基于Swin Transformer的岸基监控船舶分类方法

刘继祥[1]、孙文丽[1(□)]、高旭[2]

[1] 大连海事大学航海学院，辽宁大连 116026
itslab@dlmu.edu.cn
2 大连海事大学国家航海系统工程研究中心，辽宁大连 116026

摘要。船舶图像分类技术是智能海事监控系统的核心技术之一，准确识别船舶及其类型是分析和理解海上场景的基础。近年来，基于Transformer的模型在自然语言处理领域取得突破，并在图像分类任务中超越卷积神经网络，其中Swin Transformer表现最为突出。该模型在多头自注意力机制基础上构建了分层金字塔结构和移位窗口方案，显著降低了模型复杂度，成为计算机视觉领域的通用骨干网络。本研究采用知名船舶图像数据集Seaships验证Swin Transformer的有效性，发现其分层金字塔结构、多头自注意力机制和移位窗口方案在船舶图像分类中起关键作用。实验结果表明，Swin Transformer在船舶图像分类中准确率达到93.5%，优于典型卷积网络和Vision Transformer。

关键词：Swin Transformer - 船舶监控 - 图像分类 - 注意力机制 - 深度学习

1 引言

人工智能(AI)的快速发展推动了船舶行业的崛起。然而，大多数海事监控系统仍采用传统人工模式，难以持续有效地聚焦船舶目标[1,2]。因此，实现海事监控系统的智能化升级刻不容缓。运用AI技术分析处理船舶目标，既能降低人员劳动强度，又可提升识别精度，故AI在海事交通与走私监管中具有重要作用。

由于准确识别船舶及其类型是海上作业的重要前提，首先需要解决船舶分类问题[2]。在计算机视觉领域，为解析和理解场景，同样需识别物体类别。但图像分类在尺度、光照等不同层面仍存在诸多难点，物体外观会因此发生显著变化。

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

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

在实例层面存在视角变化、形变和遮挡问题；类别层面存在显著的类内差异、类间模糊性和背景干扰；语义层面则呈现多重稳定性。此外，海量数据和图像类别也增加了图像分类的难度[3]。

深度学习中，图像分类方法主要分为基于卷积神经网络(CNN)和基于Transformer的两类。CNN方法因采用固定尺寸窗口提取图像特征，难以捕捉长距离依赖关系。而基于Transformer的方法通过注意力机制对图像关键区域特征赋予更高权重，从而整合全局信息[4]。因此在ImageNet、COCO和ADE20K等数据集的图像分类任务中，Transformer模型较CNN模型展现出更优性能[5,6]。

在基于Transformer的模型中，由Vision Transformer改进而来的Swin Transformer具有最佳性能。该模型采用移位窗口机制，将自注意力计算限制在非重叠的局部窗口内，同时允许跨窗口连接，从而显著提升效率。此外，Swin Transformer采用的分层结构不仅使模型具备多尺度建模的灵活性，还将计算复杂度控制在图像尺寸的线性范围内，因此可作为通用计算机视觉的骨干网络。鉴于Swin Transformer在图像分类任务中兼容多种视觉任务且超越先前最先进模型[7]，我们将其应用于船舶图像分类领域。

本研究的主要贡献如下。

(1) 在船舶图像分类任务中，我们采用Swin Transformer在知名船舶数据集Seaships上实现93.5%的分类准确率，优于典型卷积神经网络。

(2) 我们研究了Swin Transformer金字塔结构的作用，实验表明该结构可使分类准确率提升27.4%。
(3) 为探究Swin Transformer注意力机制与其处理图像全局信息能力的关系，实验表明窗口移位策略对性能提升具有贡献。

本文其余部分组织结构如下：第2节回顾船舶图像分类与Swin Transformer的现有研究；第3节提出应用Swin Transformer处理船舶图像的方法论；第4节介绍数据集、评估指标与超参数并分析结果；第5节总结全文。

2 相关工作

2.1 船舶图像分类

当卷积神经网络（CNN）作为主流深度学习模型时，大量研究将其应用于船舶图像分类领域。Leclerc等人在海事船舶分类数据集Marvel上训练CNN模型，取得了显著改进

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

相较于当时的最先进成果[8]，Xu 等人提出改进版VGG-19以解决小目标过度池化问题，该模型显著提升了小型舰船识别能力[9]。Milicevic团队将CNN与水平翻转、裁剪、缩放、旋转及RGB通道值变换等多种数据增强技术应用于舰船分类，使准确率提升6.5%[10]。

在图像分类任务中，基于Transformer的模型较CNN更具优势：其一，Transformer采用注意力机制，其注意力距离随网络深度递增，优于CNN的感知野[11]；其二，Transformer建模长程交互时采用高效计算，使其架构更具普适性，适用于自然语言处理、计算机视觉及多模态学习等领域；最后，相比CNN的局部模式，Transformer学习的表征关系更具鲁棒性和泛化性[12]。因此，我们利用这些优势解决舰船图像分类问题。

2.2 Swin Transformer

Swin Transformer是Vision Transformer的改进版本。其核心思想源于自然语言处理中的自注意力机制。原始Transformer凭借其注意力机制对序列数据长程依赖关系的建模能力，最初被应用于自然语言处理领域[13]。鉴于其在NLP中的卓越表现，Dosovitskiy等人将Transformer架构引入计算机视觉领域，提出了Vision Transformer。该模型经过大规模数据集训练后展现出优异性能[11]。但由于Vision Transformer采用全局自注意力计算，其复杂度与图像尺寸呈二次方关系，因此不适用于通用骨干网络[7]。为解决这一问题，Liu等人提出Swin Transformer，通过分层金字塔结构和连续自注意力层间的移动窗口机制，将模型复杂度降低至与图像尺寸呈线性关系，使其成为适用于多种视觉任务的通用骨干网络[7]。

由于性能卓越，Swin Transformer已被拓展至众多领域并取得优异成果。Hong等人将其应用于分心驾驶图像分类任务，准确率达95.72%[14]；Xie团队利用其强大特征提取能力识别医学图像中的黑色素瘤，显著提升识别精度[15]；Xu研究组以Swin Transformer为骨干网络开发遥感图像分割模型，平均交并比较原最优模型提升2.46%[16]。在船舶图像分类领域，Qiao等2021年开展的深度学习海事视觉态势感知综述表明，该问题尚未有研究采用Swin Transformer解决[17]。为此，我们针对船舶图像分类任务开展Swin Transformer的深度应用研究。

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 方法

Swin Transformer针对不同规模问题提供四个版本：从Tiny到Large，分别称为Swin-T、Swin-S、Swin-B、Swin-L，如表1所示。四个版本的区别在于第三阶段的层数和第一阶段隐藏层的通道数，这两个参数决定了模型的规模和计算复杂度。当数据集规模介于 $0.1\mathrm{M}$ 到 $1\mathrm{M}$ 时通常使用Swin-B或Swin-L，而船舶图像分类任务的数据集大多在 $1\ \mathrm{K}$ 至 $10\ \mathrm{K}$ 范围内，更适合采用Swin-T或Swin-S。

表1. Swin Transformer四个版本的规格参数

版本	微型	小型	基础	大型
各阶段层数	2, 2, 6, 2	2, 2, 18, 2	2, 2, 18, 2	2, 2, 18, 2
隐藏层通道数	96	96	128	192
计算复杂度	0.25	0.5	1	2

3.1 Swin Transformer整体架构

图1展示了Swin-S中船舶图像特征分辨率的变化情况，其整体架构可分为四个阶段。输入特征图的分辨率逐阶段降低，而感受野则逐层扩大。由于这种层次化表征特性，Swin Transformer非常适合作为各类视觉任务的主干网络。

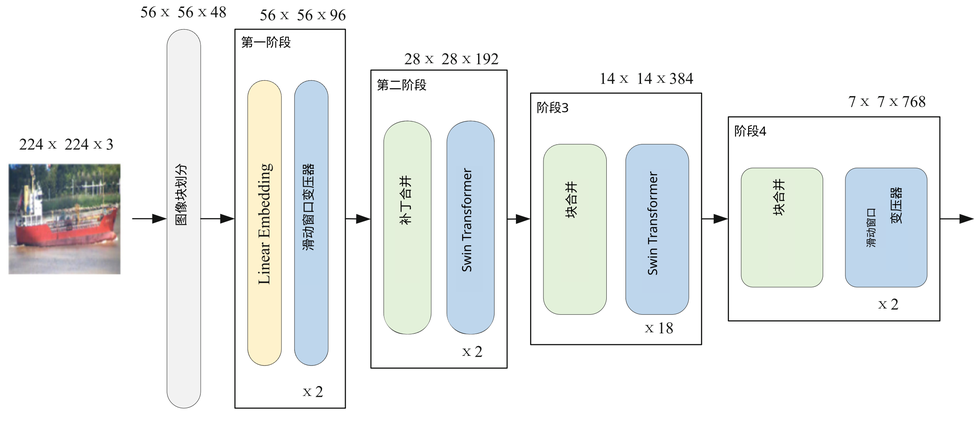

图1. Swin-S整体架构

补丁划分与线性嵌入。在分类船舶图像时，首先将船舶图像尺寸调整为 $224\times\ 224\ \times\ 3$ 并输入补丁划分层。

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.

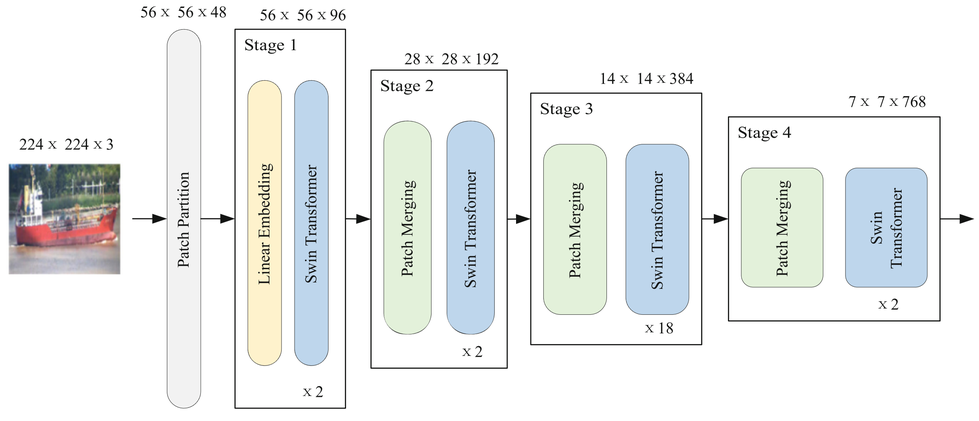

Fig. 1. Overall architecture of Swin-S.

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

图像分块层将图像划分为 4×4 个不重叠的区块，每个区块尺寸为 56×56，因此特征图尺寸变为 $56 \times 56 \times 48$。线性嵌入层进一步将图像调整为 $56 \times 56 \times 96$，并为图像向量生成新的空间表征（类似词嵌入）。
补丁合并。在第二阶段中，特征图的尺寸为 $56 \times 56 \times 96$。该层的功能类似于CNN中的池化层，将特征图中相邻的4个标记进行合并，并沿最后一个维度进行聚合。处理后特征图尺寸变为 $28 \times 28 \times 384$。随后通过全连接层进行线性降维，将通道数减半，输出尺寸为 $28 \times 28 \times 192$。同理，经过补丁合并处理后，第三阶段和第四阶段的特征图尺寸分别变为 $14 \times 14 \times 384$ 和 $7 \times 7 \times 768$。上述操作实现了Swin Transformer的层级金字塔结构。

3.2 Swin Transformer模块

除层级金字塔结构外，Swin Transformer模块是该模型的核心。如图2所示，Swin Transformer模块分为两层，包含多层感知机（MLP）、基于窗口的自注意力机制（W-MSA）、移位窗口自注意力机制（SW-MSA）、层归一化（LN）及残差连接。MLP靠近输出层，用于整合全局注意力。LN将所有通道的分布转换为标准正态分布。Swin Transformer模块的前向传播过程可用公式(1)表示。

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

其中 $\hat{z}^l$ 和 z^l 分别代表W-MSA层与SW-MSA层的输出特征。

多头自注意力机制。Swin Transformer基于自注意力机制，首先将船舶图像展开为多个补丁的序列 x，然后计算其自注意力。接着 x 分别与三个投影矩阵 W^q, W^k 和 W^v 相乘，得到可学习的查询矩阵、键矩阵和值矩阵。具体计算如公式(2)所示。

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

随后，通过公式(3)的计算将查询向量与一组键值对映射为关联分数，该分数表示不同补丁间的相关性。因此，关联性越强则分数越高，即注意力越集中。

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

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 $56 \times 56 \times 48$. The linear embedding layer further resizes the image to $56 \times 56 \times 96$ 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 $56 \times 56 \times 96$. 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 $28 \times 28 \times 384$. Then, the fully-connection layer is employed for linear dimension reduction, cutting the channel number in half, and the output size is $28 \times 28 \times 192$. Similarly, after patch merging processing, the feature maps in Stage 3 and Stage 4 are changed to $14 \times 14 \times 384$ and $7 \times 7 \times 768$, 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.

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

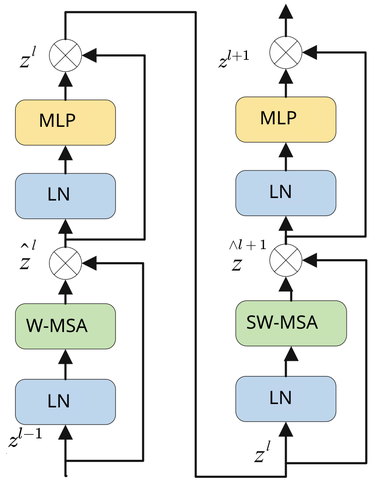

图2. Swin Transformer模块。

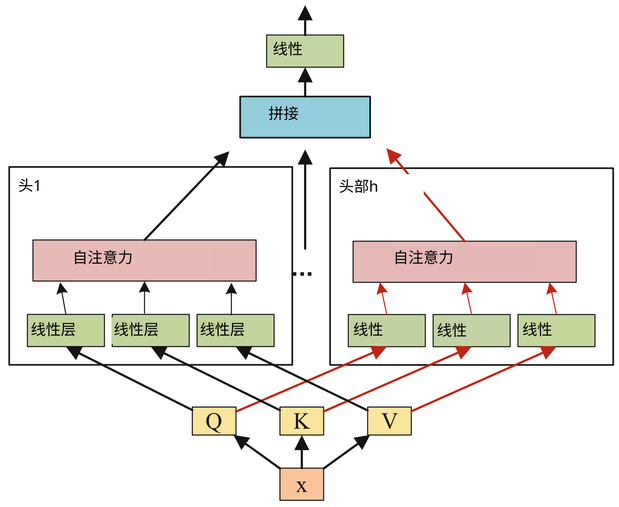

图3. MSA的计算流程。

其中 d_k 表示 K 的维度，因此 QK^T 的值需除以 $\sqrt{d_k}$，这相当于归一化处理。

图3展示了MSA的计算流程。MSA指将不同的 q 、 k 和 v 分别计算后再合并。经过全连接层后，维度被调整至与输入矩阵相同以获得输出。该计算过程可用公式(4)表示。

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

其中 head_i 代表不同的注意力头， h 是注意力头数量， W^O 为可学习的输出投影矩阵。

基于窗口偏移的自注意力机制。与MNIST和CIFAR数据集中尺寸较小的图像不同，船舶图像至少 224×224 ，因此采用全局自注意力机制的视觉Transformer并不适用。Swin Transformer将注意力计算限制在非重叠窗口内，通过在窗口内部计算注意力并进行聚合。这使得计算复杂度与窗口尺寸 M^2 (7×7) 呈正比，而非船舶图像尺寸 (224×224) 。显然，该方法显著降低了计算复杂度。

非重叠窗口的自注意力计算降低了模型复杂度，但窗口间缺乏关联性，影响了建模效果。为此，Swin Transformer采用图4(a)所示的窗口划分方案，在不增加计算量的前提下建立窗口间联系。首个模块(W-MSA)采用常规窗口划分，从左上角开始将特征均匀分割为 2×2 个窗口；后续模块(SW-MSA)则采用非均匀划分策略，通过 $\lfloor \mathrm{M}/2, \mathrm{M}/2 \rfloor$ 像素向下取整的窗口进行分割，并合并先前未建立连接的新窗口。

尽管窗口划分方案增强了窗口间联系，但同时也增加了计算量。因此引入了左上角循环移位机制

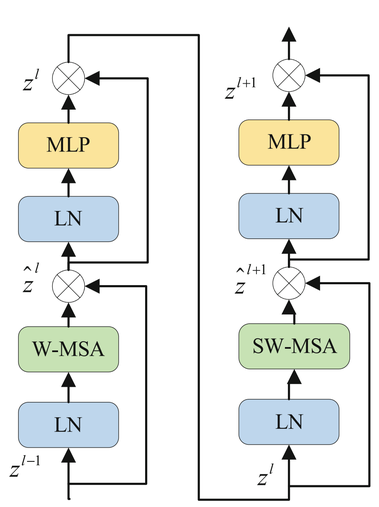

Fig. 2. Swin Transformer block.

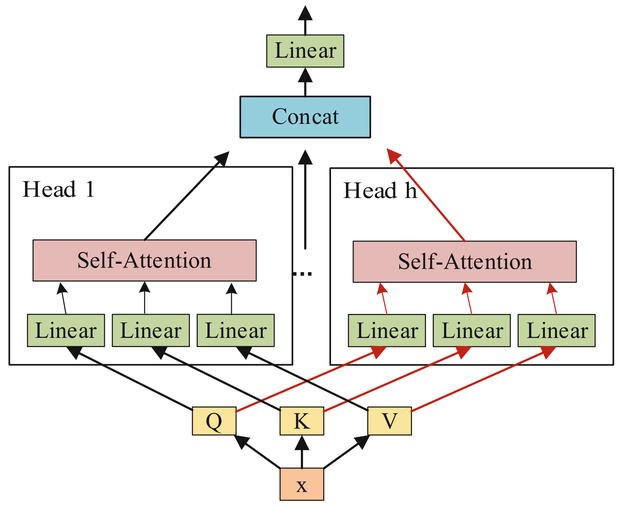

Fig. 3. The calculation process of MSA.

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 $\lfloor \mathrm{M}/2, \mathrm{M}/2 \rfloor$ 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

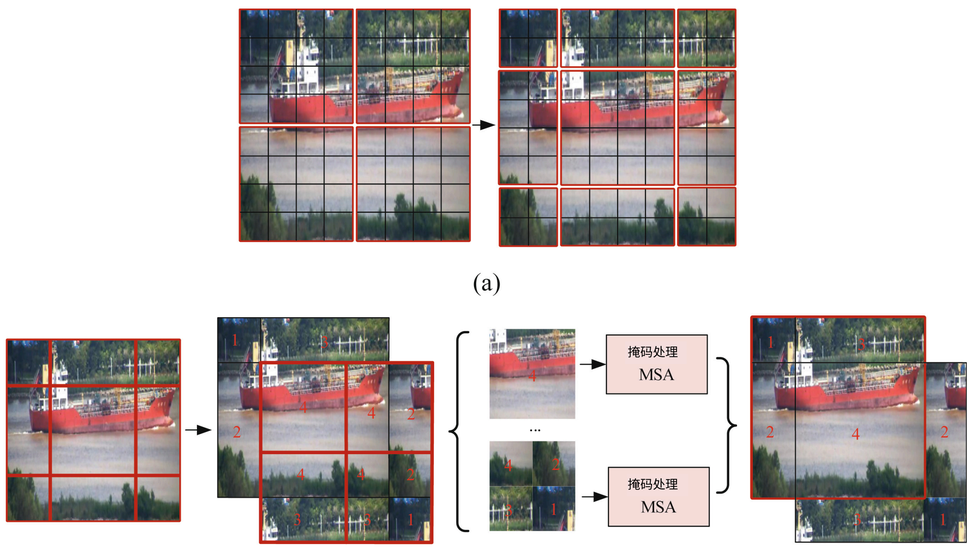

图4. 船舶图像中的循环移位过程：(a) 滑动窗口法；(b) 循环移位过程

方向如图4(b)所示。该循环移位方法既保证了非重叠窗口间的关联性，又未增加窗口数量，从而避免了额外计算。

4 实验

4.1 数据集描述

本研究采用的数据集为Seaships[18]，这是一个知名的大规模船舶数据集，用于训练和评估船舶目标检测算法。该数据集目前公开了7000张图像，涵盖六种常见船舶类型（矿砂船、散货船、杂货船、集装箱船、渔船和客船），如图5所示。数据集考虑了多种可能的成像变化因素，如不同尺度、船体部位、光照条件、视角、背景及遮挡情况。

在本研究中，为进行船舶图像分类，我们剔除了包含多艘船舶的图像以确保船型唯一性。最终保留5254张图像用于船舶分类。训练集、验证集和测试集的划分比例调整为1:1:2（而非8:1:1），这符合Seaships发布的要求。具体类别划分情况如表2所示。

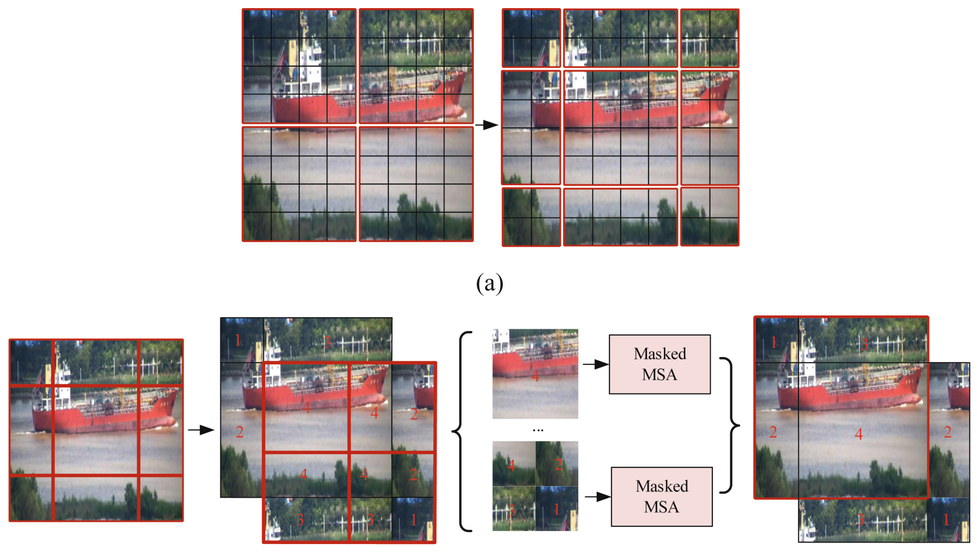

Fig. 4. Cyclic-shifting process in a ship image: (a) shifted window approach; (b) cyclic-shifting process

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.

图5. Seaships数据集图像示例：(a)矿石船；(b)散货船；(c)杂货船；(d)集装箱船；(e)渔船；(f)客船。

表2. 船舶分类与划分

编号	类别	训练集	验证集	测试集	总计
1	散货船	264	250	526	1040
2	集装箱船	190	161	325	676
3	渔船	165	176	357	698
4	杂货船	294	266	539	1099
5	矿石运输船	341	390	678	1409
6	客轮	72	89	171	332
总计		1326	1332	2596	5254

4.2 评估指标

交叉熵损失函数。交叉熵损失(CL)函数通常用于分类任务中计算损失值，输出值需归一化至0到1之间。CL值反映预测值与实际值的差异。神经网络通过最小化CL值完成反向传播，从而提高分类准确率。CL函数如公式(5)所示。

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

其中 N 表示训练样本数量， M 表示类别数量， y_{ij} 表示样本的真实分布（当 i 属于 $j, p_{ij} = 1$ 类时）。训练过程中 y_{ij} 为常量。p_{ij} 表示 i 属于 j 类的预测概率。

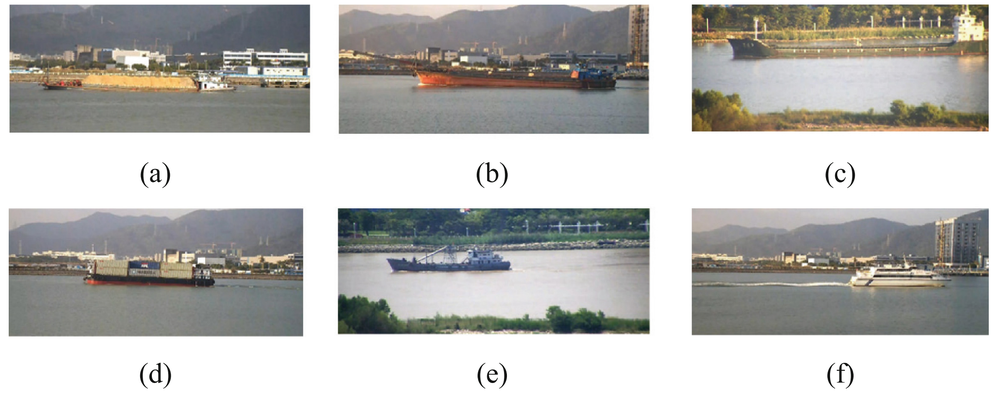

Fig. 5. Sample of images from Seaships: (a) ore carrier; (b) bulk cargo carrier; (c) general cargo ship; (d) container ship; (e) fishing boat; (f) passenger ship.

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.

分类准确率。分类准确率（CA）是指正确预测样本数占总样本数的比例，是评估分类模型性能的核心指标，其计算公式如式(6)所示。

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

式中 N_{true} 表示测试集中被正确预测的样本数量， N_{all} 为样本总数。

4.3 实验说明与设置

本研究实验了Swin Transformer在船舶图像分类中的效果。我们训练了不同版本的Swin Transformer、CNN和ViT模型，并比较了不同移位窗口和补丁窗口配置带来的变化。所有实验均在Seaships数据集上进行。为适应船舶类别数量，输出层采用六神经元全连接层，激活函数为Softmax。权重采用全零初始化，训练周期设为300（设置早停策略），批量大小为20。使用Adam优化器，学习率设为0.0001。模型在配备Intel i5 2.6 GHz CPU和NVIDIA RTX 3080 GPU的计算机上训练。此外，采用随机旋转、水平垂直平移、随机缩放、切片及像素填充等数据增强方法防止过拟合。

4.4 实验结果与分析

Swin Transformer与基于CNN模型的对比。四种SwinTransformer版本与四种典型CNN模型的分类结果如表3所示。可以看出，Swin Transformer具有更高的分类准确率(CA)，其中Swin-S版本表现最优。

表3. Swin Transformers与典型CNN模型的分类结果

模型	散货船	集装箱船	渔船	通用货船	矿石运输船	客船	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

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

在测试集上准确率(CA)达到93.5%，明显优于典型CNN模型。这表明在船舶图像分类任务中，Swin Transformer比卷积网络具有更强的特征提取能力。图6(a)和图6(b)分别展示了训练过程中这些模型在训练集和验证集上的CA变化曲线。此外我们观察到，由于训练样本较少，"客船"类别的分类效果普遍不佳，但Swin-T仍取得91.8%的CA值，显著优于其他基于CNN的模型。

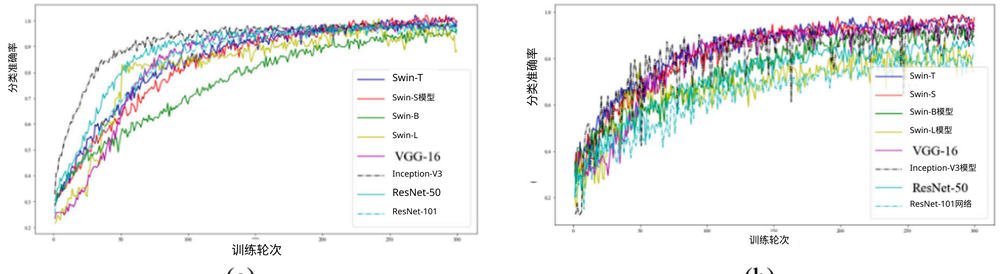

图6. Swin Transformer与典型CNN模型训练效果对比：(a)训练集分类准确率；(b)验证集分类准确率。

Swin Transformer与ViT对比。我们还在Seaships数据集上对比了Swin Transformer与ViT，以研究分层结构的作用。结果如表4所示，Swin Transformer的分类准确率较ViT提升了27.4%，这表明其在船舶图像分类中的优异表现主要归功于分层金字塔结构。

表4. 不同版本Swin Transformer与ViT的分类准确率对比

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

移位窗口的效果分析。我们在Swin-T和Swin-B模型上对比了SW-MSA与W-MSA在建立全局关系时的性能差异。表5数据显示，移位窗口机制使两个模型的分类准确率分别提升了2.5%和0.5%。

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.

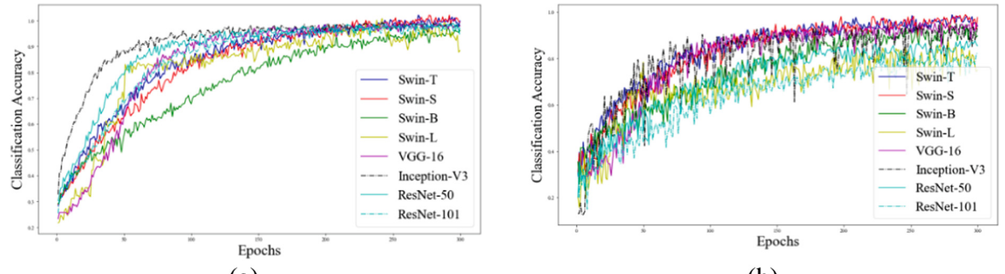

Fig. 6. Comparison of Swin Transformers and typical CNN-based models training effects: (a) CA on training sets; (b) CA on validation sets.

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.

表5. 不同注意力机制的CA值

注意力机制	Swin-T	Swin-B
带窗口位移	0.929	0.930
无窗口位移	0.904	0.925

补丁窗口尺寸的影响。如表6所示，在Swin-B和Swin-L上将补丁窗口尺寸从7调整到12后，Swin-B的分类准确率下降3.6%，而Swin-L则提升 1% 。这表明补丁窗口需与模型规模和输入图像尺寸相匹配。当模型规模较大时，采用更大的补丁窗口可提高分类精度；反之，较小窗口更为适用。

表6. 不同补丁窗口尺寸的分类准确率

补丁窗口尺寸	图像尺寸	Swin-B	Swin-L
7	224 × 224	0.930	0.920
12	384 × 384	0.894	0.930

5 结论

本研究提出采用最先进的Swin Transformer解决船舶图像分类问题。该模型通过注意力机制融合船舶图像的全局信息，其特征提取能力优于CNN。实验表明，Swin-S在Seaships数据集上达到93.5%的分类准确率，性能超越传统CNN模型。

研究表明，分层金字塔结构和移位窗口机制具有显著优势：相比无分层结构的ViT，Swin Transformer分类准确率提升27.4%；其移位窗口设计还带来2.5%的准确率增益。此外，我们发现模型规模与输入图像尺寸、分块窗口大小呈正相关，这为Swin Transformer在高分辨率图像中的应用提供了可能，值得未来深入研究。

参考文献

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)

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)

Ship Classification Using Swin Transformer for Surveillance on Shore 785

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)