IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-2010-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IGARSS52108.2023.10283157

USING SQUEEZE-AND-EXCITATION VISION TRANSFORMER WITH LOCAL FEATURE FUSION FOR SHIP CLASSIFICATION IN SAR IMAGES

Yuhang Qi[1], *Lu Wang*[1,4,*], *Chunhui Zhao*[1,4], *Ning Wang*[2] *and Jikang Chen*[2,3]

[1]College of Information and Communication Engineering, Harbin Engineering University, Harbin, China
[2]College of Shipbuilding Engineering, Harbin Engineering University, Harbin, China
[3]Frontiers Science Center for Wave Field of Extreme Ocean Environment, Harbin, China
[4]Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin, China

ABSTRACT

The categorization of synthetic aperture radar (SAR) ships primarily focuses on large ships with distinct features, but accurately identifying SAR ships remains challenging due to limited samples in certain ship categories. In this study, we propose a compressed and excited Vision Transformer model based on local feature fusion. This model leverages local feature fusion and channel modeling through the squeezing-and-excitation (SE) mechanism to effectively balance the contributions of each feature. By incorporating better local information, we are able to extract deeper features even from small datasets. To evaluate the efficacy of our model, we trained it on the three-category OpenSARShip 2.0 dataset and conducted experiments. The results demonstrate that our proposed model achieves superior classification accuracy compared to existing methods.

Index Terms— Synthetic aperture radar, Deep learning, SAR ship classification

1. INTRODUCTION

A large-scale picture can be observed without being impacted by the weather, thanks to synthetic aperture radar (SAR), which possesses all-weather working capabilities. Consequently, SAR is indispensable for long-term, ongoing, real-time monitoring of marine areas. However, classifying ships in SAR photos still presents considerable difficulties. Firstly, due to the distinctive properties of SAR images, conventional algorithms perform poorly during sample training, as they fail to simulate crucial local structures such as edges and lines between adjacent pixels. Secondly, when the SAR ship classification dataset is insufficient, the characteristics of training samples are relatively sparse, thereby hindering the model from achieving high classification accuracy.

While some research has been conducted on ship categorization in SAR photos, Wang *et al.* [1] present a novel data enhancement method combined with transfer learning to address the issue of overfitting, which often arises when training with limited labeled data sets. Xiong *et al.* [2] propose a dual-polarized SAR ship target recognition method based on a feature and loss fusion depth network. This method effectively eliminates the impact of data imbalance and background noise on recognition accuracy, thus enhancing the generalization ability of deep learning networks in general. Huang *et al.* [3] propose a group-based feature extraction method for sparse connection convolution networks, achieving high-precision network classification by utilizing fewer parameters through dynamic channel feature recalibration. However, rather than exploring deeper aspects of the restricted data itself, these technologies primarily extract typical ship features through various data augmentation or data combination methods.

This paper introduces a novel automatic classification method for SAR ship images based on a vision transformer (VIT). We propose the utilization of local feature fusion technology in SAR ship categorization to enhance the incorporation of local information. Firstly, employing soft split, the image is divided into multiple overlapping local images, from which relevant feature and location data are extracted. Secondly, nearby feature vectors are combined into a single vector by incorporating local prior information. As a result, our method not only captures the local structure information of neighboring feature vectors but also improves the classification accuracy of SAR images by gradually reducing the number of feature vectors. Furthermore, we implement the squeezing-and-excitation (SE) mechanism after the multi-head attention mechanism of the backbone network to further enhance the classification accuracy.

The study makes three significant contributions: Firstly, the utilization of soft split and feature fusion modules enables the recovery of local structural information from nearby feature vectors. By progressively reducing the number of

This work was supported in part by the Fundamental Research Funds for the Central Universities, the National Natural Science Foundation of China under Grant 61971153, and Natural Science Foundation of Heilongjiang Province(YQ2022E016). *Corresponding author: Lu Wang.

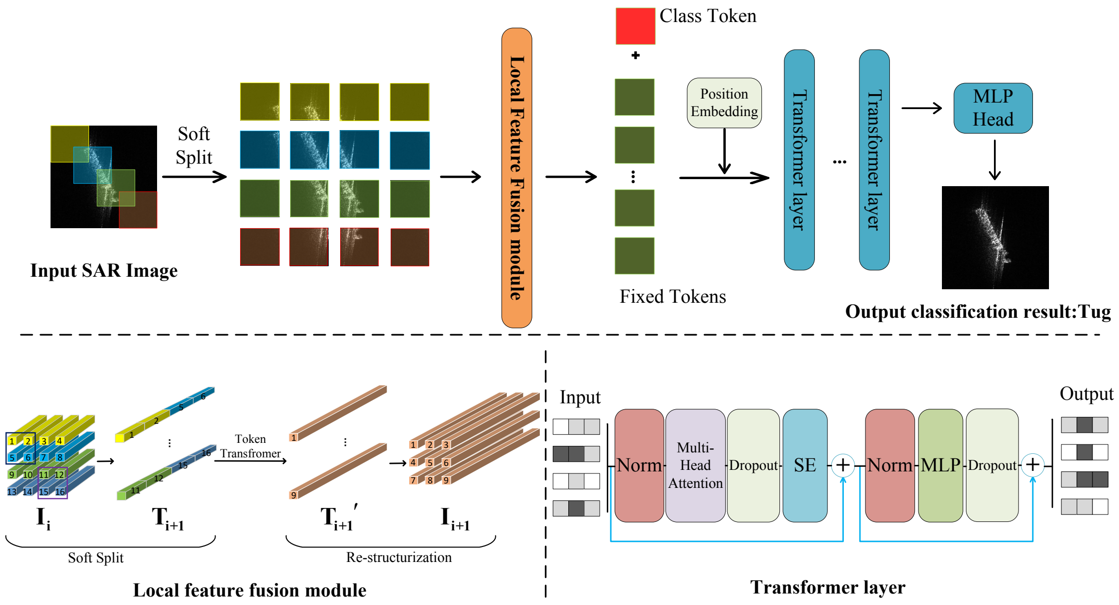

Fig. 1: The structure diagram of the proposed method.

feature vectors, the accuracy of SAR image classification is enhanced. Secondly, an SE attention mechanism is implemented in the backbone network following the multi-head attention mechanism. This mechanism adaptively determines the weight of each feature mapping, emphasizing significant features and further improving classification accuracy. Thirdly, this study pioneers the use of ViT as the feature extractor for remote sensing ship image classification, instead of the traditional CNN approach. The proposed method has been successfully applied to remote sensing ship datasets and has demonstrated significantly better performance compared to CNN-based methods.

2. PROPOSED METHOD

The structure diagram of the proposed method is illustrated in Fig. 1, and enhancements to the original method can be categorized into three components: image soft split, local feature fusion module, and SE module. Firstly, the input undergoes a soft split, where tokens are extracted from the patch. These tokens are then fed into the local feature fusion module to fuse the local structural information of neighboring feature vectors. Subsequently, the processed input, after incorporating the SE module, is passed through the Transformer Encoder. Finally, the resulting output is then sent to the MLP Head for the classification of the input image.

2.1. Image soft split

To incorporate local structural information, we employ soft split after acquiring the input image. More specifically, we split the image into blocks with overlapping sections to enhance the correlation during token generation. As a result, each image block shares the same portion as the surrounding blocks, establishing a prior that emphasizes stronger correlation among the subsequently generated tokens.

In the subsequent steps of local feature fusion, we apply soft split to the reconstructed image $\mathbf{I}$ to capture local structural information. By splitting it into overlapping patches, we establish a prior that emphasizes stronger correlation among the surrounding tokens. The tokens within each segmented patch are connected as a single token, enabling the aggregation of local information from neighboring pixels and patches.

During the soft split process, each patch is of size $k \times k$, with s overlaps and p fills on the image. The parameter $k-s$ is comparable to the step size used in convolutional operations. Hence, for the reconstructed image $\mathbf{I} \in \mathbb{R}^{h\times w\times c}$, the length of the output marker $\mathbf{T_0}$ after soft split is denoted as l_0,

$$l_0 = \left[\frac{h+2p-k}{k-s}+1\right] \times \left[\frac{w+2p-k}{k-s}+1\right]. \tag{1}$$

The size of each split patch is $k \times k \times c$. We flatten all patches on the spatial dimension to tokens $\mathbf{T_0} \in R^{l_o \times ck^2}$. The output token is fed for the subsequent local feature fusion model after soft split.

2.2. Local feature fusion model

The local feature fusion module integrates and combines the feature vectors within the local receptive field through soft split. Soft split and reconstruction are the initial two phases in each feature fusion module. Detailed illustration of the feature fusion module can be found in Fig. 2.

For input image $\mathbf{I_i}$, we convert it into tokens through soft split:

$$\mathbf{T_{i+1}} = \mathrm{SS}(\mathbf{I_i}). \tag{2}$$

After, the classic transformer encoder transformation generates $\mathbf{T_{i+1}}'$:

$$\mathbf{T_{i+1}}' = \mathrm{MLP}(\mathrm{MSA}(\mathbf{T_{i+1}})), \tag{3}$$

where MSA is the layer normalized multi head self attention operation, and MLP is the layer normalized multilayer

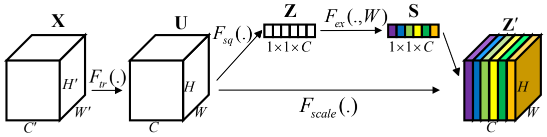

Fig. 2: The structure diagram of SE model.

perceptron in the standard transformer. Then, reshape these symbols into images $\mathbf{I_{i+1}}$ in the spatial dimension:

$$\mathbf{I_{i+1}} = \mathrm{Reshape}(\mathbf{T_{i+1}}'), \tag{4}$$

where $\mathrm{Reshape}$ reorganizes tokens $\mathbf{T_{i+1}}' \in R^{l\times c}$ into $\mathbf{I_{i+1}} \in R^{h\times w\times c}$, where l is the length of $\mathbf{T_{i+1}}'$, h, w and c are height, width and channel, and $l = h \times c$.

2.3. SE model

The multi-head attention is a crucial component of the Vision Transformer. By utilizing the multi-head attention mechanism, the model can comprehend the relationships between various points in the input sequence, capturing the internal structure and semantic information more accurately. Following the multi-head attention, the SE module is implemented to emphasize key aspects. The structure diagram of the SE model is illustrated in Fig. 2.

To compress global spatial information into the channel descriptor vector $Z \in \mathbb{R}^C$, the squeezing operation generates channel-level information through global average pooling denoted as $F_{sq}(.)$. Each element of the channel descriptor vector Z, which represents the global features of each channel in U, can be seen as a collection of local features. Specifically, we denote $Z = [z_1, z_2, \cdots, z_c]$, which is obtained by compressing the feature maps $U = [u_1, u_2, \cdots, u_c]$. The calculation of the c-th element of Z takes into account the spatial dimensions $H \times W$ of U and can be expressed as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j). \tag{5}$$

After compressing the information, excitation is employed to capture the inter-channel relationships in U comprehensively. Each element of the descriptor vector Z represents the global feature of the corresponding channel in U. To model the nonlinear correlation between each individual element of Z, two fully connected layers are introduced as the mapping function $F_{ex}(.)$. These parameters are then activated by the sigmoid activation function to determine the channel weight at the pixel level of U. The excitation equation can be expressed as follows:

$$S = F_{ex}(Z, W) = \sigma(g(Z, V)) = \sigma(V_2\delta(V_1 Z)), \tag{6}$$

where σ is the sigmoid function; δ is the rectified linear unit activation function; $V_1 \in R^{C/R\times C}$ and $V_2 \in R^{C/R\times C}$ represent the weight matrices of the full-connectivity layer; and C/R is the reducing dimension gravity of the full-connectivity layer. The model's attention to every channel of the feature maps U is represented by each element of $S \in R^C$, whose values range from 0 to 1.

The final output of the SE block obtained by rescaling U with the activation S is

$$x_c{}' = F_{scale}(u_c, s_c) = u_c s_c, \tag{7}$$

where $X' = [x_1{}', x_2{}', \cdots, x_c{}']$ and $F_{scale}(u_c, s_c)$ are the per-channel multiplication between the indicator quantity s_c and the feature map $u_c \in R^{H\times W}$. Clearly, the output X' of the SE block is obtained by adjusting the channel weights on U. During the task learning process, the channel weight associated with the traffic status is increased, thereby enhancing the expressive power of the features.

3. EXPERIMENT AND ANALYSIS

3.1. Datasets

To demonstrate the effectiveness of the proposed method, the OpenSARShip 2.0 dataset [4] is chosen as the training dataset. The OpenSARShip 2.0 dataset consists of approximately 40,000 SAR ship images extracted from 87 Sentinel-1 SAR images. It encompasses 14 categories of ship images, including cargo ships, cruise ships, passenger ships, law enforcement vessels, and fishing vessels. Considering the issue of imbalanced sample data within the dataset, we specifically selected three ship types as our classification targets: cargo ships, fishing vessels, and tugboats. From each category, we randomly selected 320 samples, resulting in a total of 960 samples for training and testing. The dataset was split into a 4:1 ratio, with 80

3.2. Evaluation indicators

Similar to most target classification models, we selected Accuracy (%) as the primary index to measure the classification accuracy. It is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. Additionally, this paper also utilizes Recall (%), Precision (%), and F_1-score (%) to evaluate the classification performance of the proposed model.

3.3. Experimental results and analysis

The results are presented in Table 1, where the achieved accuracy (Acc) is 76.25%, surpassing other methods. A comparison with the suboptimal algorithm, mini Hourglass Net [2], reveals an improvement of 0.81% in Acc by the proposed method. Additionally, the proposed method exhibits notable enhancements in other secondary indicators such as recall,

Table 1: Three-Category Results on OpenSarShip 2.0 Dataset. Results are represented as mean ± standard deviation. To clearly visualize, the highest-scoring item in each column is indicated in bold.

Model	Recall (%)	Precision (%)	F_1 (%)	Acc (%)
LeNet [5]	65.15 ± 1.12	60.54 ± 2.47	62.73 ± 1.52	65.74 ± 1.50
AlexNet [6]	68.51 ± 3.04	65.52 ± 1.23	66.94 ± 1.51	70.22 ± 0.68
VGG [7]	68.61 ± 5.33	64.63 ± 2.52	66.53 ± 3.77	70.05 ± 1.35
GoogLeNet [8]	69.73 ± 2.70	68.80 ± 1.81	69.21 ± 1.19	73.80 ± 1.32
ResNet [9]	71.67 ± 1.71	66.79 ± 1.27	69.13 ± 1.04	72.82 ± 0.75
MobileNet [10]	67.23 ± 1.59	61.85 ± 1.69	64.42 ± 1.41	66.71 ± 0.87
SqueezeNet [11]	67.42 ± 4.67	65.67 ± 1.87	66.45 ± 2.61	70.89 ± 1.11
DenseNet [12]	71.40 ± 1.80	68.83 ± 1.50	70.07 ± 1.00	74.31 ± 0.76
Inception [13]	69.26 ± 3.16	67.43 ± 2.39	68.28 ± 1.97	72.44 ± 0.70
Xception [14]	71.56 ± 3.00	68.60 ± 1.67	70.00 ± 1.29	73.74 ± 0.86
VGG16-FT [1]	57.72 ± 1.37	58.72 ± 4.76	58.12 ± 2.67	69.27 ± 0.27
mini Hourglass Net [2]	73.87 ± 1.16	71.50 ± 3.00	72.67 ± 2.04	75.44 ± 2.68
GSESCNNs [3]	74.74 ± 1.60	69.56 ± 2.38	72.04 ± 1.60	74.98 ± 1.46
Proposed method	**75.36 ± 2.03**	**72.61 ± 2.63**	**73.96 ± 2.13**	**76.25 ± 1.36**

Table 2: Classification Confusion Matrix of the Proposed Method on the Three-Category Dataset.

Predicted True	Cargo	Fishing	Tug
Cargo	54	4	6
Fishing	9	43	12
Tug	6	11	47

precision, and F_1, reaching optimal performance compared to other algorithms. These improvements can be attributed to the incorporation of soft split and local feature fusion based on VIT, along with the addition of the SE attention mechanism, enabling better extraction of deep features from the limited dataset.

The classification confusion matrix for the method in the Cargo, Fishing, and Tug categories is presented in Table 2. It can be observed that the recall for Cargo reaches 84.37%, while the recall for Fishing and Tug reaches 67.19% and 73.43%, respectively. The proposed method demonstrates the ability to accurately classify the majority of ships, with diagonal values mostly higher than those of other ships within the same row. These results indicate that the proposed method effectively utilizes the limited SAR ship data, extracting informative features and achieving superior accuracy (Acc) compared to other methods in SAR ship classification.

4. CONCLUSION

This paper presents an enhanced Vision Transformer model for SAR ship recognition tasks, incorporating a local feature fusion module. The proposed method leverages the local feature fusion module to incorporate local prior information, merging neighboring feature vectors into a single feature vector, thereby extracting deep features. Furthermore, it utilizes the SE module to adjust channel weights and emphasize crucial features. The results of three classification experiments conducted on the OpenSARShip 2.0 dataset demonstrate the superiority of the proposed method over other approaches in ship classification.

5. REFERENCES

[1] Y. Wang, C. Wang, and H. Zhang, “Ship Classification in High-Resolution SAR Images Using Deep Learning of Small Datasets,” *Sensors*, vol. 18, no. 9, pp. 2929–2944, Sep. 2018.

[2] G. Xiong, Y. Xi, D. Chen, and W. Yu, “Dual-Polarization SAR Ship Target Recognition Based on Mini Hourglass Region Extraction and Dual-Channel Efficient Fusion Network,” *IEEE Access*, vol. 9, pp. 29078–29089, Feb. 2021.

[3] G. Huang, X. Liu, J. Hui, Z. Wang, and Z. Zhang, “A novel group squeeze excitation sparsely connected convolutional networks for SAR target classification,” *International Journal Of Remote Sensing*, vol. 40, no. 11, pp. 4346–4360, Jun. 2019.

[4] B. Li *et al.*, “OpenSARShip 2.0: A large-volume dataset for deeper interpretation of ship targets in Sentinel-1 imagery,” in *Proc. 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*, Dec. 2017, pp. 1–5.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Jan. 2012, pp. 1097–1105.

[7] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *Proc. International Conference on Learning Representations (ICLR)*, May. 2015, pp. 1–14.

[8] C. Szegedy, W. Liu, and Y. Jia, “Going Deeper with Convolutions,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1–9.

[9] K. He, X. Zhang, S. Ren, and J. Sun,, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[10] A. Howard *et al.*, “Searching for MobileNetV3,” in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 1314–1324.

[11] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size,” in *Proc. International Conference on Learning Representations (ICLR)*, Apr. 2017, pp. 1-9.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261-2269.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818-2826.

[14] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1800-1807.