Ship Classification Using Swin Transformer for Surveillance on Shore

Jixiang Liu[1], Wenli Sun[1(✉)], and Xu Gao[2]

[1] Navigation College, Dalian Maritime University, Dalian 116026, Liaoning, China
itslab@dlmu.edu.cn

[2] National Engineering Research Center of Maritime Navigation System, Dalian Maritime University, Dalian 116026, Liaoning, China

Abstract. Ship image classification technology is one of the core technologies for intelligent maritime surveillance system. It is fundamental that ships and their types are accurately identified for analysing and understanding in maritime scenes. Recently, the transformer-based model successfully applied in the field of natural language processing, and they have surpassed convolutional neural networks in image classification tasks, with Swin Transformer as the leader. Swin Transformer builds a hierarchical pyramid structure and a shifted window scheme on the basis of multi-head self-attention mechanism. These qualities reduce the complexity of models, and makes it as a general backbone for computer vision. In this study, we use the well-known ship image dataset called Seaships to investigate the effectiveness of Swin Transformer. We find that its hierarchical pyramid structure, multi-head self-attention mechanism and shifted window scheme play a key role in ship image classification. The results show that Swin Transformer achieves an accuracy of 93.5% in ship image classification, and outperforms typical convolutional networks and Vision Transformer.

Keywords: Swin Transformer · Ship surveillance · Image classification · Attention mechanism · Deep learning

1 Introduction

The rapid development of artificial intelligence (AI) has led to the rise of ship industry. However, most maritime surveillance systems still adopt the traditional manual mode. People can not focus on ship targets consistently and effectively [1, 2]. Therefore, it is imperative to improve the maritime surveillance system intelligently and to upgrade technology. The use of AI method to analyze and process ship targets not only reduces staff labor intensity, but also improves the accuracy. Therefore, AI plays an important role in maritime traffic and illegal smuggling surveillance.

Since it is essential to identify ships and their types correctly as a prerequisite at sea, the problem of classifying ships should be solved firstly [2]. In the field of computer vision, it is also necessary to identify the class of objects, in order to analyze and understand the scene. Yet there are many difficulties in image classification at different aspects. Significant changes in the appearance of objects due to scale, lighting,

Y. S. Shmaliy and A. Abdelnaby Zekry (Eds.): CCIE 2021, LNEE 920, pp. 774–785, 2022.
https://doi.org/10.1007/978-981-19-3927-3_76

perspective, deformation and occlusion at instance aspect. Presence of large intra-class differences, inter-class ambiguity and background interference at category aspect. Multiple stability at semantic aspect. In addition, massive data and image categories also increase the difficulty of image classification [3].

In deep learning, the approaches for image classification include Convolutional Neural Network (CNN)-based and transformer-based. It is difficult to capture long distance dependencies of the image by using CNN-based approach due to the fixed size window when extracting image features. The transformer-based approach utilizes an attention mechanism that assigns higher weights to features in key regions in global image, making it capable to integrate global information [4]. Therefore, compared with CNN-based model, transformer-based model has achieved better performance in many datasets, such as ImageNet, COCO and ADE 20 k, within image classification tasks [5, 6].

Among transformer-based models, Swin Transformer, improved by Vision Transformer, has the best performance. Swin Transformer adopts a shifted window scheme that restricts self-attention calculation to non-overlapping local windows, as well as allows cross-window connections, making more efficiently. In addition, the hierarchical structure adopted by Swin Transformer not only makes the model have more flexibility for modeling at different scales, but also makes the computational complexity linearly related to the image size, so it can be used as backbone for general computer vision. Since Swin Transformer is compatible with a wide range of vision tasks and outperforms previous state-of-the-art models on image classification [7], we apply it to the field of ship image classification.

The main contributions of this study are as follows.

(1) In ship image classification task, we adopt Swin Transformer to achieve a classification accuracy of 93.5% on the well-known ship image dataset called Seaships, outperforming typical convolutional neural network.
(2) We study the effect of the pyramid structure in Swin Transformer, and experiments show that it can improve the classification accuracy by 27.4%.
(3) To investigate the relationship between attention mechanism of Swin Transformer and its capability of processing global information in images, the experiments show that shifted window scheme contributes to the performance.

The rest of this paper is organized as follows. The Sect. 2 reviews the previous research on ship image classification and Swin Transformer. Section 3 gives a methodology for applying Swin Transformer to process ship images. Section 4 presents the dataset, evaluation metrics and hyperparameters as well as analysis the results. Section 5 concludes the paper.

2 Related Work

2.1 Ship Image Classification

When CNNs were the dominant deep learning models, many studies applied them to the field of ship image classification. Leclerc et al. conducted a CNN training on the maritime ship classification dataset called Marvel and obtained a significant improvement

compared to state-of-the-art results at that time [8]. An improved VGG-19 was proposed by Xu et al. to solve the problem of over-pooling for small targets. This model improved the small ships recognition capability [9]. CNN and various data argumentation such as horizontal flipping, clipping, scaling, rotation and changing RGB channel values were applied to ship classification by Milicevic et al., and improved the accuracy by 6.5% [10].

In the task of image classification, the transformer-based model had more advantages than the CNN-based model. Firstly, Transformer adopted an attention mechanism with attention distance extending as the network depth increases. It was better than the perceptual field in CNN [11]. Secondly, efficient calculations were adopted by Transformer when modeling long-distance interactions, This made the Transformer architecture more general and suitable for Natural Language Processing (NLP), computer vision, multimodal learning and so on. Finally, the representation relations learned by transformer-based model were more robust and general compared to the local patterns in CNN [12]. Therefore, we exploit these advantages to solve the ship image classification problem.

2.2 Swin Transformer

Swin Transformer was a modification of Vision Transformer. The idea of Swin Transformer was derived from the self-attention mechanism in NLP. The original transformer was applied to the NLP due to the attention mechanism for modelling the long-term dependency in the sequential data [13]. Because of its excellent performance in NLP, Dosovitskiy *et al.* adopted the Transformer architecture for computer vision, and proposed Vision Transformer. After trained with large-scale datasets, the model showed excellent results [11]. However, Vision Transformer was not suitable for general backbone network, as it calculated self-attention globally, so that its complexity had a quadratic relationship with image size [7]. To solve this problem, Liu et al. proposed Swin Transformer, which used a hierarchical pyramid structure and moved windows between continuous self-attention layers, reducing the complexity of model to a linear relationship with image size. This made Swin Transformer as a general backbone for various visual tasks [7].

Due to the excellent performance, Swin Transformer has been extended to many fields and achieved good results. Hong et al. employed Swin Transformer in a distracted driver image classification task with an accuracy of 95.72% [14]. Xie *et al.* make use of the powerful feature extraction ability of Swin Transformer to identify melanoma in medical images, and improves the recognition accuracy [15]. Xu *et al.* developed an efficient transformer model for remote sensing image segmentation by using Swin transformer as backbone. Compared with the previous best model, the average intersection of union increased by 2.46% [16]. In ship image classification, a survey on deep learning-based approaches for maritime visual situational awareness was carried by Qiao *et al.* in 2021. The findings showed that Swin Transformer has not been employed to solve the ship image classification problem [17]. Therefore, we conduct an intensive study on the application of Swin Transformer for ship image classification.

3 Method

Swin Transformer provides four versions for different scales of problems, from Tiny to Large, called Swin-T, Swin-S, Swin-B, Swin-L, as shown in Table 1. The difference between four versions is the layer number in the third stage and the channel number of hidden layers in the first stage. These two parameters represent the scale and computational complexity of the model. Swin-B or Swin-L is generally used when the dataset is in the range of 0.1 M to 1 M. However, the datasets of ship image classification tasks are mostly ranging from 1 K to 10 K, Swin-T or Swin-S is more suitable.

Table 1. Specifications for four Swin Transformer versions.

Version	Tiny	Small	Base	Large
Layer number in each stage	2, 2, 6, 2	2, 2, 18, 2	2, 2, 18, 2	2, 2, 18, 2
Channel number of the hidden layer	96	96	128	192
Computational complexity	0.25	0.5	1	2

3.1 Overall Architecture of Swin Transformer

Figure 1 illustrates the variation in feature resolution of a ship image within Swin-S, and the overall architecture can be divided into four stages. The resolution of input feature map reduces at each stage, and the receptive field expands layer by layer. It is suitable for Swin Transformer to be a backbone for various visual tasks due to hierarchical representation.

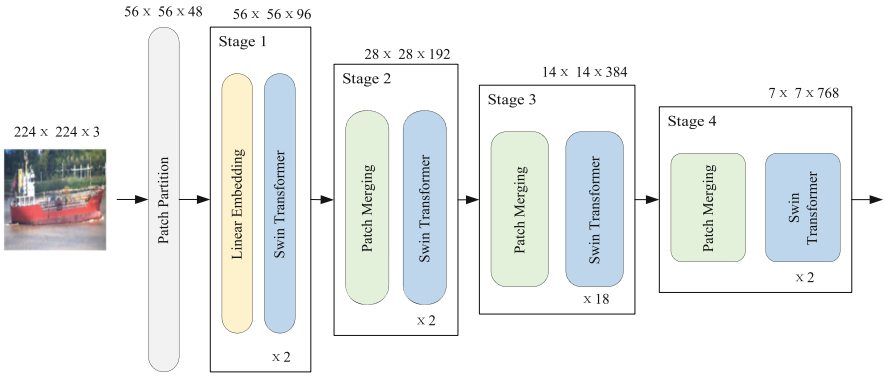

Fig. 1. Overall architecture of Swin-S.

Patch Partition and Linear Embedding. The ship image is firstly resized to 224 × 224 × 3 and input to the patch partition layer when classifying the ship image. The

patch partition layer divides the image into 4 × 4 non-overlapping patch. And the size of each patch is 56 × 56, so the size of feature map becomes 56 × 56 × 48. The linear embedding layer further resizes the image to 56 × 56 × 96 and generates a new spatial representation for the image vector, like word embedding.

Patch Merging. For the patch merging layer in Stage 2, the size of feature map is 56 × 56 × 96. The patch merging layer is similar to the pooling layer of CNN. It merges the 4 adjacent tokens in the feature map, and then merges along the last dimension. After this, the size of feature map becomes 28 × 28 × 384. Then, the fully-connection layer is employed for linear dimension reduction, cutting the channel number in half, and the output size is 28 × 28 × 192. Similarly, after patch merging processing, the feature maps in Stage 3 and Stage 4 are changed to 14 × 14 × 384 and 7 × 7 × 768, respectively. The above operations implement the hierarchical pyramid structure of Swin Transformer.

3.2 Swin Transformer Block

In addition to the hierarchical pyramid structure, Swin Transformer block is the core of the model. A Swin Transformer block is divided into two layers, as shown in Fig. 2. It includes multi-layer perceptrons (MLP), window-based self-attention (W-MSA), shifted window-based self-attention (SW-MSA), Layer Normalization (LN) and residual connection. MLP is close to the output layer, and this layer is used to integrate global attention. LN changes the distribution of all the channels into a standard normal distribution. The forward propagation of Swin Transformer block can be expressed by Eq. (1).

$$\begin{aligned} \hat{z}^l &= W\_MSA(LN(z^{l-1})) + z^{l-1}, \\ z^l &= MLP(LN(\hat{z}^l)) + \hat{z}^l, \\ \hat{z}^{l+1} &= SW\_MSA(LN(z^l)) + z^l, \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{1}$$

where $\hat{z}^l$ and z^l represent the output features of W-MSA layer and SW-MSA layer respectively.

Multi-head Self-attention. Swin Transformer is based on the self-attention mechanism, and it first unfolds the ship image into a sequence x of multiple patches before calculating its self-attention. Then x is multiplied by three projection matrices W^q, W^k and W^v, respectively, to obtain three learnable matrices of query, key and value. The calculation is shown in Eq. (2).

$$Q^i = W^q x^i, K^i = W^k x^i, V^i = W^v x^i \tag{2}$$

After that, the query and a set of key-value pairs are mapped to a score through the calculation of Eq. (3), and it represents the association between different patch. Therefore, the greater relationship, the higher score, i.e. the stronger attention.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

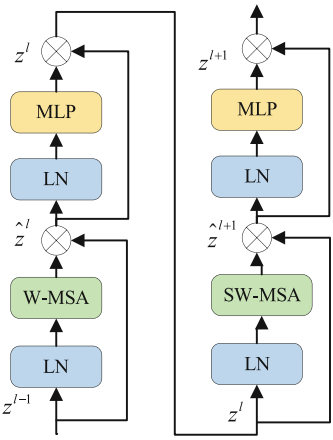

Fig. 2. Swin Transformer block.

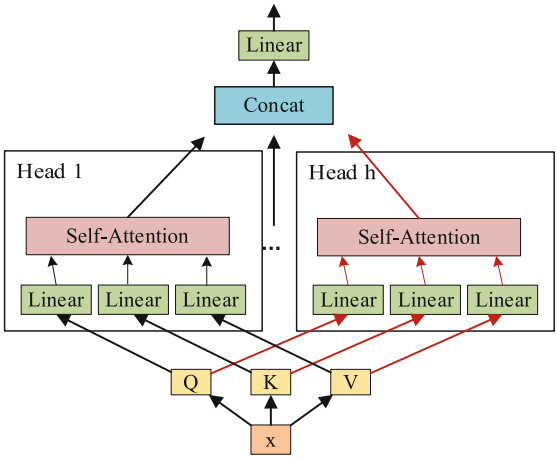

Fig. 3. The calculation process of MSA.

where d_k represents the dimension of K, so the value of QK^T is divided by $\sqrt{d_k}$, which is equivalent to normalization.

Figure 3 shows the calculation process of MSA. MSA means that the different q, k and v are calculated separately and then combined together. After a fully-connected layer, the dimensions are adjusted to the same as input matrix to obtain the output. The calculation process can be expressed by Eq. (4).

$$MSA(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{4}$$

where $head_i$ represents different attention heads, h is the number of attention heads, and W^O is a learnable output projection matrix.

Shifted Window-Based Self-Attention. Unlike the images in MNIST and CIFAR datasets are small size, the images of ships are at least 224 × 224, so it is not suitable for Vision Transformer due to its global self-attention mechanism. Swin Transformer restricts the calculation of attention to non-overlapping windows, calculating attention within the window and summing it. This makes it proportional to the window size M^2 (7 × 7), rather than the ship image size (224 × 224). Obviously, the computational complexity is reduced.

The self-attention calculation in non-overlapping window reduces the complexity of models. But there is a lack of connection between the windows. It affects the modeling. Therefore, Swin Transformer adopts a window division scheme as shown in Fig. 4(a) to establish the relationship between windows without increasing the computation. In the first module (W-MSA), regular window division is adopted, and it starts from the top-left direction and evenly divides features into 2 × 2 windows. In the next module (SW-MSA), unequal division strategy is adopted, which divides with windows rounded down by $\lfloor M/2, M/2 \rfloor$ and merges new windows that have not been connected before.

Although the window partition scheme increases the connection between windows. It also increases computation. Therefore, cyclic-shifting is introduced to the top-left

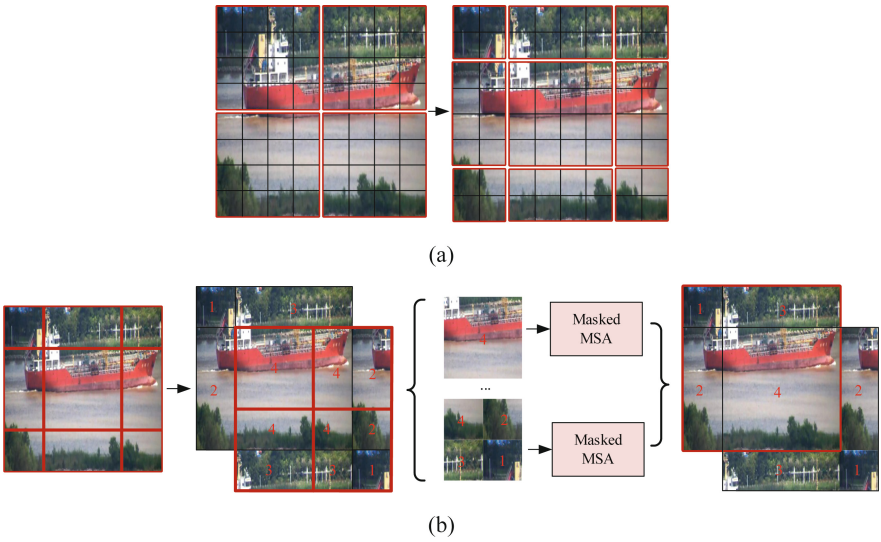

Fig. 4. Cyclic-shifting process in a ship image: (a) shifted window approach; (b) cyclic-shifting process

direction as shown in Fig. 4(b). The cyclic-shifting method not only ensures the connection between non-overlapping windows, but also does not increase windows, thus avoiding extra computation.

4 Experiments

4.1 Dataset Description

The dataset adopted in this study is Seaships [18], a well-known large scale ship dataset, and it is used to train and evaluate ship object detection algorithms. The dataset currently publishes 7000 images covering six common types of ships (ore carrier, bulk carrier, general cargo ship, container ship, fishing boat and passenger ship), as shown in Fig. 5. It takes into account many possible imaging changes, such as different scales, hull parts, lighting, viewpoint, background, and occlusion.

In this study, in order to classify ship images, we remove images containing multiple ships to ensure the uniqueness of ship types. Finally, 5254 images are retained for ship classification. The classification of training set, validation set and test set is 1:1:2 instead of 8:1:1, and this is in accordance with the requirement published by Seaships. The specific categories and classification are shown in Table 2.

Fig. 5. Sample of images from Seaships: (a) ore carrier; (b) bulk cargo carrier; (c) general cargo ship; (d) container ship; (e) fishing boat; (f) passenger ship.

Table 2. Classification and division of seaships.

Number	Category	Training set	Validation set	Test set	Total
1	bulk cargo carrier	264	250	526	1040
2	container ship	190	161	325	676
3	fishing boat	165	176	357	698
4	general cargo ship	294	266	539	1099
5	ore carrier	341	390	678	1409
6	passenger ship	72	89	171	332
Total		1326	1332	2596	5254

4.2 Evaluation Metrics

Cross-Entropy Loss Function. The cross-entropy loss (CL) function is typically used in classification tasks to calculate the loss value, and the output value needs to be normalized between 0 and 1. The value of CL indicates the difference between the predicted value and the actual value. Neural network is trained by minimizing the value of CL to accomplish back-propagation, to improve the classification accuracy. The CL is presented in Eq. (5).

$$CL = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij}) \tag{5}$$

where N represents the number of training samples, M represents the number of categories, y_{ij} represents the real distribution of samples, when i belongs to class j, $p_{ij} = 1$. During the training, y_{ij} is constant. p_{ij} denotes the predicted probability that i belongs to class j.

Classification Accuracy. Classification accuracy (CA) is the proportion that correctly predictions among the total samples. It is the main metric to evaluate the performance of classification models. CA is expressed by Eq. (6).

$$CA = \frac{1}{M}\frac{N_{\text{true}}}{N_{\text{all}}} \tag{6}$$

where N_{true} represents the number of correct samples predicted in the test set, and N_{all} is the total number.

4.3 Experiment Instructions and Settings

In this study, the effect of Swin Transformer on ship image classification is experimented. We have trained different Swin Transformer versions, CNNs and ViT. And we have compared the changes brought by different shifted windows and patch windows configuration. All the experiments are carried out on Seaships. In the study, in order to adapt to the number of ship classes, the fully-connection layer of six neurons is used in the output layer, and the activation function is Softmax. The all-zero initialization weight is used, and the training epochs is set to 300 (setting the early stop strategy). The batch size is set to 20. The Adam optimizer is used, and the learning rate is set to 0.0001. The models are training on a computer with Intel i5 2.6 GHz CPU, NVIDIA RTX 3080 GPU. In addition, we use some data arguments of random rotation, horizontal and vertical translation, random scaling, slicing and pixel filling to prevent over-fitting.

4.4 Experimental Results and Analysis

Swin Transformer vs. CNN-based Model. The classification results of four Swin Transformer versions and four typical CNN-based models are shown in Table 3. It can be seen that Swin Transformer has higher CA. Especially, Swin-S has the highest

Table 3. Classification results of Swin Transformers and typical CNN-based models.

Model	Bulk cargo carrier	Container ship	Fishing boat	General cargo ship	Ore carrier	Passenger ship	CA
Swin-T	0.876	0.923	0.933	0.944	0.960	0.918	0.929
Swin-S	0.907	0.938	0.891	0.981	0.954	0.883	0.935
Swin-B	0.876	0.948	0.882	0.963	0.968	0.912	0.930
Swin-L	0.890	0.923	0.913	0.955	0.932	0.848	0.920
VGG-16	0.893	0.982	0.924	0.911	0.972	0.860	0.930
Inception-V3	0.916	0.972	0.910	0.950	0.907	0.860	0.923
ResNet-50	0.867	0.951	0.863	0.939	0.942	0.865	0.911
ResNet-101	0.878	0.917	0.950	0.944	0.963	0.825	0.925

CA, reaching 93.5% on the test set. It is obviously superior to the typical CNN. This indicates that Swin Transformer has stronger feature extraction ability than convolution network in ship image classification. Figure 6(a) and Fig. 6(b) show the CA changes of these models on the training set and the validation set during the training process, respectively. In addition, we observe that for the "passenger ship" class, many models do not work well due to the small training set, but Swin-T archives a CA of 91.8%, which is much better than other CNN-based models.

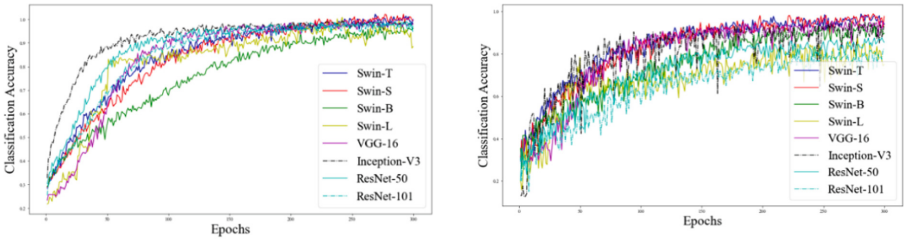

Fig. 6. Comparison of Swin Transformers and typical CNN-based models training effects: (a) CA on training sets; (b) CA on validation sets.

Swin Transformer vs. ViT. We also compare the Swin Transformer with the ViT on the Seaships to study the role of hierarchical structure. The results are shown in Table 4. As a result, the CA of Swin Transformer is improved by 27.4% compared to the ViT. It indicates that Swin Transformer outperforms in the ship image classification, mainly owing to the hierarchical pyramid structure.

Table 4. CA for different versions of Swin Transformer and ViT.

Model	Swin-T	Swin-S	Swin-B	Swin-L	ViT-B	ViT-L
CA	0.929	0.935	0.930	0.920	0.662	0.722

The Effect of Shifted Window. We compared the effect of SW-MSA and W-MSA when establishing global relationship on Swin-T and Swin-B. The results are given in Table 5. It is evident that shifted window scheme brings a 2.5% and 0.5% improvement in CA for two models, respectively.

Table 5. CA for different attention schemes.

Attention scheme	Swin-T	Swin-B
w. shifted window	0.929	0.930
w/o shifted window	0.904	0.925

The Effect of Patch Window Size. As shown in Table 6, after adjusting the patch window size from 7 to 12 on Swin-B and Swin-L, the CA of Swin-B decreases by 3.6%, while that of Swin-L increases by 1%. This indicates that the patch window may need to match the model scale and input image size. When the model scale is large, by using a larger patch window can improve the classification accuracy. Conversely, a smaller one is more sufficient.

Table 6. CA for different patch window sizes.

Patch window size	Image size	Swin-B	Swin-L
7	224×224	0.930	0.920
12	384×384	0.894	0.930

5 Conclusion

In this study, we propose to apply the state-of-the-art Swin Transformer for solving the classification problem of ship images. Swin Transformer combines global information in ship images by using attention mechanism and it has enhanced feature extraction capability than CNN. The results show that, Swin-S achieves classification accuracy of 93.5% on Seaships. The performance is better than other classical CNN.

And it is believed that hierarchical pyramid structure and shifted window scheme are significant. Compared with ViT without hierarchical pyramid structure, the classification accuracy of Swin Transformer is increased by +27.4%; The shifted window in Swin Transformer also outperforms +2.5% CA. Furthermore, we also reveal that the model scale is proportional to the input image size and patch window size. This offers the possibility of applying Swin Transformer to high-resolution images, which could be a future research.

References

1. Shao, Z.F., Wang, L.J., Wang, Z.Y., Du, W., Wu, W.J.: Saliency-aware convolution neural network for ship detection in surveillance video. IEEE Trans. Circ. Syst. Video Technol. **30**(3), 781–794 (2019)

2. Wang, K., Qu, Z., Shi, X.D., Chen, Q.S.: Application of intelligent video surveillance system in offshore oil field. Tianjin Sci. Technol. **48**(02), 55-56+61 (2021)
3. Huang, K.Q., Ren, W.Q., Tan, T.N.: A Survey of image object classification and detection algorithms. Chin. J. Comput. **37**(6), 1225–1240 (2014)
4. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, S.F., Shah, M.: Transformers in vision: a survey. arXiv: 2101.01169 (2021)
5. Liu, Y., Zhang, Y., Wang, Y.: A survey of visual transformers. arXiv: 2111.06091 (2021)
6. Zhou, H., Lu, C., Yang, S., Yu, Y.: ConvNets vs. Transformers: whose visual representations are more transferable? arXiv: 2108.05305 (2021)
7. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv: 2103.14030 (2021)
8. Leclerc, M., Tharmarasa, R., Florea, M.C., Boury-Brisset, A.C., Kirubarajan, T., Duclos-Hindié, N.: Ship classification using deep learning techniques for maritime target tracking. In: 2018 21st International Conference on Information Fusion, FUSION, pp. 737–744 (2018)
9. Xu, Z.J., Sun, J.W., Huo, Y.H.: Target recognition method of fine-grained ship Image based on multi-feature regions. Comput. Eng. Appl., 1–10 (2021)
10. Milicevic, M., Zubrinic, K., Obradovic, I., Sjekavica, T.: Data augmentation and transfer learning for limited dataset ship classification. WSEAS Trans. Syst. Control **13**, 460–465 (2018)
11. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv: 2010.11929 (2020)
12. Xu, Y., et al.: Transformers in computational visual media: a survey. Comput. Vis. Media **8**(1), 33–62 (2021). https://doi.org/10.1007/s41095-021-0247-3
13. Aswani, V., et al.: Attention is all you need. arXiv: 1706.03762 (2017)
14. Koay, H.V., Huang, C.J., Chow, C.O.: Shifted-window hierarchical vision transformer for distracted driver detection. In: 2021 IEEE Region 10 Symposium, TENSYMP, pp. 1–7 (2021)
15. Xie, J., Wu, Z., Zhu, R., Zhu, H.: Melanoma detection based on swin transformer and SimAM. In: 2021 IEEE 5th Information Technology. Networking, Electronic and Automation Control Conference, ITNEC, pp. 1517–21. IEEE Press, Xi'an (2021)
16. Xu, Z., Zhang, W., Zhang, T., Yang, Z., Li, J.: Efficient transformer for remote sensing image segmentation. Remote Sens. **13**(18), 3585 (2021)
17. Qiao, D., Liu, G., Lv, T., Li, W., Zhang, J.: Marine Vision-based situational awareness using discriminative deep learning: a Survey. J. Mar. Sci. Eng. **9**(4), 395 (2021)
18. Shao, Z.F., Wu, W.J., Wang, Z.Y., Du, W., Li, C.Y.: SeaShips: a large-scale precisely-annotated dataset for ship detection. IEEE Trans. Multimedia **20**(10), 2593–2604 (2018)