IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 March 2024, accepted 18 March 2024, date of publication 10 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387314

RESEARCH ARTICLE

Hybrid Conv-Attention Networks for Synthetic Aperture Radar Imagery-Based Target Recognition

JISEOK YOON[1], (Member, IEEE), JEONGHEON SONG[2], (Member, IEEE), TANVEER HUSSAIN[3], (Member, IEEE), SUNDER ALI KHOWAJA[4], (Member, IEEE), KHAN MUHAMMAD[5], (Senior Member, IEEE), AND IK HYUN LEE[1,6], (Member, IEEE)

[1]IKLAB Inc., Siheung-Si 15073, South Korea
[2]Korea Aerospace Research Institute, Daejeon 34133, South Korea
[3]Department of Computer Science, Edge Hill University, L39 4QP Ormskirk, U.K.
[4]Department of Telecommunication Engineering, University of Sindh, Jamshoro 76080, Pakistan
[5]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul 03063, South Korea
[6]Department of Mechatronics Engineering, Tech University of Korea, Siheung-Si 15073, South Korea

Corresponding authors: Khan Muhammad (khanmuhammad@g.skku.edu) and Ik Hyun Lee (ihlee@tukorea.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by Korean Government [Ministry of Science and ICT (MSIT)] under Grant 2021R1F1A1050022; and in part by the Gyeonggi-do Regional Research Center (GRRC) Program of Gyeonggi Province, Development of an Intelligent Inspection System and an Autonomous Navigation System for the Transportation of Multi-Material Parts, under Grant GRRC TUKorea2023-B03.

ABSTRACT In this study, we propose hybrid conv-attention networks that combine convolutional neural networks (CNNs) and transformers to recognize targets from synthetic aperture radar (SAR) images automatically. The proposed model is designed to obtain robust features from global and local patterns in the SAR image, utilizing the weights of a pre-trained backbone model with self-attention structures. Furthermore, we adopted pre-processing and training methods optimized for transfer learning to enhance performance. By comparing and analyzing the performance between the proposed model and conventional models using the OpenSARShip and MSTAR dataset, we found that our system significantly outperforms conventional approaches, with a performance improvement of 24.06%. This considerable enhancement is attributed to the ability of the model to leverage the 2D kernel-based approach of CNNs and the sequence vector-based approach of transformers, offering a comprehensive method for SAR image target recognition.

INDEX TERMS Synthetic aperture radar (SAR), target recognition, deep learning (DL), transfer learning, convolutional neural networks (CNNs), transformers.

I. INTRODUCTION

Synthetic aperture radar (SAR) imaging is a technique that captures the reflection of transmitted electromagnetic waves in 2- and 3-dimensional forms [1], [2]. It is highly resistant to environmental factors like haze, smog, and clouds. SAR is known for its ability to quickly capture large areas of land using airborne or satellite-mounted devices, making it an ideal choice for surveillance applications in both military and civilian sectors [2], [3]. SAR is crucial in various applications, including security surveillance, disaster response, and environmental monitoring. However, monitoring vast areas solely through human efforts can be challenging. Therefore, an urgent need for automation has driven researchers to develop advanced detection and recognition systems.

Recent advancements in artificial intelligence have significantly enhanced systems utilizing SAR images, streamlining tasks like automatic target recognition (ATR), land cover classification, segmentation, and image noise reduction [4], [5], [6], [7], [8], [9]. The research presented in this paper contributes to this progressive field, specifically focusing on refining the classification aspect integral to ATR systems [2], [3], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. This endeavor is crucial as classification forms the

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

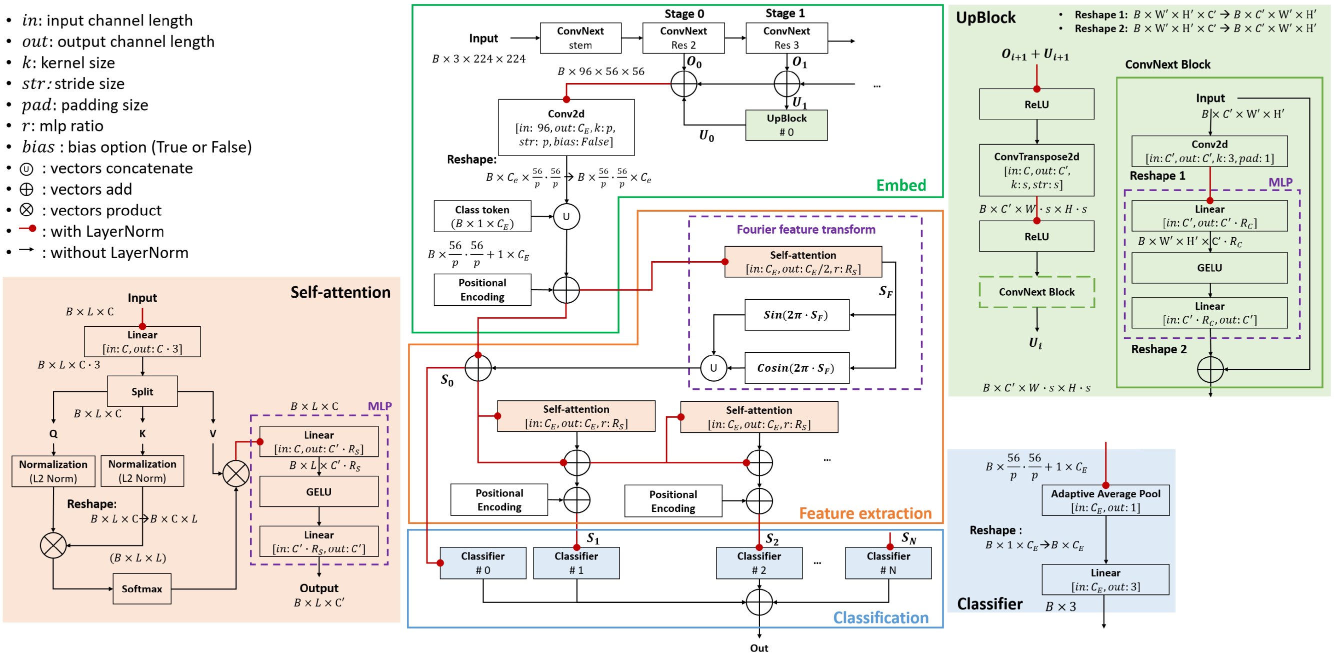


FIGURE 1. The proposed model consists of three main parts: embed, feature extraction, and classification. (zoom in for more clarity).

bedrock of efficient and accurate target recognition in varied real-world scenarios.

In automated systems, the spotlight has shifted to deep learning (DL) methodologies, celebrated for their efficacy and streamlined approach, negating the need for intricate feature vector designs characteristic of traditional methods [11], [12]. A popular tactic within DL is transfer learning, where pre-trained classifiers, initially devised for generic object classification, are repurposed [13], [14]. Despite the widespread adoption of this strategy, it is challenging to pinpoint a model that's intricately tailored for SAR images, as most are honed for publicly available data sets rather than the unique characteristics of SAR imagery. Recognizing this gap, several researchers have proposed models meticulously optimized for SAR image classification, aiming to enhance the precision and reliability of automated systems in processing and interpreting SAR imagery [3], [10], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

Previous research has focused on the overfitting problem of existing convolutional neural networks (CNNs). For example, Chen et al. [10] proposed A-ConvNets composed of sparse layers to solve this issue. Moreover, Lin et al. [21] introduced the CHU-Net, a robust SAR data classification network integrating convolutional highway layers, max pooling, and dropout for efficiency with limited data. The architecture achieved notable accuracy, effectively addressing SAR's processing challenges and illustrating the network's proficiency in DL with scarce labeled data. Zhang et al. [3] proposed a domain adaptation-based method using heterogeneous features to overcome the vulnerability of noise from the A-ConvNets method. Deng et al. [15] developed an amplitude-phase CNNs (AP-CNN), utilizing amplitude and phase data from sparse images, enhanced by a bi-iterative soft thresholding (BiIST) algorithm to improve quality. This approach improved classification accuracy more than traditional amplitude-based methods, especially with limited training data. In another approach, Li et al. [16] proposed a metric learning method to obtain robust performance even in insufficient dataset environments. They designed a two step training process; first, they designed CNNs to extract the trainable features, and second, the features were used for metric learning processes.

Other research groups proposed hybrid models by combining the CNNs with other DL models. Wang et al. [17] presented a novel approach by combining CNNs and long short-term memory (LSTM) networks in a multiview framework. The methodology involved using multiple CNNs modules to extract deep features from single-view SAR images and then employing a spatial attention module to focus on relevant target details and suppress noise. Subsequently, the LSTM module performed feature fusion using the correlations of features from adjacent azimuths, enhancing recognition performance with multiview image inputs. Feng et al. [23] developed a network based on the attributed scattering center (ASC) model with an electromagnetic scattering feature (ESF) module, utilizing convolutional layers and Bi-LSTM for enhanced SAR ATR. This design improved target recognition accuracy and interpretability by integrating SAR characteristics into DL. Zhang et al. [18] focused on using the traditional features efficiently, proposing a DL model combined with the histogram of oriented gradient (HOG) features called HOG-ShipCLSNet. This model uses the CNNs structure, feature pyramid networks, feature ensemble, and attention mechanism. Wang et al. [19] proposed a convolutional transformer (ConvT) tailored to few-shot learning scenarios, integrating convolutional layers with transformers to construct a hierarchical feature representation that captures both local and global dependencies. It introduced hybrid loss and auto augmentation strategies to optimize the model for limited SAR images, enhancing its recognition capability and generalization performance without additional SAR target images in training. Wang et al. [20] introduced a multiscale attention super-class CNNs (MSA-SCNN). The method enhanced SAR target feature representation through multiscale feature fusion coupled with channel and spatial attention modules, emphasizing the importance of different scale features. Additionally, introducing super-class labels helped increase feature differences between categories, aiming to improve the network's fine classification ability and generalization across various SAR images. Feng et al. [22] developed the part attention module (PAN), enhancing SAR vehicle target recognition with a focus on interpretability by integrating electromagnetic scattering characteristics. The network, featuring a PAN, adapted to the importance of different target parts, ensuring precise and transparent recognition and showcasing robust performance in diverse conditions. Wen et al. [24] proposed a multimodal framework integrating phase-history, scattering topology, and image data for SAR ATR, enhancing feature extraction and fusion. The paper showcased its high recognition accuracy on the MSTAR dataset, particularly with limited data, highlighting its potential to advance SAR ATR capabilities significantly. Zhang et al. [25] proposed the MGSFA-Net, integrating SAR scattering features with DL through a hybrid structure utilizing multi-scale feature enhancement and graph convolution networks (GCN) for SAR ship target recognition. This method enriches feature diversity and accuracy, improving performance in few-shot scenarios on FUSAR-Ship and OpenSARShip datasets.

Conclusively, this study aims to solve data scarcity and the need for adaptability in SAR imagery by introducing a hybrid conv-attention network. This network combines the strengths of CNNs and transformers [26], [27], [28], two technologies that have shown remarkable performance in recent AI research. The hybrid model's unique ability to fuse local and global features lies at the crux of our architecture. This comprehensive representation is crucial for effectively recognizing targets in SAR images. Our approach distinguishes itself from existing algorithms by providing a balanced and sophisticated analysis of SAR imagery designed to capture the intricate patterns inherent in such data. The proposed model leverages pre-trained models to develop a robust system adaptable to varying types of SAR images against the limitations of small-scale SAR datasets. It consists of three core modules: an embedder using a backbone CNNs model [29] with proposed trainable upscale structures to merge the outputs of each stage, features extraction by employing self-attention (SA), which extracts global features from the 1D feature sequence, and a classifier using the ensemble structure.

This paper aims to identify three types of ships - bulk carrier, container ship, and tanker - from the OpenSARShip dataset [30], [31]. Additionally, it strives to recognize different categories of military vehicles using three distinct scenarios from the moving and stationary target acquisition and recognition (MSTAR) dataset [32]. The study will be conducted in detail and organized as follows. Section II explains the proposed methodology, including the proposed system processes and proposed model architecture. The results and comparisons of our experiments are detailed in III. Finally, the overall study is concluded in Section IV with a discussion regarding useful research in SAR imaging in real-world applications.

II. METHODOLOGY

A. THE PROPOSED SYSTEM

The proposed system is a transfer learning-based methodology consisting of pre-processing and classification models to recognize targets using SAR images automatically. The pre-processing of this system transforms the input image into a form suitable for the proposed model, which consists of algorithms that optimize the characteristics of the employed SAR image used for transition learning. Because of the acquisition process of the images, they have a massive range of pixel values, various image widths & height sizes, and a single channel.

Therefore, three processes were adopted for the normalization of pixel values, image resizing, and channel upsampling. i) First, the pixel values are normalized in the $[-1, 1]$ range using z-score values with $\mu = 0.5$ and $\sigma = 0.5$. ii) Then, using bilinear interpolation, the input images are resized as $1 \times 224 \times 224$. iii) Finally, concatenating three times from one input image, the input image is upsampled as $3 \times 224 \times 224$.

Our proposed method utilizes the ConvNext model and the SA of the transformers as a hybrid model fused with some layers of the pre-trained model and newly designed layers (Figure 1). A detailed structural description is shown in subsection B.

To utilize the system configured above, it is necessary to design a training process optimized for the data used to achieve the best results. Therefore, we have established strategies for small data and converging the global minimization of training losses. The first strategy is to augment the small training datasets using physical transformation. The data augmentation algorithm consists of flip, thin plate spline, affine transform, perspective transform, and random crop. The SAR image is acquired using a satellite, and parts that can be deformed into the shape of an object during the acquisition process are reflected in the augmentation algorithm.

Additionally, for the random crop, data augmentation and the detection result of the OpenSARShip dataset can be corrected during the training process [30], [31]. Then, each algorithm transforms the images based on probability. For the second strategy, the weights of the new layers of the

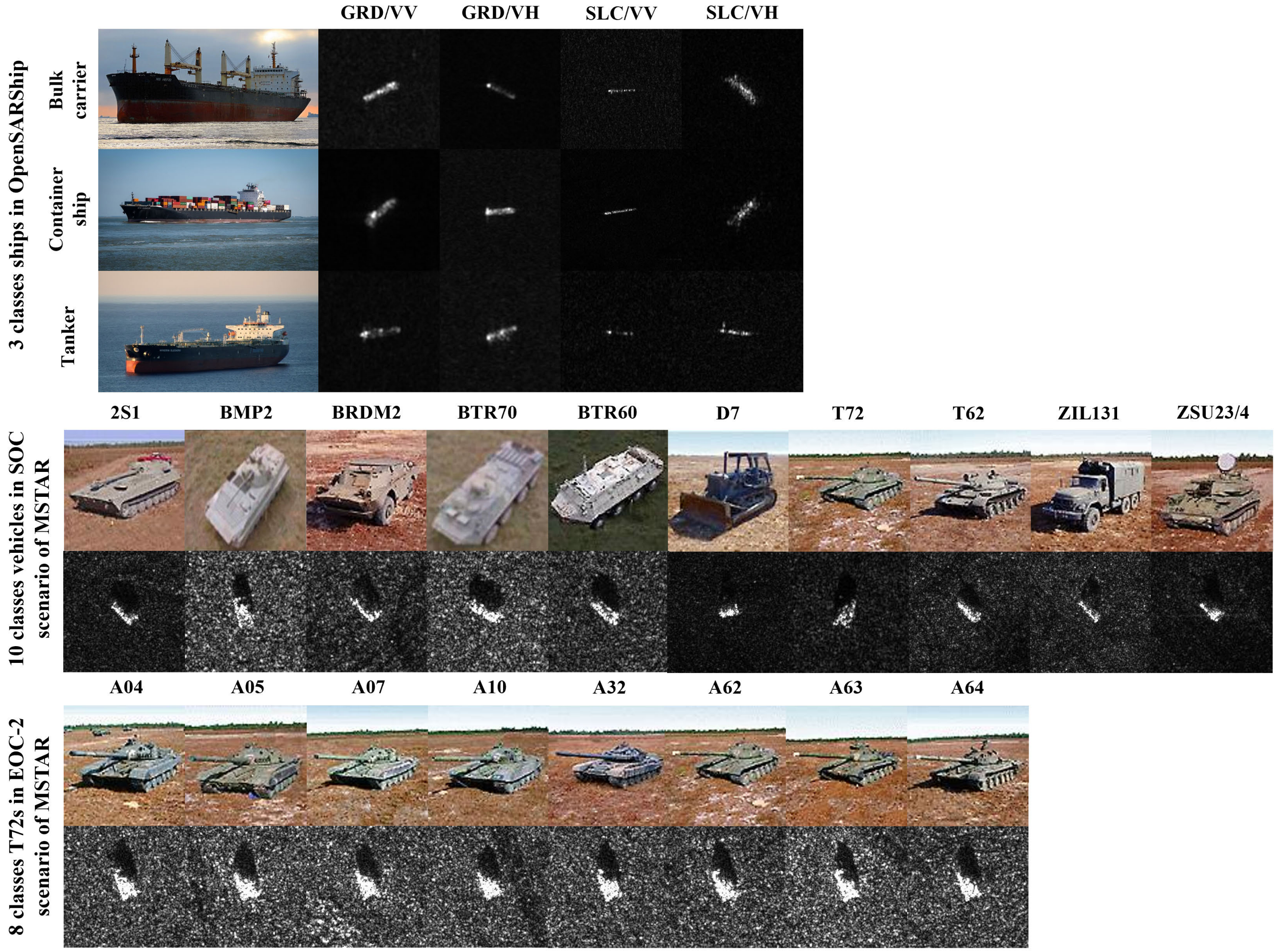


FIGURE 2. Target SAR images of the proposed system in the OpenSARShip and MSTAR dataset.

model that cannot use pre-trained weights are initialized based on the normal distribution. In addition, a scheduler that changes the learning rate in the form of a cosine wave, called the cosine annealing learning rate, was used to prevent convergence to the local minima [33]. When using this scheduler, the learning results continue to change for each epoch; the validation dataset was constructed to store the model with the best performance.

B. THE PROPOSED MODEL ARCHITECTURE

The proposed hybrid model combines CNNs, which can acquire local patterns on input images, and the SA structure of the transformers, which can extract global patterns. The main parts include an embedder, feature extractors, and a classifier (Figure 1). The embedder transforms an input image into a 1D sequence form with C channel size. The proposed embedder transforms 2D feature maps into an optimized feature sequence; in this case, a pre-trained ConvNext model is utilized. The feature extractor is used to acquire features from a sequence using an SA structure. The classifier is based on an ensemble structure that uses average scores.

1) EMBEDDER

The proposed embedder extracts a 96 × 56 × 56 2D feature vector using a 3 × 224 × 224 image and transforms the 2D feature vector into a $C_E \times 56/p \cdot 56/p$ 1D sequence again, where p is the patch size, and C_f is the factor of the output channels of the embedded.

First, the 2D feature vector is extracted by using the output of the intermediate layers of pre-trained ConvNext and the proposed upblock. The ConvNext consists of a single input filter (stem), four stages, and a classifier, where the partial outputs of the stages outputs were used for making the 2D feature vector. In our study, we used the first to third outputs of the stages. Each stage output of the pre-trained model has half the sizes of W and H compared to the previous output, and the channel length is doubled. In other words, the feature vector size of O_i is upsampled, and the length of the channels decreases. Therefore, the ConvTranspose2d layer was adopted, which flexibly selects the size of the scale (s) and the length of output channels [34]. The modified ConvNext block (CNB) [29] smooths the upsampled feature vector. In our case, the filter and padding size of the convolution layer and the hidden layer size of the multi-layer perceptron (MLP) were modified except for the whole structure flow. Here, the MLP structure consists of two linear layers with a single Gaussian error linear unit (GELU) [35]. The kernel size is 3, the padding size is 1, and the hidden layer size is designed to increase as much as the input channel length multiplied by the MLP ratio R_C. Moreover, the inputs of ConvTranspose2d and CNB are activated and normalized using a rectified linear unit (ReLU) [36] and LayerNorm to improve learning stability [37]. In this process, the LayerNorm is based on the z-score, as mathematically given in Equation 1.

$$x' = \frac{x - \mu_x}{\sqrt{\sigma_x + \epsilon}} \cdot \alpha + \beta, \tag{1}$$

where α and β are trainable factors, and ϵ is used to prevent that sigma becoming zero. Through the approaches, a 2D feature map is acquired from the merged local features obtained from each stage.

Next, the obtained $C_E \times 56 \times 56$ sized 2D feature map is transformed into a 1D sequence. For this process, the 2D feature map is downsampled by the patch size (p), of which the result is $C_E \times 56/p \times 56/p$. Then, the feature map is reshaped as a $56/p \cdot 56/p \times C_E$ sized 1D sequence. Before using the sequence as input of the feature extractor, the class token as the positional encoding is injected to increase the performance. The class token is a representation vector for the sequence [27]: the trainable parameter. Here, we use a $1 \times 1 \times 96 \cdot p$ vector. This class token is attached in front of the sequence vector; therefore, the final sequence shape is $(56/p \cdot 56/p + 1) \times C_E$. The final process of the embedder is positional encoding. The proposed model uses SA in the proposed feature extractor. In this case, positional encoding is performed because the positional information between sequences is required. The proposed method is the sinusoidal position, as given in Equation 2.

$$\begin{aligned} P_{(i,2j)} &= sin(i \cdot l^{-\frac{2j}{C_E}}) + \gamma \\ P_{(i,2j+1)} &= cos(i \cdot l^{-\frac{2j}{C_E}}) + \gamma, \end{aligned} \tag{2}$$

where i is an element of position vector $\{0, \ldots, (56/p \cdot 56/p)\}$, j is channel index, l is fixed weight of the positional encoding, and γ is trainable bias which has $1 \times 56/p \cdot 56/p \times C_E$. For our model, the weight l was 6273. After all of the above processes, the sequence vector is normalized by the LayerNorm.

2) FEATURE EXTRACTOR

The feature extractor is for extracting the global patterns, and it is designed in a form capable of learning residual components and utilizes the SA structure. In addition, a Fourier feature transformation (FFT) structure [38] was used to include the high-frequency components in the structure utilizing linear layers where the FFT is a combination of channel-wise cosine & sine waveforms.

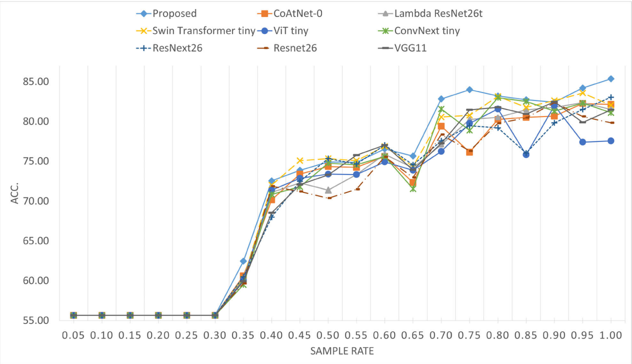


FIGURE 3. Accuracy performances using various sampled training data in the GRD/VV of OpenSARShip.

Next, the transformed vector propagates through other SA-based feature extractors to obtain an optimized feature sequence vector. However, a positional encoder was added again in each SA output because positional information can be blurred when using the residual structure in multiple SA-based feature extractors. Herein, the extractors were repeated N times, and their number varied depending on the data types. The SA structure used in these processes is shown in Figure 1, which aims to express similarity in a sequence probabilistically in the value vector. For this process, the query, key, and values vectors are obtained by splitting the first output vector of the linear layer. Then, a probability vector can be obtained through softmax output, in which the input is the product of the query and key vectors. Finally, the optimal feature information in the sequence is obtained by reflecting this in the value vector using the multiplied probability vector. In this structure, the query and key are normalized by the L2 norm, as shown mathematically in Equation 3.

$$x' = \frac{x}{\|x\|_2}. \tag{3}$$

The obtained feature sequence passes through an MLP structure, a feedforward network, and is finally transformed into one optimized feature sequence. In the MLP structure, the hidden layer size is also a changeable parameter. Therefore, the hidden layers of different sizes were adopted according to data types, and the proposed model was designed to adjust the ratio of input layers.

3) CLASSIFIER

The proposed classification process is an ensemble-based method using multiple classifiers. A classifier is attached to each proposed feature extractor to obtain each output, and the output is summed to obtain a final classification result. Each classifier has a structure in which an adaptive average of the acquired feature sequence is obtained for each channel information, and each class is expressed through a weighted sum of the acquired channel information. In this process, the adaptive average is the weighted sum of the input sequence, and the LayerNorm is used to normalize the input feature sequence.

TABLE 1. Key parameters of the model in each scenario.

Parameters	GRD/VV	GRD/VH	SLC/VV	SLC/VH	SOC	EOC-1	EOC-2
# of stage	3	3	3	3	3	3	3
MLP ratio in CNB (R_C)	4	4	4	4	4	4	4
Patch size (p)	4	4	4	2	4	4	2
# of output channels in embed (C_E)	24	48	48	48	12	24	24
# of feature extractor	2	2	1	4	1	1	3
MLP ratio in SA (R_S)	8	1	1	1	1	1	1

TABLE 2. Number of data in each case.

OpenSARShip					
		Bulk	**Container**	**Tanker**	
GRD	**Tr.**	338(Val.=50)	338(Val.=50)	338(Val.=50)	
	Te.	1054	624	216	
SLC	**Tr.**	242(Val.=36)	242(Val.=36)	242(Val.=36)	
	Te.	361	515	105	
MSTAR					
SOC		**2S1**	**BMP2**	**BRDM2**	
	Tr.(17°)	299(Val.=45)	233(Val.=35)	298(Val.=45)	
	Te.(15°)	274	195	274	
		BTR70	**BTR60**	**D7**	
	Tr.(17°)	233(Val.=35)	256(Val.=38)	299(Val.=45)	
	Te.(15°)	196	195	274	
		T62	**T72**	**ZIL131**	**ZSU23/4**
	Tr.(17°)	299(Val.=45)	232(Val.=35)	299(Val.=45)	299(Val.=45)
	Te.(15°)	273	196	274	274
EOC-1		**2S1**	**BRDM2**	**T72(A64)**	**ZSU23/4**
	Tr.(17°)	299(Val.=44)	298(Val.=44)	299(Val.=44)	299(Val.=44)
	Te.(30°)	288	287	288	288
EOC-2		**A04**	**A05**	**A07**	**A10**
	Tr.(17°)	299(Val.=45)	299(Val.=45)	299(Val.=45)	296(Val.=44)
	Te.(15°)	274	274	274	271
		A32	**A62**	**A63**	**A64**
	Tr.(17°)	298(Val.=45)	299(Val.=45)	299(Val.=45)	299(Val.=45)
	Te.(15°)	274	274	274	274

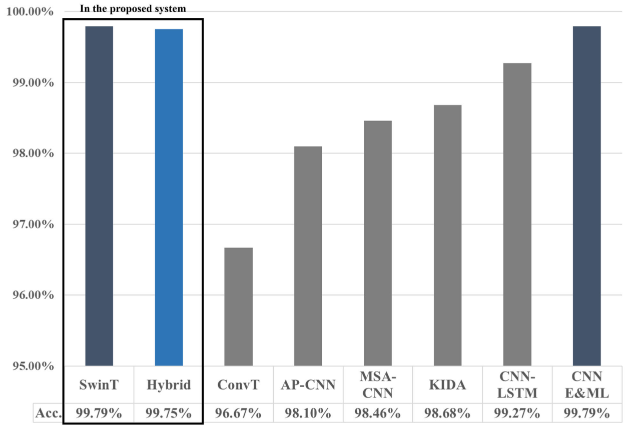


FIGURE 4. Accuracy performances compared in the SOC scenario of MSTAR.

III. EXPERIMENTAL RESULTS

A. DATASET

1) OpenSARShip

The OpenSARShip dataset (Figure 2), obtained from Sentinel-1 SAR images by the Shanghai Key Laboratory of Intelligent Sensing and Recognition, has two types: ground range detected (GRD) and single look complex (SLC). Each type is based on dual-polarization using VV (vertical transmit and vertical receive) and VH (vertical transmit and horizontal receive). Here, four types of GRD datasets were used separately (GRD/VV, GRD/VH, SLC/VV, and SLC/VH) to check the flexibility of the proposed system. Moreover, the dataset has 17 labels for ship images, which is a strongly imbalanced number. However, some ship images are unsatisfactory for evaluating the systems owing to the number of images in some labels. Therefore, we used only three labels for our experiment: bulk carrier (Bulk), container ship (Container), and Tanker (Table 2).

2) MSTAR

The MSTAR dataset (Figure 2), crafted by Sandia National Laboratories using the STARLOS sensor, achieves a detailed resolution of 1 foot (0.3m). For an in-depth classification performance evaluation, the dataset is examined under three scenarios: standard operating conditions (SOC) and extended operating conditions (EOC-1 and EOC-2) (Table 2). In the SOC scenario, the model is trained on data captured at a 17° depression angle and tested on images at 15°, encompassing all ten vehicle classes, including diverse vehicles such as the 2S1 rocket launcher, BMP2, BRDM2, BTR60, BTR70, D7 bulldozer, T62, T72 tanks, ZIL131 truck, and the ZSU23/4 air defense unit. These classes are represented under various conditions, marked by changes in aspect and depression angles, alongside unique serial numbers.

Under the EOC-1 scenario, the focus is narrowed to 4 specific targets: 2S1, BRDM2, T72, and ZSU23/4 from the SOC. For these, training occurs at 17°, with testing expanded to 30° to assess performance under more varied conditions. EOC-2 scenario introduces a specialized subset featuring 8 variants of the T72 tank (labeled A04, A05, A07, A10, A32, A62, A63, and A64), maintaining the training at 17° and testing at 15° to evaluate the algorithm's ability to discern closely related models. These rigorous and diverse conditions are designed to thoroughly test and validate the robustness of classification algorithms.

B. EXPERIMENTAL ENVIRONMENT

The proposed system was implemented on MATLAB ® 2021a and Pytorch 2.1.0 with CUDA toolkit 12.1 in the Ubuntu 20.04.6 LTS 64-bit OS. In addition, the hardware environment used a GeForce RTX 4090 and an AMD ® Ryzen threadripper pro 5975wx 32-cores x 64 CPU and 256 GB RAM. The augmentation algorithms were implemented via Kornia [39], and the pre-trained model was loaded from PyTorch Image Models [40].

C. IMPLEMENTATION DETAILS

1) MODELS

The hyperparameters of the proposed models were modified to fit a specific data type. By modifying the parameters shown in Table 1, a model optimized for data characteristics can be obtained. The parameters were obtained heuristically.

TABLE 3. Comparison results using various types of DL models in the proposed system.

Data type		Model / Metric	VGG11	Resnet26	ResNext26	ViT tiny	ConvNext tiny	Swin Transformer tiny	Lambda ResNet26t	CoAtNet-0	Proposed
OpenSARShip	GRD/VV	Acc.	0.8147	0.7983	0.8305	0.7756	0.8110	0.8205	0.8152	0.8215	**0.8537**
		Spec.	0.9073	0.8992	0.9153	0.8878	0.9055	0.9102	0.9076	0.9108	**0.9269**
		Pres.	0.8247	0.8166	0.8343	0.8031	0.8273	0.8371	0.8233	0.8340	**0.8540**
		Rec.	0.8147	0.7983	0.8305	0.7756	0.8110	0.8205	0.8152	0.8215	**0.8537**
		F1.	0.8168	0.8025	0.8318	0.7794	0.8142	0.8240	0.8171	0.8244	**0.8533**
	GRD/VH	Acc.	0.7761	0.7540	0.7624	0.7878	0.7920	0.7788	0.7777	0.8004	**0.8405**
		Spec.	0.8881	0.8770	0.8812	0.8939	0.8960	0.8894	0.8889	0.9002	**0.9203**
		Pres.	0.7962	0.8005	0.7967	0.8045	0.8129	0.8008	0.8022	0.8067	**0.8423**
		Rec.	0.7761	0.7540	0.7624	0.7878	0.7920	0.7788	0.7777	0.8004	**0.8405**
		F1.	0.7790	0.7571	0.7662	0.7913	0.7950	0.7817	0.7812	0.8021	**0.8410**
	SLC/VV	Acc.	0.7727	0.7931	0.7890	0.7778	0.7951	0.8043	0.7706	0.7737	**0.8338**
		Spec.	0.8863	0.8965	0.8945	0.8889	0.8976	0.9021	0.8853	0.8869	**0.9169**
		Pres.	0.7846	0.7925	0.7886	0.7763	0.8004	0.8092	0.7767	0.7903	**0.8376**
		Rec.	0.7727	0.7931	0.7890	0.7778	0.7951	0.8043	0.7706	0.7737	**0.8338**
		F1.	0.7745	0.7927	0.7888	0.7762	0.7956	0.8054	0.7725	0.7769	**0.8348**
	SLC/VH	Acc.	0.7717	0.7757	0.7778	0.7757	0.8236	0.7870	0.7757	0.7788	**0.8400**
		Spec.	0.8858	0.8879	0.8889	0.8879	0.9118	0.8935	0.8879	0.8894	**0.9200**
		Pres.	0.7775	0.7780	0.7815	0.7792	0.8292	0.7930	0.7767	0.7881	**0.8456**
		Rec.	0.7717	0.7757	0.7778	0.7757	0.8236	0.7870	0.7757	0.7788	**0.8400**
		F1.	0.7725	0.7766	0.7776	0.7761	0.8243	0.7881	0.7754	0.7800	**0.8410**
MSTAR	SOC	Acc.	0.9880	0.9905	0.9942	0.9938	0.9951	**0.9979**	0.9889	0.9918	0.9975
		Spec.	0.9987	0.9989	0.9994	0.9993	0.9995	**0.9998**	0.9988	0.9991	0.9997
		Pres.	0.9881	0.9908	0.9943	0.9939	0.9951	**0.9979**	0.9890	0.9919	0.9975
		Rec.	0.9880	0.9905	0.9942	0.9938	0.9951	**0.9979**	0.9889	0.9918	0.9975
		F1.	0.9880	0.9905	0.9942	0.9938	0.9951	**0.9979**	0.9889	0.9917	0.9975
	EOC-1	Acc.	0.8375	0.8054	0.8332	0.7941	0.8593	0.9331	0.8132	0.9123	**0.9852**
		Spec.	0.9458	0.9351	0.9444	0.9314	0.9531	0.9777	0.9377	0.9708	**0.9951**
		Pres.	0.8967	0.8836	0.8977	0.8810	0.9074	0.9455	0.8925	0.9272	**0.9858**
		Rec.	0.8375	0.8054	0.8332	0.7941	0.8593	0.9331	0.8132	0.9123	**0.9852**
		F1.	0.8297	0.7854	0.8144	0.8010	0.8556	0.9349	0.7885	0.9119	**0.9853**
	EOC-2	Acc.	0.9630	0.9698	0.9630	0.9772	0.9735	**0.9886**	0.9557	0.9831	0.9872
		Spec.	0.9947	0.9957	0.9947	0.9967	0.9962	**0.9984**	0.9937	0.9976	0.9982
		Pres.	0.9631	0.9702	0.9638	0.9774	0.9752	**0.9887**	0.9563	0.9834	0.9872
		Rec.	0.9630	0.9698	0.9630	0.9772	0.9735	**0.9886**	0.9557	0.9831	0.9872
		F1.	0.9627	0.9697	0.9629	0.9771	0.9736	**0.9886**	0.9557	0.9831	0.9872

TABLE 4. Confusion matrix results of the proposed model in OpenSARShip.

Data type	GRD/VV			GRD/VH			SLC/VV			SLC/VH		
Pred. / True	Bulk	Container	Tanker	Bulk	Container	Tanker	Bulk	Container	Tanker	Bulk	Container	Tanker
Bulk	942	82	30	905	119	30	297	49	15	307	44	10
Container	113	490	21	105	498	21	63	432	20	77	425	13
Tanker	21	10	185	11	16	189	8	8	89	12	1	92
Acc.	0.8701	0.8807	0.9567	0.8601	0.8622	0.9588	0.8624	0.8573	0.9480	0.8542	0.8624	0.9633
Spec,	0.8405	0.9276	0.9696	0.8619	0.8937	0.9696	0.8855	0.8777	0.9600	0.8565	0.9034	0.9737
Pres.	0.8755	0.8419	0.7839	0.8864	0.7867	0.7875	0.8071	0.8834	0.7177	0.7753	0.9043	0.8000
Rec.	0.8937	0.7853	0.8565	0.8586	0.7981	0.8750	0.8227	0.8388	0.8476	0.8504	0.8252	0.8762
F1	0.8845	0.8126	0.8186	0.8723	0.7924	0.8289	0.8148	0.8606	0.7773	0.8111	0.8629	0.8364

TABLE 5. Confusion matrix results of the proposed model in SOC scenario of MSTAR.

Pred. / True	2S1	BMP2	BRDM2	BTR70	BTR60	D7	T62	T72	ZIL131	ZSU23/4
2S1	273	0	0	0	0	0	1	0	0	0
BMP2	0	194	0	0	0	0	0	1	0	0
BRDM2	0	0	274	0	0	0	0	0	0	0
BTR70	0	0	0	196	0	0	0	0	0	0
BTR60	0	1	0	1	193	0	0	0	0	0
D7	0	0	0	0	0	273	0	0	1	0
T62	0	0	0	0	0	1	272	0	0	0
T72	0	0	0	0	0	0	0	196	0	0
ZIL131	0	0	0	0	0	0	0	0	274	0
ZSU23/4	0	0	0	0	0	0	0	0	0	274
Acc.	0.9996	0.9992	1.0000	0.9996	0.9992	0.9992	0.9992	0.9996	0.9996	1.0000
Spec.	1.0000	0.9996	1.0000	0.9996	1.0000	0.9995	0.9995	0.9996	0.9995	1.0000
Pres.	1.0000	0.9949	1.0000	0.9949	1.0000	0.9964	0.9963	0.9949	0.9964	1.0000
Rec.	0.9964	0.9949	1.0000	1.0000	0.9897	0.9964	0.9963	1.0000	1.0000	1.0000
F1	0.9982	0.9949	1.0000	0.9975	0.9948	0.9964	0.9963	0.9975	0.9982	1.0000

2) TRAINING DETAILS

All data augmentation algorithms had 0.5 probabilities. The augmentation process for random crops was utilized only in the OpenSARShip dataset, except in the MSTAR dataset. For the optimization process, cross-entropy loss [41] and AdamW optimizer were adopted [42]. The main parameters of the AdamW optimizer are the learning rate and the weight decay. Here, the decay was set to 10^{-5}, and the learning rate was changed using the cosine annealing scheduler-based learning rate, in which the range of the values is $[0, 0.001]$ and the maximum number of iterations is one epoch. The batch and maximum epoch sizes were 48 and 200. The model with

TABLE 6. Confusion matrix results of the proposed model in EOC-1 scenario of MSTAR.

True \ Pred.	2S1	BRDM2	T72	ZSU23/4
2S1	288	0	0	0
BRDM2	7	279	0	1
T72	0	0	287	1
ZSU23/4	8	0	0	280
Acc.	0.9870	0.9931	0.9991	0.9913
Spec.	0.9826	1.0000	1.0000	0.9977
Pres.	0.9505	1.0000	1.0000	0.9929
Rec.	1.0000	0.9721	0.9965	0.9722
F1	0.9746	0.9859	0.9983	0.9825

the best validation data (15% of training data) performance between 150 – 200 epoch was selected as the optimal model.

D. RESULTS AND DISCUSSION

For this study, we selected eight models for comparison: the early CNN models VGG [43] and ResNet [44], the latest CNN models ResNext [45] and ConvNext [29], the transformer models ViT [27] and Swin Transformer [28], and the hybrid models LambdaNet [46] and CoAtNet [47]. All systems, including the proposed system, were implemented using only tiny-size models (VGG11, ResNet26, ResNext26ts, ViT tiny, ConvNext tiny, SwinTransformer tiny, LambdaResNet26, and CoAtNet-0), and they utilized the same experimental environments as the proposed one. All of the models were analyzed by receiver operating characteristic (ROC). Thus, they were evaluated through the accuracy (Acc.), specificity (Spec.), precision (Prec.), recall (Rec.), and F1 score (F1.) [48]. In addition, a comparative analysis was conducted to evaluate the performance of each model in a limited training environment using training data sampled at a ratio of 5% to 100%. (Figure 3). Furthermore, the proposed model was analyzed precisely using the confusion matrix with ROC analysis. Also, the inference speed of the proposed system was compared and analyzed using frame-per-second (FPS) units. Finally, we compared the proposed system with other systems in the SOC scenario of the MSTAR dataset. Specifically, we compared the performance of our system with methods proposed in other papers using only accuracy (Figure 4).

1) ROC ANALYSIS IN THE OpenSARShip

The evaluation results in Table 3 show that the proposed model had the highest performance in the OpenSARship datasets. The model was improved by 3.57 – 11.49% compared to the lowest-performance models and by 0.90 – 5.01% compared to the highest-performance model. The models were selected using F1 scores and evaluated in terms of each metric. Moreover, the hybrid, latest CNNs, and transformer models performed remarkably in every dataset except for the proposed model. If the results were explained in detail, the best models in each dataset were as follows: ResNext26 in GRD/VV, CoAtNet-0 in GRD/VH, SwinTransformer tiny in SLC/VV, and ConvNext tiny in SLC/VH. The second-best models in each case were as follows: CoAtNet-0 in GRD/VV and GRD/VH, ConvNext tiny in SLC/VV, and SwinTransformer tiny in SLC/VH.

Furthermore, the proposed model was analyzed precisely using the confusion matrix with ROC analysis (Table 4). As seen in the F1 results, the proposed model shows remarkable classification performance of the bulk carrier of GRD and the container ship of SLC. Conversely, it was found that the model has a relatively lower classification accuracy in the container ship of GRD, the Tanker of SLC/VV, and the bulk carrier of SLC/VH.

2) ROC ANALYSIS IN THE MSTAR

The results of the comparative analysis for eight models in the proposed system in the MSTAR data are as follows (Table 3).

First, based on accuracy, the proposed system in the SOC scenario had an overall good performance of 0.9880–0.9979. The proposed model slightly decreased by less than 0.01 – 0.04% in each metric compared to the Swin transformer tiny model, which had the second-best performance within the system. It also showed performance improvements of 0.09 – 0.87% compared to the lowest performance model.

In the EOC-1 scenario, the proposed system had high performance and deviation (0.7941 – 0.9852). Compared to the Swin transformer tiny model, which had the best and second-best performance, the proposed model in this scenario showed performance improvements of 1.78 – 5.58% and 6.84 – 24.06% performance improvements compared to the lowest-performing model in each metric.

The proposed system in the EOC-2 scenario had a performance range of 0.9557 – 0.9886 based on accuracy, but there was some variation in performance. Like the SOC scenario, the proposed model had minor decreases of less than 0.02 – 0.15% in each metric compared to the best-performing model within that system. It also showed 0.45 – 3.30% performance improvements compared to the lowest-performing model.

Moreover, we analyzed the confusion matrix for each scenario in the proposed model-based system. In the SOC scenario, we found that BRDM2 and ZSU23/4 had 100% classification performance, while BMP2 had the lowest classification performance (Table 5). In the EOC-1 scenario, T72 had the best classification performance, while 2S1 had a relatively low classification performance (Table 6). Finally, in the EOC-2, we confirmed that the A07, A10, and A64 T72 tanks had 100% classification performance, while the classification performance for A32 was insufficient (Table 7).

3) ANALYSIS OF INFERENCE TIME

In this section, we have evaluated the speed at which each dataset can be processed on the proposed system. The performance has been measured in FPS to determine the system's ability to operate in real-time.

On the OpenSARShip dataset, we obtained the following results: 574.14 FPS (GRD/VV), 560.76 FPS (GRD/VH), 522.90 FPS (SLC/VV), and 473.45 FPS (SLC/VH). On the MSTAR dataset, the inference time results are as follows:

TABLE 7. Confusion matrix results of the proposed model in EOC-2 scenario of MSTAR.

Pred. True	A04	A05	A07	A10	A32	A62	A63	A64
A04	265	0	0	0	8	0	0	1
A05	0	273	0	0	0	1	0	0
A07	0	0	274	0	0	0	0	0
A10	0	0	0	271	0	0	0	0
A32	12	0	0	0	261	1	0	0
A62	0	1	0	0	1	271	0	1
A63	0	0	0	0	0	0	274	0
A64	0	0	0	0	0	2	0	272
Acc.	0.9904	0.9991	1.0000	1.0000	0.9900	0.9968	1.0000	0.9982
Spec.	0.9937	0.9995	1.0000	1.0000	0.9953	0.9979	1.0000	0.9990
Pres.	0.9567	0.9964	1.0000	1.0000	0.9667	0.9855	1.0000	0.9927
Rec.	0.9672	0.9964	1.0000	1.0000	0.9526	0.9891	1.0000	0.9927
F1	0.9619	0.9964	1.0000	1.0000	0.9596	0.9872	1.0000	0.9927

556.65 FPS (SOC), 567.03 FPS (EOC-1), and 566.91 FPS (EOC-2).

Our experiments have shown that the proposed system can easily exceed the real-time processing threshold of 30 FPS in all modes and datasets. The lowest recorded FPS was 473.45, indicating that the system is robust enough to handle even the most demanding datasets in real-time.

4) COMPARATIVE ANALYSIS OF MODEL PERFORMANCES ACROSS DIFFERENT SAMPLING RATIO

In this session, we compared and analyzed the performance of eight different models in a limited training environment (Figure 3). Each model was trained on a proposed system, and the training data was randomly sampled from 5% to 100% (5% interval) per label in 1,014 provided training data samples. The validation data set was fixed at 150 samples.

The analysis revealed that the accuracy of all models surpassed 80% when the training data was sampled at 70%. Notably, the proposed model demonstrated superior performance, even in situations with slightly less data, indicating that the proposed system can ensure a certain level of performance. When compared to other models, it was evident that the proposed model exhibited the highest performance. However, it was observed that when the training data was sampled at less than 35%, all models, including the proposed model, struggled to learn effectively, highlighting the challenges of training with very tiny datasets.

5) COMPARATIVE ANALYSIS OF SYSTEM PERFORMANCES

A comparative analysis was conducted to evaluate the effectiveness of different system architectures for SAR target recognition in the SOC scenario of the MSTAR dataset (Figure 4). In this proposed system environment, we compared a proposed hybrid model (Hybrid) and a swin transformer tiny (SwinT) against several established methods, including ConvT, AP-CNN, MSA-CNN, CNN-LSTM, the Knowledge Integration framework for Domain Adaptive SAR target recognition (KIDA), and the CNN embeddings and metric learning (CNN E&ML) approach.

In the SOC scenario of the MSTAR dataset, our comparative analysis showed that Hybrid and SwinT achieved remarkable accuracies (99.75% and 99.79%, respectively), matching the performance of the CNN E&ML model. This demonstrates the effectiveness of advanced models like hybrids and transformers in SAR image analysis, suggesting their potential for enhanced accuracy in real-world applications.

IV. CONCLUSION

This study proposes a DL-based system for automatically recognizing targets in SAR images. The model designed in this system uses transfer learning and hybrid conv-attention networks that fuse the partial model of pre-trained ConvNext and the SA structure of the transformers to utilize the local and global patterns in the image for this task. Furthermore, pre-processing and training processes were designed considering transfer learning utilization and input data characteristics. As indicated by the experimental results, the proposed hybrid model had outstanding performances in all data environments, even compared to the pre-trained models. Therefore, the proposed system can be used as an element technology in an automated surveillance system using satellites or unmanned aircraft, even if only small amounts of data are used.

Regardless, in the OpenSARShip dataset, since the proposed system used only 3 labels (bulk carrier, container ship, and tanker) among the 17 labels due to the limitation of data composition, additional research on system development that is robust to more types of labels should be conducted for system research that can perform sufficiently in the real world environment. In addition, leveraging knowledge distillation methods [49] to efficiently reduce models' size or exploring self-supervised learning [50] that are robust in small-scale dataset environments could lead to a more optimized system design. Furthermore, the utilization of generative AI models such as variational autoencoders [51], generative adversarial networks [52], and diffusion models [53] can effectively augment training data, thereby enhancing the performance of classification models. Given the reality of minimal training data in some practical settings, pursuing research based on few-shot learning methods [54] is imperative. These approaches can address the challenges of scarce data availability by enabling models to perform well even when presented with few or no labeled examples.

ACKNOWLEDGMENT

(Jiseok Yoon and Jeongheon Song are co-first authors.)

REFERENCES

[1] L. Ferro-Famil and E. Pottier, "Synthetic aperture radar imaging," in *Microwave Remote Sensing of Land Surface*, N. Baghdadi and M. Zribi, Eds. Amsterdam, The Netherlands: Elsevier, 2016, ch. 1, pp. 1–65. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9781785481598500013

[2] A. Passah, S. N. Sur, B. Paul, and D. Kandar, "SAR image classification: A comprehensive study and analysis," *IEEE Access*, vol. 10, pp. 20385–20399, 2022.

[3] Y. Zhang, X. Guo, L. Li, and N. Ansari, "Deep knowledge integration of heterogeneous features for domain adaptive SAR target recognition," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108590. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322000711

[4] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[5] X. X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler, "Deep learning meets SAR: Concepts, models, pitfalls, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 143–172, Dec. 2021.

[6] Z. Sun, M. Dai, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, 2021.

[7] A. H. Oveis, E. Giusti, S. Ghio, and M. Martorella, "A survey on the applications of convolutional neural networks for synthetic aperture radar: Recent advances," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 5, pp. 18–42, May 2022.

[8] W. Shi, Z. Hu, H. Liu, S. Cen, J. Huang, and X. Chen, "Ship detection in SAR images based on adjacent context guide fusion module and dense weighted skip connection," *IEEE Access*, vol. 10, pp. 134263–134276, 2022.

[9] J. Li, J. Chen, P. Cheng, Z. Yu, L. Yu, and C. Chi, "A survey on deep-learning-based real-time SAR ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3218–3247, 2023.

[10] S. Chen, H. Wang, F. Xu, and Y.-Q. 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] Z. Jianxiong, S. Zhiguang, C. Xiao, and F. Qiang, "Automatic target recognition of SAR images based on global scattering center model," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3713–3729, Oct. 2011.

[12] H. Liu and S. Li, "Decision fusion of sparse representation and support vector machine for SAR image target recognition," *Neurocomputing*, vol. 113, pp. 97–104, Aug. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231213002178

[13] D. Zhang, J. Liu, W. Heng, K. Ren, and J. Song, "Transfer learning with convolutional neural networks for SAR ship recognition," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 322, Mar. 2018, Art. no. 072001, doi: 10.1088/1757-899x/322/7/072001.

[14] M. Al Mufti, E. Al Hadhrami, B. Taha, and N. Werghi, "SAR automatic target recognition using transfer learning approach," in *Proc. Int. Conf. Intell. Auto. Syst. (ICoIAS)*, Mar. 2018, pp. 1–4.

[15] J. Deng, H. Bi, J. Zhang, Z. Liu, and L. Yu, "Amplitude-phase CNN-based SAR target classification via complex-valued sparse image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5214–5221, 2022.

[16] Y. Li, X. Li, Q. Sun, and Q. Dong, "SAR image classification using CNN embeddings and metric learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[17] C. Wang, X. Liu, J. Pei, Y. Huang, Y. Zhang, and J. Yang, "Multiview attention CNN-LSTM network for SAR automatic target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12504–12513, 2021.

[18] T. Zhang, X. Zhang, X. Ke, C. Liu, X. Xu, X. Zhan, C. Wang, I. Ahmad, Y. Zhou, D. Pan, J. Li, H. Su, J. Shi, and S. Wei, "HOG-ShipCLSNet: A novel deep learning network with HOG feature fusion for SAR ship classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5210322.

[19] C. Wang, Y. Huang, X. Liu, J. Pei, Y. Zhang, and J. Yang, "Global in local: A convolutional transformer for SAR ATR FSL," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

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

[21] 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.

[22] S. Feng, K. Ji, F. Wang, L. Zhang, X. Ma, and G. Kuang, "PAN—Part attention network integrating electromagnetic characteristics for interpretable SAR vehicle target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5204617.

[23] 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.

[24] Z. Wen, Y. Yu, and Q. Wu, "Multimodal discriminative feature learning for SAR ATR: A fusion framework of phase history, scattering topology, and image," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5200414.

[25] X. Zhang, S. Feng, C. Zhao, Z. Sun, S. Zhang, and K. Ji, "MGSFA-Net: Multiscale global scattering feature association network for SAR ship target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4611–4625, 2024.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[30] L. Huang, B. Liu, B. Li, W. Guo, W. Yu, Z. Zhang, and W. Yu, "OpenSARShip: A dataset dedicated to Sentinel-1 ship interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 195–208, Jan. 2018.

[31] B. Li, B. Liu, L. Huang, W. Guo, Z. Zhang, and W. Yu, "OpenSARShip 2.0: A large-volume dataset for deeper interpretation of ship targets in Sentinel-1 imagery," in *Proc. SAR Big Data Era: Models, Methods Appl. (BIGSARDATA)*, Nov. 2017, pp. 1–5.

[32] Sandia National Laboratory. (2005). *Moving and Stationary Target Acquisition and Recognition (MSTAR) Dataset*. Accessed: Jan. 1, 2024. [Online]. Available: https://www.sdms.afrl.af.mil/index.php?collection=mstar

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[34] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016, *arXiv:1603.07285*.

[35] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[36] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.

[37] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[38] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. NeurIPS*, 2020, pp. 7537–7547.

[39] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: An open source differentiable computer vision library for PyTorch," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3674–3683.

[40] R. Wightman. (2019). *PyTorch Image Models*. [Online]. Available: https://github.com/rwightman/pytorch-image-models

[41] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham, "Uses and abuses of the cross-entropy loss: Case studies in modern deep learning," 2020, *arXiv:2011.05231*.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016, *arXiv:1611.05431*.

[46] J. Wei, M. Goyal, G. Durrett, and I. Dillig, "LambdaNet: Probabilistic type inference using graph neural networks," 2020, *arXiv:2005.02161*.

[47] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," 2021, *arXiv:2106.04803*.

[48] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," 2020, *arXiv:2008.05756*.

[49] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.

[50] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," 2022, *arXiv:2208.00173*.

[51] C. Chadebec and S. Allassonnière, "Data augmentation with variational autoencoders and manifold sampling," in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*. Cham, Switzerland: Springer, 2021, pp. 184–192.

[52] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," 2017, *arXiv:1711.04340*.

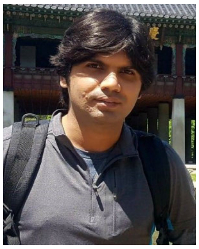
[53] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," 2023, *arXiv:2302.07944*.

[54] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, no. 13, pp. 1–40, Jul. 2023, doi: 10.1145/3582688.

JISEOK YOON (Member, IEEE) received the B.E. degree in electronics engineering from the Kumoh National Institute of Technology, Gumi, South Korea, in 2012, and the M.S. and Ph.D. degrees in mechatronics from Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2014 and 2021. From 2021 to 2023, he was a Postdoctoral Researcher with the Vision and Image Processing Laboratory, Tech University of Korea. Since 2023, he has been the CTO of IKLAB Inc. His research interests include signal and image processing with machine learning, emphasizing image-to-image translation, image classification, physical layer authentication of wireless communication, and energy prediction.

JEONGHEON SONG (Member, IEEE) received the B.S. degree from the Department of Social and Environmental Systems Engineering, Yonsei University, South Korea, in 2003, and the M.S. degree from the Department of Civil and Environmental Engineering, Seoul National University, South Korea, in 2009. Since 2003, he has been a Senior Researcher with Korea Aerospace Research Institute. His research interest includes satellite remote sensing.

TANVEER HUSSAIN (Member, IEEE) received the Ph.D. degree in software convergence from Sejong University, Seoul, South Korea. He is currently a Lecturer with the Department of Computer Science, Edge Hill University. Previously, he was a Postdoctoral Research Fellow with the Institute for Transport Studies, University of Leeds, Leeds, U.K. He has filed/published several patents and papers in peer-reviewed journals and conferences in reputed venues, including CVPRW, IEEE Transactions on Image Processing, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Industrial Informatics, *Pattern Recognition* (Elsevier), *Neurocomputing*, *Pattern Recognition Letters*, *ACM Computing Surveys*, and *Multimedia Tools and Applications* (Springer). His major research interests include multimedia data analysis, including video summarization, action and activity recognition, saliency detection, scene understanding for autonomous driving, resource-constrained programming, and fire/smoke detection and segmentation. He is providing professional review services in various reputed journals. He is serving as an Editorial Board Member for the *Journal of Artificial Intelligence and Systems* and a Review Editor for *Frontiers in Artificial Intelligence*. For further activities and implementations of his research, visit: https://github.com/tanveer-hussain.



SUNDER ALI KHOWAJA (Member, IEEE) is currently an Associate Professor with the Department of Telecommunication Engineering, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan. He received a Postdoctoral Fellowship in industrial computer vision with the Tech University of Korea. He has more than 50 papers in international and national publication venues. He has an academic experience of more than 12 years along with three years of industrial experience in the capacity of Network and RF Engineer. His research interests include computer vision, deep learning, privacy preservation machine learning, and data analytics. He is also a two-time Runner-Up Winner in UG2+ Challenges held in conjunction with IEEE CVPR, from 2022 to 2023. He has also been included in the Top 10 teams for various competitions held at CVPR, from 2022 to 2023. He is serving as an Associate Editor for *PLoS One* journal. He has been serving as the Guest Editor for top-tier journals, including IEEE Transactions on Vehicular Technology, IEEE Transactions on Consumer Electronics, IEEE Transactions on Green Communications and Networking, *Journal of King Saud University-Computer and Information Sciences*, *Sustainable Energy Technologies and Assessment*, and *Computers and Electrical Engineering*.

KHAN MUHAMMAD (Senior Member, IEEE) received the Ph.D. degree in digital content from Sejong University, Seoul, Republic of Korea, in February 2019. From March 2019 to February 2022, he was an Assistant Professor with the Department of Software, Sejong University. He is currently the Director of the Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Seoul, and an Assistant Professor (Tenure-Track) with the Department of Applied AI, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul. He has registered ten patents and contributed 240 papers in peer-reviewed journals and conference proceedings in his research areas. His research interests include intelligent video surveillance, medical image analysis, information security, video summarization, multimedia data analysis, computer vision, the IoT/IoMT, and smart cities. He is among the highly cited researchers according to the Web of Science (Clarivate), in 2021, 2022, and 2023. He is an associate editor/editorial board member of more than 14 journals.

IK HYUN LEE (Member, IEEE) received the B.S. degree in control and instrument engineering from Korea University, South Korea, in 2004, and the M.S. and Ph.D. degrees from the School of Information and Mechatronics, Gwangju Institute of Science and Technology, South Korea, in 2008 and 2013, respectively. He was a Postdoctoral Researcher with the Media Laboratory, Massachusetts Institute of Technology, and later as a Senior Researcher with Korea Aerospace Institute of Research. He is currently an Associate Professor with the Department of Mechatronics Engineering, Tech University of Korea, South Korea. His research interests include image registration, fusion, object tracking, aiming, and medical image processing.