Hierarchical Disentanglement-Alignment Network for Robust SAR Vehicle Recognition

Weijie Li, Wei Yang, Wenpeng Zhang, Tianpeng Liu, Yongxiang Liu, *Member, IEEE*, and Li Liu, *Senior Member, IEEE*

***Abstract*—Vehicle recognition is a fundamentale problem in synthetic aperture radar (SAR) image interpretation. However, robustly recognizing vehicle targets is a challenging task in SAR due to the large intraclass variations and small interclass variations. In addition, the lack of large datasets further complicates the task. Inspired by the analysis of target signature variations and deep learning explainability, this article proposes a novel domain alignment framework, named the hierarchical disentanglement-alignment network (HDANet), to achieve robustness under various operating conditions. Concisely, HDANet integrates feature disentanglement and alignment into a unified framework with three modules: domain data generation; multitask-assisted mask disentanglement; and the domain alignment of target features. The first module generates diverse data for alignment, and three simple but effective data augmentation methods are designed to simulate target signature variations. The second module disentangles the target features from background clutter using the multitask-assisted mask to prevent clutter from interfering with subsequent alignment. The third module employs a contrastive loss for domain alignment to extract robust target features from generated diverse data and disentangled features. Finally, the proposed method demonstrates impressive robustness across nine operating conditions in the MSTAR dataset, and extensive qualitative and quantitative analyses validate the effectiveness of our framework.**

***Index Terms*—Automatic target recognition (ATR), deep learning, domain alignment, robustness, synthetic aperture radar (SAR).**

I. Introduction

THANKS to its attractive imaging capabilities in nearly all weather and illumination conditions, synthetic aperture radar (SAR) has become an indispensable means of information acquisition in Earth observation. In recent years, SAR imaging techniques [1], [2] have been rapidly developing, and high-resolution SAR images can be accessed more easily than before, enabling a wide field of applications. As a result, the amount of SAR image data is growing rapidly, which requires the development of intelligent SAR image interpretation techniques. As a fundamental problem in SAR image interpretation, SAR vehicle recognition [3], [4] aims to classify a vehicle into one of the predefined categories. It has various civilian and military applications, including transportation management, automatic driving, concealment detection, and military reconnaissance [5], [6], [7], [8], [9]. Therefore, it has been an active area for several decades.

Through decades of effort, SAR automatic target recognition (ATR) has witnessed significant progress. Especially in the past several years, deep learning has injected a new vitality in this field and brought great success [3], [4]. For instance, many existing methods have achieved over 99% accuracy [10], [11], [12], [13] on the widely used moving and stationary target acquisition and recognition (MSTAR) [14] dataset with ten categories of ground vehicles under standard conditions. Despite several decades' research in SAR ATR, most approaches have not been, however, capable of performing at a level sufficient for open, real-world applications [3], [4], [15], [16]. Robust SAR vehicle recognition for practical applications is still far from being solved. *What makes the problem of robust SAR vehicle recognition in the open world so challenging?* The main fundamental challenges of robust SAR vehicle recognition are summarized in Fig. 1 and discussed briefly in the following.

1) *High robustness to large intraclass variations:* As shown in Fig. 1, challenges [3], [4], [15], [16] in achieving high recognition accuracy stem from: 1) the vast range of natural and adversary-induced difficult deployment conditions; and 2) the interclass ambiguities between potential fine-grained target categories. Natural intraclass variations include at least three types (see Fig. 1): sensor operating conditions; target operating conditions; and environmental operating conditions. The scatter characteristics of the SAR target are highly sensitive to the aforementioned operating conditions, and therefore, robust SAR ATR requires features that are highly robust to numerous possible operating conditions.
2) *High distinctiveness to small interclass variations:* The interclass ambiguities of some target categories, especially fine-grained target categories, demand great discrimination power from the features to distinguish between subtly different interclass variations.

Manuscript received 7 October 2023; accepted 8 October 2023. Date of publication 12 October 2023; date of current version 25 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100800, in part by the National Natural Science Foundation of China under Grant 62376283, Grant 61871384, Grant 61901487, Grant 61901498, and Grant 61921001, and in part by the Science and Technology Innovation Program of Hunan Province under Grant 2022RC1092. *(Corresponding authors: Wei Yang; Tianpeng Liu; Yongxiang Liu.)*

The authors are with the College of Electronic Science and Technology, National University of Defense Technology Changsha, 410073, China (e-mail: lwj2150508321@sina.com; yw850716@sina.com; zhangwenpeng@hotmail.com; everliutianpeng@sina.cn; lyx_bible@sina.com; dreamliu2010@gmail.com).

Code is available online at https://github.com/waterdisappear/SAR-ATR-HDANet.

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2023.3324182, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2023.3324182

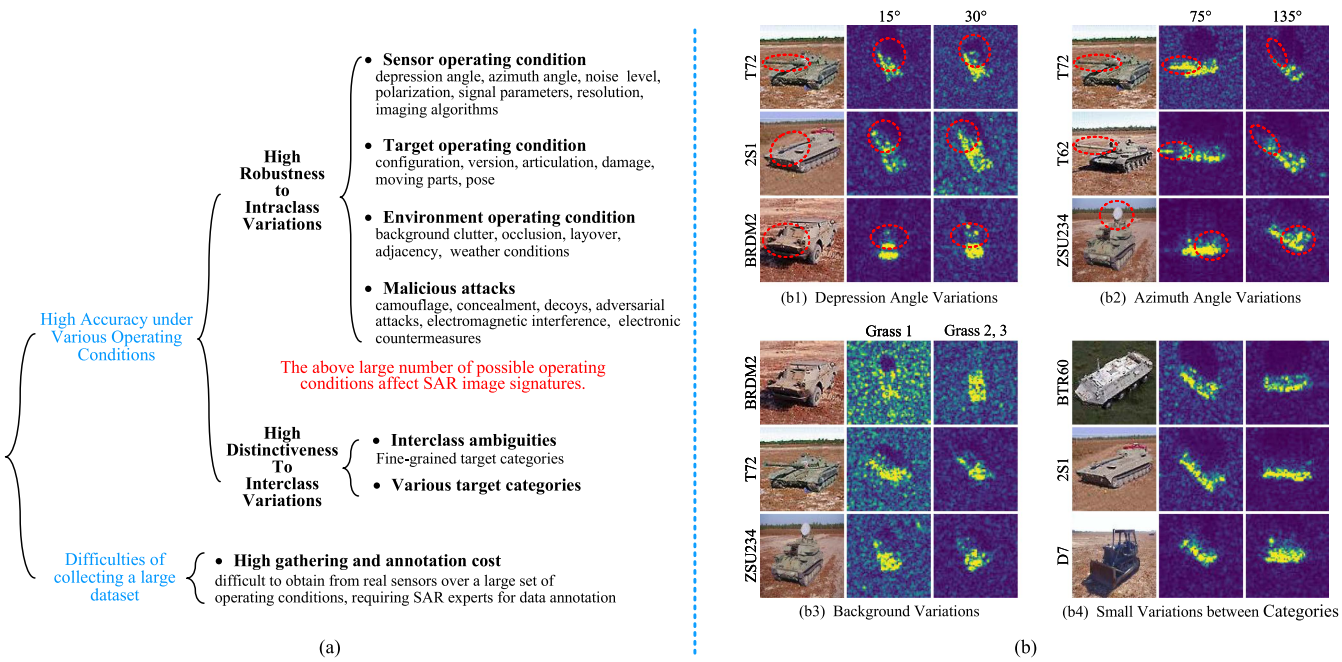

Fig. 1. Main challenges of robust SAR vehicle recognition. (a) Provides a taxonomy of these challenges brought by intraclass variations, interclass variations, and data collection. The right subfigure illustrates typical variations with SAR images in the MSTAR dataset. On the right, (b1), (b2), and (b3) show large intraclass variations due to the sensitivity to operating conditions. (b1) Contains target signatures and shadow variations in the red dashed line with different depression angles. (b2) Displays the variation of target partial structures in the red dashed line with azimuth angles, such as the T-72 tank gun barrel is most visible in the vertical line-of-sight direction. (b3) Illustrates that the intensity variation of different background clutter affects the adjacent target signatures. Therefore, SAR images of the same category have a large intraclass variation across operating conditions. (b4) Showcases small interclass variations between fine-grained vehicle target categories. Visual differences in SAR target signatures are much smaller than in natural images. The SAR images appear similar but are three different target categories. (a) Main Challenges of Robust SAR Vehicle Recognition. (b) The changes in SAR target signatures with variations in operating conditions.

3) *Lack of large target datasets:* It is highly difficult to obtain SAR images over a large set of operating conditions from real sensors, and collecting and annotating large-scale SAR vehicle datasets is clearly more timeconsuming and costly than in the natural images. Currently, methods for SAR vehicle recognition are mainly evaluated on the small MSTAR dataset collected under very constrained operating conditions. Many ATR methods have achieved near-perfect accuracy on MSTAR, suggesting the presence of strong bias in a small dataset. Therefore, the lack of large benchmark SAR vehicle datasets limits the power of deep learning methods requiring large amounts of training data and greatly impedes the development of SAR ATR technique.

Early recognition systems utilized template matching [17] and model-based [18] methods, relying on numerous samples or accurate electromagnetic (EM)-scattering models. Another technique based on machine learning depended on expert-designed features and suitable classifiers [4]. In recent years, deep learning methods have shown their superiority over traditional methods by learning features in a data-driven manner [3], [4], [16]. However, the small dataset hampers the potential of deep learning, and many data augmentation methods [19], [20], [21] have been employed to increase the richness of the dataset and enhance the robustness to extra specific operating conditions. Based on data augmentation, domain alignment methods [22], [23] were introduced to enhance features' invariance to intraclass variations further.

However, the potential of deep learning still needs to be more adequately exploited under the aforementioned fundamental challenges. Specifically, data biases, such as background correlation, [3] in a small dataset interfere with domain alignment and data-driven models. This problem results in the model using background clutter for recognition and reduced robustness [24]. Training and test set shifts due to SAR imaging sensitivities are complex and do not satisfy the independent and identically distributed assumption. These issues need domain alignment to achieve robust recognition under different operating conditions on single-domain datasets (i.e., the training sets often contain a small number of operating conditions with data bias).

In this article, to alleviate these challenges, a novel domain alignment framework, named the HDANet, is proposed to achieve robust SAR vehicle recognition under various operating conditions. The novelty of our framework stands on careful consideration of the complex variation in target signatures and the clutter interference hidden by small data, and HDANet performs feature disentanglement and alignment through three steps. First, three data augmentation methods are designed considering complex variations in target local signatures, which generate diverse domain data for alignment. Second, the multitask-assisted mask disentanglement module separates target and clutter regions at the feature layer because our previous work [24] on the

explainability of deep learning illustrated that deep learning overfits background clutter to reduce training errors. To enhance the target region weight in the mask, the segmentation task and l_1 regularization are applied as auxiliaries to add target location and sparse constraints. Third, the robustness is increased under various operating conditions. We treat operating conditions as domains and improve robustness by performing domain alignment of target features (DATF). A contrastive loss and corresponding designed structures achieve domain alignment. In the end, HDANet shows impressive robustness under MSTAR's nine operating conditions, and extensive qualitative and quantitative analyses demonstrate the effectiveness of our method. Moreover, the limitations of the proposed method are discussed.

The main contributions of this article are summarized as follows.

1) A novel domain alignment framework named HDANet is proposed to achieve robust recognition under various operating conditions. To better evaluate robustness, we design nine operating conditions, including a scene variation that has not been previously discussed.
2) Inspired by analysis of deep learning explainability and target signature variations, HDANet integrates feature disentanglement and alignment into a unified framework with three significant modules. The disentanglement suppresses clutter and ensures correct feature representation, and the alignment further enhances the robustness of deep learning features.
3) Compared with existing attention mechanisms, HDANet achieves feature disentanglement under data bias by adding priori constraints and improving the computational approach. Compared with existing domain alignment methods, HDANet increases the diversity of domain data generation (DDG) and considers the background clutter interference to achieve robust recognition under various operating conditions.

The rest of this article is organized as follows. Section II introduces related studies in SAR vehicle recognition. Section III introduces the framework of HDANet. Section IV conducts extensive experiments to demonstrate the robustness of our method, and numerous qualitative and quantitative analyses discuss the advantages and limitations of our framework. Finally, Section V concludes this article.

II. Related Work

This section reviews key issues in SAR ATR and focuses on robust recognition and deep learning-based approaches related to our work. Furthermore, considering our work draws on domain alignment, we briefly review this topic.

A. Key Issues in SAR ATR

Depending on the granularity of the target category, ATR can be divided into dimensions [25], such as detection (distinguishing between targets and clutter), classification (determining the type of target, e.g., tank), and identification (determining the specific type of target, e.g., T72). Studies based on MSTAR focus on identification, which is the fine-grained classification of targets, and we review the related work on the robustness, accuracy, efficiency, and explainability of SAR ATR [3], [15], [16] as follows.

1) Robustness: The main experimental settings for SAR vehicle target recognition based on the MSTAR dataset, include standard operating conditions (SOCs) and extended operating conditions (EOCs) [3]. The former is in a similar distribution and includes three or ten classes. The latter includes different variations for operating conditions. The popular EOC settings are the depression angle, configuration, and version variants in the [10]. In addition, other articles add noise [12], [26], occlusion [13], and other variations to this benchmark. Variations in operating conditions lead to complex shifts in the distribution between training and test sets, which leads to the task of robust recognition, i.e., achieving a high and stable classification performance in these cases. Early SAR ATR systems, included template-matching, model-based, and machine learning-based approaches [3], [4], [16]. In recent years, deep learning methods have also been widely applied in SAR vehicle recognition.

Template-matching methods: Collecting many samples can create a template library, and recognition depends on designed features and matching criteria. Ikeuchi et al. [17] used deformable template matching based on the invariant histogram. Tan et al. [27] applied the components of the target outline as matching templates.

Model-based methods: These methods generate different images from 3-D EM scattering or computer-aided design (CAD) models [18], and the recognition relies on accurate models and efficiency calculation. Ma et al. [28] matched the EM model predicted scattering centers with the test image scattering points' location and intensity. Ding [29] designed three similarity degrees to synthetically evaluate the similarity between the test image and the 3-D scattering center model.

Machine learning-based methods: These methods include some key parts, such as feature extraction and classifier. Previous work proposed many valuable methods, including geometric structure features [30], EM-scattering features [31], local descriptors [32], and sparse representations [33]. Classifiers, such as support vector machines and random forests, have also been absorbed into SAR vehicle recognition [4]. In recent years, deep learning-based methods have adopted an end-to-end approach to complete feature extraction and classification.

Deep learning-based methods: These methods can better learn correlations in a dataset but require large amounts of diverse samples. For example, Chen et al. [10] proposed an all-convolutional network (A-ConvNet) with random cropping. A-ConvNet shows a high accuracy under depression angle, configuration, and version variations but is not robust to random noise. Therefore, researchers are exploring various strategies to extracting robustness features. It is worth noting that the following methods can be combined to tackle the challenges in robust SAR vehicle recognition.

1) *Data augmentation:* It is a technique used to increase the size and diversity of a training dataset artificially. Many data augmentation approaches have been used to enhance deep learning's robustness to specific operating

conditions. Popular data augmentation in SAR vehicle recognition, includes translation [21], random cropping [10], affine transformation [34], elastic distortion [34], power transformation [19], rotation [19], [21], flipping [19], multiresolution [35], occlusion [35], and noise adding [19], [20], [21], [22], [35]. By creating augmented images, the training dataset becomes larger and more diverse, helping the model better generalize and handle intraclass variations. Therefore, we designed three data augmentation methods to simulate partial changes and overcome the drawback of a small dataset.

2) *Clutter suppression:* Due to background correlation interfering with the recognition, Zhou et al. [36] used masks to separate clutter and designed a large-margin Softmax batch-normalization (BN) convolution neural network (CNN). Heiligers and Huizing [37] employed CAD projection to separate target regions in input images for recognition and found that removing clutter in a similar scene may decrease the recognition rate. However, these methods are sensitive to hyperparameters or require the 3-D structure of the target and imaging parameters. Some researchers integrated segmentation methods with deep learning optimization. Li et al. [20] trained a fully connected (FC) codec with an input layer mask and sparse constraint to separate the target from the clutter. Ren et al. [11] proposed an extended convolutional capsule network (ECCNet) with a convolutional block attention module (CBAM) to suppress clutter. Moreover, channel attention was also used to address the effects of clutter [38]. Wang et al. [39] employed a multiscale CBAM for feature fusion. Therefore, the attention mechanism has been widely used in SAR target recognition. However, the question is whether the bias of a small dataset affects the attention mechanism. Our experiments found that CBAM enhances clutter in some MSTAR images to exploit background correlation better. Consequently, we use target location and sparse constraints to enhance the weight of target regions in masks.

3) *Feature extraction:* These methods that capture robust features with special structures are also being explored. Many works extract low-level and high-level features of vehicle targets. Lin et al. [40] increased the depth and width of the model using two convolutional highway layers with different kernel sizes. Shang et al. [41] developed a two-stage deep memory CNN to learn samples' spatial features. Ai et al. [42] used multikernel (MK) fusion to enhance the feature representation of CNN. Deformable convolution kernel [43], [44] was also applied to extract the scattering and morphological characteristics of the target. Inspired by the above work, we used multiscale feature maps and capsule networks to enhance the feature representation of deep learning. The hybrid feature is another popular strategy using traditionally designed features to improve the robustness of auto-extracted deep learning features. Zhang et al. [45] combined the designed multiorientation spatial features with a bidirectional long short-term memory network. Zhang et al. [12] proposed a lightweight modified VGG16 [46] (MVGGNet) with pretrained weights and combined this model with attributed scattering centers (ASC). Feng et al. [13] further used the physical features of ASC to constrain the deep learning features. Our approach addresses the shortcomings of deep learning features in SAR ATR, which can be used to improve the robustness of the deep learning modules in these hybrid methods.

4) *Transfer learning:* It has been applied to enhance robustness in small datasets due to the better diversity of large-scale datasets. In SAR vehicle recognition, transfer learning [47] involves leveraging pretrained models from other sensors or tasks, such as natural images or other remote sensing datasets, and adapting them to SAR vehicle recognition. The pretrained models capture generic midlevel features from large datasets, which can be beneficial in addressing downstream tasks [47].

5) *Domain alignment:* In SAR ATR, domain alignment addresses the differences between simulated and real data or extracts domain-invariant features for particular operating conditions. Wang et al. [48] integrated metalearning and adversarial learning for cross-domain and cross-task transfer learning from simulated to real data. Lewis et al. [49] filled the gap between simulated and real data through generative adversarial networks. Kwak et al. [22] proposed a speckle-noise-invariant network (SNINet) with l_2 regularization to align CNN feature maps after data augmentation. He et al. [23] applied a task-driven domain adaptation (TDDA) way to align the FC layer features of the simulated and real data by multikernel maximum mean discrepancy for the large depression variation. Two methods are similar to our work: SNINet [22] with l_2 contrastive loss and TDDA [23] with MK-MDD. However, these domain alignment approaches do not consider clutter interfering with feature robustness. We draw lessons from traditional detection and recognition processes to disentangle the target and clutter at the feature layer before domain alignment and recognition. Moreover, we consider partial changes in target signatures under various operating conditions rather than invariance to specific operating conditions.

2) Accuracy: In addition to a high-accuracy rate under large intraclass variations, another high accuracy requirement is for small interclass variations. Therefore, in addition to a high accuracy for existing fine-grained target categories, outlier rejection for various unknown target categories is also a concern task, including false alarm rate [10] and open-set recognition [50]. Chen et al. [10] set the confuser rejection rule of deep learning and tested operating characteristic curves with two confuser targets. Ma et al. [50] solved the open-set recognition problem by generative adversarial networks with classification and abnormal detection tasks.

3) Efficiency: SAR ATR systems need to consider development and deployment costs, such as data collection, training costs, and hardware resources [4], [15], [16]. In addition to the data augmentation and transfer learning described above, methods for efficiency problems also include data generation [51],

EM simulation [23], [48], few-shot learning [8], and model compression [52].

4) Explainability: Because of the black box problem, whether deep learning learns the correct feature representation in SAR ATR is an open question that deserves investigation. Previous explainability studies [24], [37], [53], [54] used posthoc methods to reveal the effect of target, clutter, and shadow regions on recognition, and the ablation studies [36], [54] discussed the influence of background clutter on the recognition rate. The above studies show that deep learning relies not only on target signatures but also on background clutter. This phenomenon is due to the bias in the small samples of the MSTAR dataset [24]. Since the data were collected under specific operating conditions, the background clutter and target categories are correlated. Therefore, we perform feature disentanglement to suppress background clutter and extract correct feature representations.

B. Domain Alignment

Since we consider operating conditions as domains, out-of-distribution generalization under different operating conditions is a domain generalization problem. Domain generalization [55] aims to generalize out-of-distribution using only source data, while the target domain may be difficult to obtain or even unknown. Let $\mathcal{X}$ be the input (feature) space and $\mathcal{Y}$ the output (label) space. A domain $\mathcal{S}$ is associated with a joint distribution P on $\mathcal{X} \times \mathcal{Y}$. Given K train (source) domain $\mathcal{S}^{\text{train}} = \{\mathcal{S}_k\}_{k=1}^{K}$ with different joint distributions $P_k \neq P_{k'}, 1 \leq k \neq k' \leq K$, the goal of domain generalization is to learn a model f from $\mathcal{S}^{\text{train}}$ to generalize on unseen test (target) domain $\mathcal{S}^{\text{test}}$[55] (i.e., $\mathcal{S}^{\text{test}}$ is not included in $\mathcal{S}^{\text{train}}$ and $P^{\text{test}} \neq P^{\text{train}}$)

$$\min_{f} \mathbb{E}_{(x,y)\in\mathcal{S}^{\text{test}}} \left[\mathcal{L}(f(x), y)\right] \tag{1}$$

where $\mathbb{E}$ is the expectation and $\mathcal{L}(\cdot,\cdot)$ is the loss function.

Domain alignment [55] is one of the common methods to solve the domain generalization problem. This method minimizes differences between different source domains and extracts domain-invariant features

$$\min_{f} \mathbb{E}_{(x,y)\in\mathcal{S}^{\text{train}}} [\mathcal{L}(f(x), y) + \lambda \mathcal{H}\left(P_k^X, P_{k'}^X\right)] \tag{2}$$

where λ is the tradeoff parameter to prevent a trivial solution, $\mathcal{H}(\cdot,\cdot)$ is domain distance metrics, and P^X is the marginal distribution. A common assumption is that the posterior distribution $P^{Y|X}$ remains stable, while domain shifts occur in the marginal distribution P^X. The domain distance measures [55] include moments, contrastive loss, Kullback–Leibler divergence, maximum mean discrepancy distance, and adversarial learning. This article assumes that domain shifts occur in P^X and use the contrastive loss with cosine similarity. The different source domain data are generated by data augmentation.

Moreover, from a causal perspective [56], only the alignment of target features is meaningful. Mahajan et al. [57] selected images of the same objects for domain generalization by a causal matching algorithm. Lv et al. [58] designed a causality-inspired representation learning algorithm, including amplitude intervention, factorization, and adversarial mask modules, to suppress noncausal factors, such as background, style, and viewpoint. Our work is also inspired by the causal domain alignment method in computer vision and proposes feature disentanglement and alignment for SAR recognition. Due to target features changing with operating conditions, both the causality and robustness of features are critical.

III. Methodology

This section introduces our method for robust SAR vehicle recognition. The main inspirations for designing a domain alignment framework for SAR are described in Section III-A. The holistic framework of HDANet is present in Section III-B, and its three modules are depicted in Sections III-C, III-D, and III-E, respectively.

A. Motivation

Feature disentanglement and alignment are our main concerns in adapting deep learning with domain knowledge for application to robust SAR vehicle recognition.

1) Feature disentanglement: The training data in SAR vehicle recognition are collected under restricted conditions, and the clutter of each target class has different strengths [24]. So deep learning can use these differences to reduce training errors. Generally speaking, clutter amplitudes in diverse scenes obey different statistical distributions. The strong randomness of clutter can lead to an instability of the extracted features. Therefore, the target and clutter features must be disentangled before domain alignment. Motivated by the traditional constant false alarm rate detector (CFAR) algorithm [59], we can enhance the causality of a deep learning model in SAR with an input image mask to distinguish between targets and clutter. However, deep learning models may focus on the mask's hard edges [37]. Moreover, the discontinuous strong clutter scattering point in SAR images increases learning difficulty. The input image mask learned by FC layers [20] detected strong clutter points at a 5 dB signal-to-clutter ratio. Therefore, a soft constraint mask at the middle feature layer is more suitable for deep learning in SAR. A deep network can filter noise when compressing information [60], and middle-layer features are generic for different categories [61]. Therefore, we extract masks at the feature layer to avoid noise points in the input image and to exploit the difference in middle patterns. This approach is similar to the attention mechanism. Nevertheless, deep learning models cannot suppress the correlation of clutter in MSTAR without prior constraints. The attention module [11] may overfit the background clutter, as shown in Fig. 2. Furthermore, its sigmoid activation function maps the background region's smaller values to around 0.5. Besides, the image reconstruction task [11], [20] needs to reconstruct both the target and clutter in the SAR image, which enhances the impact of clutter as well. Therefore, implementing feature disentanglement in SAR vehicle recognition, i.e., separating target and background clutter, requires constraints on the mask to overcome the adverse effects of inconspicuous target signatures and a small dataset.

In addition to the mask, another common preprocessing method is center cropping to remove background clutter. This

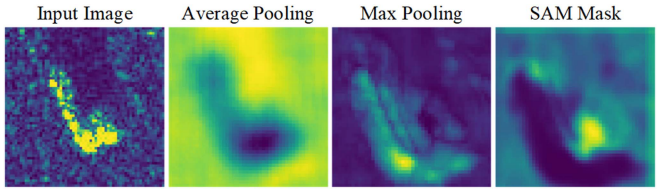

Fig. 2. Attention module results of CBAM. The spatial attention mechanism (SAM) in CBAM generates masks based on the pooling results of feature maps. SAM mask shows that the background clutter has more weight than the target, which indicates that data bias can affect mask learning.

way requires a priori knowledge of the target size. Our analysis in Table V gives that even cropping to a quarter of the original image size does not eliminate background clutter interference because the center cropping retains the background clutter near the target.

2) Feature alignment: Since SAR image properties are sensitive to operating conditions, the robustness of features under different operating conditions is critical for SAR target recognition. Extracting invariant target features under different operating conditions by domain alignment has a very attractive prospect for improving feature robustness. However, implementing domain alignment in SAR needs to address the following issues. Feature disentanglement solves the problem of extracting target features, and another question is how to generate domain data under different operating conditions. The training set in MSTAR only contains several constrained operating conditions (e.g., different azimuth angles) due to the high cost of obtaining target data under various operating conditions. Therefore, data augmentation is applied to generate domain data effectively. The proposed DDG is designed based on the assumption of local perturbations in target signatures [62]. Specifically, the overall structure of the target does not change significantly with imaging parameters, but the position and magnitude of a few scattering points change. Therefore, various operating conditions lead to partial changes in the target pixel points of SAR images. Local perturbations of target signatures in different domains are assumed to be changes in image pixel points' position, relative value, and absolute value. Based on this assumption, we simulate different domain data with three data augmentation methods. In addition, the feature representation used for domain alignment needs to be modified according to SAR image properties.

Shortly, the feature disentanglement aims to suppress clutter and extract target features for alignment, and feature alignment extract invariant target features across operating conditions. Inspired by the above two principles, we designed a novel domain alignment framework based on SAR image properties to achieve robust recognition in various operating conditions.

B. Overall Framework of HDANet

The overall framework of our method is shown in Fig. 3, and HDANet has three modules to achieve robust recognition through feature disentanglement and alignment.

1) Domain data generation: It consists of three approaches: rotation, noise perturbation, and random replacement. Moreover, these augmentation methods are applied with different parameters and probabilities to simulate complex variations. A pair of images ($\boldsymbol{x}_1'$ and $\boldsymbol{x}_2'$) can be obtained by sampling the set $\mathcal{T}$ of data augmentation methods twice (t_1 and t_2) for domain alignment. This module generates diverse domain data from the MSATR single-domain training set.

2) Multitask-assisted mask disentanglement (MMD): It includes an encoder and a decoder. The encoder consisting of ConvBlock1 extracts the target mask $\boldsymbol{z}^{\mathrm{m}}$ and concatenated feature maps $\boldsymbol{z}^{\mathrm{c}}$, and the target mask and feature maps multiply to obtain the target features for alignment and recognition. The auxiliary multitask of the mask uses the segmentation task and sparse loss to suppress clutter effectively. The segmentation task is introduced by a decoder consisting of ConvBlock2 and Conv3 with automatic pseudolabels and binary cross-entropy loss, and the sparsity constraint is added to the target mask with l_1 loss. The ConvBlock1, ConvBlock2, and Conv3 are depicted in Fig. 3(b). With extra prior constraints, this module can disentangle the target from clutter for subsequent domain alignment and recognition.

3) Domain Alignment of Target Features: After disentangling the target from clutter, the model needs to be robust to the local perturbations in target features. We convert target feature maps to capsule vectors [63]. The capsule vector $\boldsymbol{u}$ is applied as the final feature expression for domain alignment and classification because spatial relationships between features can be expressed through cosine similarity between capsule vectors. The domain alignment uses cosine similarity as the contrastive loss, and applying SimSiam [64] mitigates the conflict between the contrastive loss and classification. Consequently, with a carefully designed domain alignment module further enhancing the invariance of capsule vectors, our framework achieves robust recognition under various operating conditions.

For clarity, we briefly describe the training process of HDANet in Algorithm 1. The total loss $\mathcal{L}$ is in (3), including the classification loss l^{cls} of the capsule vectors, the contrastive loss l^{con} with the cosine similarity, the segmentation loss l^{seg} with binary crossentropy, and the sparse constraint l^{spa} with l_1 the regularization of the target mask. In the test phase, the DDG is removed, and the classification result is obtained by inputting an SAR image

$$\mathcal{L} = l^{\mathrm{cls}} + l^{\mathrm{con}} + \alpha \cdot l^{\mathrm{seg}} + \beta \cdot l^{\mathrm{spa}}. \tag{3}$$

C. Domain Data Generation

Since clutter in different scenes obeys various distributions without stable correlation with target categories, we use data augmentation to simulate target signatures rather than background clutter variations. For simplicity, we apply data augmentation to the whole image area and use the mask to separate the target region so that the clutter region does not affect the discrimination and robustness of target features. The DDG module has three methods to simulate the variations in SAR image pixels: position, relative value, and absolute value.

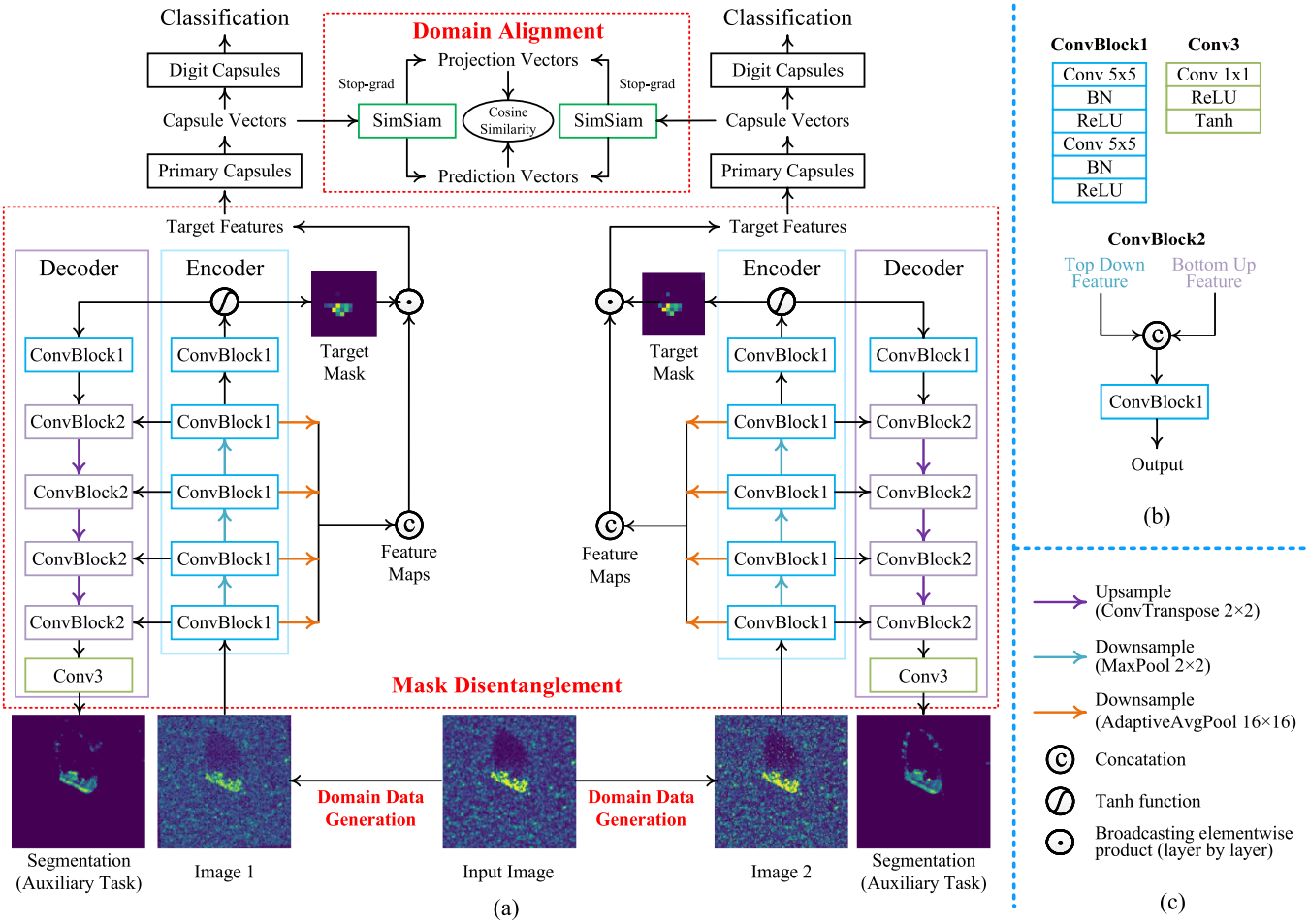

Fig. 3. Overall framework of HDANet. (a) Provides its framework with three modules: domain data generation (DDG), multitask-assisted mask disentanglement, and domain alignment of target features. Three data augmentation methods are used in DDG. Mask disentanglement includes an encoder and decoder. The feature maps and the target mask are multiplied to get target features, and the disentangled target features are used for classification and domain alignment. We use the segmentation task and the l_1 loss to assist with the target mask. In the domain alignment module, target feature maps are converted to capsule vectors, and then the contrastive loss with cosine similarity, which SimSiam calculates, is used to enhance feature robustness. The parameters and structure are the same on both sides. Right-hand side subfigure illustrates the details of (b) Convblocks (BN is batch normalization). (c) Symbols list.

Algorithm 1: HDANet: Hierarchical Disentanglement-Alignment Network.

Input: a batch of N data: images $\{\boldsymbol{x}_i\}_{i=1}^{N}$, classification labels $\{y_i^{\text{cls}}\}_{i=1}^{N}$, and segmentation labels $\{\boldsymbol{y}_i^{\text{seg}}\}_{i=1}^{N}$; models: parameters θ, encoderf_θ^{enc}, decoder f_θ^{dec}, primary capsules f_θ^{pri}, digit capsules f_θ^{dig}, and SimSiam f_θ^{sim}; data augmentation set $\mathcal{T}$; optimizer;

Output: parameters θ

for $\boldsymbol{x}_i \in \{\boldsymbol{x}_i\}$, $y_i^{\text{cls}} \in \{y_i^{\text{cls}}\}$, $\boldsymbol{y}_i^{\text{seg}} \in \{\boldsymbol{y}_i^{\text{seg}}\}$ **do**

- $t_1 \in \mathcal{T}$ and $t_2 \in \mathcal{T}$; // sample rand data augmentation
- $\boldsymbol{z}_1^{\text{c}}, \boldsymbol{z}_1^{\text{m}} \leftarrow f_\theta^{\text{enc}}(t_1(\boldsymbol{x}_i))$ and $\boldsymbol{z}_2^{\text{c}}, \boldsymbol{z}_2^{\text{m}} \leftarrow f_\theta^{\text{enc}}(t_2(\boldsymbol{x}_i))$; // compute feature maps and target mask
- $\boldsymbol{o}_1 \leftarrow f_\theta^{\text{dec}}(\boldsymbol{z}_1^{\text{c}}, \boldsymbol{z}_1^{\text{m}})$ and $\boldsymbol{o}_2 \leftarrow f_\theta^{\text{dec}}(\boldsymbol{z}_2^{\text{c}}, \boldsymbol{z}_2^{\text{m}})$; // compute segmentation results
- $\boldsymbol{u}_1 \leftarrow f_\theta^{\text{pri}}(\boldsymbol{z}_1^{\text{c}} \otimes \boldsymbol{z}_1^{\text{m}})$ and $\boldsymbol{u}_2 \leftarrow f_\theta^{\text{pri}}(\boldsymbol{z}_2^{\text{c}} \otimes \boldsymbol{z}_2^{\text{m}})$; // disentangle and convert to vectors
- $\boldsymbol{m}_1, \boldsymbol{n}_1 \leftarrow f_\theta^{\text{sim}}(\boldsymbol{u}_1)$ and $\boldsymbol{m}_2, \boldsymbol{n}_2 \leftarrow f_\theta^{\text{sim}}(\boldsymbol{u}_2)$; // compute projection and prediction vector
- $\boldsymbol{v}_1 \leftarrow f_\theta^{\text{dig}}(\boldsymbol{u}_1)$ and $\boldsymbol{v}_2 \leftarrow f_\theta^{\text{dig}}(\boldsymbol{u}_2)$; // compute classification vectors
- $l_i^{\text{cls}} = \sum_j^{1,2} l^{\text{mar}}(\boldsymbol{v}_j, y_i^{\text{cls}})$; // compute classification loss
- $l_i^{\text{con}} \leftarrow 1 - \frac{1}{2}\frac{\text{stopgrad}(\boldsymbol{m}_1)}{\|\text{stopgrad}(\boldsymbol{m}_1)\|_2} \cdot \frac{\boldsymbol{n}_2}{\|\boldsymbol{n}_2\|_2} - \frac{1}{2}\frac{\text{stopgrad}(\boldsymbol{m}_2)}{\|\text{stopgrad}(\boldsymbol{m}_2)\|_2} \cdot \frac{\boldsymbol{n}_1}{\|\boldsymbol{n}_1\|_2}$; // compute contrastive loss
- $l_i^{\text{seg}} \leftarrow \sum_j^{1,2} \text{binary cross-entropy}(\boldsymbol{o}_j, \boldsymbol{y}_i^{\text{seg}})$; // compute segmentation loss
- $l_i^{\text{spa}} \leftarrow \sum_j^{1,2} \|\boldsymbol{z}_j^{\text{m}}\|_1$; // compute sparse loss with mean l_1 loss
- $\mathcal{L}_i \leftarrow l_i^{\text{cls}} + l_i^{\text{con}} + \alpha \cdot l_i^{\text{seg}} + \beta \cdot l_i^{\text{spa}}$; // compute the loss for $\boldsymbol{x}_i$

end

$\delta\theta \leftarrow \frac{1}{N}\sum_{i=1}^{N} \partial_\theta \mathcal{L}_i$

$\theta \leftarrow \text{optimizer}(\theta, \delta\theta)$; // update parameters

1) Rotation: The position is simulated by rotating at different angles θ. It is worth mentioning that SAR imaging causes target shadows in the line-of-sight direction and the shadow above the target in the MSTAR dataset. The amplitude of the scattering point remains stable in real situations only under small viewing angle changes as well. Therefore, the rotation operation is performed within a small angle range to maintain this imaging relationship.

2) Noise Perturbation: The shift degree of the target's strong and weak scattering points is not the same across operating conditions, so noise perturbation adds Gaussian white noise $N(\mu, \sigma)$ to the original image to simulate relative value changes

$$x'_n = x_n + x^{\text{wgn}}, x^{\text{wgn}} \sim A \cdot N(\mu, \sigma) \tag{4}$$

where x'_n is the nth pixel point after augmentation, x_n is the nth pixel point before augmentation, j iterates over all pixel points of the image, A, μ, and σ are the magnitude, mean, and standard deviation (STD) of the Gaussian distribution, respectively.

3) Random replacement: It simulates absolute value changes by replacing a different proportion p of pixel points with a uniform distribution $U(0, 1)$

$$x'_n = x^{\text{uni}}, x^{\text{uni}} \sim U(0, 1) \tag{5}$$

where n iterates over the p-proportion pixel points of the image. Since the single point's amplitude in an SAR image is unstable across operating conditions, the random replacement prevents the model from being sensitive to the intensity of a single pixel point.

Based on the DDG consisting of the above three data augmentation methods, we generate a pair of image pairs (x'_1 and x'_2) from the input image x for domain alignment.

D. Multitask-Assisted Mask Disentanglement

This module includes an encoder and a decoder. The codec skeleton is the U-shaped structure of U-Net [65], commonly used in small sample tasks. As shown in Fig. 3, different modules are the variants of the U-net basic module, and we tuned the layers and parameters of the original U-Net to fit the picture size of the MSTAR dataset. In the encoder, a soft target mask $\boldsymbol{z}^{\text{m}}$ is extracted at the middle layer and multiplied with concatenated feature maps $\boldsymbol{z}^{\text{c}}$ that have multiscale details. The decoder is used to perform the segmentation task. Then, we introduce the multitask setting of this module. In order to successfully learn the target mask in a small SAR dataset, the segmentation task and sparsity constraint are auxiliary tasks to introduce position and sparse priors.

1) Segmentation task: This provides information on the target location compared with the reconstruction task, making it necessary for the model to distinguish between the spatial regions of the target and clutter. We use the decoder to apply the auxiliary segmentation task and increase the discrepancy between targets and clutter in different layers through the U-shaped structure. The segmentation loss l^{seg} is the binary crossentropy, and pseudolabels $\boldsymbol{y}^{\text{seg}}$ are multiple class saliency maps[1] [67] of the pretrained model VGG16 [46]. Moreover, we use SmoothGrad [68] to average saliency maps and remove strong clutter points. To achieve a balance between accuracy and efficiency, we design this method of automatically generating segmentation pseudolabels, but other ways to generate pseudolabels are also feasible to provide target region information.

2) Sparsity constraint: The sparsity constraint of the target mask is due to the sparsity property of the target pixel compared with the whole image pixel. In the MSTAR dataset, the vehicle range in size from 4.1 to 9.5 m long and from 2.3 to 3.6 m wide, and the image is 128 × 128 pixels (38.4 m × 38.4 m). Moreover, the target region is small under airborne or satellite-based SAR platforms. The l_1 constraint is added to the target mask as the sparse loss l^{spa} to help remove clutter further. Moreover, we use ReLU to generate a truncated Tanh activation function so that the zero value of the background region maps to zero instead of 0.5 in the sigmoid function.

We achieve feature disentanglement by the target mask with the above two auxiliary tasks, and this module enhances feature robustness by preventing clutter from being used for recognition and domain alignment.

E. Domain Alignment of Target Features

Multitask-assisted mask disentanglement module extracts target features and suppresses clutter. However, not all target features are robust under different SAR operating conditions. Therefore, we use the DATF to extract invariant features. As shown in Fig. 3, the target feature maps are converted to capsule vectors for classification and domain alignment. Specifically, we consider the robust feature representation (e.g., capsule vectors) with domain alignment in SAR target recognition.

1) Feature representation: The previous domain alignment methods used CNN feature maps [22] or FC layer features [23]. Other work showed that using capsule vector is more robust in EOCs [11] than CNN and FC. Considering the target's overall structural information is more stable under the local target perturbation, we use the capsule vector [11], [63] to preserve the spatial information between features. The target feature map is converted to a capsule vector $\boldsymbol{u}$, whose magnitude represents the probability and direction represents spatial information (i.e., cosine similarity between features). The digit capsules perform recognition based on the magnitude and direction of the classification vectors v in Fig. 3. The classification loss l^{cls} is the margin loss l^{mar} in [63] with consistent hyperparameters ($w^+ = 0.9, w^- = 0.1, \eta = 0.5$)

$$\begin{aligned} l^{\text{mar}} = & \sum_k T_k \cdot \max\left(0, w^+ - \|\boldsymbol{v}_{j,k}\|\right)^2 \\ & + \eta(1 - T_k) \cdot \max\left(0, \|\boldsymbol{v}_{j,k} - w^-\|\right)^2 \end{aligned} \tag{6}$$

$$l^{\text{cls}} = \sum_j^{1,2} l^{\text{mar}}\left(\boldsymbol{v}_j, y^{\text{cls}}\right) \tag{7}$$

where the first term of l^{mar} encourages correct prediction probability over w^+, the second term of l^{mar} penalizes incorrect prediction probability higher than w^-, $\boldsymbol{v}_{j,k}$ is the kth classification

[1]Multiple class saliency map is a variant of GradCam [66] to eliminate category information.

vector of $\boldsymbol{v}_j$, $T_k = 1$ if the classification label y^{cls} is class k, and η prevents the initial learning from shrinking the magnitude of all digit capsules.

2) Domain alignment: Correspondingly, the domain alignment module uses the cosine similarity as contrastive loss and SimSiam [64] structure. According to the SimSiam structure, the contrastive loss l^{con} is shown below:

$$D(\boldsymbol{m}, \boldsymbol{n}) = \frac{1}{2} - \frac{1}{2} \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|_2} \cdot \frac{\boldsymbol{n}}{\|\boldsymbol{n}\|_2} \tag{8}$$

$$l^{\text{con}} = D(\text{stopgrad}(\boldsymbol{m}_1), \boldsymbol{n}_2) + D(\text{stopgrad}(\boldsymbol{m}_2), \boldsymbol{n}_1) \tag{9}$$

where $\boldsymbol{m}$ and $\boldsymbol{n}$ are the projection and prediction vectors in Fig. 3, respectively. The stopping gradient and asymmetric structure of SimSiam increase the interclass distance and avoid trivial constant solutions impairing discrimination. Other methods used the hyperparameter tradeoff [22], [23] or negative samples [55] to solve the problem of identical representation in domain alignment.

As discussed above, we designed a robust feature representation with domain alignment. Eventually, with the aforementioned DDG and mask disentanglement modules, robust features under various operating conditions are obtained by the DATF. In the proposed integrated framework of feature disentanglement and alignment, we mainly address the two problems in SAR vehicle recognition methods: the existing mask methods overfit the background clutter, and the current domain alignment methods achieve recognition for several specific operating conditions.

IV. Experiments

In this section, we evaluated the robustness of the proposed method on the MSTAR dataset and analyzed the strengths and weaknesses of HDANet. We first described the experimental setting and implementation details in Sections IV-A and IV-B. The robustness of HDANet is evaluated in the MSTAR dataset's nine operating conditions compared with other methods in Section IV-C. We then performed extensive quantitative and qualitative analyses in Section IV-D to demonstrate the effectiveness of our method. Ultimately, we explored the limitations in Section IV-E.

A. Dataset and Experimental Settings

1) Dataset description: Sandia National Laboratory collected and released the popular MSTAR dataset [14] with a 10-GHz X-band spotlight SAR sensor. This dataset contains ten categories of ground military targets: infantry vehicle (BMP2), patrol car (BRDM2), personnel carrier (BTR60, BTR70), tank (T62, T72), howitzer (2S1), bulldozer (D7), truck (ZIL131), and anti-aircraft (ZSU234). The resolution of each SAR image is 0.3 m $\times$ 0.3 m, and MSTAR data are acquired at full azimuth angles from 0° to 360°. However, only some targets have different depression angles (e.g., 15°, 17°, and 30°) and scenes (grasslands in New Mexico, Northern Florida, and Northern Alabama [15]). We used the official tool to convert original SAR data into JPEG format with 128 $\times$ 128 pixels by linear transformation and automatic contrast enhancement.

2) Experimental settings: We discussed nine operating conditions based on the MSTAR dataset in Table I to evaluate the performance of our method comprehensively. The standard operating condition means that the imaging conditions of the training and test sets are similar. In EOCs, different operating conditions lead to complex variations in target signature and background clutter [15], [16]. We built an extensive EOC setting, including sensor (depression angle, azimuth angle, and noise level), target (configuration and version), and environment (occlusion and scene). The EOC-Gaussion/random/occlusion are simulation settings, and others are measured data. The detailed settings are discussed below.

3) Standard Operating Condition: The difference between the training and test sets is minor under SOC, and the main challenge is the small interclass differences. As given in Table I, the training set's depression angle under SOC is 17Â°, and the test set is 15Â°. Ten categories of targets, include BMP2 (C21, 9563, 9566), BRDM2, BTR60, BTR70, T62, T72 (132, 812, S7), 2S1, D7, ZIL131, and ZSU234.

4) EOC-Depression (Depression Angle Variation): The intensity of pixel points in SAR images is related to the depression angle, and a large depression angle can change target signatures and enhance background clutter. Therefore, robustness to depression angle variation is critical to the sensor setup. Following the setting in [10], four targets are selected to discuss the performance from 17Â° to 30Â° in Table I. These targets include BRDM2, T72 (A64), 2S1, and ZSU234.

5) EOC-Azimuth (Azimuth Angle Variation): Due to the anisotropic scattering of the different structures in targets, local signatures vary at different azimuth angles. Since the SAR images of the MSTAR dataset are collected at different azimuth angles, the robustness to azimuth angle variation is tested by reducing training data from 90% to 10% under SOC in Table I.

6) EOC-Gaussion (Gaussian Noise Corruption): Sensor noise can significantly affect the properties of the measured SAR images, but additive noise in MSTAR is below -30 dB, which is far from realistic situations. As given in Table I, the signal-to-noise ratio (SNR) of additive Gaussian white noise [12], [26] is from 10 to -10 dB based on the SOC setting.

7) EOC-Random (Random Noise Corruption): Another noise setting is randomly replacing the original pixel values with a uniform distribution noise [10], [27]. This approach simulates the degree of random strong clutter point interference. Based on the SOC setting, random noise given in Table I has a replacement rate of 5%–25%.

8) EOC-Configuration (Configuration Variant): The vehicle configurations often change depending on different actual needs, such as fuel containers and other accessories fixed to the vehicle. Following the setting in [10], the configuration variant has BMP2 (9563), BRDM2, BTR70, and T72 (132) as training set, and T72 (A32, A62, A63, A64, S7) as five test variants.

9) EOC-Version (Version Variant): Similar to the configuration variant, various vehicle versions are produced for different needs with a similar global structure and different local details. Following the setting in [10], the configuration variant has BMP2 (9563), BRDM2, BTR70, and T72 (132) as the training set. The

TABLE I
EXPERIMENTAL SETTING UNDER SOC AND EOCS OF MSTAR

Operating conditions	Setting of training set (test set)			
	# Categories	# Samples in total	Depression angle	Scene
SOC	10 (10)	3671 (3203)	17° (15°)	Grass 1 and HB (Grass 1 and HB)
EOC-Depression	4 (4)	1195 (1151)	17° (30°)	
EOC-Azimuth	10 (10)	downsampling of 3671 (3203)	17° (15°)	
EOC-Gaussian	10 (10)	3671 (3203 with Gaussian noise)	17° (15°)	
EOC-Random	10 (10)	3671 (3203 with random noise)	17° (15°)	
EOC-Configuration	4 (5 variants)	996 (2710)	17° (17°and15°)	
EOC-Version	4 (7 variants)	996 (3569)	17° (17°and 15°)	
EOC-Occlusion	10 (10)	3671 (3203 with occlusion)	17° (15°)	
EOC-Scene	3 (3)	1772 (743)	45°and 30° (45°and 30°)	Grass 1 (Grass 2 and 3)

Note: Different parameters of EOC-Azimuth/Gaussian/Random/Occlusion control the level of variations compared with SOC (see Sec. IV-A for detailed presentation). Detailed information is available in the Supplementary Materials and on our GitHub.
Robustness to EOCs includes sensor (depression angle, azimuth angle, and noise level), target (configuration and version), and environment (occlusion and scene).

TABLE II
DATA AUGMENTATION DETAILS

Approach	Parameter	Probability
Rotation	$\theta \sim U(-5°, 5°)$	0.3
Noise Perturbation	$A \cdot N(0.1, 0.1), A \sim U(0.5, 1.5)$	0.2
Random Replacement	$p \sim U(0, 5\%)$	0.2

BMP2 (9566, C21) and T72 (A04, A05, A07, A10, 812) are seven test variants.

10) EOC-Occlusion (Occlusion Interference): Objects between the target and sensor, such as trees and buildings, can obscure or attenuate the signal strength reflected from the target. Therefore, occlusion can severely eliminate or diminish some of the target signatures. According to Feng et al.[13], the square of different pixel sizes (5×5, 10×10, and 15×15) in the target and the adjacent area (64×64 at picture center) are randomly set to zero in Table I.

11) EOC Scene (Scene Variation): Background clutter in different scenes can vary significantly and affect the scattering points of adjacent target parts. Although the scenes in MSTAR are all flat grassland from different locations, the grass's height, sparsity, and water content affect its EM scattering [69], [70]. BRDM2, T72 (A64), and ZSU234 in Grass 1 are the training set. ZSU234 in Grass 2 and BRDM2 and T72 (A64) in Grass 3 are the test data in Table I.

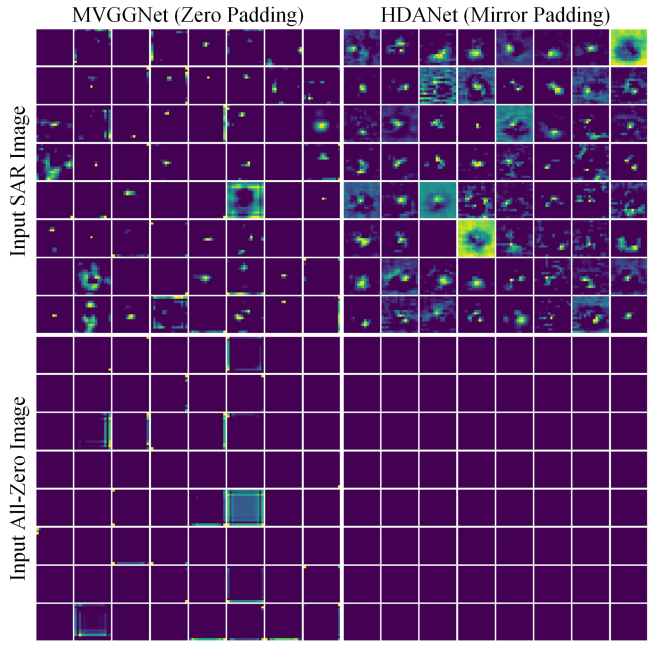

Fig. 4. Feature maps of different padding methods (feature maps on the left are MVGGNet [12] with zero padding, and on the right are feature maps of HDANet (ours) with mirror padding; input from the top is an MSTAR image, and from the bottom is an all-zero image). Models with zero padding produce target-independent edge and center artifacts in the left feature maps [71].

B. Implementation Details

1) Hyperparameter settings: Data augmentation details in this article are given in Table II. The data augmentation methods used different probabilities so that the results have a 50% of being the original image. Noise perturbation was adjusted to $A \cdot N(0, 0.8)$, $A \sim U(0, 1)$ under EOC-Gaussion. The loss hyperparameter α was 1e-1 as an auxiliary task, and β was 1e-2 to control mask sparsity. NAdam algorithm [72] was employed as the optimizer with an initial learning rate of 3e-4, weight decay of 5e-4, and an exponential learning rate decline of 0.98. The batch size was 64, and the epoch was 100. Notably, we used mirror padding instead of zero padding to eliminate feature-map artifacts in Fig. 4.

2) Compared Methods: We used A-ConvNet [10], ECCNet [11], MVGGNet [12], SNINet [22], and TDDA [23] as compared methods. A-ConvNet is one of the classical deep learning methods successfully applied to recognizing SAR vehicles. An ECCNet is a method that uses the capsule network and the attention module. We combined CFAR with them to verify the effectiveness of traditional mask methods. MVGGNet with pretraining weight has a good feature extraction ability by transfer learning. SNINet and TDDA are similar domain alignment methods in SAR vehicle recognition. All methods

TABLE III
PERFORMANCES OF DIFFERENT METHODS UNDER DIFFERENT OPERATING CONDITIONS

Method	SOC	EOC							
		Depression	Azimuth	Gaussian	Random	Configuration	Version	Occlusion	Scene
A-ConvNet	98.12±0.23	89.50±1.46	63.95+4.20	71.36±1.11	91.82±1.03	96.69±0.97	96.29±0.72	74.79±1.18	69.13±2.28
A-ConvNet-CFAR	94.91±0.39	93.55±0.95	62.60+2.63	89.61±0.44	93.76±0.41	91.41±1.25	93.44±0.61	80.54±0.62	85.25±1.12
ECCNet	98.73±0.14	94.07±0.68	50.19+1.83	73.32±0.74	90.96±1.57	97.29±0.42	93.60±0.50	84.02±0.04	76.92±1.22
ECCNet-CFAR	96.66±0.21	94.34±0.77	60.59+0.73	90.90±0.52	95.69±0.24	90.79±2.36	88.57±1.75	85.27±0.59	83.41±1.95
MVGGNet	98.34±1.20	91.04±3.30	57.74+2.26	71.63±3.40	76.42±2.58	96.81±0.72	95.05±1.09	79.90±2.64	73.50±2.57
SNINet	95.69±1.42	88.41±2.74	69.06+1.82	67.47±2.27	88.94±1.73	95.97±1.81	95.53±1.84	69.06±2.71	67.01±2.52
TDDA	98.64±0.27	93.05±1.35	54.34+4.99	79.85±1.52	96.82±0.37	97.66±0.22	97.96±0.30	81.64±1.06	73.05±1.98
HDANet (ours)	**99.64±0.13**	**96.26±0.44**	**76.75+3.46**	**92.72±0.46**	**97.38±0.86**	**98.49±0.51**	**98.36±0.22**	**86.51±1.61**	**94.78±0.65**

Note: The **bold** number denotes the best result. All results are the mean OA (%) ± STD over five runs. EOC-Azimuth/Gaussian/ Random/Occlusion results are under the most challenging parameter in our settings (see Fig. 6 for detailed accuracy curves under different parameters).

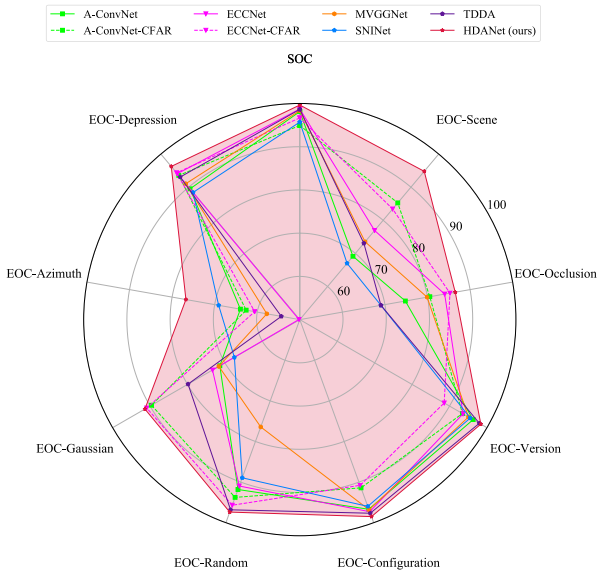

Fig. 5. Radar charts of experimental results (see Table III for detailed numbers). Our method performs more robustly than others under various operating conditions.

used the same data augmentation methods. Five repetitions of experiments were used to calculate the overall accuracy (OA) and the STD.

C. Results Under SOC and EOCs

In this section, we evaluated the robustness of HDANet in the nine operating conditions of MSTAR compared with other methods. The recognition results of these methods are given in Table III and Fig. 5, and the detailed results are given as follows.

1) Result Under SOC: Because of a little difference between training and test set distribution under SOC, deep learning-based approaches effectively learn correlations in the dataset, nearing saturation performance. Removing clutter and shadows from input images by CFAR causes a slight decrease in the accuracy rate (3.21% for A-ConvNet and 2.07% for ECCNet). The decline is due to the background correlation [3], [24] in the MSTAR dataset and the shadows containing the target's structural properties [24], [54], [73]. Previous work [37] pointed out that using a hard constraint mask for segmented images causes deep learning to pay attention to boundaries rather than target signatures, so a soft learnable constraint mask was developed for our deep learning-based framework. Good performance can also be achieved under SOC with transfer learning (98.34% for MVGGNet), demonstrating the generality of deep learning's underlying feature extraction. CNN feature maps used by SNINet (95.69%) perform lower than other methods (98.64% for TDDA and 99.64% for HDANet), which indicates the l_2 loss ignores the 2-D structural information in the feature maps. Therefore, we use capsule vectors to represent feature structure information with 99.64% performance. The corresponding cosine similarity and SimSiam structure mitigate domain alignment to impair feature discrimination.

2) Result Under EOC-Depression: The increased depression angle changes the target signature and enhances the clutter, so accuracy rates improve (4.05% for A-ConvNet and 0.27% for ECCNet) with CFAR. Transfer learning cannot handle distribution shifts in a small dataset, achieving 91.04% for MVGGNet. Domain alignment relies on a powerful encoder, with SNINet (88.41%) and TDDA (93.05%) performing below ECCNet (94.07%). Therefore, we apply domain alignment to an encoder consisting of CNN and capsule layers to achieve a performance of 96.26%.

3) Result Under EOC-Azimuth: Reducing the training set data loses some target signatures under partial azimuth. As shown in Fig. 6, the design of a powerful encoder and domain alignment alone cannot solve the robustness problem of azimuth variations. A powerful encoder (A-ConvNet, ECCNet, and MVGGNet) can overfit small data, reducing generalization performance. Domain alignment (SNINet and TDDA) may learn the same feature representation more easily with fewer data, reducing feature representation discrepancy. Data augmentation simulates local perturbations and are not effective with large variations in very few samples. However, domain alignment (TDDA, HDANet) can improve the effectiveness of data augmentation for EOC-Azimuth. HDANet strikes a balance between robustness and feature discrepancy. HDANet extracts robustness target features by domain alignment. Moreover, the SimSiam structure reduces domain alignment impairment on feature discrepancy.

4) Result Under EOC-Gaussion: Additive Gaussian noise weakens target signatures and changes the relative values of the

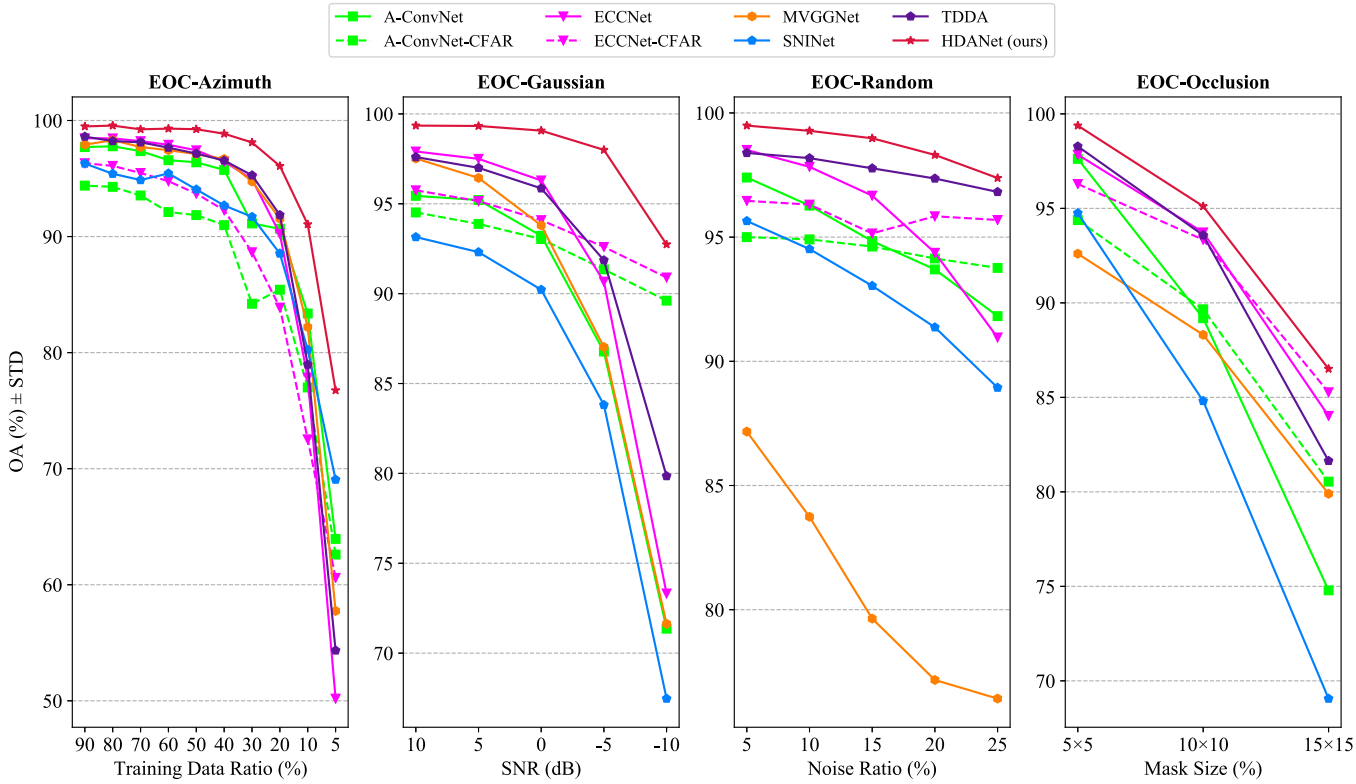

Fig. 6. Experimental results of different parameters under EOC-azimuth, EOC-Gaussion, EOC-random, and EOC-occlusion. The accuracy decreases as the perturbation intensity increases, and our method maintains better robustness under different parameters than others.

strong and weak scattering points. Fig. 6 shows that the curve decreases slowly after removing clutter with CFAR because it extracts the target region blurred by noise. On the other hand, domain alignment (TDDA) is also better than other methods at low SNR. Our method combines mask and domain alignment with the best 92.72% accuracy, using mask disentanglement to extract target features and domain alignment to enhance robustness. However, target feature extraction is difficult at low SNR (–10 dB), and data augmentation parameters need to be adjusted for the learnable target mask. It can be found that our method becomes significantly degraded below −5 dB. The CFAR method also has this problem in region extraction, but we used the masks of CFAR from the original images for simplicity under any SNR, and therefore, the trend of the CFAR method is more stable below −5 dB.

5) Result Under EOC-Random: Random noise simulates a different proportion of strong clutter point interference. As shown in Fig. 6, CFAR can effectively reduce clutter interference. ECCNet performs well at low proportions, but when interference increases, the attention module in ECCNet cannot effectively remove a large proportion of strong clutter. The use of FC layers with domain alignment (TDDA) can also reduce this interference. However, using domain alignment with CNN features (SNINet) is more likely to receive strong clutter point interference, which may be because multilayer CNNs have an aggregated information from larger receptive fields with strong clutter points from different locations. MVGGNet also has this problem. Our method suppresses clutter and extracts robust target features with the best performance.

6) Result Under EOC-Configuration: Although the T-72 configuration in the test set differs from the training set, the impact of this variant on deep learning is small at a resolution of 0.3 m, and almost all methods achieve over 95% accuracy. However, the performances of A-ConvNet and ECCNetet decrease after extracting the target region by CFAR due to removing the T-72 gun barrel signature in the shadow. Therefore, we use a soft-constrained mask and enhance robustness with domain alignment to achieve an accuracy of 98.49%. The learned mask can extract shadow edge information (see Fig. 10).

7) Result Under EOC-Version: Version variants impact recognition similarly to configuration variants. Therefore, the analysis for such variants is similar to the configuration variants. However, there is a 3.69% decrease in the accuracy of ECCNet, which may be due to the attention module in ECCNet misfitting the unstable background correlation. The scene of four T72 variants in the test set has a small different from the training set.

8) Result Under EOC-Occlusion: The occlusion setting affects the pixel values of the target signature and adjacent regions. As shown in Fig. 6, the occlusion of sizes from five to 15 impacts target signatures significantly due to the small vehicle target size. Convolutional features (A-ConvNet and SNINet) are susceptible to occlusion, while the FC layer (MVGGNet) is less susceptible. The least is using a capsule network (ECCNet)

for classification. Therefore, we use the capsule network as a classifier and design a corresponding domain alignment method, achieving 86.51% accuracy. This article [13] used EM scattering features to enhance the deep learning features with an 82.39% accuracy under a single look in the same setting. Our proposed method improves the robustness of deep learning to occlusion as well. However, it is challenging to achieve robust recognition algorithms with heavily obscured target signatures. We argue that using multiple-look images is a more effective way to deal with this interference since the occlusion is closely related to the relative positions of the target and the sensor.

9) Result Under EOC Scene: The clutter in different backgrounds significantly differs and affects the adjacent target signature. Data-driven models, such as deep learning, have the potential to misuse the features of strong clutter. Compared with the simulated noise setting, EOC scene used the measured data from different scenes to further research this problem. As given in Table III, the robustness to target signature changes of A-ConvNet and ECCNet after removing clutter needs improvement. Pretraining model MVGGNet also cannot handle the data bias and distribution shifts in this downstream task. Other domain alignment methods cannot address this problem since they ignore background clutter interfering with the final features. Our method performs the DATF with mask disentanglement. Therefore, our approach achieves robustness under this new experimental setting with a 94.78% accuracy.

As discussed above, our method has significant robustness than other methods across various experimental settings given in Table III and Figs. 5 and 6. Compared with deep learning models (A-ConvNet and ECCNet) with CFAR or transfer learning (MVGGNet), our approach improves robustness through the DATF. We achieve feature disentanglement and design a novel domain alignment framework for SAR recognition compared with other domain alignment methods (SNINet and TDDA).

D. Analysis

In this section, we conducted numerous qualitative and quantitative analyses to validate the effectiveness of our approach in robust recognition and clutter suppression. These analyses encompassed the ablation study, qualitative research, and the analysis of the clutter effect.

1) Ablation Study: Since the target signature and the background clutter change significantly under EOC scene, we performed the ablation study and hyperparameter discussion under EOC scene given in Table IV and Fig. 7.

Baseline: It used a modified CNN encoder of U-Net and a capsule classifier, and its performance is 86.61%. Compared with ECCNet, we removed the attention module and image reconstruction task to avoid enhancing the overfitting for background clutter. The baseline improvement illustrates the need for successful modules in natural images to be refined for challenges in SAR images. However, the large STD (7.31%) indicates that deep learning still overfits the background clutter.

Effects of DDG: With data augmentation on the baseline in Table IV, the accuracy is improved to an 87.13% with a smaller STD (4.10%). Then, we discussed the detailed effect of different data augmentation methods with other modules under EOC scene in Fig 7. Specifically, different data augmentation methods improve domain alignment, but the high probability rate can cause alignment centers to deviate from the original image, which reduces recognition accuracy. We advise setting each probability to [0.1, 0.4] and the total probability close to 0.5 for domain alignment. To the best of authors knowledge, atleast ten data augmentation methods have been proposed for SAR vehicle target recognition, and our method aims to simulate the local variations of target signatures for domain alignment. Although we validated the effectiveness of the proposed method under various experimental settings, it is still worth exploring how to use data augmentation and generation more effectively for complex operating conditions and small datasets.

TABLE IV
ABLATION STUDY UNDER EOC SCENE

Baseline	DDG	MMD	DATF	OA (%) ± STD ↑
✓				86.61±7.31
✓	✓			87.13±4.10
✓	✓	✓		92.29±2.46
✓	✓	✓	✓	**94.78±0.65**

Note: The bold number denotes the best result. All results are the mean OA (%) ± STD over five runs. The baseline is a CNN encoder and a capsule classifier. DDG is domain data generation. MMD is multitask-assisted mask disentanglement. DATF is domain alignment of target features.

Effects of MMD: We used the multitask setting to extract the feature layer mask, which further suppresses clutter and achieves an accuracy of 92.29% ± 2.46 in Table IV. The learnable masks can be better integrated with deep learning and exploit the good properties of the middle layer [61], which avoids strong clutter points in the input image. Moreover, the multitask setting introduces the target location and sparse priors to solve the background correlation in a small dataset. As shown in Fig 7, we discussed the hyperparameter setting of mask disentanglement under EOC scene. Specifically, α and β control the impact of the segmentation task and sparse loss, respectively. We advise setting α and β to [1e-1, 1e-3]. Hyperparameters below these ranges reduce the auxiliary task effect, and above these ranges impair the primary recognition task and discrimination of the target features. The background interference in small SAR vehicle datasets is more severe compared to large datasets in computer vision because SAR images are collected under specific operating conditions. Therefore, this module achieves feature disentanglement to ensure correct feature representation.

Effects of DATF: Considering the connection between causality and invariance [56], HDANet uses DATF further enhance the robustness of features. Its performance is impressive, reaching 94.78% with the smallest STD (0.65%). In the domain alignment module, we used capsule vectors to preserve feature space information and cosine similarity as the contrastive loss. The SimSiam structure increases interclass distance and mitigates the conflict between contrastive loss and classification loss. By realizing feature disentanglement and alignment through the above three modules, we established a new domain alignment framework for robust SAR target recognition, which ensures the causality and robustness of feature representations.

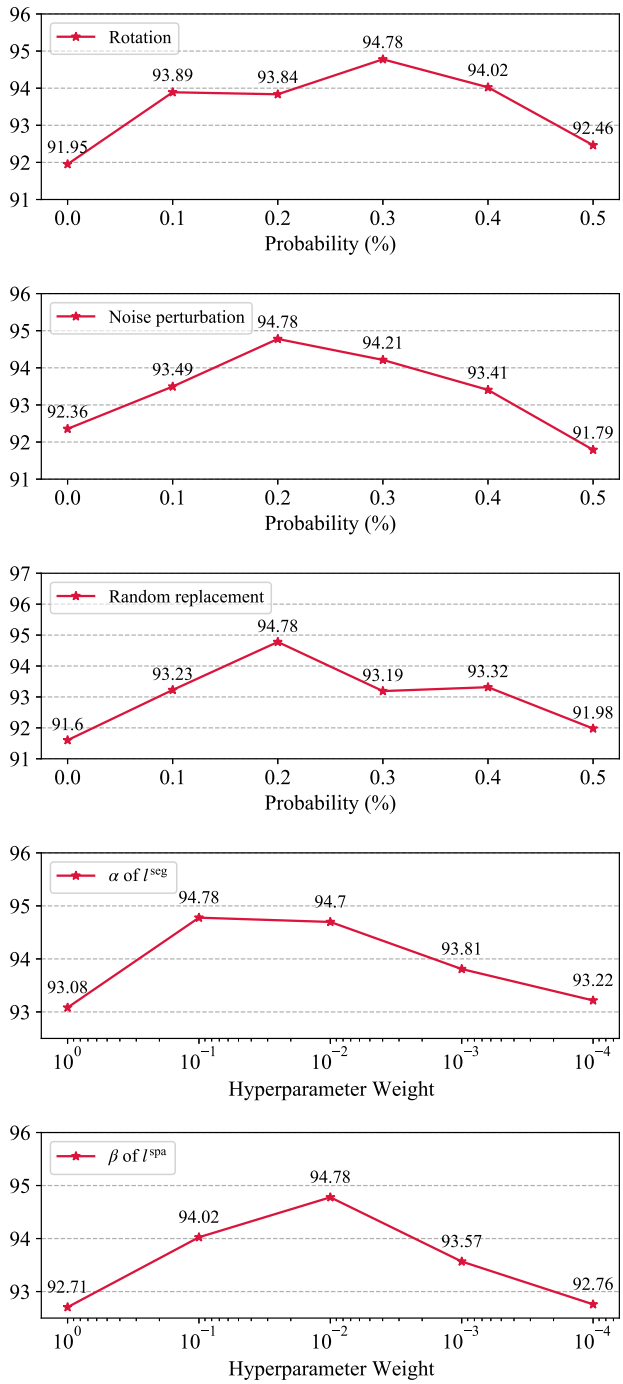

Fig. 7. Impacts of the hyperparameters under EOC scene. From top to bottom are the probabilities of the three data augmentation methods (rotation, noise perturbation, and random replacement) in domain generation and the hyperparameters of the two auxiliary tasks (segmentation task and sparse loss) in mask disentanglement. These methods are not highly sensitive to hyperparameters and are effective in relatively large intervals.

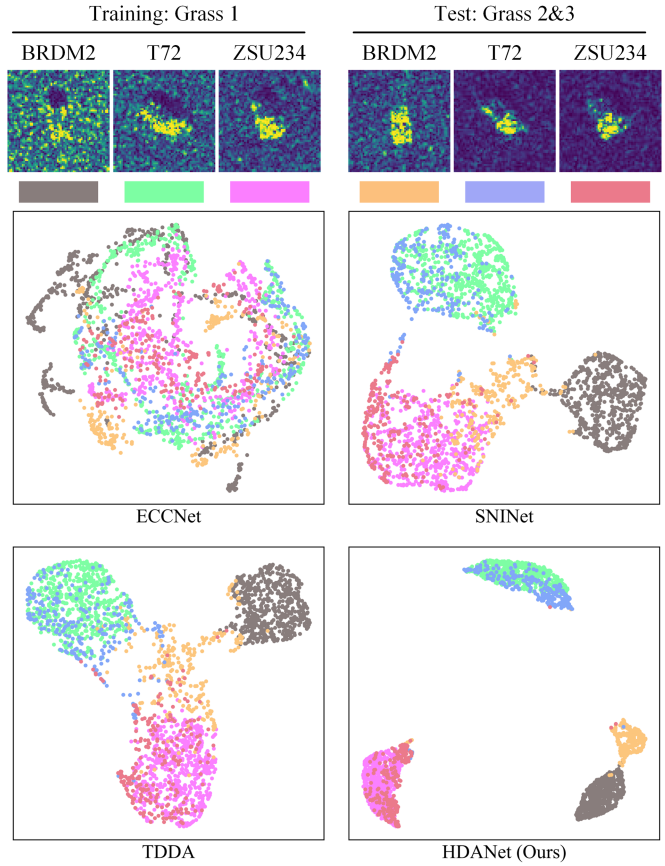

Fig. 8. Visualization of feature separability across models under EOC scene. We used uniform manifold approximation and projection [74] to visualize the penultimate layer features of each model. We can see that the domain alignment (SNINet and TDDA) method has a good clustering effect than ECCNet in the training set, and our method further improves the robustness of domain alignment methods to distribution shifts in SAR using feature disentanglement.

2) Qualitative Research: We illustrated the effectiveness of our proposed method with visualization results and methods, including feature separability, intermediate results, and segmentation results.

Feature separability: Using the EOC scene as an example, we visualized the features of the penultimate layer of different models by uniform manifold approximation and projection [74]. From Fig. 8, we can see that the domain alignment (SNINet and TDDA) methods produce good clustering results than ECCNet, resulting in larger interclass distances for improved robustness. However, the features extracted by the domain alignment methods contain unstable clutter due to background interference in the SAR images, which reduces the robustness of these methods to complex distribution shifts in SAR. As a result, these domain alignment methods yielded large interclass distances, but incorrect feature representations resulted in lower accuracy than ECCNet. Our method extracts target features by mask disentanglement, suppressing unstable clutter from interfering with the final features and improving the robustness using DATF. Despite the success of our framework, the feature shift in BRDM2 visualization results shows that there are still opportunities for further enhancements.

Intermediate masks of different methods in Fig. 9 show that our methods solved the problem of background overfitting caused by data bias. Threshold methods, such as CFAR, are susceptible to strong clutter interference in the input image, failing to detect the target in the strong clutter region. Furthermore, existing threshold methods [36], [73] combine morphological

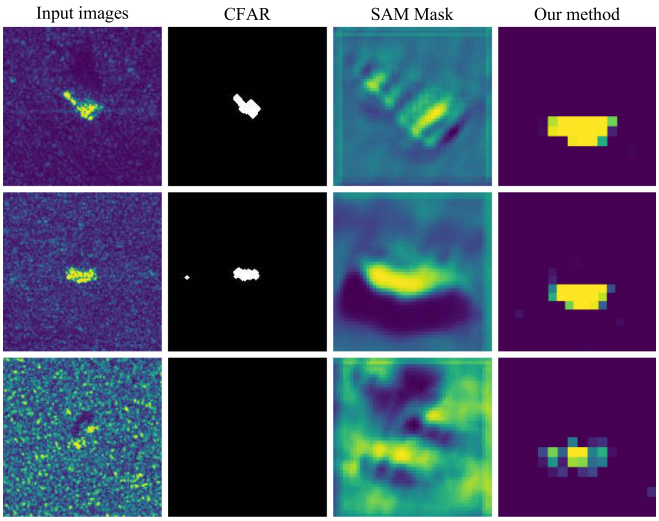

Fig. 9. Intermediate masks of different methods. We can observe that the feature layer mask avoids strong clutter interference in the input image, while our method solves the problem of SAM overfitting background clutter.

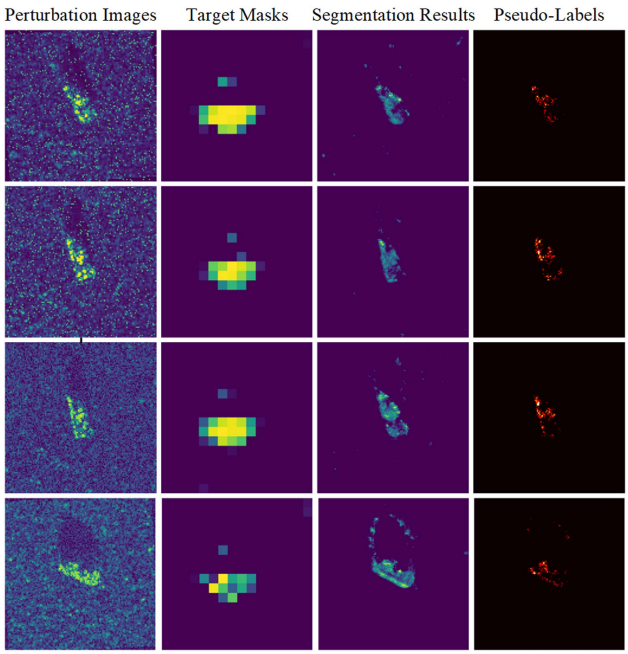

Fig. 10. Intermediate results visualization. We can observe that the target masks of the feature layer in the second column ignore the background clutter that occupies most of the images and the segmentation results in the third column also separate the target from the clutter. Our results and pseudolabels retain some shadow edges that reflect the target signatures.

operations to eliminate strong clutter points but blur fine target contours. The attention mechanism extracts the target mask at the feature layer to avoid the interference of strong clutter points. However, the spatial attention mechanism masks do not effectively suppress the background clutter due to data bias and activation function. In contrast, our method suppresses the background correlation by adding priori constraints and improves the computation of the target mask.

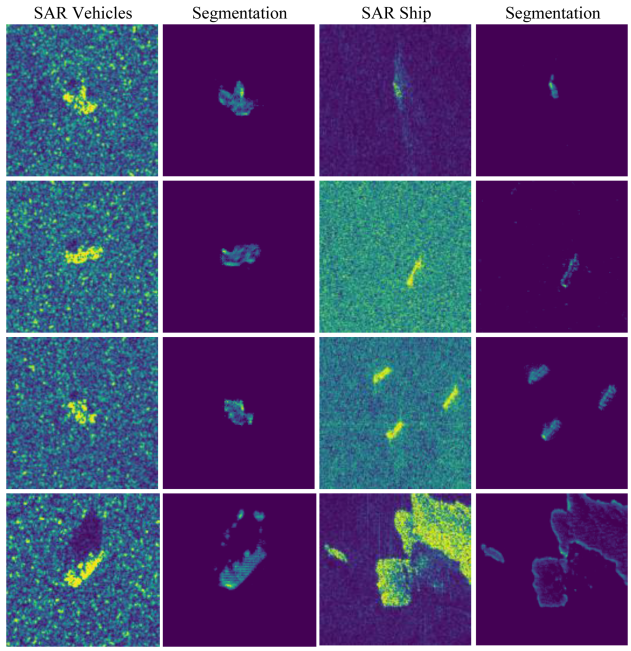

Fig. 11. Segmentation results (from left to right are SAR vehicle images in MSTAR [14], ground segmentation results, SAR ship images under complex backgrounds [75], and sea segmentation results). HDANet distinguishes targets from different ground/sea clutter.

Intermediate results in Fig. 10 show that our target mask in the feature layer separates the target region from the background, and the edge information of the target shadow is contained above the target region in the mask. Our method suppresses clutter better than CBAM's masks in Fig. 2. The segmentation results also illustrate the effectiveness of our method in clutter suppression. Our way of generating pseudolabels balances accuracy and efficiency. The weight of random clutter is reduced in the saliency maps by two averaging operations of multiple class saliency maps and SmoothGrad, and our pseudolabels extract the valuable targets and shadow regions for recognition. Since the target mask is in the middle feature layer, the automatically generated coarse labels are sufficient for the segmentation task to play an auxiliary role with l_1 loss, allowing the mask to distinguish between background and target discrepancies and obtain the correct feature representation.

Segmentation results on different datasets: The segmentation task was used as an auxiliary task to improve the robustness of the recognition task. Therefore, we used pseudolabels and did not pursue precise segmentation results, which balance accuracy and efficiency. Although not as accurate as manual annotation, the rough labels allow the model to learn the difference between the target and the background. In Fig. 11, we trained an HDANet on the MSTAR dataset and visualized the segmentation results of vehicles and ships [75]. Our method identifies target regions in ground/sea clutter, demonstrating its segmentation ability to generalize to different environments. Although the MSTAR dataset targets are all centrally placed, HDANet distinguishes between targets and clutter in multitarget SAR ship slices with

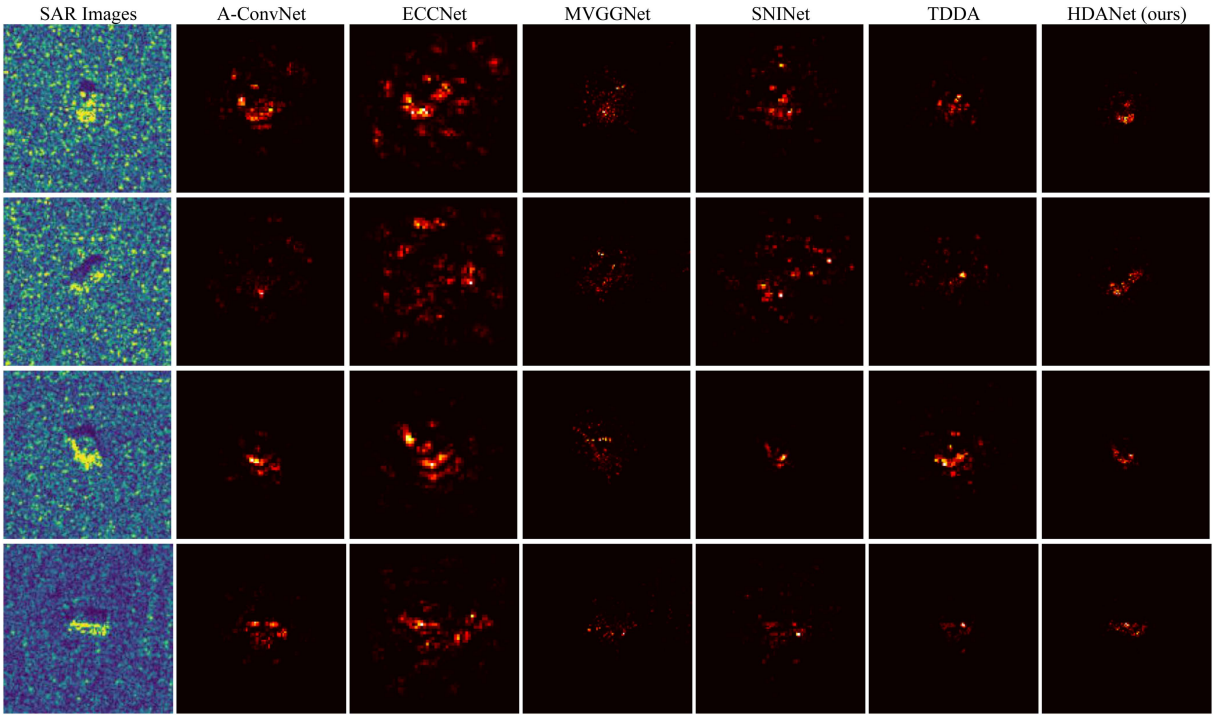

Fig. 12. Saliency maps of each model (from left to right are the SAR images and saliency maps of A-ConvNet, ECCNet, MVGGNet, SNINet, TDDA, and HDANet). The clutter intensity of the input image decreases in order from the top to the bottom. Saliency maps reflect the dependency of a model for each pixel point during recognition. From a visual perspective, our method focuses more on the target region, and others may use clutter.

256 × 256 pixels. However, the pseudolabel with the saliency map makes HDANet focus on the structural edges of targets, similar to edge detection [76], causing large scene edges preserved in Fig. 11.

3) Analysis of the Clutter Effect on Deep Learning: Clutter suppression is important for robust SAR vehicle recognition due to blurry target signatures. Therefore, it is necessary to analyze whether deep learning uses the correct features for recognition, i.e., using only the target signatures rather than clutter correlations. However, it is difficult to capture the negative impact of overfitting clutter on accuracy in experiments with similar backgrounds. Consequently, we used saliency maps and the Shapley value below to qualitatively and quantitatively analyze the degree of deep learning overfitting background clutter, i.e., the feature causality.

Saliency maps: We first qualitatively analyzed causality through the saliency map [77]. The results of different upsampling saliency maps are shown in Fig. 12. We can find that the models (A-ConvNet, ECCNet, and SNINet) using images with different input sizes overfit the background clutter around the target, so the center cropping does not solve the problem of clutter interference. ECCNet, with image reconstruction and attention module, focuses on clutter. Other domain alignment methods also contain some clutter regions. These results indicate that the methods successfully applied in natural images need to be improved in SAR images. Due to our novelty improvement for mask and domain alignment methods in SAR, our method focuses more on the target region than background clutter. Since the saliency map calculates the importance of a single image at the pixel level, it only responds qualitatively to causality and lacks the quantitative statistical metrics of the whole region. We next used the Shapley value as the quantitative metric of the clutter effect.

Shapley value is the unique solution satisfying the four properties and calculates the contribution to the cooperative game fairly [78]. According to our previous research [24], we segment the target and clutter regions in the input image based on CFAR as two players $\{0, 1\}$ and estimate their contribution to recognition based on the Shapley value

$$\mathrm{Sh}_i = \sum_{S \subseteq I \setminus \{i\}} \frac{|S|! \cdot (|I| - |S| - 1)!}{|I|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{10}$$

where Sh_i is the Shapley value of ith player, I is a set of all players, $|\cdot|$ is the number of elements in the set, and $f(\cdot)$ is classification score corresponding to the true class before softmax. The baseline values of the inputs are set to 0. We average the Shapley values for all the images and calculate the proportions of clutter Shapley values for different models to roughly estimate the degree of the overfitting for clutter.

As given in Table V, due to the texture bias of CNN [79], the texture of the clutter region is exploited by models with varying degrees. Nevertheless, our method has the least clutter effect (15.41%) due to feature disentanglement and alignment. Comparing the different image sizes, center cropping can attenuate the effect of clutter but cannot solve this overfitting. Although ECCNet uses the smallest image size (64 × 64), image

TABLE V
CLUTTER CONTRIBUTION RATIO OF DIFFERENT MODELS UNDER EOC SCENE

Model	Avg. (%) ± STD ↓	Input size	Params
A-ConvNet	27.69±2.80	88 × 88	0.30 M
ECCNet	46.71±2.77	64 × 64	7.99 M
MVGGNet	55.71±3.68	128 × 128	16.81 M
SNINet	25.99±5.69	88 × 88	0.35 M
TDDA	34.45±1.44	128 × 128	0.80 M
Our method	**15.41±2.41**	128 × 128	12.28 M

Note: The bold number denotes the best result. All results are the mean and STD of clutter contribution ratio over five runs. Although our method uses the original image size, the effect of background clutter is minimal. This metric reflects the contribution of clutter to correct recognition results in the training set. A lower value indicates that a model is less dependent on clutter for recognition.

reconstruction significantly increases clutter impact (46.71%), and its attention module does not solve the problem. Clutter has the greatest effect on MVGGNet (55.71%).[2] This number also shows that the complex model (16.81 M) is more likely to overfit clutter than A-ConvNet (0.30 M and 27.69%). Other domain alignment methods also cannot overcome overfitting clutter (25.99% for SNINet and 34.45% for TDDA), resulting in domain alignment instability with clutter interference. Although the Shapley value is a rough estimate for the effect of clutter on recognition, our analysis is sufficient to show that domain alignment in SAR needs to disentangle targets and clutter features.

E. Limitations

While our method demonstrates satisfactory performance across a wide range of operating conditions, it is important to acknowledge its limitations. In this section, we present the limitations along with proposed solutions. Our method refers to the traditional detection and recognition process, treating the target and background clutter as separate entities solved by two different modules. While this separation reduces task difficulty, further integrating these two tasks has the potential to improve robustness and eliminate the current reliance on pseudolabels and hyperparameters. In addition, our approach heavily relies on data augmentation methods to expand the diversity of the single-domain dataset. However, the insufficient data still poses a constraint on our approach. To address this limitation, considering the increasing number of data from different SAR sensors, we plan to leverage a self-supervised learning approach to effectively extract features from a large volume of real-world SAR data. This way would enable us to mitigate the issues related to data bias and distribution shifts in small datasets.

V. CONCLUSION

This article proposes a novel domain alignment framework, named HDANet, for robust SAR vehicle recognition. Its primary objective is to achieve robust recognition with feature disentanglement and alignment, and the framework comprises three essential modules: DDG, MMD, and the DATF. Extensive experimental results conducted on the MSTAR dataset substantiate the effectiveness of the proposed approach in achieving robust recognition. The advantages and limitations of this approach are thoroughly discussed through comprehensive quantitative and qualitative analyses. Furthermore, future research directions involve investigating self-supervised learning with large amounts of SAR data from open sources to extract powerful features and address downstream task problems.

[2]Pretraining helps MVGGNet, and clutter contribution without pretraining is 65.35%.

REFERENCES

[1] G.-C. Sun, Y. Liu, J. Xiang, W. Liu, M. Xing, and J. Chen, "Spaceborne synthetic aperture radar imaging algorithms: An overview," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 161–184, Mar. 2021.

[2] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 1, pp. 6–43, Mar. 2013.

[3] O. Kechagias-Stamatis and N. Aouf, "Automatic target recognition on synthetic aperture radar imagery: A survey," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 3, pp. 56–81, Mar. 2021.

[4] J. Li, Z. Yu, L. Yu, P. Cheng, J. Chen, and C. Chi, "A comprehensive survey on SAR ATR in deep-learning era," *Remote Sens.*, vol. 15, no. 5, 2023, Art. no. 1454.

[5] V. Gagliardi et al., "Satellite remote sensing and non-destructive testing methods for transport infrastructure monitoring: Advances, challenges and perspectives," *Remote Sens.*, vol. 15, no. 2, 2023, Art. no. 418.

[6] M. Rizzi et al., "Navigation-aided automotive SAR imaging in urban environments," in *Proc. Int. Geosci. Remote. Sens. Symp.*, 2021, pp. 2979–2982.

[7] P.-O. Frolind, A. Gustavsson, M. Lundberg, and L. M. Ulander, "Circular-aperture VHF-band synthetic aperture radar for detection of vehicles in forest concealment," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1329–1339, Apr. 2012.

[8] L. Zhang et al., "Domain knowledge powered two-stream deep network for few-shot SAR vehicle recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5215315.

[9] B. Peng, B. Peng, J. Zhou, J. Xie, and L. Liu, "Scattering model guided adversarial examples for SAR target recognition: Attack and defense," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5236217.

[10] S. Chen, H. Wang, F. Xu, and Y. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[11] H. Ren, X. Yu, L. Zou, Y. Zhou, X. Wang, and L. Bruzzone, "Extended convolutional capsule network with application on SAR automatic target recognition," *Signal Process.*, vol. 183, 2021, Art. no. 108021.

[12] J. Zhang, M. Xing, and Y. Xie, "FEC: A feature fusion framework for SAR target recognition based on electromagnetic scattering features and deep CNN features," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2174–2187, Mar. 2021.

[13] S. Feng, K. Ji, F. Wang, L. Zhang, X. Ma, and G. Kuang, "Electromagnetic scattering feature (ESF) module embedded network based on ASC model for robust and interpretable SAR ATR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5235415.

[14] "The air force moving and stationary target recognition database," Accessed: May 21, 2022. [Online]. Available: https://www.sdms.afrl.af.mil/datasets/mstar/

[15] T. D. Ross, J. J. Bradley, L. J. Hudson, and M. P. O'connor, "SAR ATR: So what's the problem? An MSTAR perspective," in *Proc. 6th SPIE Conf. Algorithms SAR Imagery*, 1999, pp. 662–672.

[16] J. Fan and J. Liu, "The challenges and some thinking for the intelligentization of precision guidance ATR," in *Proc. SPIE Artif. Intell. Mach. Learn. Def. Appl.*, 2019, pp. 209–217.

[17] K. Ikeuchi, T. Shakunaga, M. D. Wheeler, and T. Yamazaki, "Invariant histograms and deformable template matching for SAR target recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1996, pp. 100–105.

[18] J. R. Diemunsch and J. Wissinger, "Moving and stationary target acquisition and recognition (MSTAR) model-based automatic target recognition: Search technology for a robust ATR," in *Proc. 5th SPIE Conf. Algorithms SAR Imagery*, 1998, pp. 481–492.

[19] M. Yang, X. Bai, L. Wang, and F. Zhou, "Mixed loss graph attention network for few-shot SAR target classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5216613.

[20] C. Li, L. Du, S. Deng, Y. Sun, and H. Liu, "Point-wise discriminative auto-encoder with application on robust radar automatic target recognition," *Signal Process.*, vol. 169, 2020, Art. no. 107385.

[21] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[22] Y. Kwak, W.-J. Song, and S.-E. Kim, "Speckle-noise-invariant convolutional neural network for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 549–553, Apr. 2019.

[23] Q. He, L. Zhao, K. Ji, and G. Kuang, "SAR target recognition based on task-driven domain adaptation using simulated data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 4019205.

[24] W. Li, W. Yang, L. Liu, W. Zhang, and Y. Liu, "Discovering and explaining the non-causality of deep learning in SAR ATR," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 4004605.

[25] D. E. Dudgeon and R. T. Lacoss, "An overview of automatic target recognition," *Lincoln Lab. J.*, vol. 6, no. 1, pp. 3–10, 1993.

[26] Y. Han and N. Yu, "Synthetic aperture radar target recognition based on joint classification of selected monogenic components by nonlinear correlation information entropy," *J. Appl. Remote Sens.*, vol. 15, no. 2, 2021, Art. no. 026502.

[27] J. Tan et al., "Target recognition of SAR images by partially matching of target outlines," *J. Electromagn. Waves Appl.*, vol. 33, no. 7, pp. 865–881, 2019.

[28] C. Ma, G. Wen, B. Ding, J. Zhong, and X. Yang, "Three-dimensional electromagnetic model-based scattering center matching method for synthetic aperture radar automatic target recognition by combining spatial and attributed information," *J. Appl. Remote Sens.*, vol. 10, no. 1, 2016.

[29] B. Ding, "Model-driven automatic target recognition of SAR images with part-level reasoning," *Optik*, vol. 252, 2022, Art. no. 168561.

[30] G. Gao, "An improved scheme for target discrimination in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 277–294, Jan. 2011.

[31] T. Li and L. Du, "SAR automatic target recognition based on attribute scattering center model and discriminative dictionary learning," *IEEE Sens. J.*, vol. 19, no. 12, pp. 4598–4611, Jun. 2019.

[32] G. Dong, H. Liu, and J. Chanussot, "Keypoint-based local descriptors for target recognition in SAR images: A comparative analysis," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 139–166, Mar. 2021.

[33] B. Hou, B. Ren, G. Ju, H. Li, L. Jiao, and J. Zhao, "SAR image classification via hierarchical sparse representation and multisize patch features," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 33–37, Jan. 2016.

[34] S. A. Wagner, "SAR ATR by a combination of convolutional neural network and support vector machines," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 6, pp. 2861–2872, Dec. 2016.

[35] Y. Yan, "Convolutional neural networks based on augmented training samples for synthetic aperture radar target recognition," *J. Electron. Imag.*, vol. 27, no. 2, 2018, Art. no. 023024.

[36] F. Zhou, L. Wang, X. Bai, and Y. Hui, "SAR ATR of ground vehicles based on LM-BN-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7282–7293, Dec. 2018.

[37] M. Heiligers and A. Huizing, "On the importance of visual explanation and segmentation for SAR ATR using deep learning," in *Proc. IEEE Radar Conf.*, 2018, pp. 0394–0399.

[38] L. Wang, X. Bai, and F. Zhou, "SAR ATR of ground vehicles based on ESENet," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1316.

[39] D. Wang, Y. Song, J. Huang, D. An, and L. Chen, "SAR target classification based on multiscale attention super-class network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 9004–9019, 2022.

[40] Z. Lin, K. Ji, M. Kang, X. Leng, and H. Zou, "Deep convolutional highway unit network for SAR target classification with limited labeled training data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1091–1095, Jul. 2017.

[41] R. Shang, J. Wang, L. Jiao, R. Stolkin, B. Hou, and Y. Li, "SAR targets classification based on deep memory convolution neural networks and transfer parameters," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 8, pp. 2834–2846, Aug. 2018.

[42] J. Ai, Y. Mao, Q. Luo, L. Jia, and M. Xing, "SAR target classification using the multikernel-size feature fusion-based convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5214313.

[43] Z. Wang, C. Wang, J. Pei, Y. Huang, Y. Zhang, and H. Yang, "A deformable convolution neural network for SAR ATR," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2639–2642.

[44] Y. Zhao, L. Zhao, Z. Liu, D. Hu, G. Kuang, and L. Liu, "Attentional feature refinement and alignment network for aircraft detection in SAR imagery," *IEEE Trans Geosci. Remote Sen.*, vol. 60, pp. 1–16, 2022. [Online]. Available: https://arxiv.org/abs/2201.07124

[45] F. Zhang, C. Hu, Q. Yin, W. Li, H.-C. Li, and W. Hong, "Multi-aspect-aware bidirectional LSTM networks for synthetic aperture radar target recognition," *IEEE Access*, vol. 5, pp. 26880–26891, 2017.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[47] Z. Huang, Z. Pan, and B. Lei, "What, where, and how to transfer in SAR target recognition based on deep CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2324–2336, Apr. 2020.

[48] K. Wang, G. Zhang, and H. Leung, "SAR target recognition based on cross-domain and cross-task transfer learning," *IEEE Access*, vol. 7, pp. 153391–153399, 2019.

[49] B. Lewis, O. DeGuchy, J. Sebastian, and J. Kaminski, "Realistic SAR data augmentation using machine learning techniques," in *Proc. 26th SPIE Conf. Algorithms SAR Imagery*, 2019, pp. 12–28.

[50] X. Ma, K. Ji, L. Zhang, S. Feng, B. Xiong, and G. Kuang, "An open set recognition method for SAR targets based on multitask learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 4014005.

[51] Z. Luo, X. Jiang, and X. Liu, "Synthetic minority class data by generative adversarial network for imbalanced SAR target recognition," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2459–2462.

[52] F. Zhang, Y. Liu, Y. Zhou, Q. Yin, and H.-C. Li, "A lossless lightweight CNN design for SAR target recognition," *Remote Sens. Lett.*, vol. 11, no. 5, pp. 485–494, 2020.

[53] W. Li, W. Yang, Y. Liu, and X. Li, "Research and exploration on the interpretability of deep learning model in radar image (in chinese)," *Scientia Sinica Inform.*, vol. 52, pp. 1114–1134, 2022.

[54] C. Belloni, A. Balleri, N. Aouf, J.-M. Le Caillec, and T. Merlet, "Explainability of deep SAR ATR through feature analysis," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 1, pp. 659–673, Feb. 2021.

[55] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[56] P. Bühlmann, "Invariance, causality and robustness," *Statist. Sci.*, vol. 35, no. 3, pp. 404–426, 2020.

[57] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *Proc. Int. Conf. Machin. Learn.*, 2021, pp. 7313–7324.

[58] F. Lv et al., "Causality inspired representation learning for domain generalization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8046–8056.

[59] H. Rohling, "Radar CFAR thresholding in clutter and multiple target situations," *IEEE Trans. Aerosp. Electron. Syst.*, no. 4, pp. 608–621, Jul. 1983.

[60] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *J. Mach. Learn. Res*, vol. 19, no. 1, pp. 1947–1980, 2018.

[61] L. Wang et al., "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 136–145.

[62] B. Ding, "Research on automatic target recognition of synthetic aperture radar images under extend operating conditions (in chinese)," Ph.D. dissertation, Nat. College Electron. Sci. Technol., State Key Lab. Complex Electromagn. Environ. Effects Electron. Inf. Syst., Changsha, China, 2018.

[63] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017pp. 3859–3869.

[64] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Machin. Learn.*, 2020, pp. 1597–1607.

[65] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[67] K. Fu, W. Dai, Y. Zhang, Z. Wang, M. Yan, and X. Sun, "MultiCAM: Multiple class activation mapping for aircraft recognition in remote sensing images," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 544.

[68] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017. Accessed: Feb. 10, 2021. [Online]. Available: https://arxiv.org/abs/1706.03825

[69] F. Ulaby, M. C. Dobson, and J. L. Álvarez-Pérez, *Handbook of Radar Scattering Statistics for Terrain*. Norwood, MA, USA: Artech House, 2019.
[70] S. Huang, C. Tong, C. Tong, T. Wang, and W. Liu, "Studies on the EM scattering from the earth surface covered with grass (in chinese)," *J. Microw.*, vol. 33, no. 6, pp. 90–96, 2017.
[71] B. Alsallakh, N. Kokhlikyan, V. Miglani, J. Yuan, and O. Reblitz-Richardson, "Mind the pad–CNNs can develop blind spots," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–36.
[72] S. Ruder, "An overview of gradient descent optimization algorithms," 2016. Accessed: Mar. 25, 2020. [Online]. Available: https://arxiv.org/abs/1609.04747
[73] J.-H. Choi, M.-J. Lee, N.-H. Jeong, G. Lee, and K.-T. Kim, "Fusion of target and shadow regions for improved SAR ATR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5226217.
[74] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018. Accessed: Feb. 18, 2021. [Online]. Available: https://arxiv.org/abs/1802.03426
[75] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 765.
[76] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9505–9515.
[77] N. Kokhlikyan et al., "Captum: A unified and generic model interpretability library for pytorch," 2020. Accessed: Feb. 7, 2022. [Online]. Available: https://arxiv.org/abs/2009.07896
[78] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *Proc. Int. Conf. Machin. Learn.*, 2020, pp. 9269–9278.
[79] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–22.

Weijie Li was born in 1998. He received the B.E. degree in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2019. He is currently working toward the Ph.D. degree in information and communication engineering with the National University of Defense Technology, Changsha, China.

He has authored or coauthored papers in respected journals, including the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, *Scientia Sinica Informationis*, and *Journal of Radars*. His research interests include radar target recognition and deep learning.

Wei Yang was born in 1985. He received the B.E. degree in information and communication engineering from Wuhan University, Wuhan, China, in 2006, and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2008 and 2012, respectively.

He is currently an Associate Professor with the National University of Defense Technology. He has authored or coauthored numerous papers in respected journals, including the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, *IET Signal Processing*, and *Journal of Systems Engineering and Electronics*. His research interests include cognitive radar target detection and recognition and radar target tracking.

Wenpeng Zhang was born in 1989. He received the B.E., M.S., and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2012, 2014, and 2018, respectively.

He is currently an Assistant Professor with the National University of Defense Technology. He has authored or coauthored numerous papers in respected journals, including the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, IEEE SIGNAL PROCESSING LETTERS, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. His research interests include radar signal processing, parameter estimation, and sparse signal representation.

Tianpeng Liu received the B.E., M.E., and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2008, 2011, and 2016 respectively.

He is currently an Associate Professor with the College of Electronic Science and Technology, Chengdu, China. He has authored or coauthored numerous papers in respected journals, including the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS and International Conference on Signal Processing. His research interests include radar signal processing, electronic countermeasure, and cross-eye jamming.

Yongxiang Liu (Member, IEEE) received the B.S. and Ph.D. degrees in information and communication engineering from the College of Electronic Science, National University of Defense Technology, Changsha, China, in 1997 and 2004, respectively.

He is currently a Full Professor with the National University of Defense Technology. He has authored or coauthored numerous papers in respected journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include radar imaging, SAR image interpretation, and artificial intelligence.

Li Liu (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2012.

She is currently a Full Professor with NUDT. She has held visiting appointments with the University of Waterloo, Waterloo, ON, Canada, The Chinese University of Hong Kong, Hong Kong, and University of Oulu, Oulu, Finland, respectively. Her research interests include computer vision, pattern recognition, and machine learning.

Dr. Liu was a Co-Chair of many International Workshops along with major venues, such as CVPR and ICCV. She was the Leading Guest Editor of the special issues for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and *International Journal of Computer Vision*. She was also the Area Chair of several respected international conferences. She is currently an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Pattern Recognition*. Her papers currently have more than 10 000 citations, according to Google Scholar.