combinations for artificial targets in urban areas [5], and geometric and tracking-based features for road targets [6]. While these methods benefit from strong physical interpretability by designing features based on inherent properties of different targets, their performance is limited by two key factors. First, handcrafted features have limited capacity to capture the nonlinear and high-dimensional discriminative information that distinguishes targets from backgrounds in complex scenes [7]. Second, the judgment threshold of the detector relies on prior knowledge, lacking the capability to adaptively learn the optimal decision boundary directly from data [8]. As a result, when applied to diverse and previously unencountered scenarios, these traditional methods often fall short in meeting the increasingly stringent performance requirements in terms of both robustness and detection accuracy [9].

Deep learning, as a powerful technique originating from computer vision, has demonstrated remarkable ca-

This work was supported by the National Natural Science Foundation of China (U2241202) and Science and Technology Development Fund (FDCT) of Macau Special Administrative Region. (Corresponding author: Xianxun Yao.)

Zihan Zhou, Huaping Xu, and Xianxun Yao are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: zhouzihan98@buaa.edu.cn; xuhuaping@buaa.edu.cn; xianxun.yao@buaa.edu.cn). Wei Liu is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: wei2.liu@polyu.edu.hk). Xianghua Liu is with Houtang Technology Development Co., Ltd., Macao, China (e-mail: lewis.liu@houtang.net).

0018-9251 © IEEE

https://mc.manuscriptcentral.com/taes

pability in learning complex features and nonlinear relationships through end-to-end training. It has been widely applied to single-band SAR target detection [10], achieving significant performance gains compared to traditional non-learning methods. However, the effectiveness of deep learning methods is highly dependent on large-scale and high-quality labeled datasets, which poses a significant challenge in multi-band SAR target detection. Specifically, while background samples are generally abundant, high-value target multi-band samples are extremely scarce and costly to annotate [11], which renders multi-band SAR target detection a typical imbalanced few-shot problem. Consequently, in such a few-shot multi-band scenario, deep learning models struggle to learn robust and generalizable representations, resulting in overfitted models with poor performance [12].

For the few-shot learning problem, data augmentation has become a vital strategy for expanding the dataset to improve model generalization and robustness, ultimately enhancing detection performance. Existing augmentation methods can be broadly categorized into two types. The first type is based on geometric transformations, including rotation, flipping, noise addition, and hybrid manipulations [13], which augment data by applying pixel-level operations to existing samples. While simple and efficient, these methods are inherently limited to rule-based and shallow transformations without learning the underlying scattering structures or feature distributions of the real data. Consequently, they can only modify the apparent appearance of the samples and fail to generate new semantic content or essential discriminative features, thereby constraining the diversity of the generated samples and limiting the overall improvement in model discriminability [14]. The second type is based on deep generative models, which learn complex mappings from a latent space to the real data distribution through neural networks [15]. In contrast to the pixel-level manipulations of geometric transformation based augmentation methods, generative models create new samples at the data distribution level. The generated samples not only appear visually realistic but also possess discriminability and diversity, effectively addressing overfitting, poor robustness, and limited generalization associated with few-shot problems of deep learning, and improving detection performance [16] [17].

Variational Autoencoder (VAE) is one of the earliest generative models that established a fundamental connection between variational inference and deep neural networks [18] . By introducing a latent variable distribution constraint into the autoencoder, the VAE achieves explicit modeling of the data generation process, enabling the sampling of new samples from the latent space. Unlike conventional autoencoders, VAE benefits from an interpretable and continuous latent space, enabling the generation of diverse samples. However, the generated results of VAE often lack fine details and appear overly smooth, since the model is trained mainly through reconstruction error minimization. Various modifications have been proposed to enrich the details of generated images: β-VAE [19], which introduces a regularization coefficient to strengthen the disentanglement of latent variables; CVAE [20], which incorporates class conditions into the encoding and decoding processes; perceptual loss [21], which constrains the reconstruction in a high-level feature space to preserve semantic consistency and mitigate the blurring artifacts of generated images. Although these VAE variants alleviate the blurring and detail-loss issues to some extent, they still struggle to generate high-fidelity structures for complex multi-band SAR imagery.

In contrast, the Generative Adversarial Network (GAN) approximates the real data distribution implicitly through an adversarial training between a generator and a discriminator [22]. GANs is excellent at recovering high-frequency details and textures of images, producing visually superior results compared to VAE. However, the training process of GAN is inherently unstable, often suffering from mode collapse and gradient instability [23], which can impede convergence, especially in few-shot learning scenarios. Consequently, many variants have been developed to improve model stability and controllability. Wasserstein GAN [24] replaces the original JS-divergence with the Wasserstein distance, substantially improving training stability and mitigating mode collapse; InfoGAN [25] introduces mutual information constraints to disentangle the latent space, thereby enhancing the interpretability and controllability of the generation process; Conditional GAN (CGAN) [26] utilized label information to guide the generation of multi-class targets; Auxiliary Classifier GAN (ACGAN) [27] further enhances CGAN by incorporating an classifier to improve the inter-class separability of the generated samples. Considering that the stability of VAE and the visual realism of GAN are complementary, hybrid frameworks like VAE-GAN [28] and VAE-ACGAN [29] have been proposed, which integrate the latent space modeling of VAE with GAN-based adversarial discrimination to achieve the stable generation of high-quality samples, offering a promising approach to addressing few-shot learning challenges.

Although generative models have made significant progress in single-band SAR target detection tasks [30] [31] [32], no prior work has explored their extension to multi-band SAR, leaving a clear research gap. These methods face several fundamental limitations when directly applied to multi-band SAR. First, due to the different imaging parameters across multiple frequencies, multi-band SAR imagery exhibit substantial variations in appearance, including difference in texture patterns, structural expression and intensity distributions. However, most existing generative models are fundamentally designed for single-channel imagery and therefore lack the ability to accommodate such cross-band diversity. When directly extended to multi-band SAR data, these models tend to learn an averaged representation across bands rather than preserving band specific properties, resulting in generated samples that fail to retain the intrinsic inter-band variability present in real multi-band SAR imagery.

https://mc.manuscriptcentral.com/taes

Second, the scattering difference across bands in multi-band SAR imagery reflects the frequency-dependent characteristics of target, proving important information for distinguishing targets from complex background. However, existing single-channel generative models lack mechanisms to enforce cross-band consistency in these characteristics. Consequently, the generated samples may exhibit distorted or mismatched inter-band relationships. Moreover, because existing generative models are not specifically designed for SAR imagery, the generated samples, even when visually plausible, may still fail to preserve critical SAR specific structural or semantic features that can distinguish the targets from the complex background. These limitations collectively hinder existing generative models from effectively addressing the imbalanced few-shot learning problem in multi-band SAR target detection.

To address the above challenges, a Modified Variational Autoencoder Auxiliary Classifier Generative Adversarial Network (MVAE-ACGAN) framework for multi-band SAR data augmentation is proposed in this paper. Three key improvements are incorporated into the VAE-ACGAN framework to generate multi-band SAR images for multi-band SAR target detection. First, a data pre-augmentation process based on geometric transformations is employed to preliminarily expand the limited target sample set. Second, the VAE module is redesigned with a band-weighted pixel reconstruction loss and a statistical reconstruction loss, enabling the model to better preserve band-specific structural variability and maintain the inter-band relationships of multi-band target imagery. Third, a feature extraction interface is incorporated into the classification branch of ACGAN to compute a feature perceptual loss, which is progressively integrated through a staged training strategy to ensure stable optimization and more faithful reconstruction of deep semantic features.

Furthermore, a multi-band SAR dataset for vehicle target detection in urban environments is constructed in this work, using airborne SAR images acquired at P-, S-, C- and Ka-bands. Experiments on this dataset demonstrated that, compared to existing data augmentation methods, the proposed MVAE-ACGAN framework significantly enhances target detection performance under imbalanced few-shot dataset, effectively validating its feasibility and effectiveness in addressing the few-shot problem in multi-band SAR target detection tasks.

The rest of this article is organized as follows. Section II introduces the foundational background on VAE, GAN and VAE-ACGAN. Section III details the proposed MVAE-ACGAN framework and its three key improvements, including pre-augmentation, the modified reconstruction loss of VAE, and the feature perceptual mechanism with staged training strategy in an ACGAN-based discriminator. In Section IV, effectiveness of the proposed framework is demonstrated by experimental results on a self-built multi-band SAR target detection dataset. Conclusions are drawn in Section V.

II. Related Work

In this section, the fundamental generative models relevant to this study are briefly reviewed. The basic principles of VAE and GAN are introduced, followed by their hybrid architecture, the VAE-ACGAN.

A. Variational Autoencoder

VAE is a generative model that combines deep learning with variational inference, whose fundamental principle is to reconstruct and generate data through an encoding-decoding process. Specifically, the encoder $E(\cdot)$ maps an input sample x to the distribution parameters of a latent variable z, denoted as $E(z|x) = N(z; \mu(x), \sigma^2(x))$, where $N(z)$ represents a Gaussian distribution parameterized by the mean and variance defined by $\mu(x)$ and $\sigma^2(x)$. The decoder then samples from the latent space to reconstruct the input data, expressed as $\hat{x} = R(x|z)$. The loss function of the VAE consists of a reconstruction loss and a Kullback-Leibler (KL) divergence loss, and is formulated as follows [18]:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(E,R) &= \mathcal{L}_{\text{rec}}(E,R) + \mathcal{L}_{\text{KL}}(E,R) \\ &= \mathbb{E}_{E(z|x)}[\log R(x|z)] - \text{KL}[E(z|x)\|P_Z(z)]\end{aligned} \tag{1}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator, and $KL[\cdot\|\cdot]$ represents the KL divergence between two distributions. The reconstruction loss $\mathcal{L}_{rec}(E,R)$ constrains the reconstructed sample $\hat{x}$ to be consistent with the input real sample x in the pixel space. The KL divergence loss $\mathcal{L}_{KL}(E,R)$ encourages the distribution of the latent variable z produced by the encoder to approximate a prior distribution $P_Z(z)$, thereby ensuring continuity of the latent space.

To promote consistency between the generated samples and the real samples in feature space, the Feature Perceptual Loss is introduced in [21], which computes the loss based on the feature differences between the real and reconstructed samples extracted by a feature extraction operator $F(\cdot)$. It can be formulated as follows:

$$\mathcal{L}_{fea}(E,R) = \|F(x) - F(\hat{x})\|_2^2 \tag{2}$$

Then, the loss of VAE is the sum of the reconstruction loss, KL divergence loss and feature perception loss and can be denoted as:

$$\mathcal{L}_{\text{VAE}}(E,R) = \mathcal{L}_{\text{rec}}(E,R) + \mathcal{L}_{\text{KL}}(E,R) + \mathcal{L}_{\text{fea}}(E,R) \tag{3}$$

B. Generative Adversarial Network

The GAN consists of a generator and a discriminator and achieves data generation through adversarial training. The generator $G(z)$ learns a mapping function to transform an input random noise $z \sim P_z(z)$ vector into the sample space. The discriminator $D(x)$ distinguishes whether an input sample is a real sample $x \sim P_{data}(x)$ or one generated by the generator. The adversarial training

https://mc.manuscriptcentral.com/taes

between G and D can be formulated as the following min-max game [22]:

$$\min_G \max_D \mathcal{L}_{GAN}(G,D) \tag{4}$$

where

$$\begin{aligned}&\mathcal{L}_{GAN}(G,D) = \\ &\mathbb{E}_{x\sim P_{data}}[\log D(x)] + \mathbb{E}_{z\sim P_Z(z)}[\log(1-D(G(z)))]\end{aligned} \tag{5}$$

The first and second terms of (5) correspond to the discriminator losses for real data and for generated samples from random noise, respectively. During training, the discriminator D aims to accurately distinguish real samples from generated ones by assigning labels of 1 and 0, respectively, and therefore seeks to maximize $\mathcal{L}_{GAN}(G,D)$. In contrast, the generator G attempts to minimize the second term of $\mathcal{L}_{GAN}(G,D)$ so that the generated samples can resemble real ones and deceive the discriminator into labeling them as 1. Through this adversarial training process, the generator gradually learns the underlying distribution of real samples, enabling it to produce diverse and realistic samples from random noise to augment the dataset.

ACGAN is a variant of GAN designed for multi-class target generation, which extends the aforementioned GAN framework by introducing a class-conditional variable and an additional classification branch [26]. The generator is reformulated into a conditional generation model $G(z,c)$, which takes both random noise and class labels as input to guide the generation of class-specific samples. The discriminator incorporates both an adversary branch $D_{adv}(x)$ for authenticity assessment and a classification branch $D_{cls}(c|x)$ for category prediction. The overall loss function of ACGAN can be expressed as:

$$\begin{aligned}\mathcal{L}_{ACGAN}(G,D) &= \mathcal{L}_{adv}(G,D) + \lambda\mathcal{L}_{cls}(G,D)\\ \mathcal{L}_{adv}(G,D) &= \mathbb{E}_{x\sim P_{data}(x)}[\log D_{adv}(x)]\\ &\quad + \mathbb{E}_{z\sim P_Z(z),c\in C}[\log(1-D_{adv}(G(z,c)))]\\ \mathcal{L}_{cls}(G,D) &= \mathbb{E}_{x,c\sim P_{data}(x,c)}[-\log D_{cls}(c|x)]\\ &\quad + \mathbb{E}_{z\sim P_Z(z),c\in C}[-\log D_{cls}(c|G(z,c))]\end{aligned} \tag{6}$$

where $\mathcal{L}_{adv}(G,D)$ and $\mathcal{L}_{cls}(G,D)$ denotes the loss functions of $D_{adv}(x)$ and $D_{cls}(c|x)$, respectively.

C. VAE-ACGAN

To integrate the training stability of the VAE with the high visual fidelity and class-conditional generation capability of the ACGAN, the VAE-ACGAN framework was proposed [28], which replaces the standard ACGAN generator with a VAE model and trains the encoder, decoder and discriminators jointly through an adversarial objection. The loss function of VAE-ACGAN can be formulated as:

$$\mathcal{L}_{VAE\text{–}ACGAN}(E,R,D) = \mathcal{L}_{VAE}(E,R) + \mathcal{L}_{ACGAN}(E,R,D) \tag{7}$$

where $\mathcal{L}_{ACGAN}(E,R,D)$ represents the adversarial loss of VAE introduced by incorporating a discriminator D with an adversary branch and a classification branch.

These generation models offer promising solutions to the few-shot problem in multi-band SAR target detection. Nevertheless, directly applying these methods to multi-band SAR data is not effective in solving the few-shot learning problem. Existing single-channel models tend to generate average representations across different bands, ignoring the inherent diversity between them. Moreover, they lack dedicated mechanisms to enforce consistency in cross-band characteristics. Consequently, they fail to produce multi-band SAR samples that are authentic both inter-band and intra-band. These limitations motivate the development of a specialized generation framework tailored to the imbalanced few-shot learning challenge in multi-band SAR target detection.

III. MVAE-ACGAN

To address the limitation of existing generative models when applied to the multi-band SAR imbalanced few-shot target detection task, a Modified Variational Autoencoder Auxiliary Classifier Generative Adversarial Network (MVAE-ACGAN) framework is proposed in this section. The proposed framework incorporates three key improvements over the VAE-ACGAN paradigm, including a geometric transformation-based pre-augmentation module, a modified reconstruction loss of VAE with band-weighted pixel loss term and statistical loss term, and a staged training feature-perceptual mechanism implemented by a deep feature interface in classification branch. These improvements collectively guide the proposed MVAE-ACGAN to generate multi-band SAR samples that exhibit authenticity both inter- and intra-band at the pixel-level, statistical-level and the deep feature-level.

The architecture of the proposed method is illustrated in Fig. 1. First, the training data set is divided into target and background sets. The target set undergoes the geometric transformation-based pre-augmentation process to initially expand the dataset. The augmented target samples are then fed into the VAE for multi-band target sample reconstruction and generation, during which the modified reconstruction loss and the KL divergence loss are computed between the reconstructed and input samples. Next, the generated target samples, along with real target samples and real background samples, are input into the ACGAN-based discriminator. Its adversarial branch receives both generated and real target samples, outputs a real/fake judgment, and computes the adversarial loss. The classification branch receives all target and background samples, performs a target/background classification, and yields the classification loss. At a later training phase, when the classification branch has gradually converged, an additional feature perceptual loss is calculated based on the deep features extracted by the feature interface of the classification branch from real and reconstructed target samples. Throughout training, these losses jointly guide the learning of the generator, ultimately producing generated target samples that exhibit authenticity both inter- and intra-band.

https://mc.manuscriptcentral.com/taes

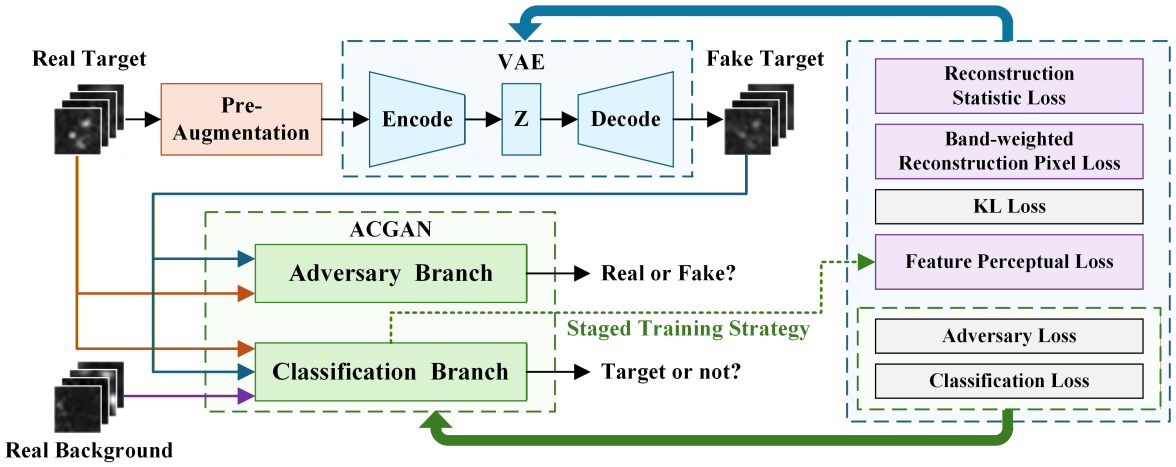


Fig. 1. Overall architecture of the proposed MVAE-ACGAN framework.

The details of three improvements in the proposed MVAE-ACGAN framework are described as follows.

A. Pre-Augmentation Module

For extremely limited samples, directly training a complex generative network often leads to mode collapse or unstable convergence. To alleviate this issue and improve robustness of the training process, a data pre-augmentation module based on geometric transformation is added prior to the generative model.

This module takes the multi-band SAR training target samples $\mathbf{x}$ as input and performs a series of transformations, including rotation $Rot(\mathbf{x})$, flipping $Flip(\mathbf{x})$, noise injection $Noi(\mathbf{x})$ and samples mixing $Mix(\mathbf{x})$ to construct an augmented target training set as:

$$\mathbf{X}_{aug} = \{\mathbf{x}, Rot(\mathbf{x}), Flip(\mathbf{x}), Noi(\mathbf{x}), Mix(\mathbf{x})\} \quad (8)$$

This pre-augmentation module efficiently expands the number of target samples by five times, which mitigated the instability in network training under few-shot conditions. Nevertheless, as the operations are restricted to pixel-level geometric transformations without modeling the intrinsic intra- and inter-band characteristics in multi-band SAR data, the diversity of generated samples remains bounded by the input sample distribution. Therefore, to achieve deeper and more representative generative enhancement, a VAE with modified reconstruction loss is introduced to learn the underlying distribution of multi-band SAR target samples.

B. VAE Module with Modified Reconstruction Loss

To effectively capture both the inter- and intra-band characteristics of multi-band SAR targets, a multi-channel VAE with a modified reconstruction loss is designed. The input to the encoder is the augmented multi-band SAR target sample $\mathbf{x}_{aug} \in \mathbb{R}^{B\times H\times W}$, where B, H and W denote the number of bands, height and width of sample patch, respectively. The encoder is implemented using a multi-channel convolutional neural network (CNN), in which shared convolutional kernels are employed to extract cross-band features and to estimate the distribution parameters of multi-dimensional latent variables $\mathbf{z}$ as follows:

$$\boldsymbol{\mu}, \boldsymbol{\sigma}^2 = E(\mathbf{x}_{aug}) \quad (9)$$

where $\boldsymbol{\mu} \in \mathbb{R}^K$ and $\boldsymbol{\sigma}^2 \in \mathbb{R}^K$ denote the mean vector and the variance vector of $\mathbf{z} \in \mathbb{R}^K$, and K denotes the dimensional of the latent space.

Then, reparameterization is employed to obtain the multi-dimensional latent variables $\mathbf{z}$:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}^2 \odot \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, \mathbf{I}) \quad (10)$$

where $\odot$ denotes the Hardamard product, $N(0, \mathbf{I})$ represents the $K-$ dimensional normal distribution and $\mathbf{I}$ is the identity matrix.

Subsequently, the decoder reconstructs the target sample from the latent space, which can be expressed as:

$$\hat{\mathbf{x}} = R(\mathbf{z}) \quad (11)$$

The original loss function of the VAE, as in (1), consists of the reconstruction loss and the KL divergence loss. For multi-band SAR images, system resolutions and the target scattering characteristics vary among different bands, leading to cross-band variations in images. These cross-band variations, which exhibit distinct regularities in target and background, serve as an important source of discriminative information for multi-band SAR target detection. However, simply extending the existing VAE model to multiple channels for generating multi-band SAR sample leads to indiscriminate learning across bands due to the uniform reconstruction loss, ignoring cross-band variability and regularity. Therefore, in this work, the reconstruction loss is redefined as a band-weighted reconstruction pixel loss term and a reconstruction statistical loss term to constrain the cross-band variability and regularity of the generated multi-band SAR target samples.

https://mc.manuscriptcentral.com/taes

1. Band-weighted Reconstruction Pixel Loss

The reconstruction loss of VAE refers to the pixel-level difference between the generated and the real samples, which can be regarded as the reconstruction pixel loss. Considering the varying roles of different bands in image generation, a weight coefficient is introduced for the reconstruction pixel loss of each band, differentially emphasizing their respective contributions to target characteristics at the pixel level. Specifically, high-resolution band, like Ka-band, which contains more spatial information and is crucial for target structure and texture, is assigned a higher weight. In contrast, low-resolution band, like P-band, which provides richer statistical features but weaker spatial information, is assigned a lower weight. Therefore, the band-weighted reconstruction pixel loss in VAE can be expressed as:

$$\mathcal{L}_{pixel} = \sum_{b=1}^{B} \lambda_b \left\| x_b - \hat{x}_b \right\|_2^2 \tag{12}$$

where, x_b and $\hat{x}_b$ represent the input and reconstructed samples of the b-th band, respectively, and λ_b denotes the weight for the b-th band. By introducing the band weight, higher attention is given to high-resolution bands during pixel-level reconstruction, thereby guiding the VAE to generate multi-band SAR target images that preserve the cross-band structural discrepancies present in real target sample.

2. Reconstruction Statistical Loss

The statistical properties of multi-band SAR imagery, such as mean and variance, describe the scattering energy distribution of the target and background in multiple bands. These statistics exhibit band-dependent variation patterns that are closely related to the physical characteristics of an object, serving as important discriminative features for effectively distinguishing target from complex background. Therefore, preserving the regularity of cross-band statistics in multi-band reconstructed samples is particularly crucial for generating target samples with inter-band authenticity. Accordingly, a reconstruction statistical loss is introduced as follows, focusing on the means and variances of multi-band SAR images.

$$\mathcal{L}_{stat} = \sum_{b=1}^{B} \left(\left\| \mu_b - \hat{\mu}_b \right\|_2^2 + \left\| \sigma_b^2 - \hat{\sigma}_b^2 \right\|_2^2 \right) \tag{13}$$

where μ_b and σ_b^2 represent the mean and variance of the input sample at the b-th band, while $\hat{\mu}_b$ and $\hat{\sigma}_b^2$ denote the mean and variance of the reconstructed sample at the same band. By enforcing consistency in inter-band statistics, this loss guides the VAE to generate samples that better preserve the cross-band energy distribution of real targets.

Integrating these loss terms, the modified loss function for the VAE can be expressed as:

$$\mathcal{L}_{VAE} = \alpha_1 \mathcal{L}_{pixel} + \alpha_2 \mathcal{L}_{stat} + \alpha_3 \mathcal{L}_{KL} \tag{14}$$

where $\alpha_i, i = 1, 2, 3$ is the coefficient that balances the contributions of the band-weighted pixel reconstruction loss, reconstruction statistical loss and KL divergence loss in guiding the learning direction of the network. This modified VAE module prevents the multi-channel generative network from treating images cross bands uniformly. Instead, it enables band-dependent preservation of spatial structural details of targets according to the imaging parameters, while simultaneously maintaining the regularity of statistical properties across multiple bands.

C. Feature Perceptual Mechanism with a Staged Training Strategy in ACGAN-based Discriminator

This work focuses on the multi-band SAR target detection task under imbalanced few-shot learning problem, aiming to separate targets from complex background. Therefore, an ACGAN-based discriminator, which incorporates an adversarial branch D_{adv} and a classification branch D_{cls}, is selected to assist the VAE generator in producing target samples with enhanced realistic details and discriminative feature through adversarial training. The adversarial branch D_{adv} takes both the real target samples $\mathbf{x}$ and the generated target samples $\hat{\mathbf{x}}$ from VAE as inputs, and outputs the probability that a given sample is real. The loss function of D_{adv} can be expressed as $\mathcal{L}_{adv}$ in (6), where the generator $G(z, c)$ is replaced by $R(\mathbf{z})$.

The classification branch D_{cls} takes real target sample $\mathbf{x}$, generated target sample $\hat{\mathbf{x}}$, and real background sample $\mathbf{y}$ as inputs, and performs a binary classification task to distinguish targets from background without differentiating between real and generated samples. Its loss function can be expressed as a binary cross-entropy loss:

$$\begin{aligned} \mathcal{L}_{cls} = & \mathbb{E}_{\mathbf{x}\sim\mathbf{P_{tar}(x)}}[-\log D_{cls}(1|\mathbf{x})] \\ & + \mathbb{E}_{\mathbf{y}\sim\mathbf{P_{back}(y)}}[-\log D_{cls}(0|\mathbf{y})] \\ & + \mathbb{E}_{\mathbf{z}\sim\mathbf{P_Z(z)}}[-\log D_{cls}(1|R(\mathbf{z})] \end{aligned} \tag{15}$$

where $P_{back}(\mathbf{y})$ denotes the distribution of multi-band SAR background sample $\mathbf{y}$ which is labeled as 0, and the target sample is labeled as 1.

In such an ACGAN-based discriminator, which performs both real or generated target discrimination and target or background classification, a feature perceptual mechanism is introduced to enhance the discriminative feature in generated multi-band target sample and facilitate separation from the background. Unlike previous approaches that compute feature perceptual loss using feature extracted from the adversarial of GAN [28], a feature extraction interface D_{cls}^{fea} is incorporated into the intermediate layers of the classification branch that distinguishes targets from backgrounds, as illustrated in Fig. 2. This modification ensures that the extracted features are deep representations specifically tailored for the target detection task and exhibit substantial distinction from background features.

This interface extracts high-level semantic features from the intermediate layer of D_{cls}, which effectively distinguish targets from background and exhibit strong discriminability. Accordingly, the feature perceptual loss

https://mc.manuscriptcentral.com/taes

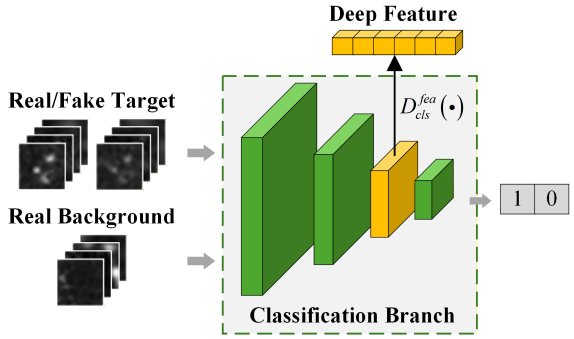


Fig. 2. Classification branch with feature extraction interface.

of the VAE under this mechanism can be formulated as:

$$\mathcal{L}_{fea} = \left\| D_{cls}^{fea}(\mathbf{x}) - D_{cls}^{fea}(\hat{\mathbf{x}}) \right\|_2^2 \tag{16}$$

By introducing this feature perceptual mechanism, the distribution of the generated target samples in the feature space is further constrained. Specifically, the discriminative features relevant for target detection are emphasized during the VAE reconstruction process. Since the VAE generation is now guided by features derived from a network trained jointly on both target and background samples, the generated target samples maintain high separability from the background even though the VAE generator is trained only on target samples. This high separability is particularly crucial for the imbalanced few-shot learning problem in multi-band SAR target detection, where target samples are extremely scarce compared to abundant background samples.

Since the classification branch D_{cls} is dynamically trained and its feature extraction capability is unstable at the early training stage, applying the feature perceptual loss too early may lead to unstable feature distributions and interfere with the joint optimization of the VAE and discriminator. Therefore, a staged training strategy is adopted for this feature perceptual loss to ensure effective learning of both the VAE and the discriminator.

Specifically, in the initial stage, which occupies the first η proportion of the total training epochs, the proposed framework is trained without the feature perceptual loss, and the generator and discriminator are optimized adversarially to improve their respective capabilities. During this stage, the VAE is guided by dual objective. Internally, it learns to reconstruct the cross-band variability and regularity of multi-band target samples through the band-weighted reconstruction pixel loss, the statistical reconstruction loss, and the KL divergence loss. Externally, it is driven by the adversarial loss and the classification loss from the ACGAN-based discriminator, which together ensure that the generated target samples are visually realistic and correctly classifiable as targets. The overall optimization objective at this stage is expressed as:

$$\begin{aligned}\min_{E,R,D_{cls}} \max_{D_{adv}} (&\alpha_1 \mathcal{L}_{pixel} + \alpha_2 \mathcal{L}_s tat + \alpha_3 \mathcal{L}_{KL} \\ &\varphi_1 \mathcal{L}_{adv} + \varphi_2 \mathcal{L}_{cls})\end{aligned} \tag{17}$$

where φ_1 and φ_2 are weighting coefficients that balance the contributions of the adversarial loss term and the classification loss term.

As the training progresses, the classification branch D_{cls} gradually converges and learns highly discriminative semantic features between targets and background. At this stage, which occupies the remaining $(1-\eta)$ proportion of the training epochs, the feature perceptual mechanism is activated to further constrain the reconstructed images in the feature space, ensuring that the reconstructed target samples are also close to real ones in the space of deep features that are discriminative for separating target from the complex background. Adding the feature perceptual loss with a weight φ_3, the overall optimization objective becomes:

$$\begin{aligned}\min_{E,R,D_{cls}} \max_{D_{adv}} (&\alpha_1 \mathcal{L}_{pixel} + \alpha_2 \mathcal{L}_s tat + \alpha_3 \mathcal{L}_{KL} \\ &\varphi_1 \mathcal{L}_{adv} + \varphi_2 \mathcal{L}_{cls} + \varphi_3 \mathcal{L}_{fea})\end{aligned} \tag{18}$$

This staged training strategy effectively avoids the instability that may occur when complex feature constraints are introduced prematurely. By progressively introducing the feature perceptual loss, the network achieves smoother convergence and a better balance between the spatial fidelity and feature-level consistency of the generated target samples.

In summary, the proposed MVAE-ACGAN first expands the training target data set through a pre-augmentation module to prevent mode collapse in subsequent generative training. It then employs three refined loss functions to enforce generated target sample authenticity at the pixel, statistical and feature levels. The band-weighted reconstruction pixel loss realizes differentiated generation across bands by allocating more weight to the reconstruction of high-resolution bands that contain richer spatial details. The reconstruction statistical loss ensures the regularity of cross-band energy distribution by enforcing consistency in statistical properties between generated and real target samples across all bands. The feature perceptual loss constrains target sample generation in a discriminative feature space that separates targets from the background, while a staged training strategy is adopted to enhance the stability and effectiveness of this mechanism during training. Collectively, these improvements enable the framework to generate target samples that are authentic both inter- and intra-band, thereby effectively addressing the imbalanced few-shot learning problem in multi-band SAR target detection.

IV. Experiment Results

To validate the proposed MVAE-ACGAN framework for few-shot multi-band SAR target detection, an airborne multi-band SAR dataset containing vehicle targets is constructed in this section. Comparative experiments, including reconstruction quality evaluation and detection performance analysis, are conducted on the self-built multi-band SAR dataset to demonstrate the effectiveness of the proposed framework in addressing multi-band

https://mc.manuscriptcentral.com/taes

target detection challenges. In addition, ablation studies are conducted to verify the contribution of each key improvements of the proposed framework.

A. Multi-Band Airborne SAR Dataset for Target Detection

In this study, a Multi-band SAR Vehicle Detection Dataset (MSVDD) is constructed using the airborne SAR data. The dataset contains HH-polarized SAR images across four bands with respective resolutions: P-band at 1m, S-band at 1m, C-band at 0.5m and Ka-band at 0.3m. It includes 163 vehicle target samples and 1790 background samples collected from diverse urban environments such as airports, residential areas, harbors, and toll stations. Since vehicles occupy only a small fraction of urban scenes and manual annotation is costly, the number of target samples is far fewer than that of background samples, resulting in a imbalanced few-shot target dataset.

The dataset has been co-registered and resampled to a uniform size, with each sample consisting of a 16×16-pixel image patch for four bands. Representative samples from the dataset are illustrated in Fig. 3, where (a)-(e) show five vehicle target examples and (f)-(j) show five background examples. From top to bottom, each row corresponds to the P-, S-, C-, and Ka-bands SAR images. As the frequency increases from the S-band to the Ka-band, the image characteristics of both targets and backgrounds exhibit evident differences. In the low-frequency P- and S- bands, the images are dominated by large-scale bright scatterers and relatively smooth background regions, with vehicle targets typically appearing as stable strong-scattering patches. In the C-band, the image texture becomes more complex, and target scattering evolves from coarse structures into multi-scale details. In the high-frequency Ka-band, the shorter wavelength introduces richer fine textures and stronger speckle effects, resulting in highly complex local structures. Overall, the low-frequency bands emphasize large-scale scattering and energy distribution, whereas the high-frequency bands provide more detailed structural and textural information.

As the first multi-band SAR detection dataset, the MSVDD constructed in this study offers a reliable foundation for research on few-shot multi-band SAR target detection.

B. Evaluation Metrics

To comprehensively evaluate the performance of the proposed framework in few-shot multi-band SAR target detection, two categories of evaluation metrics are employed, corresponding to the assessment of reconstruction quality and the detection performance of generated samples.

The reconstruction quality metrics include

1) **Multi-band Euclidean Reconstruction Error**

The multi-band Euclidean reconstruction error D_E measures the pixel-wise Euclidean distance between the reconstructed sample set $\hat{\mathbf{x}}$ and the real target sample set $\mathbf{x}$, defined as:

$$D_E = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B}\sum_{b=1}^{B}\left\|x_i^{(b)} - \hat{x}_i^{(b)}\right\|_2 \tag{19}$$

where $x_i^{(b)}$ and $\hat{x}_i^{(b)}$ denote b-th band images of the i-th real target sample $\mathbf{x}_i$ and its reconstruction $\hat{\mathbf{x}}_i$, respectively. A lower D_E indicates that the reconstructed sample is closer to the real target sample in pixel intensity and that the model achieves better reconstruction performance.

2) **Multi-band Correlation Coefficient**

The multi-band correlation coefficient (MCC) measures the structural similarity between the multi-band reconstructed samples and the real samples [33], defined as:

$$MCC = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B}\sum_{b=1}^{B}\frac{Cov(x_i^{(b)}, \hat{x}_i^{(b)})}{Std(x_i^{(b)})Std(\hat{x}_i^{(b)})} \tag{20}$$

where $Cov(\cdot)$ and $Std(\cdot)$ represent the image covariance operator and standard deviation operator, respectively. A higher MCC indicates better structural consistency between real and reconstructed samples.

3) **Fréchet Feature Distance**

Inspired by the Fréchet Inception Distance [34], the Fréchet Feature Distance (FFD) is proposed to measure the distributional discrepancy between the real sample set and the reconstructed sample set in a deep feature space. It is defined as:

$$\begin{aligned} FFD = &\left\|\mu_{real}^{fea} - \mu_{rec}^{fea}\right\|_2^2 \\ &+ Tr(\Sigma_{real}^{fea} + \Sigma_{rec}^{fea} - 2(\Sigma_{real}^{fea}\Sigma_{rec}^{fea})^{\frac{1}{2}}) \end{aligned} \tag{21}$$

where $(\mu_{real}^{fea}, \Sigma_{real}^{fea})$ and $(\mu_{rec}^{fea}, \Sigma_{rec}^{fea})$ denote the mean and covariance of deep features extracted from real and reconstructed samples, respectively. $Tr(\cdot)$ denotes the trace of a metric. A lower FFD indicates better feature-level fidelity.

The detection performance metrics include [35] [36]:

1) **Detection Probability**

The detection probability P_d, also known as the Recall, represents the proportion of true targets that are correctly detected, assessing the completeness of target coverage in the detection results. A higher P_d indicates fewer missed detections in the results.

2) **False Alarm Probability**

The false alarm probability P_f denotes the proportion of background samples that are incorrectly identified as targets, evaluating the reliability of detection results over background regions. A lower P_f indicates reduced risk of false alarms in background regions.

https://mc.manuscriptcentral.com/taes

3) **F1-Score**
The F1-Score focuses on the overall balance between accuracy and completeness in the detection task. It effectively reflects the task-level quality of detection results, especially under few-shot or imbalanced scenarios. A higher F1-Score signifies a better trade-off between missed detections and false alarms. It is the harmonic mean of Precision and Recall.

4) **Average Precision**
The Average Precision (AP) focuses on the robustness and stability of detection performance across different decision thresholds. It is equivalent to the area under the Precision-Recall curve. A higher AP value indicates that the detection method maintains high performance across varying thresholds, demonstrating more stable decision boundaries.

5) **Receiver Operating Characteristic Curve and Area Under the Curve**
The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between P_d and P_f across different thresholds. A curve closer to the top-left corner indicates better detection performance. The Area Under the Curve (AUC) quantifies the overall effectiveness of this discrimination, with a value closer to 1 indicating stronger inherent class separation capability.

In summary, the reconstruction quality metrics D_E, MCC and FFD evaluate the reconstruction ability of generator in pixel-, statistics- and feature-levels, respectively. The detection performance metrices P_d and P_f describe the detection correctness from complementary perspectives; the F1-Score reflects the balance between accuracy and completeness; the AP and AUC metrics offer global evaluations of detection robustness and separability. These metrics collectively form an assessment of the generation framework for few-shot multi-band SAR detection challenge, providing a basis for evaluating the effectiveness of the proposed MVAE-ACGAN framework.

C. Experimental Setup and Visualization of Results

The proposed MVAE-ACGAN framework is validated on the self-built dataset MSVDD. The data split and the key parameters of the network are set in TABLE I. The band weights are assigned according to the resolution of different bands, while the coefficients in the loss function are selected based on the magnitude of each term.

TABLE I
Parameters Setting

Parameter	Value
Training/Test Split	$8:2$
Number of Generated Samples	80
$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	$(0.5, 0.5, 0.8, 1)$
$(\alpha_1, \alpha_2, \alpha_3)$	$(2, 3, 1000)$
$(\varphi_1, \varphi_2, \varphi_3)$	$(0.2, 0.1, 0.1)$
η	50%

To provide an intuitive demonstration of the effectiveness of the proposed MVAE-ACGAN generation framework, Fig. 4 presents the visualized results of multi-band SAR target samples at different processing stages. The figure includes the real target samples used as network inputs (a), the pre-augmentation samples obtained through rotation (b), flipping (c), noise injection (d), and samples mixing (e), the VAE-reconstructed samples (f), and three generated samples produced by the proposed framework (g)-(i). From top to bottom, each row corresponds to the P-, S-, C-, and Ka-bands, respectively.

It can be observed that the pre-augmentation samples maintain consistent spatial structures with the real targets but produce no new scattering information, as this processing merely applied geometric transformations to the input real target samples. In contrast, the reconstructed samples from the modified VAE module successfully

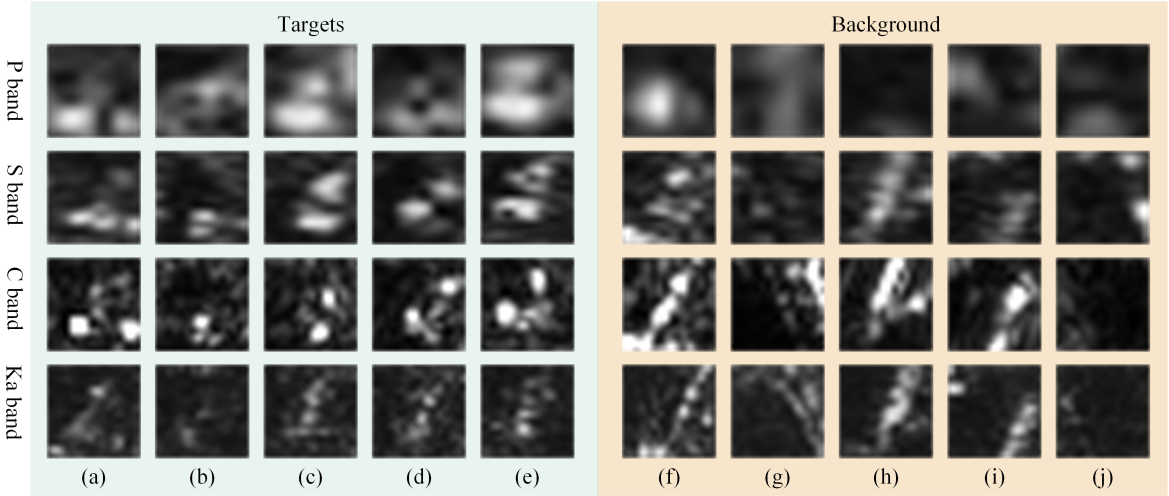


Fig. 3. Representative samples from MSVDD: (a)-(e) are five different vehicle target samples, and (c)-(d) are five background samples; from top to bottom, each row corresponds to the P-, S-, C-, and Ka-bands, respectively.

https://mc.manuscriptcentral.com/taes

reproduce the energy distribution and scattering structure across all bands, indicating that the latent space effectively captures the core features of multi-band SAR targets. Furthermore, the generated target samples exhibit inter- and intra-band representations similar to those of real target samples. Specifically, as the resolution improves, the generated samples evolve from concentrated, blurry bright spots at low-frequency bands to rich structural details with dispersed energy in higher-frequency bands, consistent with the cross-band behavior of real target samples. Moreover, rather than merely replicating the structural details of the input real target samples, the generated samples produce novel textural patterns, thereby effectively expanding the diversity of the training target set.

D. Comparative Evaluation of Reconstruction Quality

Image reconstruction plays a central role in the proposed MVAE-ACGAN generation framework, as higher quality reconstruction generally leads to better generative performance. Accordingly, this subsection evaluates the reconstruction quality against three representative generation approaches: a geometric transformations based augmentation method [13] employing geometric operations such as flipping, rotation, and noise addition (denoted as Aug); a standard VAE approach [18] (denoted as VAE); and a VAE method integrated with the ACGAN concept [29] (denoted as VAE-ACGAN).

Fig. 5 presents the real target samples and the reconstructed target samples obtained by the four methods, where each method illustrates four different target examples and each example contains the P-, S-, C-, and Ka-bands patches from top to bottom. The Aug method yields the visually closest results to the real target images, as it applies only geometric transformations to the original samples without performing genuine reconstruction. The VAE exhibits severe over smoothing across all bands, where the fine scale scattering characteristics and inter-sample differences are largely lost. The VAE-ACGAN benefits from adversarial learning and recovers limited local textures, producing reconstructed images with slight target-like variations. However, both VAE and VAE-ACGAN perform poorly in terms of inter-band variability and diversity across samples. This may be due to the lack of effective mechanisms in both methods, which prevents them from emphasizing key features when learning and generating multi-band image characteristics. Moreover, with only 130 training target samples, these models struggle to adequately learn the data distribution, thereby limiting the diversity of the generated results.

In contrast, the target samples reconstructed by the proposed MVAE-ACGAN framework more accurately restores the inter-band and intra-band characteristics of the real target images. The enhanced performance is attributed to key improvements in the proposed framework. Specifically, the pre-augmentation module initially expands the training sample set, allowing the proposed framework to be sufficiently trained and to perform targeted reconstruction for different samples rather than applying a uniform decoding process. Then, the band-weighted reconstruction loss strengthens high-frequency texture retention, and the reconstruction statistic loss regularizes inter-band contrast and dynamic range. Furthermore, the feature perception loss enforces consistency between the reconstructed output and the real sample at the deep feature level, further improving reconstruction fidelity.

A quantitative comparison of reconstruction performance of the four methods is summarized in TABLE II, where three evaluation metrics are reported: D_E, MCC and FFD.

For the pixel-level error metric D_E, Aug and the proposed framework perform poorly, yielding higher reconstruction errors than VAE and VAE-ACGAN, even though

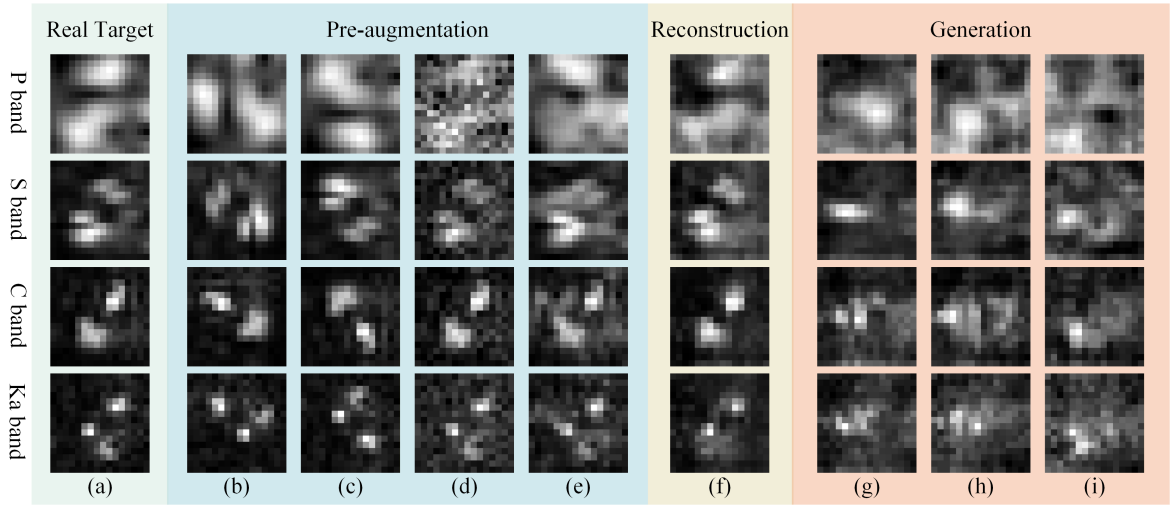


Fig. 4. Results of the proposed MVAE-ACGAN framework across different stages: (a) is a vehicle target sample, (b)-(e) are four additional samples after pre-augmentation, (f) is the reconstructed sample of VAE and (g)-(i) are three final generated samples of MVAE-ACGAN, from top to bottom, each row corresponds to the P-, S-, C-, and Ka-bands, respectively.

https://mc.manuscriptcentral.com/taes

TABLE II
Quantitative Reconstruction Evaluation

	Aug	VAE	VAE-ACGAN	Proposed
D_E	5.3590	4.1688	4.3055	5.1666
MCC	0.8431	0.3450	0.5073	0.6576
FFD	10.1424	25.6753	66.8835	11.6507

they produce the most visually faithful results in Fig. 5. This because D_E only measures pointwise intensity differences between the reconstructed sample and the real sample, without considering deeper semantic similarity. Consequently, VAE and VAE-ACGAN, which reconstruct over-smoothed images that collapse toward band-wise average intensities, achieve better scores under this metric. Moreover, D_E is mathematically identical to the original reconstruction loss used in VAE. The proposed framework alters this averaging-oriented loss, which consequently results in a comparatively poorer performance on D_E. However, this does not indicate inferior reconstruction by MVAE-ACGAN. Instead, it suggests that this classic pixel-level metric D_E is insufficient for evaluating the reconstruction quality of multi-band SAR images, which contain rich inter- and intra-band information.

The structure preserving metric MCC and deep feature preserving metric FFD provide a more reliable measure of reconstruction fidelity. The proposed MVAE-ACGAN substantially outperforms both VAE and VAE-ACGAN on these two metrics, demonstrating its capability to maintain cross-band structural characteristics and high-level semantic consistency during reconstruction. On both MCC and FFD, Aug achieves the best performance, as it merely applies geometric transformations such as rotation and flipping to the input samples, completely preserving the texture structure and deep features of the target. The proposed method ranks second, especially on the FFD, achieving 11.6507 that is notably close to 10.1424 of Aug. This further indicates that the introduced feature perception loss effectively guides the model to perform reconstruction at the deep feature level. Nonetheless, the superior reconstruction performance of Aug does not imply that the proposed framework is less effective in addressing the few-shot learning problem of multi-band SAR, as the generated samples by the Aug method lacks the diversity needed to improve detection accuracy.

E. Comparative Evaluation of Detection Performance

Following the reconstruction evaluation, the detection experiments are conducted to validate the effectiveness of the proposed MVAE-ACGAN framework in the imbalanced few-shot multi-band SAR target detection task. To ensure a fair comparison, all methods adopt the same multi-channel CNN detector. The compared methods include: a CNN baseline trained directly on the few-shot dataset; and three data augmentation methods compared in the previous section, namely Aug, VAE, and VAE-ACGAN. For each of these data augmentation methods, an equal number of target samples is generated and combined with the real target and background training set

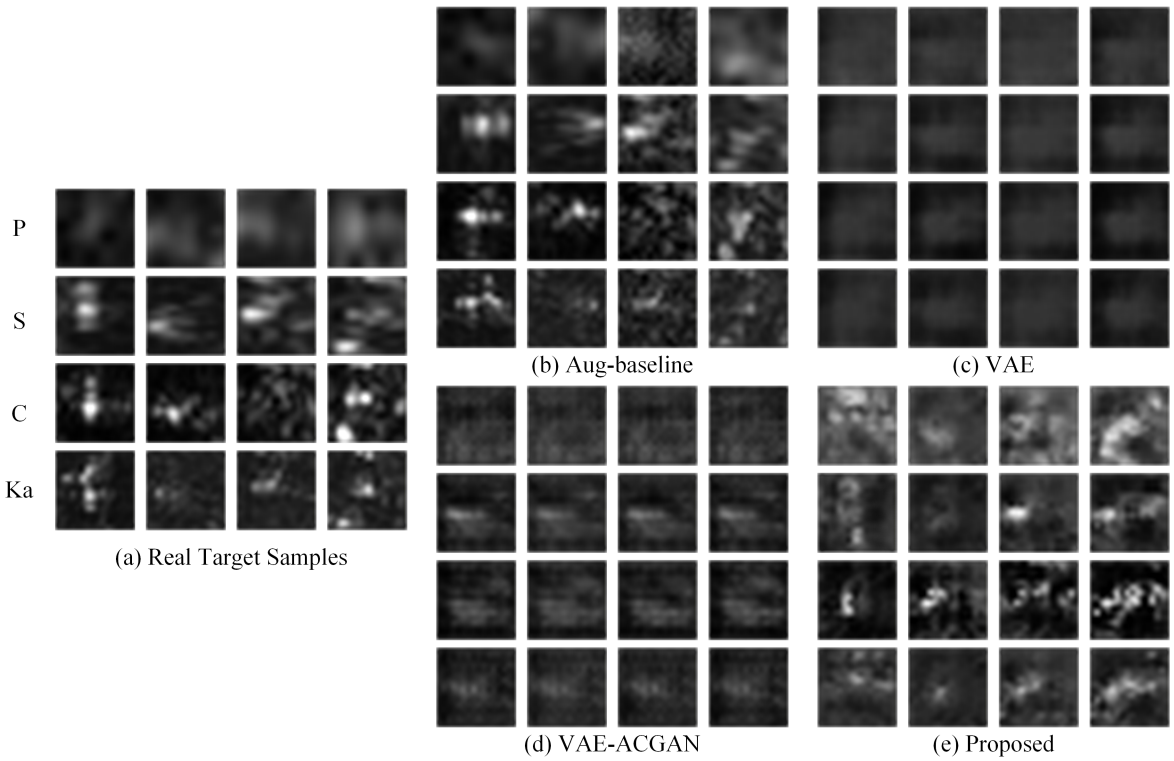


Fig. 5. Reconstruction samples of different methods corresponding to four different target samples; from top to bottom, each row corresponds to the P-, S-, C-, and Ka-bands, respectively.

https://mc.manuscriptcentral.com/taes

to form the corresponding augmented training dataset, on which the same detector is then trained, and the detection performance is evaluated.

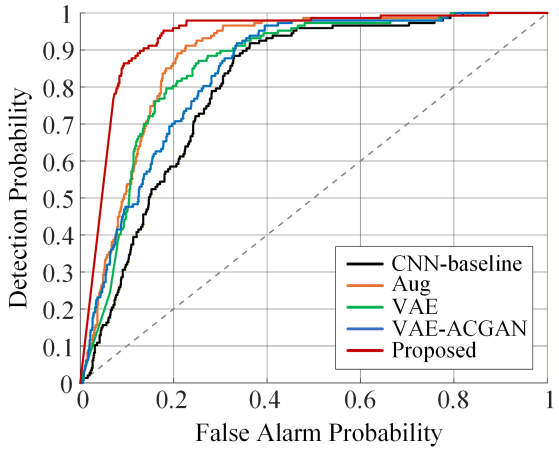


Fig. 6. Receiver operating characteristic (ROC) curves of different methods.

The ROC curves of the five methods are shown in Fig. 6. The curve of the proposed method is positioned closest to the top-left corner and consistently achieves a higher detection probability than the other methods across nearly every false alarm rate. These results indicate that the data generated by the proposed framework lead to a significant improvement in detection quality, characterized by a superior balance between false alarm suppression and target detection.

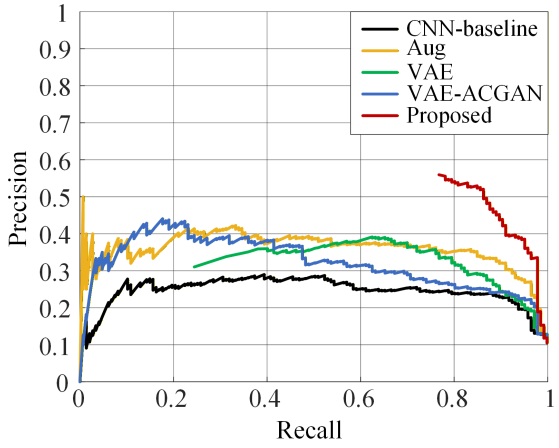


Fig. 7. Precision-recall (PR) curves of different methods.

Fig. 7 presents the Precision-Recall (PR) curves corresponding to these five methods. The PR curve of the proposed framework is positioned closer to the top-right corner and consistently dominates those of other counterparts across most recall levels, indicating that the detection model trained with data augmented by our method maintains high precision and recall stably under varying decision thresholds. In contrast, the PR curves of the other methods exhibit significant fluctuations, with precision dropping rapidly, particularly in high-recall regions. This suggests that while these compared augmentation strategies can improve recall to some extent, they fail to effectively suppress false alarms, leading to sharp degradation in precision. The proposed MVAE-ACGAN framework, however, enables a steady increase in recall while preserving high precision, demonstrating its effectiveness and superiority in enhancing target detection performance.

The detection performance of the CNN baseline and the four augmentation methods (Aug, VAE, VAE-ACGAN, and the proposed MVAE-ACGAN) is quantitatively evaluated using the five metrics including the P_d, P_f, AUC, F1-Score, and AP. The results are presented in TABLE III. The proposed framework achieves the best performance across all evaluation metrics.

TABLE III
Quantitative Detection Performance Evaluation

Method	P_d	P_f	AUC	F1	AP
CNN-baseline	88.44%	32.88%	0.8065	0.3774	0.2474
Aug	91.15%	22.59%	0.8813	0.4752	0.3638
VAE	87.07%	25.30%	0.8576	0.4324	0.3249
VAE-ACGAN	91.83%	33.52%	0.8447	0.3846	0.3228
Proposed	95.24%	17.79%	0.9324	0.5490	0.5317

Regarding the trade-off between the mutually constrained metrics P_d and P_f, the proposed framework achieves the highest P_d of 95.24% while simultaneously attaining the lowest P_f of 17.79%. In contrast, the other methods struggle to balance these two aspects. For instance, VAE-ACGAN, which has the second-best P_d, exhibits the highest P_f; meanwhile, Aug maintains a relatively low P_f but its P_d is less competitive. In terms of the comprehensive detection metric AUC, the proposed method reaches 0.9324, representing a significant improvement of 0.0511 over the Aug method. These overall detection improvements can be attributed to the ability of the proposed MVAE-ACGAN framework to generate both authentic and diverse multi-band SAR target samples. In particular, the feature perception mechanism integrated through the classification branch ensures that the generated target samples not only remain highly consistent with real targets samples at the deep feature level but also exhibit clear distinction from background samples. As a result, the detection network is supplied with discriminative training samples that possess well-defined class boundaries, effectively enhancing its generalization capability and detection robustness in complex scenarios.

For the class-imbalance-aware metrics F1-Score and AP, the proposed framework also achieves the highest scores of 0.5490 and 0.5317, respectively, significantly outperforming the other methods. This indicates that the proposed MVAE-ACGAN can effectively adapt to the highly imbalanced dataset in multi-band SAR target detection. For such imbalanced dataset, the proposed framework not only performs target generation but also

https://mc.manuscriptcentral.com/taes

involves both background and target samples in discriminator training, thereby establishing a clear target background boundary in the feature space. This design partially compensates for issues of detector bias towards the background category and reduced target recognition ability caused by the scarcity of target samples and the confusion of inter-class features.

These results collectively confirm that the proposed framework succeeds in generating highly authentic and discriminative samples, which directly leads to enhanced detection accuracy in challenging few-shot and class-imbalanced multi-band SAR scenarios.

F. Ablation Study

The above comparative results have shown clearly the superior performance of the proposed framework in few-shot multi-band SAR target detection. To further understand how the proposed improvements contribute to this effect, an ablation study is conducted on the three key improvements of the proposed framework: the pre-augmentation module, the modified reconstruction loss in VAE and the feature perceptual mechanism with staged training strategy. The baseline is the VAE-ACGAN framework [28]. The experimental configurations are as follows: (1) v1: the baseline model with the pre-enhancement stage; (2) v2: the baseline model with the modified reconstruction loss of VAE including weighted reconstruction pixel loss and reconstruction statistics loss; (3) v3: the baseline model incorporating feature perceptual loss with staged training. Additionally, the proposed MVAE-ACGAN, which combines all these modules, is also evaluated. For clarity and focus, AUC, F1-score, and AP are reported in the ablation study, as they comprehensively capture the overall detection performance.

As results shown in the TABLE IV, incorporating each individual module into the baseline framework consistently improves the detection performance. Specifically, the pre-augmentation module (v1) achieves the most significant gains, improving AUC, F1-score, and AP by 0.0697, 0.1727, and 0.1528, respectively. This indicates that preliminarily expanding the original few-shot target samples through geometric transformations before generative model can significantly enhance subsequent generation and detection performance. Notably, the F1-score of V1 (0.5573) is slightly higher than that of the proposed framework (0.5490). This is because pre-augmentation directly increases the number of target samples consistent with the real distribution, enabling the detector to achieve a better precision-recall balance near the decision threshold. As F1 is the harmonic mean of precision and recall at a specific threshold, such improvement leads to a localized improvement of this metric. However, pre-augmentation does not introduce new data diversity, as the expanded samples are limited to geometric transformations of the original target samples, leading to lower performance on threshold-independent metrics such as AUC and AP compared with the proposed framework.

The reconstruction module (v2) also contributes positively to the performance, improving AUC, F1-score, and AP by 0.0357, 0.0853, and 0.1039, respectively. Compared to the VAE-ACGAN baseline, which employs a simple pixel-wise loss for generation, the improved band-weighted reconstruction pixel loss and reconstruction statistic loss in the proposed framework jointly constrain the inter- and intra- band variability and regularity. This ensures that the generated multi-band images are realistic at both the pixel-level and the statistical-level, thereby further improving the performance of multi-band SAR target detection.

With the introduction of the feature perceptual mechanism incorporating staged training (v3), the detection metrics also show improvement, with AUC, F-Score, and AP increasing by 0.0417, 0.0678 and 0.1333, respectively. This module extracts discriminative deep features by the classification branch, ensuring the generated target samples maintain both feature consistency with real target samples and separability from the background. The enhancement in detection metrics validates the effectiveness of this improvement in generating multi-band SAR samples that are beneficial for target detection.

Finally, the proposed method, with all of these improvements jointly integrated, achieves superior overall detection performance with AUC, F1-score, and AP of 0.9324, 0.5490, and 0.5317, respectively, showing significant improvements over the baseline. This demonstrates that each improvement contributes to the overall detection performance gain, validating the effectiveness of the proposed MVAE-ACGAN framework of multi-band SAR target detection.

V. Conclusion

In this paper, a novel Modified Variational Autoencoder Auxiliary Classifier Generative Adversarial Network (MVAE-ACGAN) framework is proposed to address the imbalanced few-shot learning problem in multi-band SAR target detection. The proposed framework contains three key improvements: a pre-augmentation module; a VAE with band-weighted reconstruction pixel loss and reconstruction statistic loss; a feature perceptual mechanism with a staged training strategy in classification branch of an ACGAN-based discriminator. These improvements collectively work to generate realistic multi-band SAR images that exhibit both inter- band and intra-band variability and regularity, thereby providing high quality training data for subsequent multi-band SAR target detection networks. Experiments on a self-built Multi-band SAR Vehicle Detection Dataset (MSVDD) demonstrate that the proposed MVAE-ACGAN framework yields substantial improvements in multi-band SAR target detection. Compared with existing best-performing augmentation method, the proposed approach increases AUC by 0.0511, F1-score by 0.0738, and AP by 0.1679, achieving higher

https://mc.manuscriptcentral.com/taes

TABLE IV
Ablation Study on the Proposed MVAE-ACGAN Framework

	Pre-augmentation	Modified Reconstruction Loss	Feature Perceptual Mechanism	AUC	F1	AP
Baseline	×	×	×	0.8447	0.3846	0.3228
V1	✓	×	×	0.9144	0.5573	0.4756
V2	×	✓	×	0.8804	0.4699	0.4267
V3	×	×	✓	0.8864	0.4524	0.4561
Proposed	✓	✓	✓	0.9324	0.5490	0.5317

detection probability, lower false alarm probability, and overall superior detection performance. In summary, the proposed MVAE-ACGAN framework provides a practical and effective solution for multi-band SAR few-shot target detection.

Acknowledgement

The authors acknowledge the Aerospace Information Research Institute, China Academy of Sciences for providing the airborne multi-band SAR images utilized to construct the dataset MSVDD in this study. This work is also supported by National Natural Science Foundation of China (U2241202) and Science and Technology Development Fund (FDCT) of Macau Special Administrative Region.

REFERENCES

[1] R. Del Prete, M. D. Graziano, and A. Renga
Unified framework for ship detection in multi-frequency sar images: A demonstration with cosmo-skymed, sentinel-1, and saocom data
Remote Sensing, vol. 15, no. 6, p. 1582, 2023.

[2] J. Shang, H. McNairn, C. Champagne *et al.*
Application of multi-frequency synthetic aperture radar (SAR) in crop classification. IntechOpen, 2009.

[3] H. Zhang, D. An, J. Li *et al.*
The dual-band sar image fusion based foliage penetrating target detection method
IEEE Transactions on Geoscience and Remote Sensing, 2024, early Access.

[4] H. Chen, Y. Zhang, L. Zeng *et al.*
Target covered by camouflage netting detection based on multi-band and polarization sar image fusion
Opto-Electronics Engineering, vol. 38, no. 7, pp. 106–118, 2011.

[5] B. Zou, H. Cai, W. M. Moon *et al.*
Target detection based on l- and c-band polsar data
In *2011 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2011, pp. 397–400.

[6] Q. Deng, L. Gao, and J. Yang
Particle filter based approach to road detection in multiband sar images
In *2009 2nd Asian-Pacific Conference on Synthetic Aperture Radar*, 2009, pp. 684–688.

[7] Z. Zeng, J. Sun, X. Yao *et al.*
Sar-atr with knowledge hierarchy division and information dissemination networks
ISPRS Journal of Photogrammetry and Remote Sensing, vol. 206, pp. 242–257, 2023.

[8] S. Sun and J. Wang
Ship detection in sar images based on steady cfar detector and knowledge-oriented gbdt classifier
Electronics, vol. 13, no. 14, p. 2692, 2024.

[9] J. Li, Z. Yu, L. Yu *et al.*
A comprehensive survey on sar atr in deep-learning era
Remote Sensing, vol. 15, no. 5, p. 1454, 2023.

[10] P. Lang, X. Fu, J. Dong *et al.*
Recent advances in deep learning based sar image targets detection and recognition
IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, early Access.

[11] X. Zhou, Y. Zhang, D. Liu *et al.*
Sar target recognition with limited training samples in open set conditions
Sensors, vol. 23, no. 3, p. 1668, 2023.

[12] R. Zhang, Z. Wang, Y. Li *et al.*
FewSAR: A few-shot SAR image classification benchmark
arXiv preprint arXiv:2306.09592, 2023, preprint.

[13] C. Shorten and T. M. Khoshgoftaar
A survey on image data augmentation for deep learning
Journal of Big Data, vol. 6, no. 1, p. 60, 2019.

[14] B. Trabucco, K. Doherty, M. Gurinas *et al.*
Effective data augmentation with diffusion models
arXiv preprint arXiv:2302.07944, 2023, preprint.

[15] A. R. Flanagan, D. Dalal, and F. G. Glavin
Exploring generative artificial intelligence and data augmentation techniques for spectroscopy analysis
Chemical Reviews, 2025, early Access.

[16] V. Correa, P. Funk, N. Sundelius *et al.*
Applications of gans to aid target detection in sar operations: A systematic literature review
Drones, vol. 8, no. 9, p. 448, 2024.

[17] M. Durgadevi
Generative adversarial network (gan): A general review on different variants of gan and applications
In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, 2021, pp. 1–8.

[18] D. P. Kingma and M. Welling
Auto-encoding variational bayes
arXiv preprint arXiv:1312.6114, 2013, preprint.

[19] I. Higgins, L. Matthey, A. Pal *et al.*
beta-vae: Learning basic visual concepts with a constrained variational framework
In *International Conference on Learning Representations (ICLR)*, 2017.

[20] K. Sohn, H. Lee, and X. Yan
Learning structured output representation using deep conditional generative models
Advances in Neural Information Processing Systems, vol. 28, pp. 1–9, 2015.

[21] X. Hou, L. Shen, K. Sun, and G. Qiu
Deep feature consistent variational autoencoder
In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, 2017, pp. 1133–1141.

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*
Generative adversarial nets
Advances in Neural Information Processing Systems, vol. 27, pp. 1–9, 2014.

[23] V. Kushwaha and G. C. Nandi

https://mc.manuscriptcentral.com/taes

Study of prevention of mode collapse in generative adversarial network (gan)
In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, 2020, pp. 1–6.

[24] M. Arjovsky, S. Chintala, and L. Bottou
Wasserstein generative adversarial networks
In *International Conference on Machine Learning (ICML)*, ser. PMLR, 2017, pp. 214–223.

[25] X. Chen, Y. Duan, R. Houthooft *et al.*
Infogan: Interpretable representation learning by information maximizing generative adversarial nets
Advances in Neural Information Processing Systems, vol. 29, pp. 1–9, 2016.

[26] M. Mirza and S. Osindero
Conditional generative adversarial nets
arXiv preprint arXiv:1411.1784, 2014, preprint.

[27] A. Odena, C. Olah, and J. Shlens
Conditional image synthesis with auxiliary classifier gans
In *International Conference on Machine Learning (ICML)*, ser. PMLR, 2017, pp. 2642–2651.

[28] A. B. L. Larsen, S. K. Snderby, H. Larochelle *et al.*
Autoencoding beyond pixels using a learned similarity metric
In *International Conference on Machine Learning (ICML)*, ser. PMLR, 2016, pp. 1558–1566.

[29] F. Al Zegair, B. Betz-Stablein, M. Janda *et al.*
Identifying suspicious naevi with dermoscopy via variational autoencoder auxiliary generative classifiers
Physical and Engineering Sciences in Medicine, pp. 1–12, 2025, early Access.

[30] S. Du, J. Hong, Y. Wang *et al.*
Physical-related feature extraction from simulated sar image based on the adversarial encoding network for data augmentation
IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2021, early Access.

[31] 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, 2020, early Access.

[32] Y. Sun, W. Jiang, J. Yang *et al.*
Sar target recognition using cgan-based sar-to-optical image translation
Remote Sensing, vol. 14, no. 8, p. 1793, 2022.

[33] C. Gao and G. Xue
A matching method to reduce the influence of sar geometric deformation
The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 42, pp. 355–358, 2018.

[34] M. Heusel, H. Ramsauer, T. Unterthiner *et al.*
Gans trained by a two time-scale update rule converge to a local nash equilibrium
Advances in Neural Information Processing Systems, vol. 30, pp. 1–11, 2017.

[35] Z. Wang, L. Du, J. Mao *et al.*
Sar target detection based on ssd with data augmentation and transfer learning
IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 1, pp. 150–154, 2018.

[36] X. Geng, G. Dong, Z. Xia *et al.*
Sar target recognition via random sampling combination in open-world environments
IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 331–343, 2022.

https://mc.manuscriptcentral.com/taes