Research paper

Diffusion-augmented direct classification: A few-shot learning framework for Synthetic Aperture Radar image automatic target recognition

Zilu Ying [a,1], Wenyu Ke [a,1], Yikui Zhai [a,*,1], Xinglin Liu [a,1], Jianhong Zhou [a,b,1], Pasquale Coscia [c], Angelo Genovese [c]

[a] School of Electronics and Information Engineering, Wuyi University, 22nd Dongcheng Village, Pengjiang District, Jiangmen, 529000, Guangdong, China
[b] School of Electronics and Information Engineering, South China University of Technology, 381st Wushan Road, Tianhe District, Guangzhou, 510000, Guangdong, China
[c] Department of Computer Science, Universita Degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Milano (MI), Italy

ARTICLE INFO

Keywords:
Synthetic Aperture Radar
Automatic Target Recognition
Denoising Diffusion Probability model
Few-shot learning

ABSTRACT

Deep learning-based (DL-based) synthetic aperture radar automatic target recognition technology (SAR-ATR) has undergone extensive development, demonstrating superiority over other competitive methods. However, the intrinsic requirement of deep learning for a large labeled dataset restricts its practical application. Moreover, some DL-based few-shot SAR-ATR methods are overly complex, hindering their deployment in real-world applications. In addressing these obstacles, our solution introduces a straightforward yet efficient few-shot learning approach titled Diffusion-Augmented Direct Classification for few-shot SAR-ATR applications. The proposed method adopts a two-stage paradigm, where a diffusion model first learns from unlabeled data and then produces synthetic samples to train a recognition model. In the upstream stage, a lightweight diffusion-based image generator build upon the shuffle-residual network structure is trained on a limited number of annotated SAR images to generate artificial training samples for the downstream recognition model. In the downstream stage, a Siamese network-based recognition model and a similarity training procedure are proposed to train the model on a combination of real-world and artificial samples, thereby improving recognition accuracy. A projection expansion layer is proposed to improve the efficiency of cosine similarity loss in the downstream. Experiments conducted on the Moving and Stationary Target Acquisition and Recognition dataset demonstrated that our method outperformed other few-shot learning methods concerning recognition accuracy in SAR-ATR tasks. Specifically, our method achieves over 73% accuracy in a 5-sample-per-class scenario and over 85% accuracy in a 10-samples-per-class scenario. Source code of our paper is available at https://github.com/yikuizhai/DADC.

1. Introduction

Synthetic Aperture Radar (SAR) (Fu et al., 2024) is a type of sensor that utilizes the movement of its carrier to take high-resolution radar images of ground targets. Compared to ordinary microwave radar imaging technology and other optical-based remote sensing technology, SAR can be used to obtain high-resolution ground images day and night without being interfered with by weather or environment. Automatic Target Recognition (ATR) is a technology that utilizes computer vision technology to classify targets images into different categories. Given the inference of speckle noise (Ma et al., 2016; Hu et al., 2015) and the absence of visually discernible features in SAR imagery, SAR-ATR emerges as the principal approach for achieving precise target recognition, garnering significant attention within the remote sensing research community. Early SAR-ATR methodologies predominantly relied on plain computer vision techniques, specifically template matching (Novak et al., 1997; Jin et al., 2018) and model-based approaches (Hummel, 2000). These methods involve pixel-wise comparison of SAR images with a predefined set of templates, identifying the closest match as the classification outcome. They are thereby susceptible to noise and changes in conditions, resulting in a low level of generalization ability.

The emergence of Deep Learning (DL) technology has facilitated advancements in SAR-ATR. Leveraging the remarkable performance and generalization ability of Convolutional Neural Network (CNN) in

* Corresponding author.
E-mail addresses: ziluy@163.com (Z. Ying), kewenyu@wyu.edu.cn (W. Ke), yikuizhai@163.com (Y. Zhai), jmxlliu@163.com (X. Liu), jackychou_lab@126.com (J. Zhou), pasquale.coscia@unimi.it (P. Coscia), angelo.genovese@unimi.it (A. Genovese).

1 These authors contributed equally to this work.

computer vision tasks, they have been promptly adapted for use in SAR-ATR tasks (Ai et al., 2022; Li et al., 2022; Shang et al., 2018; Chen et al., 2016) as well. However, to attain high recognition accuracy, DL-based supervised SAR-ATR methods typically require a large number of annotated SAR image samples. This does not coincide with real-world SAR-ATR application scenarios, as acquiring and annotating a large volume of SAR images requires extensive human labor and investment. Due to the complexity of SAR data, manual annotation demands significant expert effort and remains highly labor-intensive. Additionally, some sample types are inherently scarce, given the high cost of data acquisition from platforms such as satellites or fighter jets. These challenges hinder the practical deployment of SAR-ATR methods in real-world applications. The industry urgently requires a method capable of achieving high recognition accuracy while utilizing as minimal annotated SAR images as possible.

The Few-Shot Learning (FSL) methods has garnered significant attention in research circles in addressing above problems. One such method is Meta-Learning, which achieves relatively high accuracy even with a handful of available samples by training the model alternately on positive sample pairs and negative counterparts from different support and query sets. Two such prominent Meta-Learning-based frameworks are Prototypical Network (Snell et al., 2017) and Relation Network (Sung et al., 2018a). However, as general meta learning methods, neither the Prototypical Network nor the Relation Network is specifically tailored to few-shot SAR-ATR problems, making it difficult for them to extract features of SAR images. To tackle the few-shot SAR-ATR with meta learning techniques, dedicated meta learning methods (Wang et al., 2021, 2023b) are proposed, making progress in recognition accuracy. In addition to Meta-Learning, Transfer Learning is another common techniques used in few-shot learning tasks. In addition to Meta-Learning, Transfer Learning is another common technique used in few-shot learning tasks. Transfer Learning methods (Huang et al., 2017; Malmgren-Hansen et al., 2017) for few-shot SAR-ATR typically involve pre-training a model on a large source dataset, such as optical images, to learn general feature representations. The model is then fine-tuned on a small target SAR dataset. This approach leverages the feature extraction capabilities learned from the source domain to perform target recognition. Both Meta-Learning and Transfer Learning approaches have limitations in SAR-ATR applications. Meta-Learning methods typically require sufficient labeled data across many categories to generalize, which does not align with real-world SAR-ATR scenarios that involve few categories and limited samples per category. As a result, these methods struggle to learn effective representations from such sparse data. Transfer Learning methods face challenges due to distribution shifts between the source and target datasets, making it difficult to extract optimal features and increasing the risk of overfitting. Moreover, many recent FSL models are becoming increasingly complex, often relying on elaborate training procedures. This added complexity hinders their deployment on resource-constrained platforms like satellites or missiles.

In real-world SAR-ATR tasks, only a small number of SAR images are available, and few are labeled, making it challenging to train accurate models. A promising solution is to use image generation models trained on unlabeled data to synthesize realistic samples. These synthetic images can then augment scarce classes and improve the performance of downstream recognition models. Currently, there exist methods (Qin et al., 2022; Liu et al., 2024) that utilize image generation models, primarily Generative Adversarial Networks (GANs), for few-shot SAR-ATR tasks. GANs synthesize artificial SAR images through an adversarial process where a generator produces samples to deceive a discriminator tasked with distinguishing real from fake images. While this enables GANs to generate high-quality images, the training process is often unstable, which can introduce artifacts and defects in the synthesized samples. Moreover, if the capabilities of the generator and discriminator are not well balanced, the diversity of generated images may be severely limited, reducing their effectiveness in augmenting training data for downstream recognition tasks.

To address these problems, we propose a novel two-stage generation-based method for few-shot SAR-ATR tasks. The motivation behind our method is to tackle the few-shot SAR-ATR problem using a generative model capable of learning the features of SAR images from available unlabeled samples to produce artificial training samples for the target recognition model. This naturally divides the entire process into an upstream image generation stage and a downstream target recognition stage, as depicted in Fig. 1. This approach brings an additional advantage: by addressing the few-shot problem in the upstream stage using a generative model, the downstream target recognition model can be designed to be as simple and lightweight as possible. This enhances its applicability for deployment on devices with limited computational resources, such as satellites or missiles. To overcome the stability and diversity issues commonly associated with GAN-based models, we adopt an alternative image generation paradigm, namely the Denoising Diffusion Probabilistic Model (DDPM), to implement the upstream generation model. A summary of the motivation and the overview of the main highlight of our contributions are provided below.

1. An innovative two-stage FSL framework named Diffusion-Aug mented Direct Classification (DADC) for SAR-ATR tasks is proposed. This approach decouples the few-shot learning problem from the image recognition problem, providing flexibility to choose different models with varying complexities for each stage, which makes it easily deployed to embedded devices for the recognition task.
2. A SAR image generator based on the DDPM is designed. The generator utilizes a lightweight ShuffleNet-based U-Net backbone as its backbone. Compared to other image generators, our design can synthesize artificial samples that more closely match the distribution of real-world SAR images while retaining fewer parameters.
3. Pragmatically, a Siamese Network-based image classifier is implemented and integrated into our framework as the downstream classifier to be trained on artificial samples. This classifier is lightweight and can achieve high accuracy compared to other classification models trained on artificial samples.
4. We perform comprehensive trials and confirm the superior performance of DADC in comparison to other FSL methods in SAR-ATR jobs.

The subsequent sections of this paper are structured as follows. Section 2 discusses the existing literature and development timeline concerning SAR image generation models and few-shot learning methods for SAR-ATR tasks. Section 3 presents the general workflow and key components of the proposed DADC framework, including its upstream diffusion generator and downstream Siamese classifier. Section 4 details the results of our extensive experiments, comparing our method with others and including an ablation study. Finally, Section 5 provides both a summary of our works and the outlook for further direction, which form the conclusion.

2. Related works

In this section, we will introduce current generative models utilized in SAR image generation, and current few-shot learning methods for SAR-ATR tasks.

2.1. SAR image generation methods

DL-based Image Generation Models have undergone significant development over time and have been utilized in SAR image generation tasks. Currently, the majority of SAR image generation method are based on the GAN (Goodfellow et al., 2014), a prominent image generation method grounded in the principles of Nash equilibrium and Game

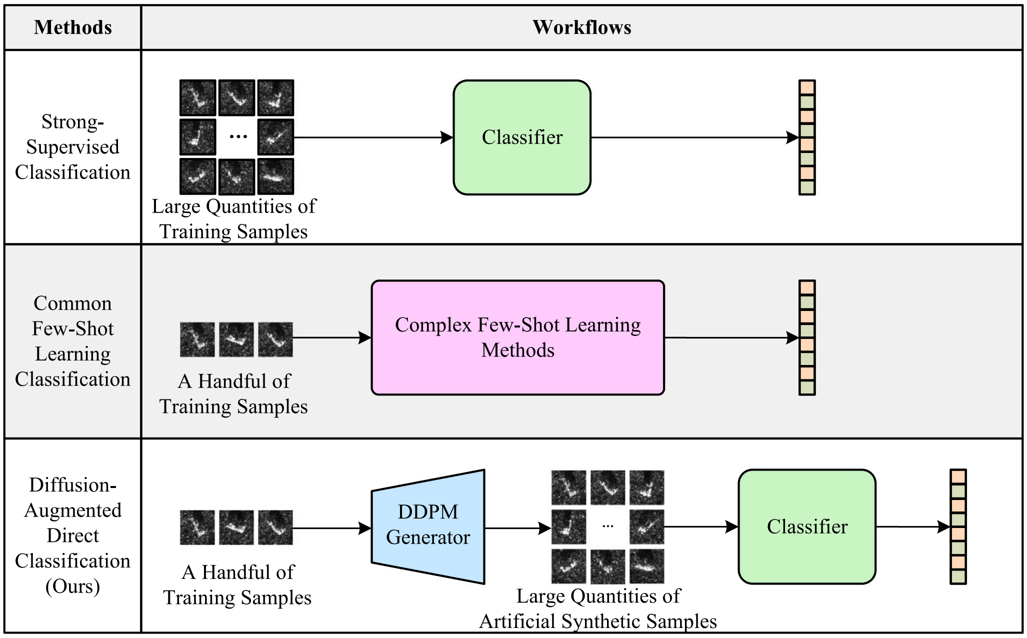

Fig. 1. Difference among traditional strong-supervised classification, common Few-Shot Learning classification, and our Diffusion-Augmented Direct Classification methods.

Theory. In a GAN, the generator produces artificial images based on an input noise vector, while the discriminator evaluates whether the input is genuine or fake. In the course of this adversarial process, the generator becomes prudent enough to synthesize artifacts that closely resemble real-world examples that is realistic enough to deceive the discriminator. Due to their ability to produce highly-realistic images, there are works (Qin et al., 2022; Liu et al., 2024; Du et al., 2022) incorporate the GAN as their basis for SAR image generation. However, due to the scarcity of image features in SAR images compared to ordinary natural images, to better make use of existing information in guiding the generation model while also acting as a conditioning mechanism for generating SAR images with given specifications, some generative models incorporate extra data as labels or feature embeddings into the GAN model. Among this extra information, the azimuth angle is one of the most crucial attributes that not only affects the SAR image generation quality but is also crucial for downstream image target recognition tasks. To generate SAR image samples with arbitrary azimuth angles, Zeng et al. (2024) and Wang et al. (2022) propose GAN-based generation models respectively, utilizing the source azimuth angle as conditioning embedding for their respective generators, making their models capable of generating SAR image samples with a given azimuth angle setting while also improving generation quality. Downstream target recognition models can therefore benefit from training data containing samples with various azimuth angle settings. An example of a method that extensively utilizes azimuth data in recognition tasks is proposed by Zhao et al. (2024), who present an azimuth-aware few-shot SAR-ATR method that significantly improves few-shot recognition accuracy in class-incremental SAR-ATR tasks by using azimuth data more effectively and solving the azimuth dependence problem. Beside the azimuth angle, Sun et al. (2023a) utilize the attributes of targets, including categorical labels, aspect angle, and auxiliary data related to a specific dataset, such as whether the ground target has armor or wheels, to improve the quality of SAR image generation when training data is scarce. However, despite the capability and effectiveness GANs can provide in SAR image generation tasks, training a GANs is challenging because of the adversarial nature of the training process. If the model capacity of the generator mismatches that of the discriminator, a phenomenon known as *mode collapse* may occur, preventing the GAN from generating diverse samples.

To overcome the aforementioned problem, a generative model that does not require the adversarial training while still capable of producing diverse and realistic samples is needed in SAR generation tasks. A feasible choice is the Denoising Diffusion Probability Model (DDPM), a generative model that is rooted in the principles of nonequilibrium thermodynamics theory. The training process of the Diffusion-based models is a Markov Chain comprising a forward noise addition process and a backward denoising process. By gradually adding noise to an image until it becomes pure noise, and then removing the noise step by step, the diffusion model learns the mapping relationship between images, noise, and timesteps, thereby enabling it to produce realistic samples from random noise. Benefiting from both generation quality and training stability, many works utilize diffusion-based generative models for SAR image generation. To achieve generation quality better than GAN-based method in SAR image generation tasks, Qosja et al. (2024) proposed SAR image generation method based on the DDPM and explored the optimal network choice and the procedure of pre-training DDPM model with clutter SAR image samples. However, this work only explores the quality of SAR images generated using a DDPM model, without investigating the application of diffusion models in SAR-ATR tasks. To address high-resolution range profile recognition tasks, Zhou et al. (2024) also proposed a two-stage method based on the DDPM model, which first pre-trains the model using a certain amount of existing samples, and then uses the fine-tuned model to generate synthetic target samples as augmentation for the training data in the downstream domain-adaptive target recognition task. The effectiveness of both methods are confirm on ground target recognition tasks using MSTAR dataset as the benchmark. For target recognition tasks other than ground targets, Zhang et al. (2024) proposed a diffusion-based model for marine target recognition. Instead of using a pre-training and fine-tuning paradigm, the method proposed an instance-to-image inpainting model to inpaint samples from other domains, such as optical images, into SAR image samples as augmentation for training the ship detection model in the downstream task. Although this work uses generated samples to train a recognition model, unlike our method, it does not provide a way to apply diffusion models in a few-shot learning scenario, where labeled samples are scarce. Meanwhile, to further improve the generation quality of SAR images, Ying et al. (2025) proposed a diffusion model that utilized the frequency characteristics of SAR

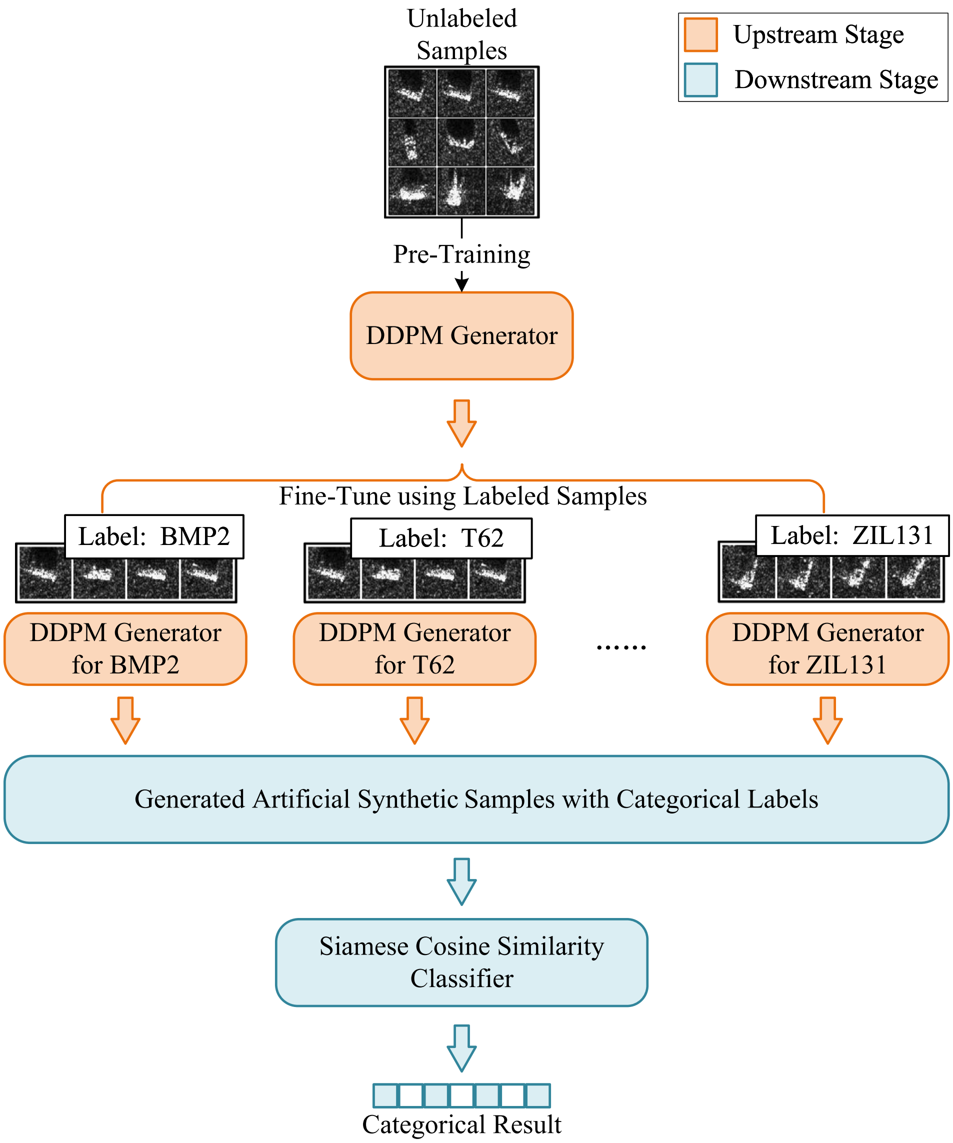

Fig. 2. The general workflow of our DADC framework comprises an upstream stage for sample generation and a downstream stage for actual classification. The DDPM generator is first trained with a limited number of unlabeled samples and then fine-tuned with a handful of labeled samples from different categories. Each fine-tuned generator will then be used to synthesize artificial samples for the training of the downstream classifier.

images by separating high-frequency and low-frequency components of the input during noise prediction, leading to improved generation performance. All methods aforementioned confirm the feasibility of utilizing a diffusion-based model to synthesize artificial samples for training recognition models in few-shot scenarios.

Given the diffusion-based image generation model's ability to synthesize highly realistic samples and its stable convergence process, we have selected it as our image generator in the upstream stage. we will utilize the diffusion-based image generator in the upstream stage to synthesize artificial SAR images for training the classifier in the downstream recognition stage.

2.2. Few-shot learning methods for SAR-ATR tasks

Since acquiring large quantities of annotated training samples for SAR-ATR is expensive and time-consuming, many attempts have been made to utilize few-shot learning methods in SAR-ATR. Besides using traditional generalization-enhancing techniques (Wang et al., 2023a) to ensure model generalization in few-shot learning scenarios, dedicated few-shot learning frameworks have also been used in SAR-ATR tasks. Meta-learning-based methods have been widely used in SAR-ATR tasks when training data is insufficient. To address the shortage of training samples in few-shot SAR-ATR tasks, Wang et al. (2021) introduced probabilistic inference with Meta-Learning in SAR-ATR tasks, which first trains the model using simulated image samples and then uses real SAR images as the query set for target recognition. Instead of using simulated samples for base training, Fu et al. (2022) applied Meta-Learning to SAR-ATR tasks in a different few-shot learning setting, using samples from part of the categories in the same dataset for base training and then utilizing the trained meta-model for recognition on targets in categories that are not part of the training data. The proposed method achieved high recognition accuracy in 1-shot and 5-shot few-shot recognition tasks in the given scenario. There are also works that adapt general-purpose meta-learning models to few-shot SAR-ATR tasks. An example is the work by Wu et al. (2024), who adapted the Prototypical Network to SAR-ATR tasks by adding a dynamically weighted mechanism to stabilize the training on SAR images, as the diversity and features of SAR images are significantly different from the natural images for which the Prototypical Network was originally designed. A similar dynamic mechanism is also used by other methods, such as the method proposed by Wang et al. (2023b), which is an entropy-aware meta-learning approach that constructs a dynamic feature space to enhance the model's effectiveness on samples from previously unseen categories. In contrast to our method, the above meta-learning-based approaches tackle few-shot learning and recognition together, leading to increased model and training complexity.

In conjunction with meta-learning-based methods, several practices have applied transfer learning methods to SAR-ATR tasks to make use of the knowledge that model can learned from existing large scale datasets. To tackle SAR-ATR tasks in scenarios where labeled data is limited, Huang et al. (2017) proposed a transfer learning-based method that utilizes a dual-pathway CNN network for knowledge transfer, which includes an autoencoder for data reconstruction. Besides transferring knowledge from data in other scenarios, Malmgren-Hansen et al. (2017) utilized transfer learning techniques to transfer features learned from simulated SAR image data for recognition tasks on real-world SAR target images. This demonstrates that simulated or synthesized data can be used for training recognition models via knowledge transfer in few-shot SAR-ATR tasks. Since SAR images differ significantly from other types of images, Huang et al. (2020) further discussed which networks and layers are more suitable for transfer learning in few-shot SAR-ATR tasks to address the knowledge discrepancy that may hinder the performance of transfer learning. However, transfer learning-based methods are affected by domain shift, as there is a substantial gap between SAR and optical images.

Class-incremental SAR target recognition tasks are another research direction within the few-shot SAR-ATR problem that requires specifically designed methods. Unlike ordinary few-shot SAR-ATR tasks, the problem that class-incremental SAR target recognition methods aim to solve is the scarcity of training samples for newly emerging target types. The typical procedure for incremental training is to first train a model with abundant samples from known and existing target types, and then adapt the model to unknown target types with a small number of training samples. Due to the scarcity of samples for the unknown targets, models are susceptible to overfitting the newly introduced samples. To mitigate this problem, Kong et al. (2025) proposed a method for class-incremental recognition tasks that utilizes orthogonal distribution features and random augmentation to alleviate overfitting. Meanwhile, during the adaptation phase, to prevent the model from forgetting learned knowledge, Gao et al. (2024) proposed a method based on strongly separable features, which successfully increases the efficiency and accuracy of incremental SAR-ATR tasks. Notably, image generation models can also be used in class-incremental tasks by pre-training the generation model with samples of known targets and then fine-tuning it using a handful of samples from unknown targets, as proposed in Zeng et al. (2024) and Wang et al. (2022).

Notably, there are also works that incorporate generative models into SAR-ATR tasks in scenarios where training data is scarce. In order to improve the diversity of SAR image samples when training data is insufficient, Song et al. (2021) proposed an adversarial-based autoencoder for SAR image generation from scratch and confirmed the feasibility of utilizing the generated samples as augmented data for FSL SAR-ATR tasks. Sun et al. (2023b) incorporated the CycleGAN model for synthesizing SAR images from optical images to supply its training data in its SAR-ATR task. To accomplish ship detection tasks when it is impossible to access more SAR sensors to acquire large quantities of SAR images, Ju et al. (2023) proposed a method based on the GAN model for limited-data ship detection tasks that capable specific target images by giving the model the location and type of the desired targets. Hampered by a lack of SAR samples at different aspect angles, Sun et al. (2020) proposed an angular rotation generative network that also accepts conditioning input, namely the aspect angle, to produce SAR training samples with given aspect angles for better target recognition results. These works demonstrate the feasibility of applying generative models, especially GANs, in few-shot SAR-ATR tasks. However, unlike our approach that uses a diffusion model for sample generation, GAN-based models often suffer from mode collapse and unstable training, and the aforementioned methods generally do not explore dedicated downstream classifiers designed to work effectively with the generated samples. Since diffusion-based generative models are more capable than GANs, applying them in few-shot SAR-ATR tasks holds greater potential to improve recognition accuracy and efficiency.

3. Proposed method

In this section, we will begin by outlining the fundamental concept and structure of our DADC framework. Subsequently, we will scrutinize the implementation details of our diffusion image generator. Finally, we will introduce the downstream Siamese Similarity Classifier and its training procedure.

3.1. Diffusion-augmented direct classification framework

As Fig. 2 illustrated, the overarching concept of the DADC framework involves employing a robust image generator to comprehend the distribution of SAR images and subsequently utilize it as a data augmenter to produce extensive quantities of image samples. This facilitates direct training of the downstream classifier on the generated samples. Consequently, our DADC framework is structured into two distinct stages: an upstream diffusion augmentation stage and a downstream classification stage.

The upstream diffusion augmentation stage can be further divided into two sequential steps. Initially, the diffusion model is trained using unlabeled SAR images to acquaint itself with the general characteristics and distribution of SAR images. This step uses unlabeled SAR image samples as input to the diffusion model, training the model to generate artificial samples through a forward noise-adding diffusion process and a backward denoising diffusion process. Given the absence of labels, this initial training step operates in an unsupervised manner. Subsequently, in the second step, a fine-tuning procedure will be carried out by training the model using limited quantities of labeled samples from each category, producing category-specific fine-tuned models that are capable of generating SAR images belonging to specific classes.

In the downstream stage, the classifier is directly trained on samples generated by the upstream diffusion generator. As the few-shot problem is addressed in the upstream stage, any general-purpose classification model can be employed in the downstream stage, affording genuine plug-and-play flexibility.

3.2. Diffusion SAR image generator

As Fig. 3 illustrates, there are two processes in the DDPM model: a procedure that introduces noise to the image named forward process and an opposite procedure that reconstructs the image from pure noise called reverse process. For a comprehensive understanding of the DDPM model and its mathematical foundation, the detailed exposition provided in Ho et al. (2020) can be referred to.

In the forward process, a real-world SAR image x_0, adhering to the distribution $q(x_0)$, undergoes the addition of Gaussian noise with different variance. The variance of added noise is controlled by the timestep t, a scalar value ranging from 1 to T, where T is the maximum number of diffusion steps. With the progress of timesteps, the variance of the Gaussian noise is increased accordingly, resulting higher level of noise and more corrupted images x_t until reaching the fully corrupted image x_T. Here, β_t represents the variance of the normal distribution at timestep t, and α_t is defined as $\alpha_t = 1 - \beta_t$ while $\bar{\alpha}_t$ is defined as $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Subsequently, a model can be trained to discern both the distribution of the source and the noise at timestep t by taking the noise-polluted samples x_t and timestep t as the input and predicting the added noise x_t'. Mean Square Error loss function can then be utilized to

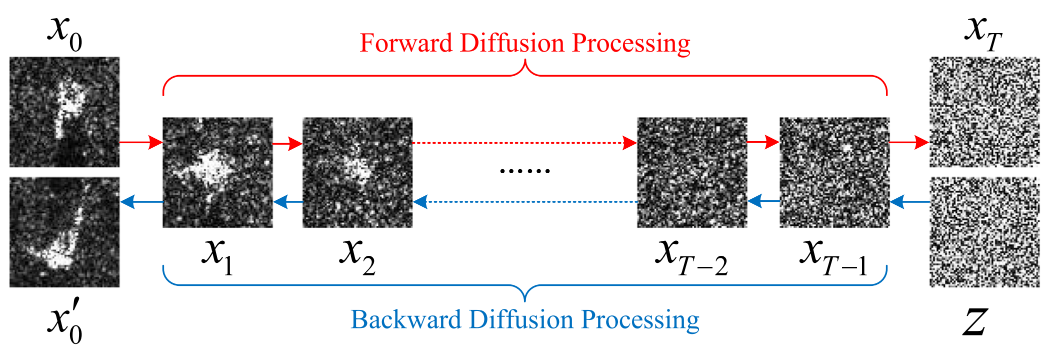

Fig. 3. The forward and backward processing of the diffusion generator. x_0 is the original image. In forward processing, noise is gradually added onto the image, and x_1 to x_{T-1} is the intermediate result while x_T is the image totally covered by noise. In backward processing, z is a random Gaussian noise which will be gradually denoised until reaching x_0', the generated artificial sample.

measure the discrepancy between predicted and ground-truth noise for model training, as per Eq. (1):

$$\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2 \quad (1)$$

ϵ represents noise following a normal distribution, while ϵ_θ denotes a function capable of predicting the original noise based on the timestep t and the noise-polluted image $\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$.

In the backward process, starting with the initial stochastic blank noise $x_T \sim \mathcal{N}(0, I)$, we can perform one step of denoising using the Eq. (2):

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z \quad (2)$$

where $z \sim \mathcal{N}(0, I)$ and σ is its standard deviation. By repeating this process until t reaches 0, an artificial synthetic SAR image will be produced. Diverse SAR images can be generated by inputting different z and x_T into the model.

In our implementation, an architecture incorporating a U-Net with multiple shuffle residual modules serves as the ϵ_θ function. This choice leverages the simplicity and efficiency inherent in a convolution-based structure. The skeleton of our U-Net is illustrated in Fig. 4.

3.2.1. Shuffle residual module

Inspired by Zhang et al. (2018), we have designed the shuffle residual module (SRM) for our U-Net, aiming to create a lightweight yet effective backbone. The SRM divides the input into two groups, each undergoing a separate depth-wised convolution operation. The outputs of these convolutions are then mixed using channel shuffling. The output of the shuffle residual module can be computed using the Eqs. (3), (4), (5), and (6):

$$x_1, x_2 = \mathrm{Trunk}(x_{\mathrm{in}}) \quad (3)$$

$$x_1' = \mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{3\times3}(x_1)) \quad (4)$$

$$x_2' = \mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{3\times3}(\mathrm{Conv}_{1\times1}(x_2))) \quad (5)$$

$$x_{\mathrm{out}} = \mathrm{ChannelShuffle}(\mathrm{Concat}(x_1', x_2')) \quad (6)$$

Please be advised that we omit some of the components in the equations above for a cleaner expression, namely the batch normalization and a sigmoid linear unit (SiLU) activation function, which is applied after each convolutional operation. The skeleton of our Shuffle Residual Module implementation is also displayed in Fig. 4.

3.2.2. Encoder blocks

Inside our U-Net, an encoder block encodes the inputs into latent space features. As illustrated in Fig. 4, there are two encoder blocks in our U-Net. Each encoder block comprises three SRMs with a stride of 1 and one SRM with a stride of 2, which downsamples the input to half resolution. The downsampled result and the residual are then outputted. It should be noted that the latent block is identical to an encoder block without the final downsampling SRM. The output of an encoder block can be computed using the Eqs. (7) and (8):

$$x_{\mathrm{residual}} = \mathrm{SRM}^{(3)}_{s=1}(x_{\mathrm{in}}) \quad (7)$$

$$x_{\mathrm{result}} = \mathrm{SRM}_{s=2}(x_{\mathrm{residual}}) \quad (8)$$

where $SRM^{(k)}_{s=n}$ denotes a Shuffle Residual Module with the stride set to n and be applied repeatedly to the input k times.

3.2.3. Decoder blocks

A decoder block resembles an encoder block, albeit with the initial step of upsampling the input to double resolution using the Bilinear interpolation algorithm. Subsequently, the residual from the corresponding encoder block is concatenated with the upsampled input. This concatenated input then undergoes 4 SRMs, and the output can be computed using the Eqs. (9), (10), and (11):

$$x_{\mathrm{up}} = \mathrm{Upsample}(x_{\mathrm{in}}) \quad (9)$$

$$x_{\mathrm{merged}} = \mathrm{Concat}(x_{\mathrm{up}}, x_{\mathrm{residual}}) \quad (10)$$

$$x_{\mathrm{out}} = \mathrm{SRM}^{(4)}_{s=1}(x_{\mathrm{merged}}) \quad (11)$$

3.2.4. Timesteps embedding

Timesteps, denoted as t, are integrated into the U-Net by adding t to the inputs of every encoder and decoder block in the U-Net. The numerical value of t is firstly converted into a tensor whose size matches that of the input tensor of the shuffle block. Subsequently, these two tensors are added together. The output of a timestep embedding can be computed using the Eqs. (12) and (13):

$$f_t = \mathrm{MLP}(\mathrm{Embedding}(t)) \quad (12)$$

$$f_{\mathrm{in}}' = f_{\mathrm{in}} + f_t \quad (13)$$

where t is the numeric value of the timestep; f_t is the embedded timesteps vector; f_{in} is the input of a Encoder or Decoder.

3.3. Siamese similarity classifier

The classifier is responsible for executing the actual recognition task in the downstream stage. In theory, since the few-shot problem has been resolved by the diffusion generator in the upstream stage, we could employ arbitrary general-purpose classification models. However, the SAR-ATR task presents unique challenges: SAR images are inherently noisy and low-resolution, with differences between two SAR images often subtle and difficult to discern even for human observers. Therefore, instead of instructing a model to discern the pixel features of SAR samples, we have developed a Siamese Similarity Classifier (SSC) to teach the model to differentiate between SAR images based on their

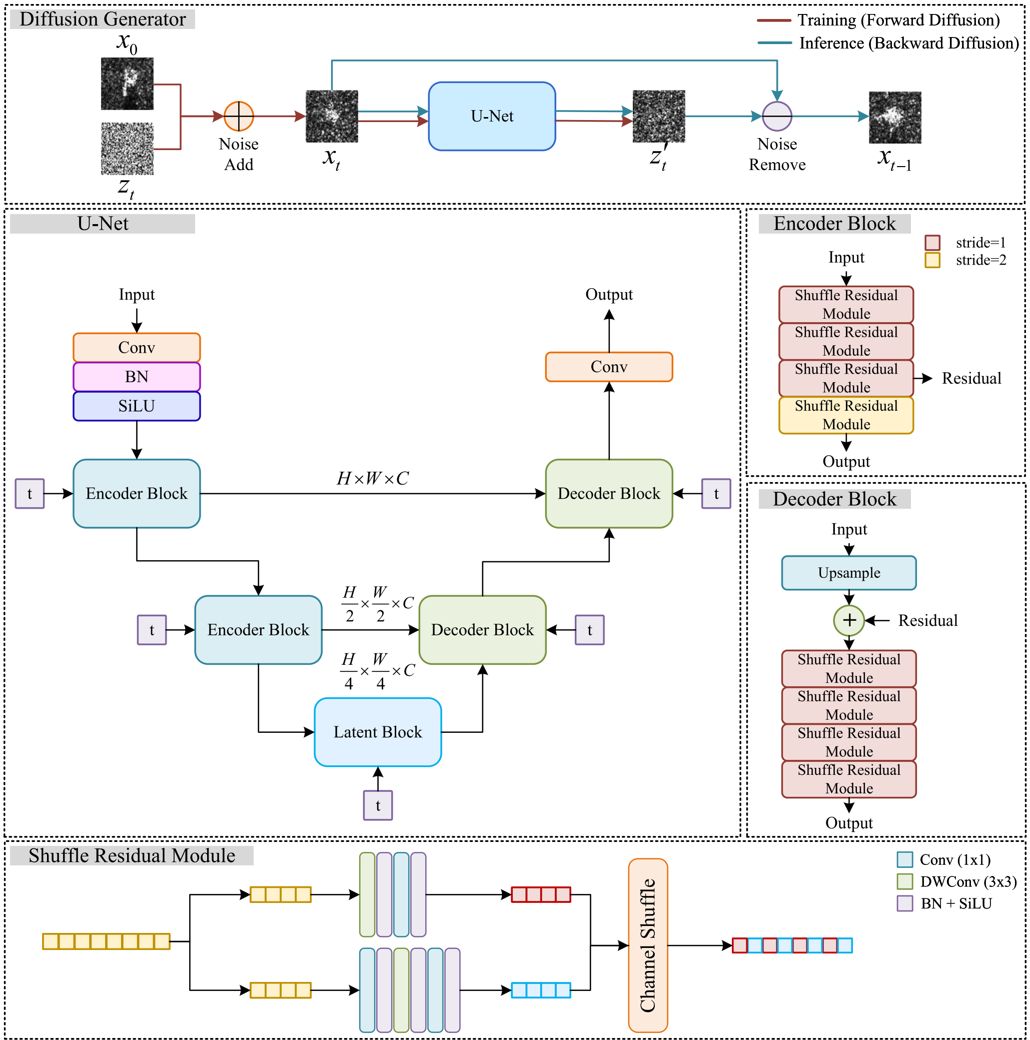

Fig. 4. An U-Net is utilized in our generator to predict the noise at timestep t. The structure of the U-Net contains two Encoder Blocks and two Decoder blocks. Each Encoder Block and Decoder Block contains four Shuffle Residual Modules. The last Shuffle Residual Module in an Encoder Block will downsample the input to half resolution by setting the stride to 2. The Latent Block is identical to Encoder blocks except it will not downsample the input. Decoder Blocks will first upsample the input and concatenate it with the residual before passing it to Shuffle Residual Modules. And t is timesteps embedding which will be simply added to the inputs of every Encoder Blocks and Decoder Blocks.

dissimilarities. The complete architecture of our SSC is illustrated in Fig. 5.

Our SSC is lightweight and straightforward, consisting of just three components: a feature extractor, a cosine similarity measurement layer, and a projection layer. Amid the training procedure, we arbitrarily fetch pairs of samples from the training dataset. Pairs containing samples from the same category are labeled as positive pairs, while pairs containing samples from different categories are labeled as negative pairs. Both images in a pair are fed into the same feature extractor to obtain two image feature vectors as in Eq. (14):

$$f_1, f_2 = \text{FeatureExtractor}(x_1, x_2) \tag{14}$$

Subsequently, the cosine similarity layer calculates the similarity between the two feature vectors as in Eq. (15):

$$\text{similarity} = \frac{f_1 \cdot f_2}{\max(||f_1||_2 \cdot ||f_2||_2, \epsilon)} \tag{15}$$

The sigmoid function is utilized to map the similarity value to a predicted positive or negative label. Given that the similarity score falls within the range of [−1, 1], which is too narrow for direct input into a sigmoid function, an additional projection layer is introduced to expand the range before passing it to the sigmoid function as in Eq. (16):

$$x = \frac{1}{1 + e^{-\text{Projection(similarity)}}} \tag{16}$$

Finally, the Binary Cross Entropy function is utilized as the criterion to verify the correctness of the predictions as in Eq. (17):

$$L_n = -w_n(y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)) \tag{17}$$

in which x_n is the predicted label and y_n denotes the ground truth label.

During the inference process, both the input image and images from a reference set are fed through the feature extractor to acquire their respective feature vectors. Subsequently, the cosine similarity function will be used to reckon the similitude between the input and all feature representation coming from the reference set. The category label associate with the reference set vector exhibiting the highest similarity is chosen as the ultimate classification result as in Eq. (18):

$$y_{\text{out}} = \text{argmax}\left(\frac{xQ^T}{||x||_2 \cdot ||Q^T||_2}\right) \tag{18}$$

Among the equation, x represents the feature representation of the input image, while Q is all feature vectors from the reference set.

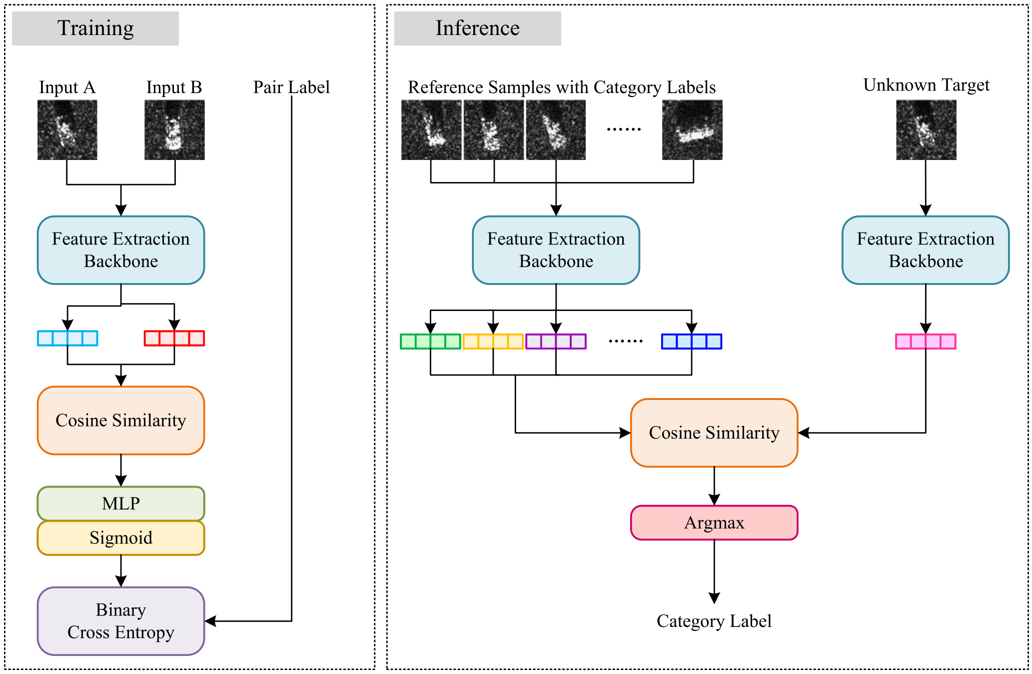

Fig. 5. The structure of the Siamese Similarity Classifier. In the training phase, pairs of samples will be converted to feature vectors by the same Feature Extraction Backbone, and the cosine similarity function will be used to measure the distance between two vectors. The value range of the cosine similarity function will be expanded via the combination of an MLP and a sigmoid function before being inputted into the Binary Cross Entropy loss function. Pair Label is an integer either 0 or 1, indicating whether two samples belong to the same category. In the inference phrase, an unknown target's feature vector will be compared with vectors from the reference samples set. The label of the sample in the reference set with maximum similarity will be selected as the classification result.

4. Experiments

We will undertake extensive experiments to assess the efficacy of both the proposed framework and its individual components in this section.

4.1. Dataset and experiment settings

In this section we will describe the dataset we used, the preparation of the dataset, and the experiment settings.

4.1.1. Metrics

To evaluate the quality and diversity of the generated images in the upstream stage, we use the Fréchet Inception Distance (FID) as our metric, which can be calculated by Eq. (19):

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \tag{19}$$

where μ_r and μ_g denotes the mean feature vectors derived from real images and synthesized images respectively, while Σ_r and Σ_g denote the covariance matrices of the feature maps obtained from both real and synthesized images.

The metric used to evaluate the performance of the downstream classification model and the general DADC framework is the classification accuracy and macro-F1 score. The formula of classification accuracy is show in Eq. (20). The definition of macro-F1 score is provided in Eqs. (21), (22), (23), and (24).

$$\mathrm{Accuracy} = \frac{\text{Correct Prediction Num}}{\text{Total Sample Num}} \tag{20}$$

$$P_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{21}$$

$$R_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{22}$$

$$\mathrm{F1}_i = 2 \times \frac{P_i \times R_i}{P_i + R_i} \tag{23}$$

$$\text{macro-F1} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{F1}_i \tag{24}$$

4.1.2. MSTAR dataset

The MSTAR dataset comprises SAR target recognition data which includes SAR images representing ten distinct military target classes obtained from radar scans at depression angles of 15° or 17°. Sample images from the MSTAR dataset are illustrated in Fig. 6, and the number of samples for each category is provided in Table 1. To ensure model generalization, we adhere to the common practice used in many works (Zhai et al., 2022a,b), which utilize images captured at 15° of depression for testing and make use of those with 17° angles of depression for training.

Table 1
Number of samples for each category in the MSTAR dataset.

Target	Num		
	17° depression	15° depression	Total
2S1	299	274	573
BMP-2	233	587	820
BRDM-2	298	274	572
BTR-60	256	195	451
BTR-70	233	196	429
D7	299	274	573
T-62	299	273	572
T-72	232	582	814
ZIL131	299	274	573
ZSU-23-4	299	274	573
Total	2747	3203	5950

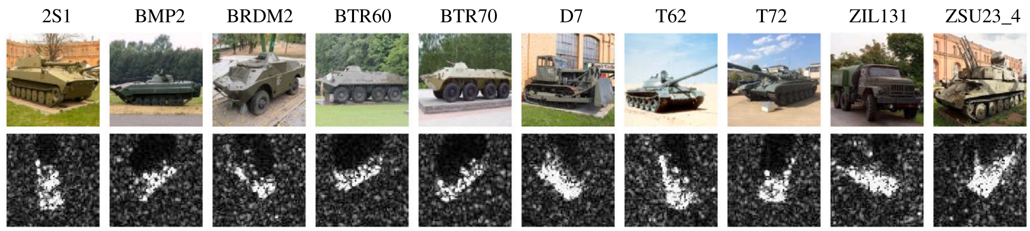

Fig. 6. Sample images from the MSTAR dataset.

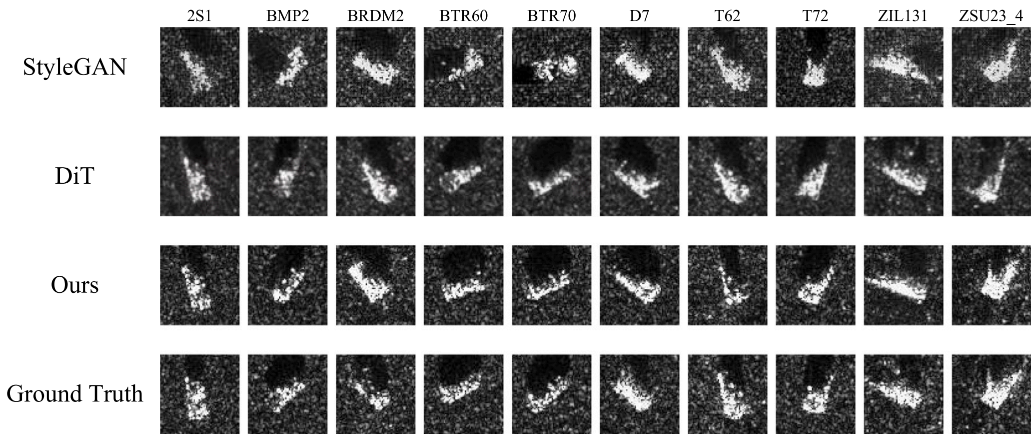

Fig. 7. Visual inspection of generated samples from StyleGAN2-Ada, DiT-S/2, and our diffusion generator. Samples generated by StyleGAN contain an unnatural "grid effect" in background noise; samples generated by DiT appear blurred; samples generated by our method look most realistic. Zoom in for a better view.

4.1.3. Few-shot database

Before conducting our experiments, it is imperative to partition the MSTAR dataset into smaller sample databases. Typically, two approaches are employed for sampling smaller databases from a larger dataset: one involves selecting a fixed proportion of images for each class, while the other entails selecting a fixed quantities of samples per category. Given the disparate sample counts across classes in the MSTAR dataset, we opt for the latter approach. To be precise, we randomly pick 5 and 10 samples from individual categories in the gross dataset to constitute our 5-samples and 10-samples small sample databases. It should be emphasized that this procedure is exclusive to the training set itself, while the whole test set of the MSTAR dataset remains intact for evaluation purposes.

During the training of the upstream diffusion image generator, only the training set from the small sample databases will be utilized, given the unsupervised nature of the upstream stage's training process. Subsequently, during the training of the downstream classifier, the training set for the classifier will incorporate images generated by the upstream diffusion image generator. The reference and test set will consist of samples from the training and test sets of the small-sample databases, respectively.

4.1.4. Experiment settings

The diffusion generator and the Siamese classifier within our DADC framework were implemented using version 2.0.1 of the PyTorch deep learning framework. Experiments were performed on a workstation with a single Nvidia GeForce RTX 3060 graphic adapter, boasting 12 GB of VRAM. In the training of the upstream diffusion generator, we designate the optimizer for the model to be AdamW. During the initialization, we empirically set the learning rate to 1e−3, and we utilized a one-cycle cosine annealing learning rate scheduling strategy. Due to VRAM constraints, the batch size had to be fixed at 64. For training the downstream Siamese classifier, we again chose to utilize the AdamW as our optimizer, with a batch size setting to 128 to fully utilize available VRAM. We employed a cosine annealing learning rate scheduling strategy with initial LR set to 1e−4. To expedite training and reduce VRAM consumption, we leveraged the automatic mixed precision (AMP) mechanism provided by the PyTorch framework in this stage. The Fine-tuned upstream generator for each category is utilized to generate 500 artificially synthesized images per category, which formed the training dataset for the downstream classifier.

Table 2
Downstream classification accuracy of different methods.

Methods	Classification accuracy (%)	
	5 samples per cls.	10 samples per cls.
DeepEMD sample	50.34 ± 0.39	52.36 ± 0.28
DeepEMD grid	52.24 ± 0.37	56.04 ± 0.31
DeepEMD	53.14 ± 0.40	59.64 ± 0.39
Relation	53.14 ± 0.40	59.64 ± 0.39
DN4	53.48 ± 0.41	64.88 ± 0.34
Prototypical	69.42 ± 0.39	78.01 ± 0.29
FTL	70.53 ± 0.37	79.46 ± 0.32
FTL-dis	72.13 ± 0.40	81.21 ± 0.31
DKTS-N	72.32 ± 0.32	84.59 ± 0.24
AADR	69.91 ± 0.29	78.77 ± 0.19
CFHFL	**74.52**	83.46
DADC (Ours)	73.31 ± 0.66	**85.67 ± 0.30**

4.2. Comparison with other FSL methods

To determine the effectiveness of our DADC framework and to benchmark its performance against other cutting-edge FSL methods, we chose the following methods or frameworks as baseline models: FTL, DeepEMD and its variants, DN4, Prototypical Network, Relation Network, DKTS-N, AADR, and CFHFL:

1. *FTL* (Zhang et al., 2023b), a two-stage few-shot SAR-ATR framework, integrates an augment-and-fine-tune concept similar to ours. It utilizes a style transfer model to augment the training

Table 3
Comparison of different upstream generators.

Models	FID↓	Downstream classification accuracy (%)	
		5 samples per cls.	10 samples per cls.
None	–	38.78	66.38
StyleGAN2-Ada	22.392	64.60	81.33
DiT-S/2	19.642	61.10	82.39
Ours	**9.467**	**73.31**	**85.67**

Table 4
Performance of different upstream generators.

Models	Parameters↓	GFLOPs↓	Training time↓ (s/epoch)
StyleGAN2-Ada	45,645,839	15.48	248.12
DiT-S/2	32,963,360	17.39	244.20
Ours	**1,107,549**	**0.66**	**41.74**

Table 5
Classification accuracy of different feature extraction backbones.

Backbone	Classification accuracy (%)		F1 score	
	5 samp. per cls.	10 samp. per cls.	5 samp. per cls.	10 samp. per cls.
ResNet18	24.51	33.38	0.2326	0.3300
ResNet10	23.77	32.86	0.2309	0.3349
ResNet50	23.21	33.14	0.2264	0.3250
ShuffleNet	24.09	33.76	0.2211	0.3279
MobileNetv2	23.91	32.35	0.2267	0.3184
ConvNext-Tiny	22.96	34.34	0.2235	0.3389
ViT-Base-16	20.69	35.83	0.1853	0.3432
SwinTransformer-Tiny	22.48	34.25	0.2129	0.3351
ResNet18 + DADC	**73.31**	85.67	**0.7135**	0.8546
ResNet10 + DADC	71.40	84.61	0.6989	0.8387
ResNet50 + DADC	69.99	86.64	0.6852	0.8659
ShuffleNet + DADC	72.46	85.58	0.7117	0.8529
MobileNetv2 + DADC	70.75	84.42	0.6916	0.8445
ConvNext-Tiny + DADC	70.84	86.48	0.6962	0.8620
ViT-Base-16 + DADC	68.41	**90.63**	0.6747	**0.9042**
SwinTransformer-Tiny + DADC	67.94	89.20	0.6549	0.8935

samples, subsequently fine-tuning the few-shot classifier on the augmented data.

2. *DeepEMD* (Zhang et al., 2023a), a FSL method that partitions images into blocks and employs Earth Mover's Distance to assess these likeness amid blocks from the support set and the query set, thereby determining the category of an image.
3. *DN4* (Li et al., 2019), a metric learning method that picks multiple local descriptive features from a sample and measures the similarity of these features against a metric image to determine the category of the image.
4. *Prototypical Network* (Snell et al., 2017) and *Relation Network* (Sung et al., 2018b). Both Prototypical Network and Relation Network are meta-learning-based few-shot learning methods. They convert the input and query images into a feature vector and select the most similar one from the query set to the input as the result.
5. *DKTS-N* (Zhang et al., 2022), a few-shot SAR target recognition method that utilize domain knowledge of the target such as the azimuth angle and phase data.
6. *AADR* (Zhang et al., 2023c), a semi-supervised few-shot SAR-ATR method that make avail of azimuth-aware discriminative representation feature.
7. *CFHFL* (Wen et al., 2024), a few-shot learning framework that extract multiple hierarchical feature from image for SAR image target recognition tasks.

The comparison result is presented in Table 2. Our DADC framework achieve comparable classification accuracy on 5-samples-per-class database and highest classification accuracy on 10-samples-per-class databases, demonstrating its capability of tackling the few-shot SAR-ATR tasks.

4.3. Ablation study

In this subsection, we will evaluate the performance of both upstream generators and downstream classifier respectively.

4.3.1. Effectiveness of the diffusion generator

In order to evaluate the effectiveness of our proposed Shuffle-U-Net-based SAR image generator, we replaced the upstream generator with other cutting-edge image generators and assessed their effectiveness based on generation quality measured by FID and final classification accuracy. Specifically, we selected two models for comparison: StyleGAN2-ada (Karras et al., 2020) and DiT (Peebles and Xie, 2023). StyleGAN2-ada is a state-of-the-art GAN-based generator. DiT is another DDPM-based generator that utilizes the self-attention mechanism. Additionally, we tested the impact on final classification accuracy when no upstream generator was used, and instead, simple data augmentation techniques (random shift and random rotation) were employed. To ensure an impartial comparison, all generators were trained with the same total sample number, either explicitly set in their arguments or adjusted using the batch-size setting couple with total epoch number to match the total sample number defined for our proposed framework. Meanwhile, the downstream architecture across all comparisons is identical, including the same Siamese-based structure, the same AdamW optimizer with the learning rate set to 1e-3, the same ResNet-18 backbone, and the same training strategy.

After training, each model was used to generate 5000 samples for evaluation of FID scores and visual inspection. The FID scores were computed on the generated samples, and visual inspection was performed. The visual inspection results are presented in Fig. 7. Notably,

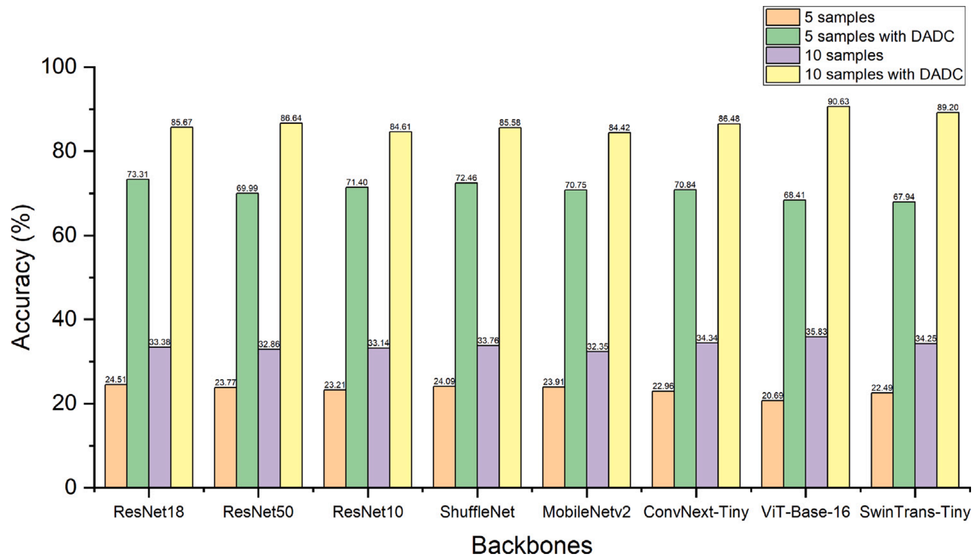

Fig. 8. Bar chart illustrating classification accuracy of classifiers using different feature extraction backbones.

Table 6
Computation complexity of different feature extraction backbones.

Backbone	Parameters	GFLOPs
ResNet18	13.280 M	0.301
ResNet10	4.922 M	0.181
ResNet50	23.508 M	0.676
ShuffleNet	5.345 M	0.097
MobileNetv2	2.224M	0.053
ConvNext-Tiny	27.810 M	0.729
ViT-Base-16	57.298 M	22.571
SwinTransformer-Tiny	18.852 M	0.486

samples generated by StyleGAN exhibited blocking patterns in the background noise, while those generated by DiT appeared blurred. In contrast, our method produced clear samples that closely matched the characteristics of real-world SAR images. Table 3 presents the FID scores and final classification accuracy results. In addition to the comparison of generation quality, we also conduct a performance comparison between the generative models, with the results shown in Table 4.

4.3.2. Backbone in the siamese similarity classifier

To determine how our DADC framework will benefit these plain-old classification models, and also investigate the impact of using different feature extraction backbones, we replaced ResNet18 with various network models and evaluated the final classification accuracy. For convolution-based networks, we selected ResNet (He et al., 2016) with varying numbers of layers, ShuffleNet (Zhang et al., 2018), MobileNet (Howard et al., 2017), and ConvNext (Liu et al., 2022) as our comparison targets. For Transformer-based networks, we chose ViT (Dosovitskiy et al., 2020) and SwinTransformer (Liu et al., 2021). We then evaluated the final classification accuracy of classifiers with different backbones. The recognition accuracy of different backbones are displayed in Table 5 and Fig. 8. The performance of different backbones are presented in Table 6.

4.3.3. Effectiveness of the Siamese Similarity structure

To assess the performance of our Siamese Similarity Classifier, we carried out experiments to analyze the consequence of removing key components from the classifier on the final classification accuracy. Firstly, we removed the final projection layer from the classifier to assess its contribution. Subsequently, we completely eliminated the Siamese network and the similarity comparison mechanism to gauge their individual effects. Utilizing ResNet18 as the backbone, we compared the performance of the original Siamese Similarity Classifier, the variant without projection layers, and the variant without the Siamese network on the small sample database. The results are presented in Table 7.

4.3.4. Effect of generated samples quantities

To investigate how the number of generated samples from the upstream diffusion generation model affects the classification accuracy of the downstream classifier, we have conducted an experiment that utilizes different total numbers of generated samples for downstream classifier training. In the experiment, a baseline Siamese similarity classifier that uses a ResNet18 as its feature extraction backbone is trained with various numbers of generated samples in a 10-reference-samples-per-category scenario. The results are presented in Table 8 and Fig. 9.

From the results, we can observe that the classification accuracy marginally benefits from an increasing number of generated samples until it reaches a plateau at 20,000 samples in total. The classification accuracy begins to drop significantly when the total number of generated samples falls below 500. The conclusion drawn from this result is that the number of generated samples must match the scale of the recognition problem, as an ordinary supervised classifier does. In our DADC workflow, any number exceeding 2000 samples per category is unnecessary, as it only increases the time required for sample generation. The minimum number required for training before accuracy degradation occurs is approximately 100 samples per category.

4.3.5. Generalization across dataset

To evaluate the generalization capability of our proposed method, we have conducted additional generalization experiments using a different dataset with SAR images samples capture in different scenario, band, and polarization mode. The dataset we used to test the generalization capability is SRSDD (Lei et al., 2021) dataset, which contains SAR image samples from 6 different marine categories. The samples from each category in the SRSDD dataset are provided in Fig. 10. To conduct the experiments, we cropped out the target chips from the SAR images in the SRSDD dataset and followed the same procedure that we used on the MSTAR dataset. The results of the experiments are provided in Table 9.

Table 7
Classification accuracy of the classifier with partial of the components enabled.

Projection layers (MLP + sigmoid)	Siamese structure	Classification accuracy (%)	
		5 samples per cls.	10 samples per cls.
✓	✓	**73.31**	**85.67**
✗	✓	72.50	84.36
✗	✗	67.31	83.23

Table 8
Donwstream classification accuracy with different numbers of generated samples.

Num in total	50	100	250	500	1000	2500	5000	10 000	20 000
Accuracy (%)	51.12	58.42	74.53	82.62	84.73	84.64	85.67	85.30	85.61

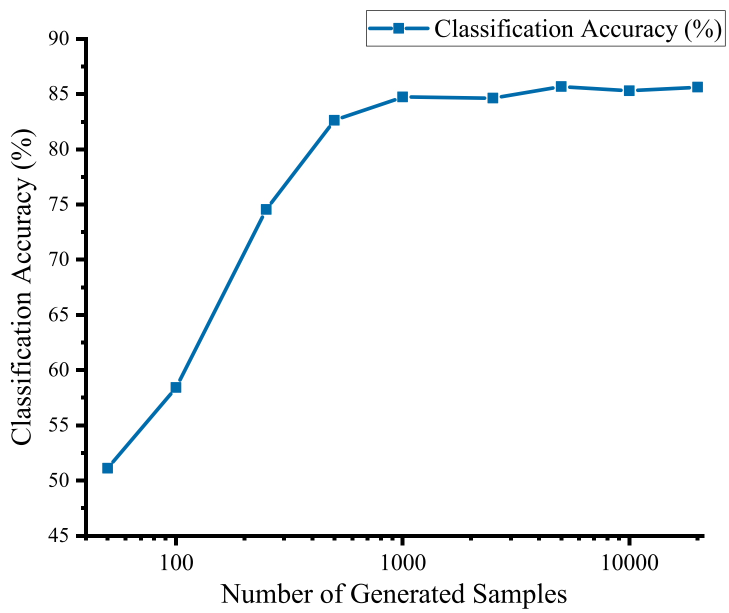

Fig. 9. The line chart of donwstream classification accuracy with model trained by different numbers of generated samples.

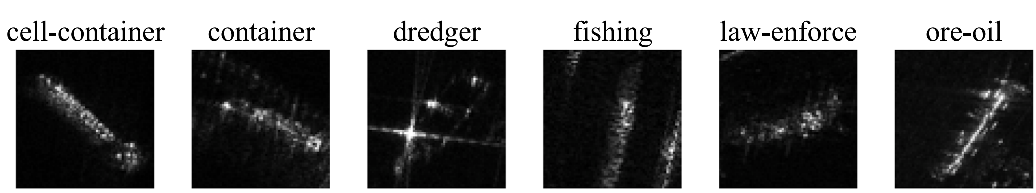

Fig. 10. Samples from each category in the SRSDD dataset.

Table 9
Final classification accuracy of methods on SRSDD dataset.

Method	Classification accuracy (%)	
	5 samples per cls.	10 samples per cls.
Resnet18	25.20	31.61
DADC with Resnet18 as backbone	58.06	62.75

From the results, we can observe that without our proposed DADC framework, the pure ResNet18 network can hardly perform the recognition tasks in both 5-shot-per-category and 10-shot-per-category scenarios. In contrast, when utilizing ResNet18 as the feature extraction backbone in our Siamese-based similarity classification model, the recognition accuracy in both 5-shot-per-category and 10-shot-per-category scenarios is significantly increased, making it competent in marine target recognition tasks, which are more challenging than ground target recognition tasks due to more interference from the ocean surface and shore obstacles.

4.3.6. Discussion

By conducting the ablation study, we assessed the significance of each element in our framework. In the upstream stage, our Shuffle-U-Net-based diffusion model achieved the lowest FID scores, indicating that the samples generated by our model closely matched the distribution of real-world SAR images. Moreover, our model achieved the highest classification accuracy in the downstream task. Importantly, using any image generation model resulted in significantly higher classification accuracy compared to no image generation at all, underscoring the effectiveness of the general DADC framework.

Table 10
Recognition accuracy when interference by various strength of speckle noise.

Noise Strength	α=0.9	α=0.8	α=0.7	α=0.6	α=0.5	α=0.4	α=0.3	α=0.2
Prototypical Network	49.92	49.50	48.68	50.64	50.14	50.46	48.66	49.79
DADC (ours)	70.22	70.75	71.12	71.59	71.31	70.90	68.65	60.19

By changing the backbone in our downstream classifier, we observed a strong correlation between backbone performance, the quantities of available training samples, and backbone complexity. In scenarios with low available training samples, such as 5-samples-per-class classification, more complex backbones tended to have lower classification accuracy. We believe this is due to the complexity of these networks, which makes them more susceptible to overfitting the artificially synthesized training samples, thereby hindering further increases in classification accuracy. Conversely, in scenarios with adequate training samples, such as 10-samples-per-class classification, backbones with more parameters tended to achieve higher classification accuracy, particularly those based on Transformers. Among all the backbones tested, ResNet18 demonstrated a balanced performance, yielding competent results across all both few-shot learning scenarios while maintaining reasonable parameter count and computational complexity.

To assess the viability of the general skeleton of our classifier, we disabled some of the components in the classifier respectively and observed the impact on the final classification accuracy. The findings underscore the importance of the projection layers, Siamese network structure, and similarity comparison mechanism in our classifier. One possible explanation for these results is that the use of the Siamese Similarity Classifier expands the total sample space. By training the model on positive and negative sample pairs, the total quantities of them, denoted as C_k^2, is much larger than the k samples used in plain classification training. The enlarged sample space makes the model less susceptible to overfitting on the training data.

4.4. Resistance to speckle noise

SAR images are susceptible to speckle noise, a multiplicative noise that causes speckles on SAR image samples. Speckle noise may significantly impair the recognition accuracy of a SAR-ATR model, particularly in the few-shot learning scenario. To evaluate the robustness of our proposed method when samples are affected by speckle noise, we conducted additional noise-resistance experiments following the test procedure described in Zhai et al. (2023). The experiment is conducted in a 5-samples-per-category few-shot learning scenario setting. During the model training phase, no noise is added to the training samples. During the test phase, test samples are corrupted with various strengths of speckle noise to evaluate whether the trained model is robust enough to resist speckle noise and maintain a reasonable recognition accuracy even when interfered with by speckle noise. As in Zhai et al. (2023), a parameter α is used to control the strength of the speckle noise in the experiment, with smaller α values yielding stronger speckle noise. We selected the Prototypical Network as a comparison baseline because its recognition mechanism is similar to that of the downstream classifier in our proposed method, both relying on the similarity between target samples and query samples. The result of the experiment is provided in Table 10.

The experimental results reveal that our proposed method can still achieve a competent recognition accuracy even when SAR image samples are contaminated with different levels of speckle noise, confirming the robustness of the proposed method. In analyzing the results, we believe that the ability to resist speckle noise in the few-shot learning scenario stems from the nature of the downstream classifier, which selects the most similar reference sample using cosine similarity rather than fitting to a specific categorical probability distribution. This makes the model less susceptible to corrupted feature vectors produced by the feature extraction backbone when processing samples contaminated with speckle noise.

5. Discussion and conclusion

In this paper, we introduce a two-stage framework named Diffusion-Augmented Direct Classification for few-shot SAR target recognition tasks, designed to better align with real-world SAR-ATR scenarios. By utilizing an image generation model to learn the features from unlabeled samples of known targets, the proposed method splits the entire few-shot SAR-ATR process into an upstream sample generation stage and a downstream target recognition stage. Innovatively, the two-stage framework decouples the few-shot problem and the recognition problem, assigning the few-shot problem to the upstream image generation model, while the downstream recognition model, which will eventually be deployed for target recognition, can be implemented to be as lightweight as possible. This also opens the possibility of reusing existing recognition models, offering greater flexibility in real-world SAR-ATR tasks. Delving into the details of our model, to overcome the convergence stability problem and the diversity problem that GAN-based image generation models may encounter, we implement the upstream image generation using the DDPM paradigm. A downstream classifier based on a Siamese cosine-similarity model is proposed to work with training samples synthesized by the upstream generation model. Experiments reveal the superiority of our DDPM-based image generation upstream in terms of generated quality and the reduction of defects and artifacts in the generated samples. Meanwhile, the effectiveness of the Siamese-based downstream recognition model is also confirmed by its classification accuracy.

Envisioning future work, several remaining challenges can be further addressed. One limitation arises from the sample generation process using the DDPM-based model, which is time-consuming since it requires each sample to be iteratively processed for up to T steps in the backward diffusion process. Efforts can be made to reduce the number of steps required for generation by exploring faster diffusion paradigms, such as implicit diffusion models, whose backward diffusion process requires only a few iteration steps. The efficiency of fine-tuning also has room for improvement, as newly emerging fine-tuning techniques, such as low-rank adaptation, are worth exploring. Regarding the downstream recognition model, the application of additional auxiliary techniques, such as pseudo-labeling, may further enhance the efficiency of utilizing synthetic samples, which is also a research direction in future works.

CRediT authorship contribution statement

Zilu Ying: Conceptualization. **Wenyu Ke:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Yikui Zhai:** Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Xinglin Liu:** Validation. **Jianhong Zhou:** Writing – review & editing, Conceptualization. **Pasquale Coscia:** Validation. **Angelo Genovese:** Validation.

Acknowledgments

Funding:This work was supported by Guangdong Basic and Applied Basic Research Foundation [grant numbers 2021A1515011576], Guangdong Science and Technology Planning Project [grant numbers 2021A0505030080, 2021A0505060011], Guangdong Higher Education Innovation and Strengthening School Project [grant numbers 2023ZDZX1029, 2024ZDZX1009], Wuyi University Hong Kong and Macao Joint Research and Development Fund [grant numbers 2022W GALH19], Guangdong Jiangmen Science and Technology Research Project [grant numbers 2220002000246, 2023760300070008390].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ai, J., Mao, Y., Luo, Q., Jia, L., Xing, M., 2022. SAR target classification using the multikernel-size feature fusion-based convolutional neural network. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2021.3106915.

Chen, S., Wang, H., Xu, F., Jin, Y., 2016. Target classification using the deep convolutional networks for SAR images. IEEE Trans. Geosci. Remote Sens. 54 (8), 4806–4817. http://dx.doi.org/10.1109/TGRS.2016.2551720.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations.

Du, S., Hong, J., Wang, Y., Qi, Y., 2022. A high-quality multicategory SAR images generation method with multiconstraint GAN for ATR. IEEE Geosci. Remote. Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2021.3065682.

Fu, K., Zhang, T., Zhang, Y., Wang, Z., Sun, X., 2022. Few-shot SAR target classification via metalearning. IEEE Trans. Geosci. Remote Sens. 60, 1–14. http://dx.doi.org/10.1109/TGRS.2021.3058249.

Fu, X., Zhou, Z., Meng, H., Li, S., 2024. A synthetic aperture radar small ship detector based on transformers and multi-dimensional parallel feature extraction. Eng. Appl. Artif. Intell. 137, 109049. http://dx.doi.org/10.1016/j.engappai.2024.109049.

Gao, F., Kong, L., Lang, R., Sun, J., Wang, J., Hussain, A., Zhou, H., 2024. SAR target incremental recognition based on features with strong separability. IEEE Trans. Geosci. Remote Sens. 62, 1–13. http://dx.doi.org/10.1109/TGRS.2024.3351636.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. http://dx.doi.org/10.48550/arXiv.1704.04861, arXiv preprint arXiv:1704.04861.

Hu, J., Guo, R., Zhu, X., Baier, G., Wang, Y., 2015. Non-local means filter for polarimetric SAR speckle reduction-experiments using TERRASAR-X data. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. II-3/W4, 71–77. http://dx.doi.org/10.5194/isprsannals-II-3-W4-71-2015.

Huang, Z., Pan, Z., Lei, B., 2017. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. Remote. Sens. 9 (9), 907. http://dx.doi.org/10.3390/rs9090907.

Huang, Z., Pan, Z., Lei, B., 2020. What, where, and how to transfer in SAR target recognition based on deep CNNs. IEEE Trans. Geosci. Remote Sens. 58 (4), 2324–2336. http://dx.doi.org/10.1109/TGRS.2019.2947634.

Hummel, R., 2000. Model-based ATR using synthetic aperture radar. In: Record of the IEEE 2000 International Radar Conference [Cat. No. 00CH37037]. pp. 856–861. http://dx.doi.org/10.1109/RADAR.2000.851947.

Jin, S., Li, X., Yang, X., Zhang, J.A., Shen, D., 2018. Identification of tropical cyclone centers in SAR imagery based on template matching and particle swarm optimization algorithms. IEEE Trans. Geosci. Remote Sens. 57 (1), 598–608. http://dx.doi.org/10.1109/TGRS.2018.2863259.

Ju, M., Niu, B., Hu, Q., 2023. SARGAN: A novel SAR image generation method for SAR ship detection task. IEEE Sensors J. 23 (22), 28500–28512. http://dx.doi.org/10.1109/JSEN.2023.3323322.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020. Training generative adversarial networks with limited data. In: Proc. NeurIPS.

Kong, L., Gao, F., He, X., Wang, J., Sun, J., Zhou, H., Hussain, A., 2025. Few-shot class-incremental SAR target recognition via orthogonal distributed features. IEEE Trans. Aerosp. Electron. Syst. 61 (1), 325–341. http://dx.doi.org/10.1109/TAES.2024.3443014.

Lei, S., Lu, D., Qiu, X., Ding, C., 2021. SRSDD-v1.0: a high-resolution SAR rotation ship detection dataset. Remote. Sens. 13 (24), 5104. http://dx.doi.org/10.3390/rs13245104.

Li, Y., Li, X., Sun, Q., Dong, Q., 2022. SAR image classification using CNN embeddings and metric learning. IEEE Geosci. Remote. Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2020.3022435.

Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J., 2019. Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7260–7268.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Liu, M., Wang, H., Chen, S., Tao, M., Wei, J., 2024. A two-stage SAR image generation algorithm based on GAN with reinforced constraint filtering and compensation techniques. Remote. Sens. 16 (11), 1963. http://dx.doi.org/10.3390/rs16111963.

Ma, X., Shen, H., Zhao, X., Zhang, L., 2016. SAR image despeckling by the use of variational methods with adaptive nonlocal functionals. IEEE Trans. Geosci. Remote Sens. 54 (6), 3421–3435. http://dx.doi.org/10.1109/TGRS.2016.2517627.

Malmgren-Hansen, D., Kusk, A., Dall, J., Nielsen, A.A., Engholm, R., Skriver, H., 2017. Improving SAR automatic target recognition models with transfer learning from simulated data. IEEE Geosci. Remote. Sens. Lett. 14 (9), 1484–1488. http://dx.doi.org/10.1109/LGRS.2017.2717486.

Novak, L., Halversen, S., Owirka, G., Hiett, M., 1997. Effects of polarization and resolution on SAR ATR. IEEE Trans. Aerosp. Electron. Syst. 33 (1), 102–116. http://dx.doi.org/10.1109/7.570713.

Peebles, W., Xie, S., 2023. Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205.

Qin, J., Liu, Z., Ran, L., Xie, R., Tang, J., Guo, Z., 2022. A target SAR image expansion method based on conditional wasserstein deep convolutional GAN for automatic target recognition. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 15, 7153–7170. http://dx.doi.org/10.1109/JSTARS.2022.3199091.

Qosja, D., Wagner, S., O'Hagan, D., 2024. SAR image synthesis with diffusion models. In: 2024 IEEE Radar Conference (RadarConf24). pp. 1–6. http://dx.doi.org/10.1109/RadarConf2458775.2024.10549257.

Shang, R., Wang, J., Jiao, L., Stolkin, R., Hou, B., Li, Y., 2018. SAR targets classification based on deep memory convolution neural networks and transfer parameters. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 11 (8), 2834–2846. http://dx.doi.org/10.1109/JSTARS.2018.2836909.

Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. vol. 30.

Song, Q., Xu, F., Zhu, X.X., Jin, Y.-Q., 2021. Learning to generate SAR images with adversarial autoencoder. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Sun, Y., Wang, Y., Hu, L., Huang, Y., Liu, H., Wang, S., Zhang, C., 2023a. Attribute-guided generative adversarial network with improved episode training strategy for few-shot SAR image generation. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 1785–1801. http://dx.doi.org/10.1109/JSTARS.2023.3239633.

Sun, Y., Wang, Y., Liu, H., Wang, N., Wang, J., 2020. SAR target recognition with limited training data based on angular rotation generative network. IEEE Geosci. Remote. Sens. Lett. 17 (11), 1928–1932. http://dx.doi.org/10.1109/LGRS.2019.2958379.

Sun, Y., Yan, K., Li, W., 2023b. CycleGAN-based SAR-optical image fusion for target recognition. Remote. Sens. 15 (5569), http://dx.doi.org/10.3390/rs15235569.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M., 2018a. Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M., 2018b. Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208.

Wang, C., Luo, S., Pei, J., Huang, Y., Zhang, Y., Yang, J., 2023a. Crucial feature capture and discrimination for limited training data SAR ATR. ISPRS J. Photogramm. Remote Sens. 204, 291–305. http://dx.doi.org/10.1016/j.isprsjprs.2023.09.014.

Wang, C., Luo, S., Pei, J., Liu, X., Huang, Y., Zhang, Y., Yang, J., 2023b. An entropy-awareness meta-learning method for SAR open-set ATR. IEEE Geosci. Remote. Sens. Lett. 20, 1–5. http://dx.doi.org/10.1109/LGRS.2023.3269620.

Wang, C., Pei, J., Liu, X., Huang, Y., Mao, D., Zhang, Y., Yang, J., 2022. SAR target image generation method using azimuth-controllable generative adversarial network. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 15, 9381–9397. http://dx.doi.org/10.1109/JSTARS.2022.3218369.

Wang, K., Zhang, G., Xu, Y., Leung, H., 2021. SAR target recognition based on probabilistic meta-learning. IEEE Geosci. Remote. Sens. Lett. 18 (4), 682–686. http://dx.doi.org/10.1109/LGRS.2020.2983988.

Wen, Y., Wang, X., Peng, L., Qiao, Y., 2024. A coarse-to-fine hierarchical feature learning for SAR automatic target recognition with limited data. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 13646–13656. http://dx.doi.org/10.1109/JSTARS.2024.3423377.

Wu, C., Yang, J., Shang, Y., Pei, J., Mao, D., Huo, W., Huang, Y., Zhang, Y., 2024. Dynamically weighted prototypical learning method for few-shot SAR ATR. IEEE Geosci. Remote. Sens. Lett. 21, 1–5. http://dx.doi.org/10.1109/LGRS.2024.3365147.

Ying, Z., Ke, W., Zhai, Y., Long, Z., Zhou, J., Zhu, H., Chen, C.L.P., 2025. Diffusar: frequency domain-aware diffusion model for SAR image generation. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 18, 11851–11866. http://dx.doi.org/10.1109/JSTARS.2025.3563798.

Zeng, Z., Tan, X., Zhang, X., Huang, Y., Wan, J., Chen, Z., 2024. ATGAN: a SAR target image generation method for automatic target recognition. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 6290–6307. http://dx.doi.org/10.1109/JSTARS.2024.3370185.

Zhai, Y., Liao, J., Sun, B., Jiang, Z., Ying, Z., Wang, W., Genovese, A., Piuri, V., Scotti, F., 2023. Dual consistency alignment based self-supervised learning for SAR target recognition with speckle noise resistance. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 3915–3928. http://dx.doi.org/10.1109/JSTARS.2023.3267824.

Zhai, Y., Zhou, W., Sun, B., Li, J., Ke, Q., Ying, Z., Gan, J., Mai, C., Labati, R.D., Piuri, V., Scotti, F., 2022a. Weakly contrastive learning via batch instance discrimination and feature clustering for small sample SAR ATR. IEEE Trans. Geosci. Remote Sens. 60, 1–17. http://dx.doi.org/10.1109/TGRS.2021.3066195.

Zhai, Y., Zhou, W., Sun, B., Li, J., Ke, Q., Ying, Z., Gan, J., Mai, C., Labati, R.D., Piuri, V., Scotti, F., 2022b. Weakly contrastive learning via batch instance discrimination and feature clustering for small sample SAR ATR. IEEE Trans. Geosci. Remote Sens. 60, 1–17. http://dx.doi.org/10.1109/TGRS.2021.3066195.

Zhang, C., Cai, Y., Lin, G., Shen, C., 2023a. DeepEMD: differentiable earth mover's distance for few-shot learning. IEEE Trans. Pattern Anal. Mach. Intell. 45 (5), 5632–5648. http://dx.doi.org/10.1109/TPAMI.2022.3217373.

Zhang, C., Dong, H., Deng, B., 2023b. Improving pre-training and fine-tuning for few-shot SAR automatic target recognition. Remote. Sens. 15 (6), 1709. http://dx.doi.org/10.3390/rs15061709.

Zhang, L., Leng, X., Feng, S., Ma, X., Ji, K., Kuang, G., Liu, L., 2022. Domain knowledge powered two-stream deep network for few-shot SAR vehicle recognition. IEEE Trans. Geosci. Remote Sens. 60, 1–15. http://dx.doi.org/10.1109/TGRS.2021.3116349.

Zhang, L., Leng, X., Feng, S., Ma, X., Ji, K., Kuang, G., Liu, L., 2023c. Azimuth-aware discriminative representation learning for semi-supervised few-shot SAR vehicle recognition. Remote. Sens. 15 (2), 331. http://dx.doi.org/10.3390/rs15020331.

Zhang, X., Li, Y., Li, F., Jiang, H., Wang, Y., Zhang, L., Zheng, L., Ding, Z., 2024. Ship-Go: SAR ship images inpainting via instance-to-image generative diffusion models. ISPRS J. Photogramm. Remote Sens. 207, 203–217. http://dx.doi.org/10.1016/j.isprsjprs.2023.12.002.

Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6848–6856.

Zhao, Y., Zhao, L., Zhang, S., Ji, K., Kuang, G., Liu, L., 2024. Azimuth-aware subspace classifier for few-shot class-incremental SAR ATR. IEEE Trans. Geosci. Remote Sens. 62, 1–20. http://dx.doi.org/10.1109/TGRS.2024.3354800.

Zhou, Q., Wang, Y., Zhang, X., Zhang, L., Long, T., 2024. Domain-adaptive HRRP generation using two-stage denoising diffusion probability model. IEEE Geosci. Remote. Sens. Lett. 21, 1–5. http://dx.doi.org/10.1109/LGRS.2024.3379275.