Dual Wavelet Attention Networks for Image Classification

Yuting Yang, *Student Member, IEEE,* Licheng Jiao, *Fellow, IEEE,* Xu Liu, *Member, IEEE,* Fang Liu, *Senior Member, IEEE,* Shuyuan Yang, *Senior Member, IEEE,* Lingling Li, *Senior Member, IEEE,* Puhua Chen, *Senior Member, IEEE,* Xiufang Li, *Student Member, IEEE,* Zhongjian Huang, *Student Member, IEEE,*

***Abstract*—Global average pooling (GAP) plays an important role in traditional channel attention. However, there is the disadvantage of insufficient information to use the result of GAP as the channel scalar. At the same time, the existing spatial attention models focus on the areas of interest using average pooling or convolutional networks, but there is a loss of feature information and neglect of the structural feature. In this paper, dual wavelet attention is proposed, which can effectively alleviate the aforementioned problems and enhance the representation ability of CNNs. Firstly, the equivalence between the sum of the low-frequency subband coefficients of 2D DWT (Haar) and GAP is proved. On this basis, the statistical characteristics of low-frequency and high-frequency subbands are effectively combined to obtain the channel scalars, which can better measure the importance of each channel. In addition, 2D DWT can effectively capture the approximate and detailed structural features. Thus, wavelet spatial attention is proposed, which can effectively focus on the key spatial structural features. Different from traditional spatial attention, it can better curve the structural and spatial attention for different channels. The experiments are verified on four natural image data sets and three remote sensing scene classification data sets, which shows the effectiveness and versatility of the proposed methods. The code of this paper will be available at https://github.com/yutinyang/DWAN.**

***Index Terms*—Attention mechanism, 2D DWT, Dual wavelet attention, Wavelet channel attention, Wavelet spatial attention.**

I. Introduction

With the continuous development of deep learning, more and more convolutional neural networks (CNNs) models have emerged and achieved great success in various fields [1], such as computer vision [2], [3], [4], natural language processing [5] and so on. Especially in the image classification task, CNNs have made continuous breakthroughs. The main reason for their success is that CNNs have a relatively significant advantage in extracting underlying features (point, line, etc.) and visual structural features [6]. In CNNs, the visual attention mechanism has become a hot topic. Inspired by the human visual system (HVS), it mainly enhances objects or regions of interest and suppresses the useless information [7]. An attention module is usually an additional neural network that strictly selects certain input parts or assigns different weights to the other areas. It is usually combined with various neural network structures to improve the interpretability of the neural network. Meanwhile, it is often effective in improving the performance of the network [8].

In addition, attention-based models have been widely developed in recent years [9], [10]. And there are many novel models appearing for different visual tasks, such as video classification [11], full-reference image quality predictions [12] and so on. The existing attention mechanism models can be divided into different types according to different division methods. The standard attention mechanisms include channel-domain, spatial-domain, and mixed-domain attention models.

The channel attention can learn the importance of each feature channel through the network and finally assign different weight coefficients to each channel. It can strengthen important

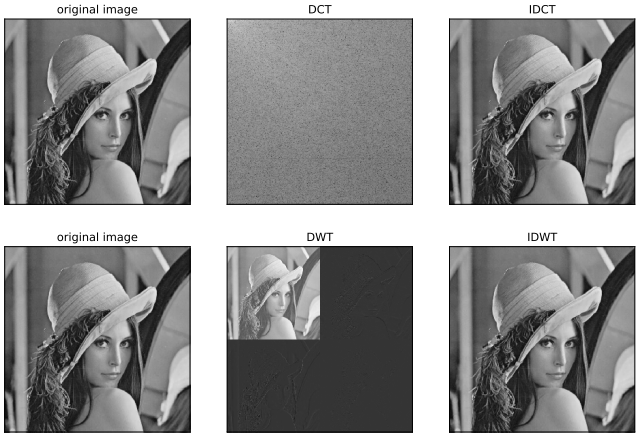

Fig. 1. The results of Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), inverse Discrete Cosine Transform (IDCT) and inverse Discrete Wavelet Transform (IDWT) of an 2D image.

Manuscript received July 25, 2022; revised September 11, 2022; accepted Octber 26, 2022. This work was supported in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the State Key Program and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (61836009, 61621005), Key Research and Development Program in Shaanxi Province of China (2019ZDLGY03-06), the Major Research Plan of the National Natural Science Foundation of China (91438201, 91438103, and 61801124), the National Natural Science Foundation of China (U1701267, 62006177, 61871310, 61902298, 61573267, and 61906150), the Fund for Foreign Scholars in University Research and Teaching Programs 111 Project (B07048), the Program for Cheung Kong Scholars and Innovative Research Team in University (IRT 15R53), the ST Innovation Project from the Chinese Ministry of Education, the National Science Basic Research Plan in Shaanxi Province of China(2019JQ-659), the China Postdoctoral Fund(2019M663641, 2017M613081), the Scientific Research Project of Education Department In Shaanxi Province of China (No.20JY023), the fundamental research funds for the central universities (XJS201901, XJS201903, JBF201905, JB211908), and the CAAI-Huawei MindSpore Open Fund. *(Corresponding author: Licheng Jiao.)*

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, International Research Center of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an, China (e-mail:lchjiao@mail.xidian.edu.cn; ytyang_1@stu.xidian.edu.cn).

feature channels and suppress unimportant feature channels. For example, SENet [13] is a typical channel attention model. It mainly proposes to use the result of global average pooling (GAP) as the channel scalar. GAP is a standard and algorithmic way of using per-channel scalars because of its simplicity and efficiency. It has gained a great success in the past, but [14] clearly points out that using average pooling leads to the loss of spatial information, which is very important for object and scene recognition. In addition, convolutional block attention module (CBAM) [15] and the style-based recalibration module (SRM) [16] further use global max pooling and global standard deviation pooling to enhance the performance of GAP. However, the simplicity of GAP makes it challenging to capture the complex information of various inputs well, and there is still the disadvantage of insufficient information. Thus, generating the proper scalars to measure the importance of different channels is still an essential challenge in channel attention.

In contrast, spatial attention often transforms spatial information to move the original image into another space and preserve the essential spatial information. In ordinary convolution, the pooling layer directly uses max pooling or average pooling methods to compress information and retain the main features of the image. It reduces the computational complexity and improves accuracy. But the pooling method is relatively violent. Consolidating the max pooling and average pooling information directly will fail to identify critical information. Thus, seeking better spatial attention that remains the structural feature kept is a good choice to improve the performance of traditional spatial attention.

The attention models above are proposed in the time domain. Later on, some scholars thought about the attention mechanism in the frequency domain. For example, FcaNet [17] treats the scalar representation of channels as a compression problem in the frequency domain. It suggests that the channel information should be compactly encoded by scalars while preserving the representation ability of the entire channel as much as possible. The discrete cosine transform (DCT) is used as the compress method to obtain the channel scalar. FcaNet shows good performance on image classification. However, it owns much spectrum and faces the problem of frequency selection.

Both the DWT and DCT transform belong to the mainstream methods of image compression [18]. As shown in Fig. 1, the inverse DCT (IDCT) and the inverse DWT (IDWT) results show that both of them own the reconstructional ability. Unlike the DCT transform, the DWT transform can capture both the image's spatial-frequency (wavelet domain) components. In addition, it is mentioned that wavelet transform can be used to extract features in polarization space and to obtain the contextual features [19], [20], [21]. Besides,[22], [21] mainly use the features extracted by 3D discrete wavelet transform as identification features, which is conducive to the accurately detecting motion in video sequences.

Unlike previous works, the dual wavelet attention networks is proposed for image classification based on rethinking visual attention in the wavelet domain. The original features are effectively compressed through the DWT decomposition, and then a unique channel scalar can be provided for each channel. Meanwhile, the spatial features of the wavelet decomposition subbands can effectively capture the critical components and structural features, thereby realizing spatial attention. The proposed model has been experimentally evaluated on the four natural benchmark data sets (CIFAR10, CIFAR100, SVHN, and ImageNet) and three remote sensing scene classification data sets (WHU-RS19, AID, and UCM). In summary, the main contributions of this paper are summarized as follows.

1) The dual wavelet attention is proposed to alleviate the problem of insufficient information or structural information loss in traditional attention modules. It can effectively enhance the characterization ability of convolutional networks.
2) The wavelet channel attention is proposed based on proving the equivalence between Haar wavelet transform and global average pooling, which can obtain better channel scalars.
3) Wavelet spatial attention is proposed, which is adaptive to obtain the spatial structural attention for different channels.
4) Extensive experiments have validated the effectiveness and versatility of our models on four natural data sets and three remote sensing classification data sets.

The remainder of this paper is organized as follows. Section II briefly introduces the related works. Section III presents the proposed dual wavelet attention networks for image classification. Section IV shows the experimental results and theoretical analysis to confirm the effectiveness of our methods. Finally, Section V gives a summary and the direction to further work.

II. Related Works

The dual wavelet attention block is a novel attention block that rethinks the attention mechanism in the wavelet domain. Thus, the attention mechanism and the wavelet domain learning in CNNs are mainly introduced as related works.

A. Attention Mechanism

There is no strict mathematical definition of the attention mechanism. The traditional local image feature extraction, saliency detection, sliding window-based methods, etc., can be regarded as attention mechanisms. In recent years, the attention mechanism has been a research hotspot of deep learning. An attention module is often regarded as an additional module to improve the interpretability and performance of neural networks. In the continuous development of deep learning, the attention mechanism has developed rapidly. Thus, many attention-related models have emerged, such as [13], [15], [23], [24], [25], [26].

The initial channel attention is proposed in SENet [13]. Its core idea is to perform global adaptive average pooling on the spatial dimension and then learn the channel weights through the fully-connection (FC) layer. SKNet [25] is an improvement to SENet. For each image, it will use different convolution kernels to calculate the channel weights dynamically. Another improved channel attention model based on SENet, ECANet [26], proposes the local cross-channel interaction strategy

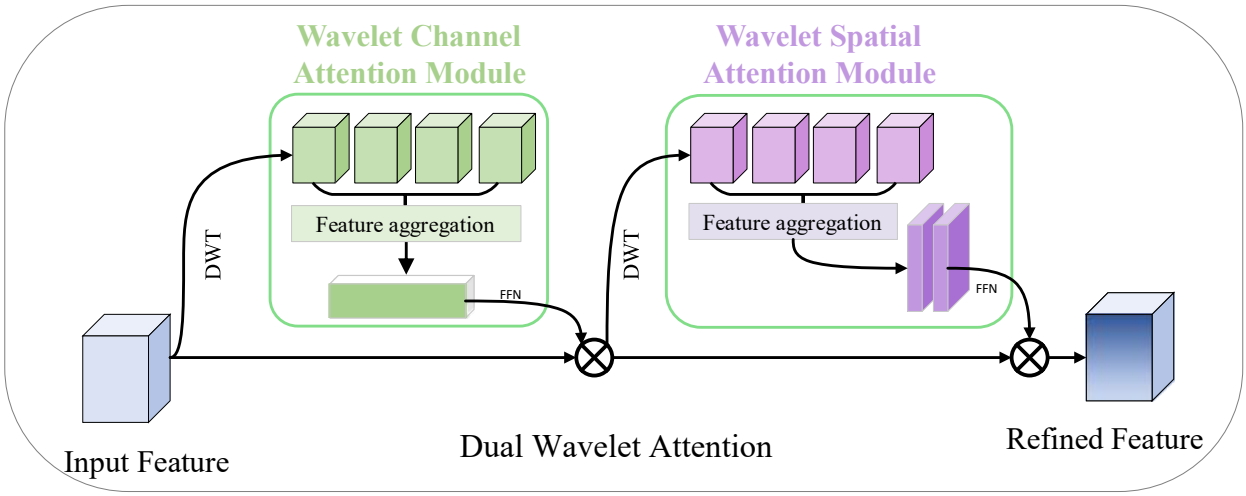

Fig. 2. The overview of dual wavelet attention block. It consists of two sequential sub-modules: wavelet channel (WC) and wavelet spatial (WS) module. FFN represents the feedforward network.

without dimensionality reduction. It can adaptively select the size of one-dimensional convolution kernels. Besides, its improved effect is obvious with little increase in complexity. The channel attention only pays attention to the importance of the channel. Besides, the channel information is directly processed globally, and the spatial information interaction is often neglected.

In addition, spatial attention has also received extensive attention. It pays attention to the spatial feature. However, it treats the image feature in each channel equally and ignores the channel information. Therefore, there are some attention mechanism researches that combine channel attention with spatial attention, such as CBAM [15], BAM [27], and CoordAttention [28].

Besides, some scholars have also begun to explain attention from different perspectives. For example, FcaNet [17] proposes to understand the channel attention mechanism from the frequency domain. It proves that GAP is a particular case of DCT. In addition, it presents the FcaNet with the multi-spectral attention module, which generalizes the existing channel attention mechanism in the frequency domain. Sufficient information preserved by introducing more frequency components makes up for the shortcomings of insufficient feature information in the present channel attention methods. Unlike FcaNet, the proposed methods take full advantage of the compression properties of the wavelet transform instead of DCT. Besides, it extends FcaNet to obtain better spatial attention. It can capture the geometric features in visual attention, which is not achievable by FcaNet.

B. Wavelet Domain Learning

Wavelet transform is an effective spatiotemporal analysis tool [29]. In recent years, many neural network models have been proposed to learn in the wavelet domain. Many researchers mainly adopt the wavelet transform characteristics for image compression [30], [31], image super-resolution [31], image fusion [32], and image restoration [33] and so on. In [34], it mainly uses the lifting scheme of wavelet transform to convert images into coefficients without losing any information and achieve lossless compression. [35] proposes a flexible wavelet-based convolutional neural network for image super-resolution. It adopts wavelet transform to capture the global information of the face and the local texture information. Besides, the super-resolution deep learning framework is designed for wavelet coefficient prediction tasks. [36] proposes the wavelet channel attention module integrated network for single image denoising.

Of course, some researchers mainly adopt the wavelet transform for the image classification task, such as [37], [38]. Among them, [37] mainly uses the features extracted by 2D DWT and combines them with convolutional neural network features for image classification. In [38], wavelet pooling is proposed and applied to the neural networks. It uses Level-2 wavelet decomposition to obtain four subbands. Then, the IDWT result of the low-frequency sub-band is taken as the result of wavelet pooling, removing the high-frequency sub-bands. It effectively reduces the feature dimension and avoids overfitting caused by max pooling. Besides, the previous wavelet channel and wavelet spatial attention [36], [39] directly integrate wavelet composition features with CNNs.

In addition, some existing works related to wavelet attention arise, including [40], [41], [36], [42]. Among them, [40] uses wavelet transform as a preprocessing method, which pre-separates low-frequency and high-frequency features and employs channel and spatial attention modules to rescale the features during training adaptively. [36] proposed a wavelet channel attention module with a fusion network for single-image rain removal. DWT and IDWT extract various frequency features to replace the downsampling and upsampling. [42] proposed to use the wavelet attention block to capture only the detailed information of the feature map in the high-

frequency components, which can guide the model to filter out the useless information in the low-frequency domain. In addition, a wavelet attention embedded network for video super-resolution is proposed. It includes a wavelet embedded network for spatial features and an attention embedded network for temporal features. In a word, wavelet transform can obtain the high-pass and low-frequency components and captures the structural features.

III. Dual Wavelet Attention Networks for Image Classification

The dual wavelet attention networks for image classification are designed based on the dual wavelet attention block. It rethinks the attention mechanism from the wavelet domain. Firstly, the sum of DWT's low-frequency subband coefficients proved equivalent to global average pooling (GAP). On the basis of the aforementioned theory, the wavelet channel attention adapts the scalar with high-frequency information kept to distribute the better weight of different channels. In addition, since the wavelet transform also captures spatial features, the proposed method extends the ideas of FcaNet to spatial attention.

A. 2D DWT and Global Average Pooling

For an image $f(x, y)$ $(M \times N)$, the discrete cosine transform (DCT) coefficients can be defined as

$$B_{mn} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cos \frac{\pi(2x+1)m}{2M} \cos \frac{\pi(2y+1)n}{2N}, \tag{1}$$

where $0 \leq x \leq M-1, 0 \leq y \leq N-1$.

For an image $f(x, y)$ $(M \times N)$, the discrete wavelet transform (DWT) coefficients after the scale function can be generated by

$$W_\varphi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cdot \varphi_{j_0,m,n}(x, y), \tag{2}$$

where $\varphi_{j_0,m,n}(x, y)$ represents the scale functions. The scale function of orthogonal wavelet basis satisfies as $\varphi(x, y) = \varphi(x)\varphi(y)$. Thus,

$$W_\varphi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cdot \varphi_{j_0,m}(x) \cdot \varphi_{j_0,n}(y), \tag{3}$$

it equals to the convolutional operations with the scale kernels.

When $p = q = 0$, $B_{00} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)$, which has been proved to be equal to a special case of global average pooling (GAP) [17]. Similarly, for the **Haar** wavelet transform, $\varphi_{j_0,m,n}(x) \cdot \varphi_{j_0,m,n}(y) = 1$ in its shifting window, which is equal to the average pooling in the sub-domain. In addition, the $\sum_{m,n \in 0,1,2,\ldots,2^{j-1}} W_\varphi(j_0, m, n)$ can be formulated as

$$\begin{aligned} \sum_{m,n \in 0,1,2,\ldots,2^{j-1}} W_\varphi(j_0, m, n) &= \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \\ &= GAP(f(x, y))\sqrt{MN}, \end{aligned} \tag{4}$$

where $W_\varphi(j_0, m, n)$ represents the low-frequency components of 2D DWT with Haar wavelet basis. Thus, it is clear that the sum of low-frequency subband feature coefficients is equivalent to GAP.

The DCT and DWT transform results of **Lena** are shown in Fig. 1. Compared with DCT, DWT can maintain spatial structural features. Inspired by this, dual wavelet attention extends the channel attention to the spatial attention in the frequency domain.

B. Dual Wavelet Attention

Based on the theoretical analysis above, we propose a novel attention mechanism in the wavelet domain named dual wavelet attention. The details of the dual wavelet attention block are shown in Fig. 2. It mainly consists of the wavelet channel attention module and the wavelet spatial attention module. The dual wavelet attention block mainly effectively utilizes the low-frequency and high-frequency features to construct the attention model.

For the input feature F, the output refined feature F'' of the dual wavelet attention block can be formulated as

$$F' = M_{Wc}(F) \bigotimes F, \tag{5}$$

$$F'' = M_{Ws}(F') \bigotimes F', \tag{6}$$

where $\bigotimes$ denotes element-wise multiplication. M_{Wc} and M_{Ws} are respectively the wavelet channel attention map and the wavelet spatial attention map.

1) Wavelet Channel Attention: The wavelet channel attention module mainly uses DWT to decompose the features into high-frequency and low-frequency parts. It uses the statistical features of high-pass and low-frequency features as channel scalars, replacing the GAP operation in SENet, as shown in Fig. 3 (b).

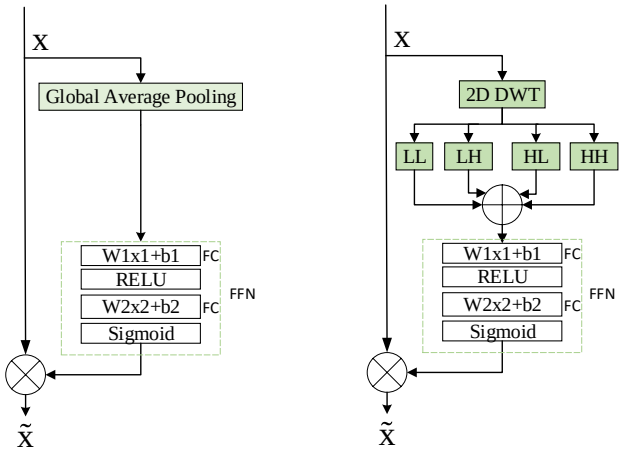

Fig. 3. Illustration of existing channel attention (SENet) and wavelet channel attention module. Wavelet channel attention uses the statistical features of high- and low-frequency features as channel scalars, replacing the global average pooling method in SENet. Different parts are marked with close green.

In detail, the proposed WC attention module can be expressed as follows. Firstly, 2D DWT can decompose the

original input into the approximation (LL), horizontal detail (LH), vertical detail (HL), and diagonal detail (HH) subbands, as shown in Fig. 1.

$$LL, LH, HL, HH = 2DDWT(F), \tag{7}$$

where LL, LH, HL, and HH are the four subbands of level-2 2D DWT decomposition. The value of each map represents the wavelet coefficients.

And then, the statistical features are obtained by aggregating the four subbands' coefficients as channel scalars. It can be formulated as

$$Wc = \sum_{i=0}^{M/2} \sum_{j=0}^{N/2} (LL + LH + HL + HH), \tag{8}$$

where $+$ represents element-wise summation.

And then, the simple feed forward networks (FFN) is adopted before the output of the wavelet channel attention module. The FFN is designed with two fully-connection layers with the RELU inserted followed by the Sigmoid function. The attention map of the wavelet channel attention can be expressed as

$$M_{Wc} = Sigmoid(FC(RELU(FC(Wc)))). \tag{9}$$

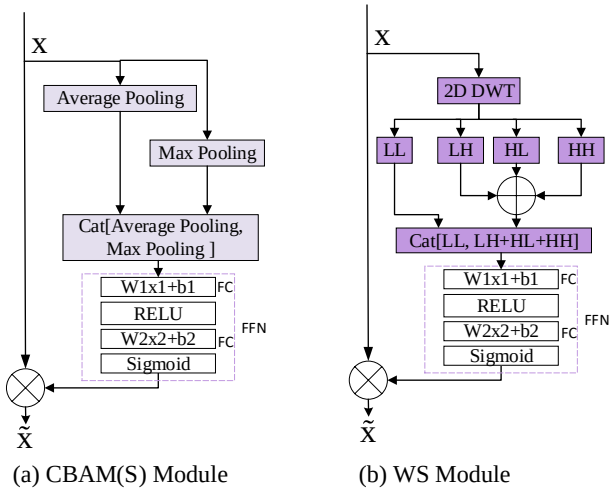

Fig. 4. Illustration of existing spatial attention module in CBAM (CBAM (S)) and wavelet spatial attention module. The wavelet spatial attention uses the high- and low-frequency features to replace the average pooling and max pooling in CBAM. Different parts are marked with close purple.

2) Wavelet Spatial Attention: The wavelet spatial attention module mainly uses the high and low-frequency feature subbands after wavelet decomposition to have the property of extracting images' spatial features. The features of image key points and structure can be obtained by wavelet decomposition. The details of our proposed wavelet spatial attention module are shown in Fig. 4 (b), compared with the CBAM's spatial attention module (shown as Fig. 4 (a)).

The proposed wavelet spatial attention framework can be expressed as follows. The low-pass and the high-pass subbands of 2D DWT are aggregated as two different spatial

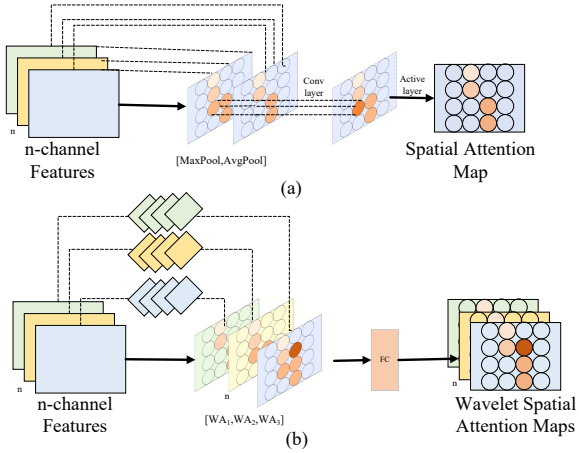

Fig. 5. The comparison of (a) spatial attention module and (b) the wavelet spatial attention module for n-channel features. The proposed wavelet spatial attention can generate adaptive spatial attention matrices $[WA_1, WA_2, WA_3]$ for different channels.

information context descriptors. The aggregation features can be formulated as

$$Ws = Cat[LL, LL + LH + HH], \tag{10}$$

where Cat represents the concatenation operation. And the attention map of the wavelet spatial attention can be expressed as

$$M_{Ws} = Sigmoid(FC(RELU(FC(Ws)))). \tag{11}$$

Compared with the original spatial attention methods with average pooling and max pooling features, aggregating low-frequency and high-frequency features can provide feature maps with more key features. In addition, the comparison of traditional spatial attention and wavelet spatial attention is shown in Fig. 5. The difference is that the wavelet spatial attention module can generate the attention maps for different channels, which can better curve the spatial attention for each channel. Besides, it can remain more abundant spatial information than the average pooling and max pooling operations to some extent from the compressed view. As Fig. 5 shows, the proposed wavelet spatial attention module (shown as Fig. 5 (b)) generates different spatial attention maps for each channel of the input feature, which is different from the initial spatial attention module (shown as Fig. 5 (a)).

C. The Architecture of the Dual Wavelet Attention Networks

The proposed wavelet channel attention, wavelet spatial attention, and dual wavelet attention can all be embedded in different feature layers of the convolutional networks. Taking ResNet as an example, three different attention modules are applied after different layers. The detailed architecture of the dual wavelet attention networks is shown in Fig. 6.

Fig. 7 (a) shows the network construction mode of ResNet with wavelet channel attention, Fig. 7 (b) presents the network construction mode of ResNet with wavelet spatial attention, and Fig. 7 (c) is the network construction mode of ResNet with dual wavelet attention. The order of wavelet channel

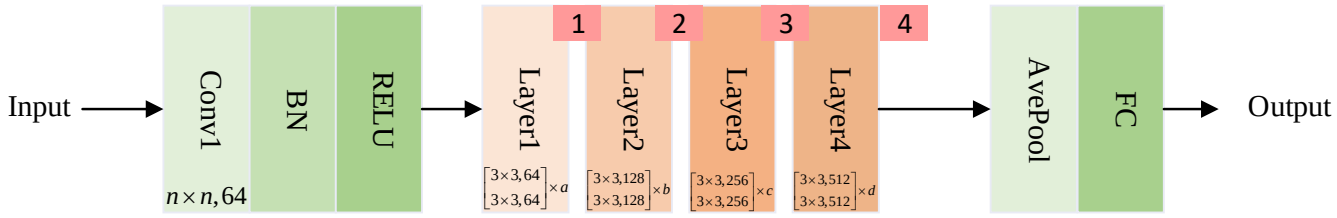

Fig. 6. An example of the architecture of the dual wavelet attention networks. When input is 32×32, $n = 3$. When input is 224×224, $n = 7$. ResNet18 is of $a = b = c = d = 2$. ResNet50 is with $a = 3$, $b = 4$, $c = 6$, $d = 3$. In addition, $1, 2, 3, 4$ in pink block represents the inserted position of the proposed wavelet attention module.

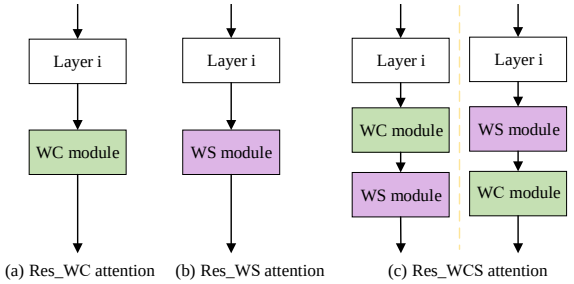

Fig. 7. An example of the proposed wavelet attention combined with CNNs networks. (a) Res_WC attention represents that the wavelet channel attention module is embedded in the network layer. (b) Res_WS attention represents the wavelet spatial attention module is embedded in the network layer. (c) Res_WCS attention represents that the dual wavelet attention module is embedded in the network layer.

attention and wavelet spatial attention can be adjusted while constructing the dual wavelet attention networks. It should be noted that the wavelet decomposition itself owns a down-sampling effect, so the conventional upsampling is adopted when computing the wavelet spatial attention map. In this way, the attention map is the same size as the original image and can be calculated point-to-point.

IV. Experiments

The experimental settings and results are mainly detailed and analysed in this section.

A. Experimental Settings

1) Data Sets: The proposed methods experiment on four benchmark data sets for nature image classification, including CIFAR10 [43], CIFAR100 [43], SVHN [44], and ImageNet [45]. The CIFAR10 and CIFAR100 include data sets of 10 and 100 categories, respectively, with 60,000 images and training set to test set ratio of 5:1. SVHN mainly contains a data set of 10 categories obtained from house numbers in Google Street View imagery. This paper adopts the standard split of training and testing. The 73,257 images were used for training and about 26,032 images for testing. Three data sets above are of 32×32 resolution. ImageNet2012 is adopted, including 1000 categories of digital images with a resolution of 224×224. A validation set of 5,0000 images (1000 categories) is used to test the model performance.

Besides, three benchmark data sets for remote sensing scene classification are used to validate the effectiveness of the proposed module, including WHU-RS19, AID, and the UC_Merced Land Use data sets (UCM). The characteristics of these three data sets are shown in Table I. There are 19, 30, and 21 categories, respectively in the three data sets above. WHU-RS19 includes about 1,005 images, AID contains about 10,000 images and the UCM data set mainly includes about 2,100 images. All three data sets above are resized 224×224 as the network's input. Randomly splitting of the data sets is adopted at the percentage of 80% for training and 20% for testing or 20% for training and 80% for testing.

In addition, data augmentation is adopted for all the data sets in our experiments. It consists of random cropping and horizontal flip operations on the aforementioned training data sets. Meanwhile, data normalization is used for all data sets.

2) Implementation Details: **End-to-end Training**. For the 1D signal, the low-frequency part s_1 and the high-frequency part d_1 are obtained after 1D DWT decomposition. Thus, 1D DWT decomposition can be expressed as $s_1 = Ls, d_1 = Hd$, where L and H are respectively the low-pass and high-pass filter matrices. The backpropagation of DWT is associated with the gradients $\frac{\partial s_1}{\partial s}$ and $\frac{\partial d_1}{\partial d}$, which are expressed as $\frac{\partial s_1}{\partial s} = L^T, \frac{\partial d_1}{\partial d} = H^T$. The forward and backward propagation of 2D DWT are slightly more complicated, but similar to 1D DWT [49]. What's more, the proposed models are trained end-to-end.

Experimental Setup. The SGD optimizer is used to train our models for image classification with a momentum of 0.9 with an initial learning rate of $lr = 0.03$. The model is trained on CIFAR10, CIFAR100, SVHN, UCM, AID, and WHU-RS19 for 300 epochs, with the learning rate (0.01) decreased at 150 and 225 epochs. The model was trained on the ImageNet data set for 90 epochs, and the learning rate was decreased at 30 and 60 epochs. The batch size is 32 for training CIFAR10, CIFAR100, SVHN, and the three remote sensing scene classification data sets, and 512 for ImageNet. All training is done on four Nvidia V100 4 GPUs and Pytorch 0.4.1. All experiments were evaluated more than three times, and the average value was taken as the final result.

3) Contrast Methods: There are mainly about ten methods selected as the contrast methods. ResNet18 is selected as the primary baseline model to validate the performance of the proposed methods. In addition, the proposed dual wavelet

TABLE I
CHARACTERISTICS OF THE BENCHMARK DATA SETS MENTIONED IN EXPERIMENTS.

Data sets	CIFAR10 [43]	CIFAR100 [43]	SVHN [44]	ImageNet [45]	UCM [46]	WHU-RS19 [47]	AID [48]
Images per class	6000	600	-	-	100	∼50	200∼400
Scene class	10	100	10	1000	21	19	30
Total images	60,000	60,000	600,000+	14,197,122	2,100	1,005	10,000
Spatial resolution(m)	-	-	-	-	0.3	up to 0.5	0.5∼0.8
Image sizes	32×32	32×32	32×32	224×224	256×256	600×600	600×600

TABLE II
PERFORMANCE COMPARISONS (%) ON THE CIFAR10, CIFAR100, AND SVHN DATA SETS. PARAM IS THE TRAINING PARAMETERS WITH THE UNIT M.

Models	CIFAR10	SVHN	Param.(M)	CIFAR100	Param.(M)
ResNet18	94.52	95.67	11.17	75.54	11.20
ResNet50	95.20	96.30	21.40	77.30	22.56
Se-ResNet18	95.33	96.20	11.32	77.29	12.47
Se-ResNet50	95.38	96.39	21.42	77.46	22.57
FcaNet	88.67	95.89	11.27	58.52	11.31
CBAM	95.76	96.22	11.32	77.55	12.47
WavePooling	80.28	91.10	-	-	-
WCNN L3	89.85	-	2.22	65.17	2.28
Scatter+WRN	92.31	-	45.58	72.26	45.67
DAWN(256init)	93.34	96.01	13.23	74.04	13.53
WaveCNet	85.02	95.85	11.69	55.08	11.7
ResNet18 + WC	95.54	96.49	11.31	77.56	12.47
ResNet18 + WS	95.37	**96.55**	11.31	77.78	12.47
ResNet18 + WC+WS	**95.78**	96.39	11.32	**77.85**	12.49
ResNet50 +WC	95.64	96.57	21.42	**78.60**	22.57
ResNet50 +WS	95.73	96.63	21.43	78.45	22.58
ResNet50 +WC+WS	**95.89**	**96.80**	21.45	78.48	22.60

attention integrates channel attention and spatial attention in the wavelet domain. Thus, the typical channel attention model SENet [13], the typical spatial attention model CBAM [15] and the novel frequency-domain attention model named FcaNet [17] are selected as the contrast methods. In addition, several wavelet transform-related methods have also been compared, including WavePooling [38], WaveCNet [49], WCNN with Level-3 DWT composition (WCNN L3) [37], the Scattering network with a hybrid configuration (Scatter+WRN [50]), and DAWN with 256 initial convolutional layers [51].

TABLE III
PERFORMANCE COMPARISONS (%) OF THE PROPOSED METHODS AND THE BASELINE MODEL RESNET18 AND RESNET50 ON IMAGENET.

Models	Top-1.err	Param.(M)
ResNet18	33.51	11.690
ResNet18+WC (L2)	32.42	11.705
ResNet18+WS (L2)	32.27	11.714
ResNet18+WS+WC (L2)	**31.25**	11.731
ResNet50	29.97	21.798
ResNet50+WC (L2)	29.01	21.814
ResNet50+WS (L2)	29.05	21.822
ResNet50+WS+WC (L2)	**28.93**	21.839

B. Experimental Results

The proposed methods are mainly evaluated on the CIFAR10, CIFAR100, and SVHN data sets, compared to other methods. And then, its effectiveness is validated to determine whether it can be expanded to the large ImageNet data set and the remote sensing data sets. In addition, the ablations of the two sub-modules and the embedded position are explored in this section.

1) Results on CIFAR10, CIFAR100 and SVHN: The accuracy comparison of different comparison models on each experimental data set is shown in Table II. Experimental results show that our proposed dual wavelet attention outperforms wavelet channel and spatial attention. Compared with SENet and CBAM, the performance of our model also performed relatively well on CIFAR10, CIFAR100 and SVHN. SENet only utilizes global average pooling as a channel scalar and loses some details. And our proposed channel attention not only contains the global average pooling results and the statistical features of the high-frequency detail components of the features. Therefore, the channel scalar can better reflect the difference between different channels.

In addition, compared with spatial attention in CBAM, our proposed wavelet spatial attention uses low-frequency and high-frequency features as aggregated features, replacing the previous aggregation methods of average pooling and max pooling. The wavelet spatial attention can capture more abundant and details features than the features extracted by the average pooling and max pooling. Thus, the proposed methods obtain better performance. Compared with the frequency-channel attention method FcaNet, the proposed methods show about 7.11% and 19.33% improvements on the CIFAR10 and the CIFAR100 data sets with comparable parameters. Besides, the performance of some wavelet-based methods are shown in Table II too, including WavePooling [38], WCNN L3 [37], Scatter+WRN [50], DAWN with the initial convolutional layers at 256 [51] and WaveCNet [49]. It shows that the proposed

methods greatly improved performance with nearly proper parameters. As for the convergency, the training loss curves of ResNet18+WC+WS (WCS) on CIFAR10, CIFAR100, and SVHN are shown as Fig. 8. It shows that the proposed method is of good convegence.

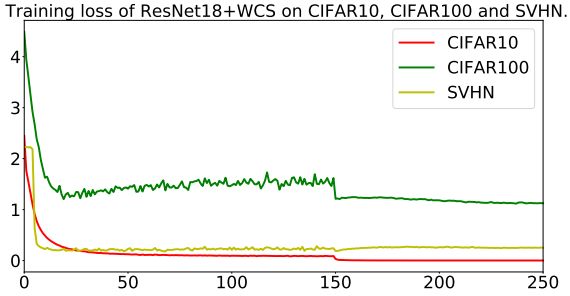

Fig. 8. Training loss curves of ResNet18+WCS model on CIFAR10, CIFAR100, and SVHN data sets.

2) Results on ImageNet Classification: On the ImageNet data set, we mainly compare the performance of the proposed wavelet channel attention, wavelet attention and wavelet composite attention with the performance of the benchmark network ResNet18 and ResNet50. The corresponding performance comparison is shown in the Table III.

It shows that the proposed three modules have improved performance on the ImageNet data set compared to the ResNet18 baseline. The final ResNet18 embedded dual wavelet attention achieves a top-1 error rate of 33.25%. It exhibits a performance improvement of 2.26% with 11.731M parameters. In addition, The final ResNet50 embedded dual wavelet attention achieves 71.07% accuracy on ImageNet, which shows about 1.04% increased precision than that of ResNet50.

3) Experiments on Remote Sensing Scene Classification: As Table IV shows, the proposed methods are validated on three benchmark data sets for remote sensing scene classification, mainly compared with the baseline ResNet18. Under relatively fair experimental conditions, it shows that the proposed methods performed well and obtained about 2.1%, 4.89%, and 6.0% improvement on AID, WHU-RS19, and the UCM data sets. Besides, it suggests that the proposed methods show better performance when the samples are of 20% for training instead of the 80% for training.

TABLE IV
PERFORMANCE COMPARISONS (%) OF THE PROPOSED METHODS AND THE BASELINE MODEL RESNET18 ON THE AID, UCM, AND WHU-RS19.

Data sets	AID		UCM		WHU-RS19	
Ratio for training	80%	20%	80%	20%	80%	20%
ResNet18	92.25	80.48	94.02	68.16	92.23	74.34
ResNet18+WC	91.95	82.51	94.05	71.55	**95.63**	75.34
ResNet18+WS	92.05	81.33	94.04	**73.04**	93.21	76.35
ResNet18+WC+WS	**93.85**	**82.64**	**94.77**	72.26	93.20	**77.97**

Besides, the confusion matrices of the proposed methods on the WHU-RS19 data set (80% for testing) are given in Fig. 9. The confusion matrix of the baseline and the baseline with the proposed methods, including the wavelet channel attention module, wavelet spatial attention module and the dual wavelet attention module, are detailed in Fig. 9 (a)-(d). The corresponding detailed categories are Airport (0), Beach (1), Bridge (2), Commercial (3), Desert (4), Farmiand (5), footballField (6), Forest (7), Industrial (8), Meadow (9), Mountain (10), Park (11), Parking (12), Pond (13), Port (14), railywayStation (15), Residential (16), River (17) and Viaduct (18). It shows that the proposed methods are effective in distinguishing the categories of the remote sensing scene, especially for the Viaduct (18), Residential (16), Port (14) and so on. Especially for Viaduct (18), the accuracy is increased a lot from the 56% to 100%.

As shown in Table II-IV, it shows that the proposed methods perform well on both natural image and remote sensing data sets. In particular, it shows much-improved accuracy on the remote sensing data sets when the training samples are fewer. Besides, using the dual wavelet attention model can achieve a relatively stable performance improvement. When wavelet channel attention or wavelet spatial attention module is used, it may achieve better performance for specific data set or task. This is related the characteristics of the specfic data set or task. In addtition, the proposed method performs well nearly without increased parameters.

C. Ablations

The proposed dual wavelet attention mechanism is mainly composed of wavelet channel attention and wavelet spatial attention. Therefore, we explore the influence of these two submodules on the final network performance. Meanwhile, the proposed attention modules can be embedded into a different position in the benchmark network. Therefore, the impact of embedded the proposed modules in each layer or all layers is ablated here.

1) The effects of wavelet channel attention and wavelet spatial attention modules: In Table V, it shows that both wavelet spatial attention and wavelet channel attention can improve the network performance compared to the ResNet18 benchmark network. Compared with the proposed wavelet channel attention, the proposed wavelet spatial attention performance is more improved. That is, the wavelet spatial attention module dominates the performance of dual wavelet attention. The main reason is that our proposed wavelet spatial attention can not only effectively aggregate low-frequency and high-frequency key features of images, but also each channel does not share spatial features. That is, each channel has its spatial attention map in a sense. Besides,

2) The effect of different embedded positions on model performance: The wavelet channel attention (WC), the wavelet spatial attention (WS), and the dual wavelet attention module (WCS or WC+WS) are all explored after adding different layers on CIFAR10, CIFAR100, and SVHN when all layers are embedded. The results are shown in Table V. All layers that represent the proposed modules are inserted behind each layer. Layer 1 (L1), Layer 2 (L2), Layer 3 (L3), and Layer 4 (L4) represent that the proposed modules are inserted behind the specific layer. It suggests that the performance is improved

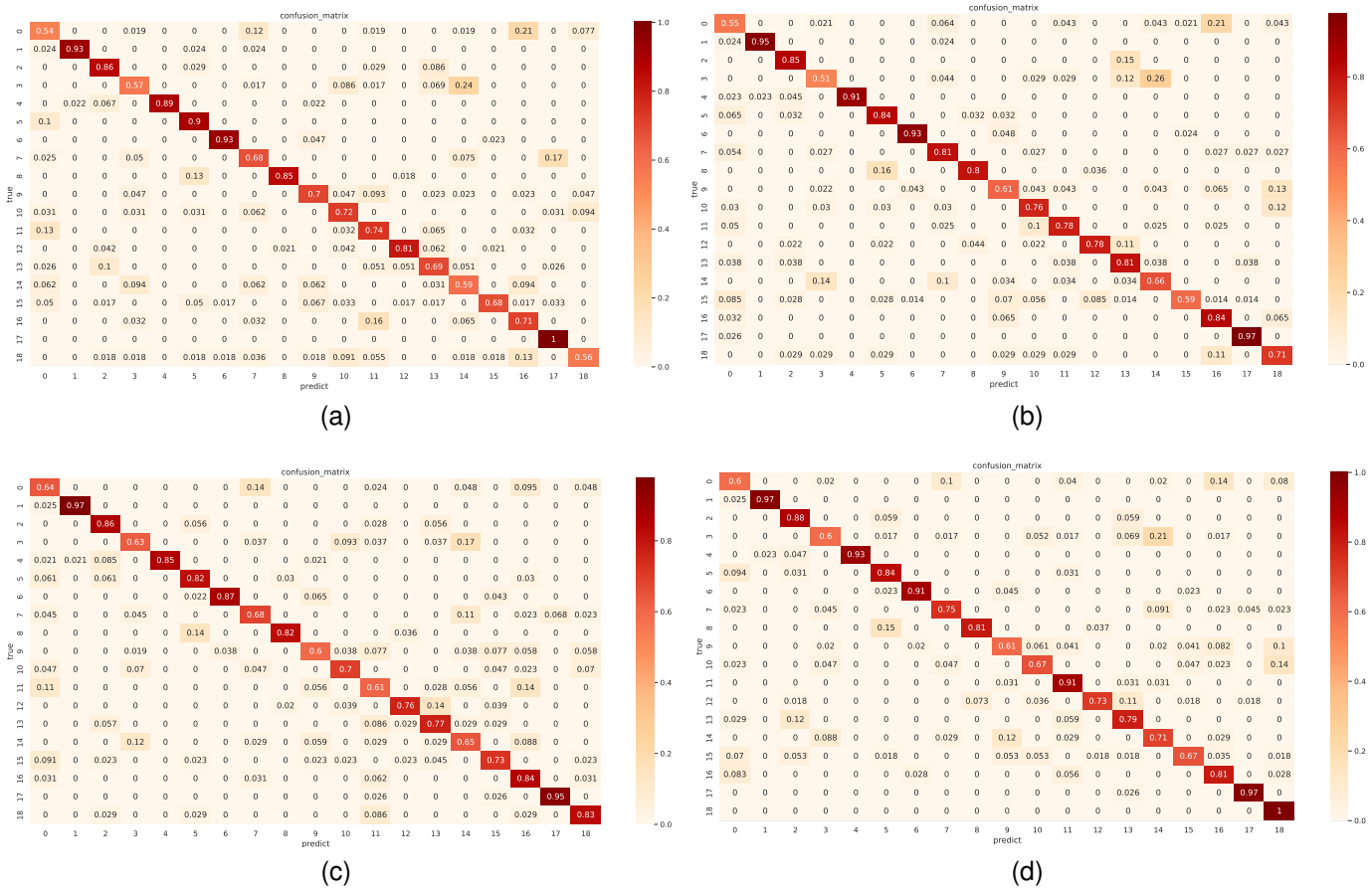

Fig. 9. The confusion matrices of the proposed methods on the WHU-RS19 data set (80% for testing). (a) the confusion matrix of ResNet18 on WHU-RS19 data set. (b) the confusion matrix of ResNet18 with wavelet channel attention on WHU-RS19 data set. (c) the confusion matrix of ResNet18 with wavelet spatial attention on WHU-RS19 data set. (d) the confusion matrix of ResNet18 with dual wavelet attention on WHU-RS19 data set.

TABLE V
COMPARE PRECISION (%) OF THE PROPOSED METHODS WITH DIFFERENT EMBEDDED POSITIONS ON THE CIFAR10, CIFAR100, AND SVHN DATA SETS.

Models	CIFAR10	CIFAR100	SVHN
ResNet18	94.52	75.54	95.67
ResNet18+WCS (all layers)	95.41	75.91	96.27
ResNet18+WCS (L1)	95.43	77.77	96.11
ResNet18+WCS (L2)	95.55	**77.85**	96.37
ResNet18+WCS (L3)	**95.78**	76.71	96.14
ResNet18+WCS (L4)	95.48	77.28	**96.39**
ResNet18+WC (all layers)	94.84	76.74	**96.49**
ResNet18+WC (L1)	95.02	77.32	95.13
ResNet18+WC (L2)	95.32	77.28	96.06
ResNet18+WC (L3)	94.97	77.4	95.85
ResNet18+WC (L4)	**95.54**	**77.56**	96.36
ResNet18+WS (all layers)	95.36	76.25	**96.55**
ResNet18+WS (L1)	95.23	**77.78**	96.17
ResNet18+WS (L2)	**95.37**	77.77	96.31
ResNet18+WS (L3)	95.30	77.75	96.12
ResNet18+WS (L4)	95.28	76.38	96.52

when the proposed modules are embedded behind one layer or after all layers, compared with the baseline network. It should be noted that not all layers are added after the proposed attention module obtains the greatest improvement. In addition, the position to insert the proposed wavelet attention module is not fixed to obtain better performance for different data sets.

D. Visualization

We provide activation maps of testing images for each class of CIFAR10, as shown in Fig. 10. The visualization includes the results for the category activation map (**Cam**), the class activation map of the backpropagated Relu model (**gb**) and the overlay effect map of the category activation map and the backpropagation activation map (**Cam-gb**) [52], [53]. Compared the Fig. 10 (a), (b), and (c), it can be seen that the proposed wavelet channel attention not only pays attention to the target and the background features of the surrounding parts of the target. The wavelet spatial attention pays more attention to the target, including the target's contour and other information, so that the attention is focused on a more comprehensive target.

Besides, three samples of WHU-RS19 are selected to show the feature visualization, including Airport (0), Residential (16), and Viaduct (18) (shown in Fig. 11). Comparing the **gb** and **Cam-gb** results shows that the proposed methods are adaptive to complex remote sensing scene data sets. It is more evident that wavelet spatial attention can keep the critical structural feature and suppress the background feature, thus improving the accuracy of complex scene classification.

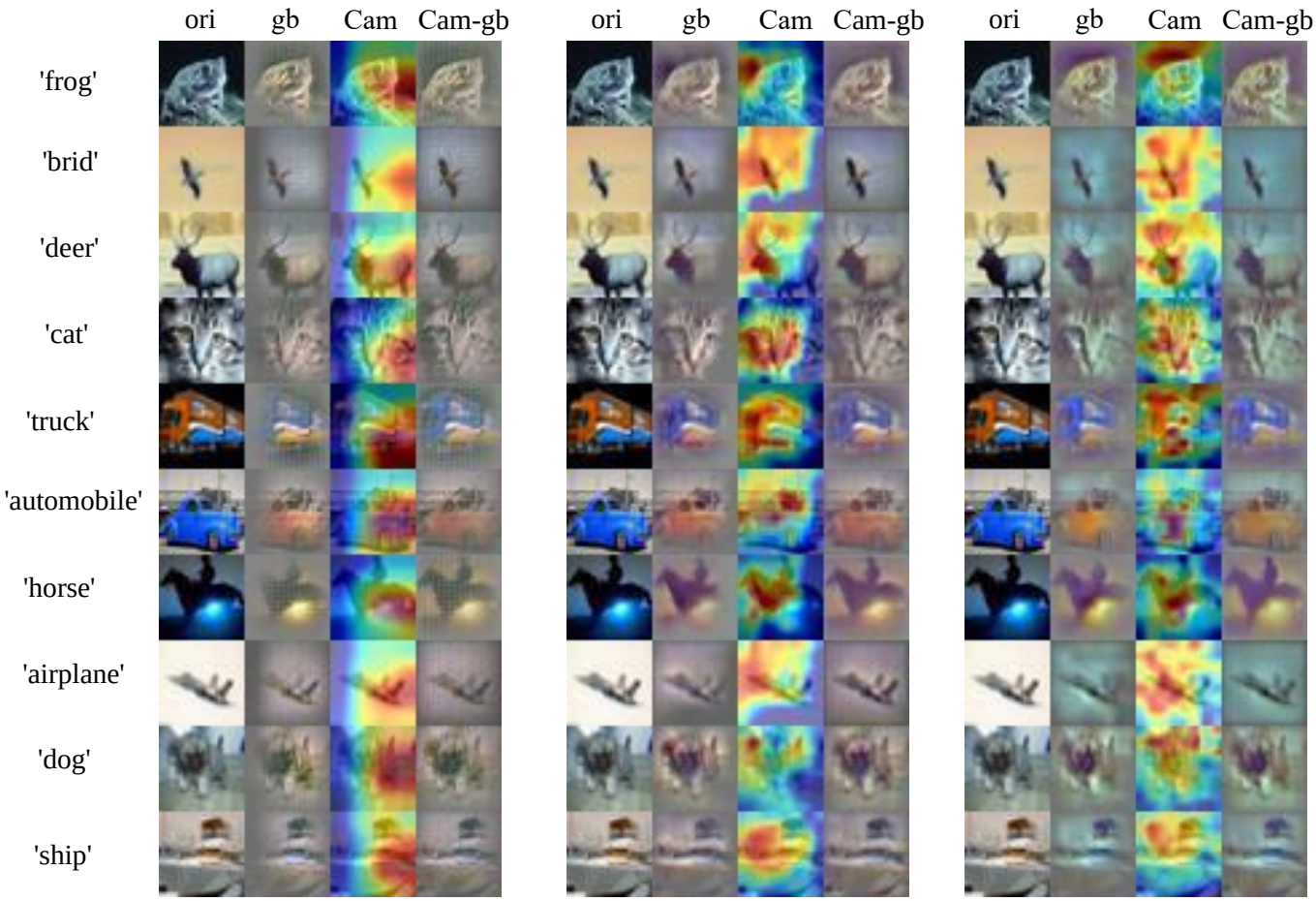

Fig. 10. Model category activation visualization for wavelet attention on CIFAR10 data set. (a) The class activation map of ResNet18. (b) The class activation map of ResNet18 with wavelet channel (WC) attention. (c) The class activation map of ResNet18 with wavelet spatial (WS) attention. Ori represents the original image of each category. Cam stands for the category activation map. gb represents the class activation map of the backpropagated Relu model. Cam_gb represents the overlay effect map of the category activation map and the backpropagation activation map. Red areas correspond to high scores kind.

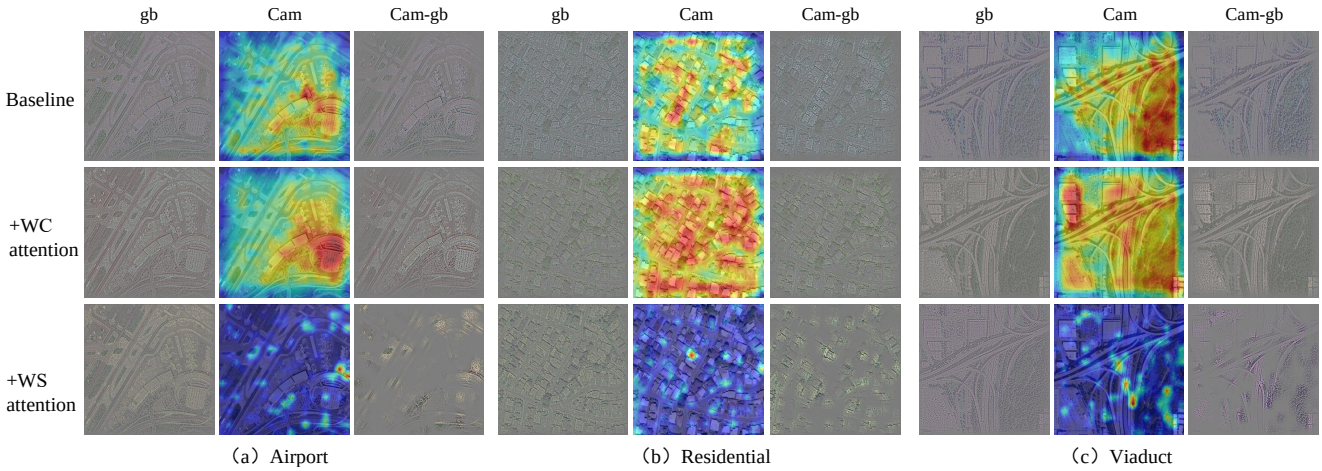

Fig. 11. Model category activation visualization for WHU-RS19 data set, including ResNet18, ResNet18 with wavelet channel (+WCand wavelet spatial (+WS) attention. The visualization results of three samples are shown, including (a) Airport (0), (b) Residential (16), and the (c) Viaduct (18).

V. Concluding Remarks

In this paper, visual attention is rethought in the wavelet domain. Furthermore, the dual wavelet attention block is proposed by exploiting the compression properties and the structural feature extraction capability of the 2D DWT.

The sum of the low-frequency subband coefficients of the Haar wavelet is proved as a specific case of GAP. Thus, the traditional channel attention only adopts the low-frequency statistical features as the channel scalars. However, the proposed wavelet channel attention considers low-frequency and high-frequency statistical features as the channel scalars. It can better characterize the importance of different channels. Besides, the proposed wavelet spatial attention can obtain the critical structural features in the attention feature maps. Compared with traditional spatial attention, wavelet spatial attention shows better performance in retaining structural features. Experimental results demonstrate that the proposed models outperform the baseline and the existing attention models. Besides, the proposed methods can be applied to the large ImageNet data set and remote sensing scene classification.

In addition, the dual wavelet attention networks show better feature representation ability. Thus, the performance of the proposed methods can be explored in many other fields. In the future, we would like to continue studying this direction to be adaptive to more visual tasks.

References

[1] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for ai," *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, 2021.

[2] M. Yang, P. Kumar, J. Bhola, and M. Shabaz, "Development of image recognition software based on artificial intelligence algorithm for the efficient sorting of apple fruit," *International Journal of System Assurance Engineering and Management*, vol. 13, no. 1, pp. 322–330, 2022.

[3] S. Lian, W. Jiang, and H. Hu, "Attention-aligned network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3140–3153, 2021.

[4] Y. Cao, H. Ji, W. Zhang, and S. Shirani, "Feature aggregation networks based on dual attention capsules for visual object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 674–689, 2022.

[5] S. Rodzin, V. Bova, Y. Kravchenko, and L. Rodzina, "Deep learning techniques for natural language processing," in *Computer Science Online Conference*. Springer, 2022, pp. 121–130.

[6] Z. Guo, C. Wang, G. Yang, Z. Huang, and G. Li, "Msft-yolo: Improved yolov5 based on transformer for detecting defects of steel surface," *Sensors*, vol. 22, no. 9, p. 3467, 2022.

[7] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021.

[8] Y. Mo, Y. Wang, C. Xiao, J. Yang, and W. An, "Dense dual-attention network for light field image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[9] Y. Hao, S. Wang, P. Cao, X. Gao, T. Xu, J. Wu, and X. He, "Attention in attention: Modeling context correlation for efficient video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[10] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1901–1914, 2021.

[11] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 773–786, 2018.

[12] S. Seo, S. Ki, and M. Kim, "A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2602–2616, 2020.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[14] S. Gao, L. Duan, and I. W. Tsang, "Defeatneta deep conventional image representation for image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 494–505, 2015.

[15] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[16] H. Lee, H.-E. Kim, and H. Nam, "Srm: A style-based recalibration module for convolutional neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1854–1862.

[17] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[18] E. Amiri, M. Rahmanian, S. Amiri, and H. Y. Praee, "Medical images fusion using two-stage combined model dwt and dct," *International Advanced Researches and Engineering Journal*, vol. 5, no. 3 (Under Construction), pp. 344–351, 2021.

[19] H. Bi, J. Yao, Z. Wei, D. Hong, and J. Chanussot, "Polsar image classification based on robust low-rank feature extraction and markov random field," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[20] C. He, S. Li, Z. Liao, and M. Liao, "Texture classification of polsar data based on sparse coding of wavelet polarization textons," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 8, pp. 4576–4590, 2013.

[21] H. Bi, L. Xu, X. Cao, Y. Xue, and Z. Xu, "Polarimetric sar image semantic segmentation with 3d discrete wavelet transform and markov random field," *IEEE transactions on image processing*, vol. 29, pp. 6601–6614, 2020.

[22] S. Yousefi, M. M. Shalmani, J. Lin, and M. Staring, "A novel motion detection method using 3d discrete wavelet transform," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3487–3500, 2018.

[23] J. Bian and Y. Liu, "Dual channel attention networks," in *Journal of Physics: Conference Series*, vol. 1642, no. 1. IOP Publishing, 2020, p. 012004.

[24] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[25] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 531–11 539.

[27] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[28] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 713–13 722.

[29] L. Jiao, Y. Yang, F. Liu, S. Yang, and B. Hou, "The new generation brain-inspired sparse learning: A comprehensive survey," *IEEE Transactions on Artificial Intelligence*, 2022.

[30] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2019.

[31] D. Mishra, S. K. Singh, and R. K. Singh, "Wavelet-based deep auto encoder-decoder (wdaed)-based image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1452–1462, 2020.

[32] J. Li, G. Yuan, and H. Fan, "Multifocus image fusion using wavelet-domain-based deep cnn," *Computational intelligence and neuroscience*, vol. 2019, 2019.

[33] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-cnn for image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 773–782.

[34] H. Ma, D. Liu, R. Xiong, and F. Wu, "iwave: Cnn-based wavelet-like transform for image compression," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1667–1679, 2019.

[35] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1689–1697.

[36] H.-H. Yang, C.-H. H. Yang, and Y.-C. F. Wang, "Wavelet channel attention module with a fusion network for single image deraining," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 883–887.
[37] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," *arXiv preprint arXiv:1805.08620*, 2018.
[38] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *International Conference on Learning Representations*, 2018.
[39] G. LIAO, Z. ZHANG, Y. NIU, and Z. SONG, "Vehicle re-identification based on multi-scale and wavelet spatial attention," *Journal of Jishou University (Natural Sciences Edition)*, vol. 42, no. 6, p. 15, 2021.
[40] S. Xue, W. Qiu, F. Liu, and X. Jin, "Wavelet-based residual attention network for image super-resolution," *Neurocomputing*, vol. 382, pp. 116–126, 2020.
[41] Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Wavelet attention embedding networks for video super-resolution," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7314–7320.
[42] X. Zhao, P. Huang, and X. Shu, "Wavelet-attention cnn for image classification," *Multimedia Systems*, vol. 28, no. 3, pp. 915–924, 2022.
[43] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[44] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.
[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
[46] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.
[47] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and remote sensing letters*, vol. 8, no. 1, pp. 173–176, 2010.
[48] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.
[49] Q. Li, L. Shen, S. Guo, and Z. Lai, "Wavecnet: Wavelet integrated cnns to suppress aliasing effect for noise-robust image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 7074–7089, 2021.
[50] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the scattering transform: Deep hybrid networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5618–5627.
[51] M. X. B. Rodriguez, A. Gruson, L. Polania, S. Fujieda, F. Prieto, K. Takayama, and T. Hachisuka, "Deep adaptive wavelet network," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3111–3119.
[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
[53] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

Yuting Yang (Student Member, IEEE) received the B.S. degree in electronic information science and technology from Northwest University, Xian, China, in 2018. She is majored in computer science and technology as a Ph.D. candidate of Xidian University, Xian, China.

She is currently a member of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, and Joint International Research Laboratory of Intelligent Perception and Computation, Xidian University, Xi'an, China. Her research interests include computer vision, the interpretability of deep learning and multiscale geometric analysis.

Licheng Jiao (Fellow, IEEE) received the B.S.degree from Shanghai Jiaotong University, Shanghai, China, in 1982 and the M.S. and PhD degree from Xian Jiaotong University, Xian, China, in 1984 and 1990, respectively.

Since 1992, he has been a distinguished professor with the school of Electronic Engineering, Xidian University, Xian, where he is currently the Director of Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. He has been a foreign member of the academia European and the Russian academy of natural sciences. His research interests include machine learning, deep learnig, natural computation, remote sensing, image processing, and intelligent information processing.

Prof. Jiao is the Chairman of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the fellow of IEEE/IET/CAAI/CIE/CCF/CAA, a Councilor of the Chinese Institute of Electronics, a committee member of the Chinese Committee of Neural Networks, and an expert of the Academic Degrees Committee of the State Council.

Xu Liu (Member, IEEE) received the B.Sc. degrees in Mathematics and applied mathematics from North University of China, Taiyuan, China in 2013. He received the Ph.D. degrees from Xidian University, Xian, China, in 2019.

He is currently a postdoctoral researcher of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, China. He is the chair of IEEE Xidian university student branch (2015-2019). His current research interests include machine learning and image processing.

Fang Liu (Senior Member, IEEE) received the B.S. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 1984 and the M.S. degree in computer science and technology from Xidian University, Xi'an, in 1995.

She is currently a Professor with the School of Computer Science, Xidian University. Her research interests include signal and image processing, synthetic aperture radar image processing, multiscale geometry analysis, learning theory and algorithms, optimization problems, and data mining.

Shuyuan Yang (Senior Member, IEEE) received the B.A. degree in electrical engineering, and the M.S. and Ph.D. degrees in circuits and systems from Xidian University, Xian, China, in 2000, 2003, and 2005, respectively.

She has been a Professor of Electrical Engineering with Xidian University. Her research interests include machine learning and multiscale geometric analysis.

Lingling Li (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2011 and 2017 respectively. Between 2013-2014, she was an exchange Ph.D. student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Leioa, Spain.

She is currently a Postdoctoral Researcher with the School of Artificial Intelligence, Xidian University. Her current research interests include quantum evolutionary optimization, machine learning, and deep learning.

Puhua Chen (Senior Member, IEEE) received the B.S. degree in environmental engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the Ph.D. degree in circuit and system from Xidian University, Xian, China, in 2016.

She is currently an associate professor with the School of Artificial Intelligence, Xidian University. Her current research interests include machine learning, pattern recognition and remote sensing image interpretation.

Xiufang Li (Student Member, IEEE) received the master's degree in safety science and engineering from Xian University of Science and Technology, Xi'an, China in 2017. She is currently pursuing the Ph.D. degree in circuit and system from Xidian University, Xi'an China.

Currently, she is a member of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, and international Research Center for Intelligent Perception and computation, Xidian University, Xi'an China. Her research interests include deep learning and images processing.

Zhongjian Huang (Student Member, IEEE) received the B.S. degree in intelligent science and technology from Xidian University, Xian, China in 2018. He is currently pursuing the PhD degree in computer science and technology from Xidian University, Xi'an China.

He is currently a member of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, and Joint International Research Laboratory of Intelligent Perception and Computation, Xidian University, Xi'an, China. His current research interests include video tracking and satellite videos analysis.