IECE

Chinese Journal of Information Fusion
http://dx.doi.org/10.62762/CJIF.2024.932552

RESEARCH ARTICLE

Explainable Classification of Remote Sensing Ship Images Based on Graph Network

Haoran Li[1,*], **Wei Xiong**[1], **Yaqi Cui**[1] **and Zhenyu Xiong**[1]

[1] Naval Aviation University, Yantai 264001, China

Abstract

Remote sensing image plays an important role in maritime surveillance, and as a result there is increasingly becoming a prominent focus on the detection and recognition of maritime objects. However, most existing studies in remote sensing image classification pay more attention on the performance of model, thus neglecting the transparency and explainability in it. To address the issue, an explainable classification method based on graph network is proposed in the present study, which seeks to make use of the relationship between objects' regions to infer the category information. First, the local visual attention module is designed to focus on different but important regions of the object. Then, graph network is used to explore the underlying relationships between them and further to get the discriminative feature. Finally, the loss function is constructed to provide a supervision signal to explicitly guide the attention maps and overall learning process of the model. Through these designs, the model could not only utilize the underlying relationships between regions but also provide explainable visual attention for people's understanding. Rigorous experiments on two public fine-grained ship classification datasets indicate that the classification performance and explainable ability of the designed method is highly competitive.

Keywords: explainable visual feature, remote sensing image, ship classification.

Citation
Li, H., Xiong, W., Cui, Y., & Xiong, Z. (2024). Explainable Classification of Remote Sensing Ship Images Based on Graph Network. Chinese Journal of Information Fusion, 1(2), 126–133.

Academic Editor:
Xingchen Zhang

Submitted: 30 July 2024
Accepted: 25 September 2024
Published: 29 September 2024

Vol. 1, **No.** 2, 2024.
10.62762/CJIF.2024.932552

***Corresponding author:**
✉ Haoran Li
rizhaolihaoran@163.com

1 Introduction

With the rapid development of remote sensing (RS) satellite technology, the quantity and quality of RS images have been significantly improved. Due to gradually realizing the significance of maritime rights and interests [1], a highly accurate and reliable RS ship images classification method is required.

The traditional RS ship images classification studies mainly focus on coarse-grained ship classification and most depend on shallow global features [2], which can only be used for simple classification task. In recent years, as the requirements for RS images classification becoming more detailed in military and civilian ocean resource utilization [3], fine-grained classification of RS ship images is becoming more and more significant. However, RS ship images can be confusing and difficult to distinguish, as the ships belonging to different category may appear very similar, and the content of RS images can be complex [4]. Fortunately, deep learning methods seem to be more effective and have made great achievement in computer vision field. In the task of fine-grained RS ship images classification, A push-and-pull network is proposed in [5], which

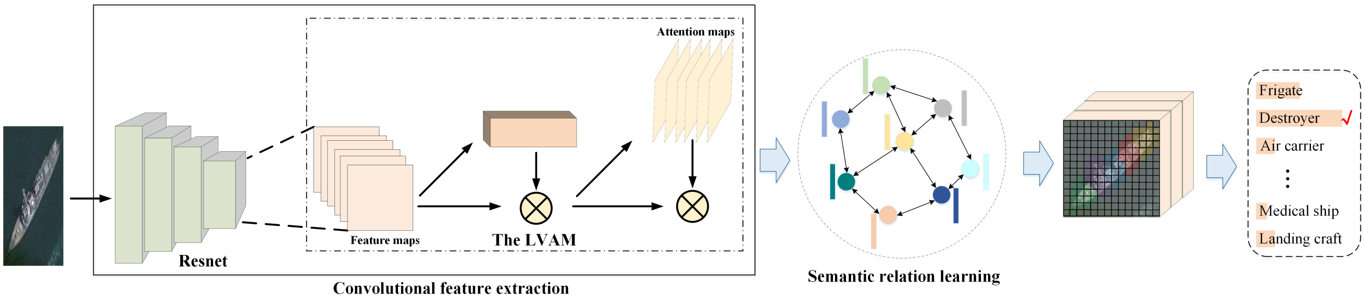


Figure 1. The framework of the proposed network.

integrated with the advances of contrastive learning to make the classification effective. Self-calibrated convolutions and class-balanced loss are introduced to the classification network by Chen et al. [6], to enrich features and overcome the class imbalance.

In addition to the powerful feature representation capability of the model, explainability is received the widespread attention in the last few years, which tries to make the decisions and mechanisms of the model more understandable to humans [7]. However, most of the existing fine-grained ship classification methods based on the end-to-end learning strategy provide infrequent interpretability, the models are only trained to match the datasets. If human could understand the region where models focus on or the mechanism how models make prediction in the task of fine-grained image classification, they may give more trust on the final predicted results. Visual attention mechanism has already been used in image classification to improve the performance of the model [8], and it also can yield heatmaps that indicate key regions for driving a model's decision. Notably, graph convolutional network (GCN) [9] becomes a hot network architecture, which could establish a graph structure by analyzing the relations between nodes [10]. Therefore, GCN could be used to capture and explore the intrinsic relationships of different regions in an image, contributing to improve the models' explainability. Some scholars have tried combining graph network with other networks to improve the performance or explainability of the model [11–13].

With the aforementioned consideration, we propose an explainable RS ship images classification framework based on graph network. In order to make people have a better understanding of the model's final prediction meanwhile not affecting the classification ability of the model, the local visual attention and graph network are combined in the proposed model to obtain the discriminative feature and explore the semantic relation of different parts of the target. As a result, the network can focus on the key regions of targets and then effectively learn the relations between them through graph structure. In contrast to alternative methodologies, our model could present the targets' parts it focused on and the relatively clear decision-making process without compromising the classification performance. The main contributions of this paper are as follows.

1) An explainable RS ship images classification network is proposed, to the best of our knowledge, we are the first to use the combination of convolutional neural network (CNN) and GCN in this field to explore the explainability of it.

2) By adding the local visual attention module (LVAM) to the feature extraction process, the network can pay attention to the key local regions which are used for final prediction. Moreover, attention maps may help users understand the decision-making progress from the human visual angle.

3) By introducing the GCN, it could help the proposed network capture the underlying association relationships between local regions that are important for the its final decision. The experimental results on two public ship classification datasets validate the classification ability and explainability of the proposed method.

2 Methodology

The overall framework proposed in this paper is shown in Figure 1, and there are mainly two parts, namely, convolutional feature extraction and semantic relation learning. The LVAM in feature extraction is to capture the important regions of the objects in input images, and then these local representations are further input to the relation learning part which utilizes GCN to exploit intrinsic relationships of the regions, providing more thorough explanations. Next, we will explain each of these components in detail.

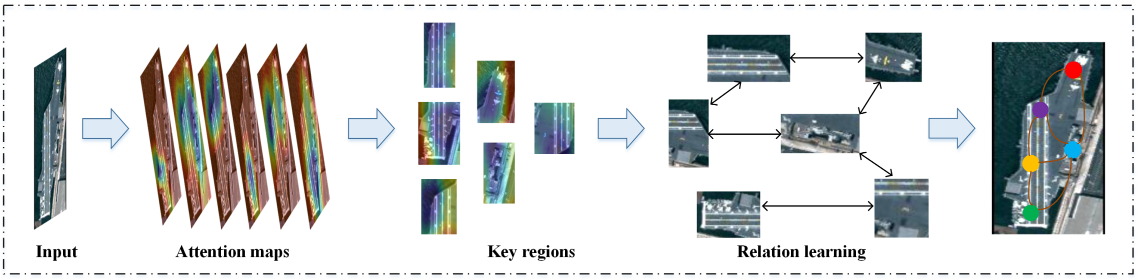


Figure 2. The schematic illustration of the learning process.

2.1 Local Visual Attention Module

Most existing deep neural network explanation methods only produce low-level attention maps, but they may not be intuitive for human to understand. The LVAM in the network is developed to capture key regions of the object and help improve model's explainability, and now we give a detailed introduction. As shown in Figure 1, when a RS ship image is input to the proposed network, the global high-level feature maps obtained from the last residual blocks of ResNet is represented as $\boldsymbol{F} \in \mathbb{R}^{W \times H \times C}$, where W, H, and C represent the weight, height, and number of channels of feature maps, respectively. First, different sizes convolutional kernels are used to obtain the fused feature maps:

$$\boldsymbol{F}_f = Conv_{1\times1}(\psi(Conv_{1\times1}(\boldsymbol{F}), Conv_{3\times3}(\boldsymbol{F}))) \tag{1}$$

where $Conv_{k\times k}$ denotes the size convolutional kernel, and $\psi()$ represents the cat operation between feature maps in the channel dimension.

Then, the LVAM adopts a simple structure to obtain local attention maps $\boldsymbol{L} = \phi(\boldsymbol{F}_f)$, $\boldsymbol{L} = \{\boldsymbol{L}_1, \boldsymbol{L}_2, \cdots, \boldsymbol{L}_R\} \in \mathbb{R}^{W \times H \times R}$, where $\phi()$ is the convolutional function with the kernel size of 3×3, and note that R is the outputting channel number which indicates the number of key regions and is set in advance. Finally, the key regions of $\boldsymbol{F}_f$ can be obtained according to the maximal value of the channel-wise position in $\boldsymbol{L}$. Furthermore, local attention loss function is designed to supervise this learning process, and we give detail introduction in Section 2.3.

2.2 Semantic Relation Learning

Based on the local attention maps, we could get the visual explanations and be aware of the important object parts which are beneficial to the final decision. In this step, the GCN is further used to mine the semantic relations between them so as to make the obtained feature representation is more discriminative. The GCN utilizes node features and neighborhoods' relations to extract advanced features. A graph is defined as $G = (V, E)$, where V is the set consisting of nodes and E represents the set of edges. Generally, the adjacency matrix $\boldsymbol{A}$ is used to represent the edge relations between nodes. About a single-layer GCN, the process of graph convolution can be defined as:

$$\boldsymbol{Y} = \sigma(\boldsymbol{A} \cdot \boldsymbol{X} \cdot \boldsymbol{W}) \tag{2}$$

where $\boldsymbol{X}$ is the input feature matrix, $\boldsymbol{Y}$ is the output feature matrix, $\boldsymbol{W}$ is the trainable matrix, and $\sigma()$ denotes the activation function.

As shown in Figure 2, in the proposed model, the regions obtained from the LVAM is regarded as the nodes to construct a graph. Specifically, according to the index of the maximal value in each local attention map, we average the fused features within each region as the graph nodes $\boldsymbol{x}_i \in \mathbb{R}^C$, which form the input feature matrix $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_R\}$. So as to capture the regions' interrelationship, the adjacency matrix $\boldsymbol{A}$ is defined as:

$$\boldsymbol{A}_{ij} = \mathrm{e}^{-d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2} \tag{3}$$

where $d()$ means the Euclidean distance metric, which is used to measure the resemblance between graph nodes. After the GCN updating nodes' features according to equation (2), a convolution function with the kernel size of 1×1 is applied to obtain the discriminative features, and the convolution's output channel number is equal to the number of semantic classes in corresponding dataset. Then, the final feature representation vector is obtained through calculating the average of each feature map, so as to get the ship images' semantic labels rapidly and precisely.

2.3 Loss Function

The loss functions for training our model can be divided into two parts: local attention loss function and total loss function.

Generally, the attention maps may not focus on the discriminative parts of the objects. In order to make the model pay attention to different and important regions after the LVAM, the local attention loss function is designed to guide the attention learning process. Specifically, the definition is as follows:

$$\ell_{\text{local}} = \mathrm{e}^{-\alpha \sum_{m=1}^{B} \sum_{i,j} s_{ij}^2} + \frac{\lambda}{B} \sum_{i=1}^{B} \mathrm{CE}(\hat{a}_i, y_i) \tag{4}$$

where CE() is the standard cross-entropy loss, $\hat{a}$ indicates the prediction from output attention features of the LVAM, s_{ij} $(i \neq j)$ is the Pearson's correlation coefficient among different local attention features, and y is the true label of the input, α and λ are the hyper-parameters, which could balance the weight of corresponding part.

About the total loss function, we use cross-entropy loss function as the model's total classification loss, which provides a supervision signal over the feature learning process. It could be expressed as follows:

$$\ell_{\text{total}} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{CE}(\hat{y}_i, y_i) \tag{5}$$

where $\hat{y}$ indicates the model's final prediction, and y is the true label of the input.

The final loss function for training is written as:

$$\ell = \ell_{\text{local}} + \ell_{\text{total}} \tag{6}$$

3 Experiments and Analysis

3.1 Datasets and Implementation Details

Two public datasets are adopted to verify the proposed model in subsequent experiments, and they are FGSC-23 [1] and FGSCR-42 [14]. The FGSC-23 dataset is a 23-category fine-grained RS ship classification dataset, and it contains totally 4080 ship images cropped from Google Earth public images and GF-1 satellite. The sizes of images range from 40 × 40 pixels to 800 × 800 pixels. The FGSCR-42 dataset is a lager fine-grained RS ship classification dataset, and it contains 42 categories and about 7776 RS images. The sizes of images range from 50 × 50 pixels to 1500 × 1500 pixels.

We conduct adequate experiments to demonstrate the effectiveness and explainability of our proposed model. In our experiments, the FGSC-23 dataset is divided into a training set and test set in a 8:2 ratio in accordance with [1], and we follow [14] to randomly select half of the images of each category from the dataset FGSCR-42 for training, while the rest for testing. Similarly, in order to get over the imbalanced sample problem, data augmentation operations are performed to supplement image samples of the fewer sample subclasses in the training set. The ResNet50 [15] pretrained on ImageNet is adopted to extract the high-level features from the input RS images in our network. All the models are trained for 80 epochs with a minibatch size of 16. The adaptive moment estimation (Adam) optimizer is employed for the training with an initial learning rate set at 5e-5.

The classification performance of the proposed method is evaluated by overall accuracy (OA), average accuracy (AA) and accuracy rate (AR) of each category. OA is the ratio of the correctly predicted images of total testing images, AA which seems more reasonable is the average of the accuracy of all categories, and AR is the ratio of correctly classified images among a category on the testing set.

3.2 Comparison With Other Methods

To verify the performance of the proposed model, we compare the classification accuracy of the proposed method with other deep learning-based methods on two datasets aforementioned.

As shown in Table 1, we make comparison with various methods. Specifically, common CNN-based methods

Table 1. Experimental results of comparison with other methods.

Dateset	Model	OA
FGSC-23	Inception-v3	83.88
	DenseNet	84.00
	MobileNet	84.24
	Xception	87.76
	ME-CNN	85.58
	FDN	82.30
	B-CNN	84.00
	SIM	86.30
	P^{2}Net	87.27
	ours	88.85
FGSCR-42	VGG19	77.36
	DenseNet	88.69
	ResNext-50	89.16
	B-CNN	89.53
	RA-CNN	91.63
	DCL	93.03
	TASN	93.51
	SIM	97.90
	ours	98.35

Table 2. Experimental results of AR for each category on FGSC-23 dataset.

Model	AR_0	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	AR_7	AR_8	AR_9	AR_{10}
ResNet50	90.72	79.41	97.92	100.00	89.66	66.67	90.00	100.00	100.00	84.06	81.82
ours	87.63	94.12	93.75	100.00	96.55	64.44	85.00	100.00	100.00	81.16	87.88
AR_{11}	AR_{12}	AR_{13}	AR_{14}	AR_{15}	AR_{16}	AR_{17}	AR_{18}	AR_{19}	AR_{20}	AR_{21}	AR_{22}
40.00	85.19	66.67	100.00	100.00	81.82	88.14	66.67	81.36	83.33	96.77	88.89
70.00	94.44	72.22	100.00	90.91	77.27	94.92	94.44	86.44	100.00	100.00	83.33

Table 3. Experimental results of AR for each category on FGSCR-42 dataset.

Model	AR_0	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	AR_7	AR_8	AR_9
ResNet50	95.27	100.00	77.14	100.00	97.47	96.02	100.00	91.67	96.77	79.14
ours	100.00	100.00	85.71	100.00	100.00	100.00	100.00	100.00	100.00	100.00
AR_{10}	AR_{11}	AR_{12}	AR_{13}	AR_{14}	AR_{15}	AR_{16}	AR_{17}	AR_{18}	AR_{19}	AR_{20}
84.28	79.41	81.82	89.52	92.00	83.10	84.62	84.95	92.59	80.65	96.04
100.00	100.00	100.00	100.00	96.00	98.59	94.87	100.00	98.41	96.77	100.00
Model	AR_{21}	AR_{22}	AR_{23}	AR_{24}	AR_{25}	AR_{26}	AR_{27}	AR_{28}	AR_{29}	AR_{30}
ResNet50	98.46	99.04	98.97	99.38	94.69	92.50	100.00	100.00	100.00	50.00
ours	100.00	100.00	99.74	99.38	100.00	100.00	100.00	100.00	91.67	50.00
AR_{31}	AR_{32}	AR_{33}	AR_{34}	AR_{35}	AR_{36}	AR_{37}	AR_{38}	AR_{39}	AR_{40}	AR_{41}
100.00	0.00	85.29	100.00	80.00	100.00	33.33	88.89	0.00	76.57	65.52
100.00	50.00	100.00	100.00	100.00	100.00	66.67	88.89	33.33	94.39	91.38

such as Inception-v3 [16], DenseNet [17], MobileNet [18], Xception [19], VGG19 [20], ResNext-50 [21] are included, in addition to B-CNN [22], RA-CNN [23], DCL [24], TASN [25], SIM [26] are fine-grained classification methods for natural images, and there are also some methods including FDN [27], ME-CNN [28], P2Net [5] which are proposed for the remote-sensing image classification. In Tables 2 and 3, we list the AR for each category of FGSC-23 and FGSCR-42 in the experiment respectively, and we compare the results of proposed method with the backbone network ResNet50. From the results in the tables, we can see that the performance of proposed method has achieved significantly enhanced compared with the representative common CNN models on both FGSC-23 and FGSCR-42 datasets and it also achieves superior results in contrast to other classification algorithms. This demonstrates the effectiveness of the model and shows the sufficiently good classification performance. In particular, the corresponding confusion matrixes of our method on FGSC-23 and FGSCR-42 datasets are displayed in Figures 3 and 4 respectively.

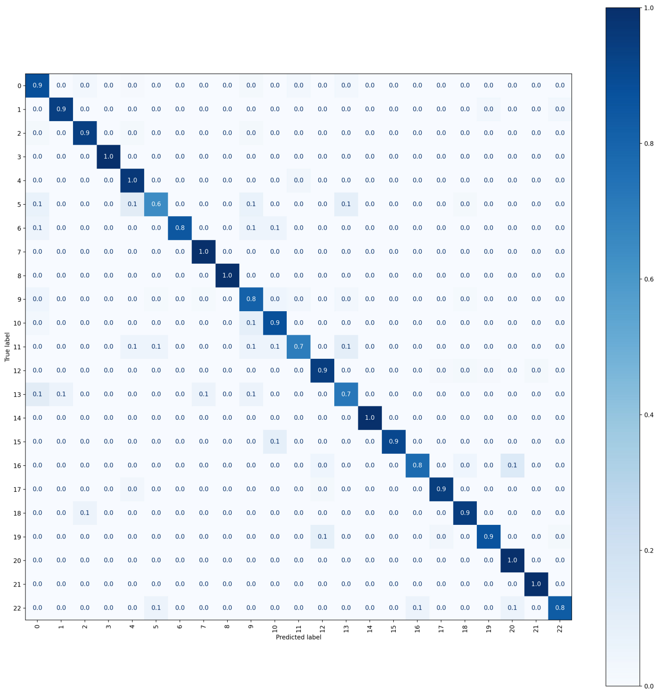


Figure 3. The confusion matrix of our method on FGSC-23.

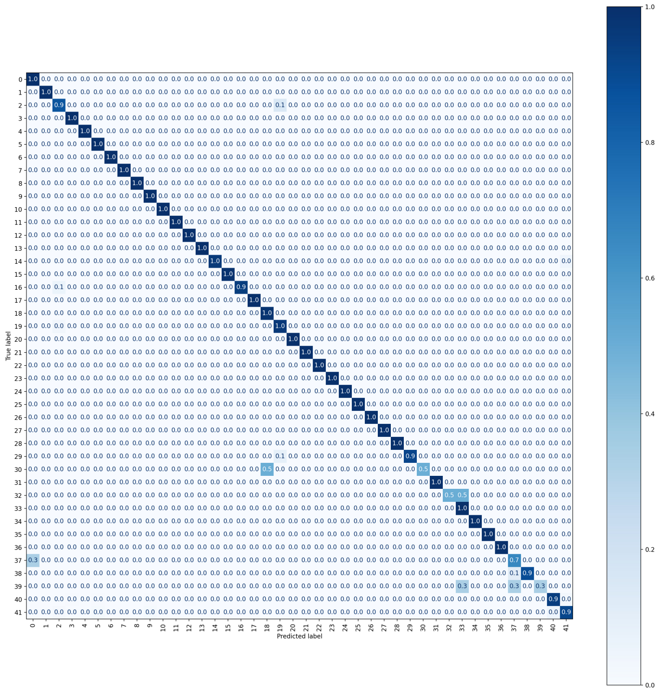


Figure 4. The confusion matrix of our method on FGSCR-42.

Table 4. Experimental results of the ablation study.

Dataset	Model	Test time	OA	AA
FGSC-23	R50	2.01 s	85.21	85.18
	R50+LVAM	2.09 s	87.27	87.55
	R50+LVAM +GCN	2.17 s	88.85	89.33
FGSCR-42	R50	7.12 s	89.83	84.41
	R50+LVAM	7.17 s	96.27	91.60
	R50+LVAM +GCN	7.31 s	98.35	93.71

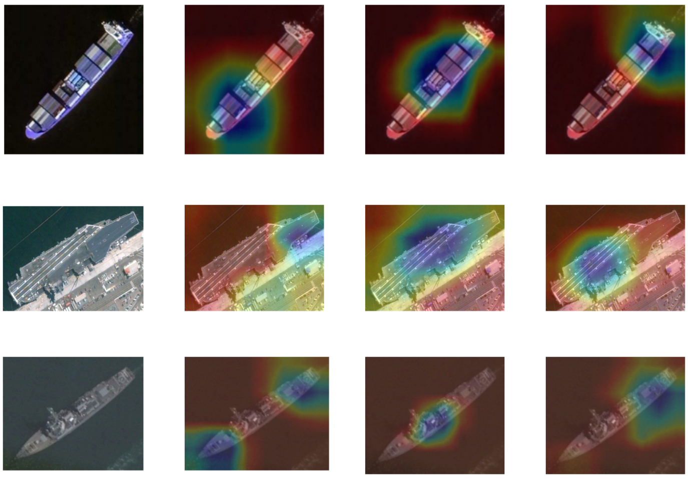


Figure 5. Visualization results of attention maps from different channels on FGSC-23 dataset.

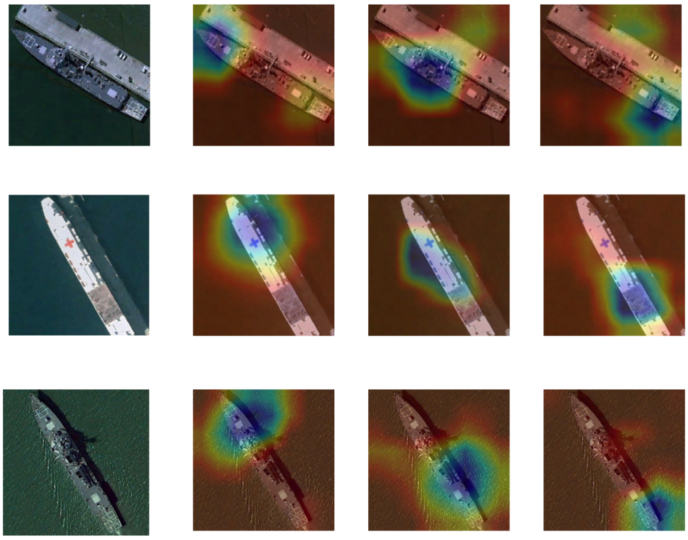


Figure 6. Visualization results of attention maps from different channels on FGSCR-42 dataset.

3.3 Ablation Study

Ablation studies are conducted to evaluate the effectiveness of each part in the proposed model and analyze the influence on the experimental results. Specifically, we set up the ablation experiments about the modules of LVAM and GCN, and at the same time give the time spent on corresponding model testing on each dataset. As shown in the Table 4, R50 represents convolutional backbone Resnet50, from the table we can see that the LVAM can help improve the learning performance notably compared with the same backbone network. This is because the LVAM focuses its attention on important parts of the ship images, which contribute to the final prediction. Furthermore, combined with GCN, the model's performance is further improved. This is mainly attributed to the semantic relationships between regions are exploited by the GCN, so that the model could learn more discriminative feature representation. Although the elapsed time for model testing has a certain degree of increase in the process of adding LVAM and GCN module, it is still very close to the convolutional backbone network. Besides, the proposed model has markedly enhanced the classification ability, which further demonstrates the effectiveness of our model.

3.4 Visualization

In this section, we simply select part representative attention maps as samples for analysis, and visualize them to verify the effectiveness of the designed LVAM, further to illustrate the effectiveness and explainability of the proposed method.

As shown in Figures 5 and 6, the first column displays input RS ship images, and different channel's attention maps are placed in the following last three columns.

It can be seen that most areas of attention are focused on discriminative parts of the ship object and other irrelevant parts catch little attention in the image. Also, different channels concentrate on diverse local visual information of the objects. And the meaningful visualization information of provided by the model could make it easier for users to know key regions that the classification model pays attention to clearly, thus helping people better understand its decision-making progress. Meanwhile, the feature representation from LVAM will input to the graph network for further relation learning among the regions, which may contribute to enhancing the discriminative ability of the model. Combined the results in Table 4, we can conclude that the proposed method could provide explainable classification results and maintain an acceptable classification performance.

4 Conclusion

In this paper, an explainable classification method for RS ship images based on graph network is proposed. The method mainly consists of local visual attention and graph network two parts. When an image is input, the LVAM could make the model focus on important regions, and then the graph network is used to exploit the underlying relationships between them and get the final feature representation. Extensive experiments and ablation studies on two public fine-grained ship classification datasets verify the effectiveness of the proposed model. Although we provide explainable attention maps visualization which may help people understand, we will seek a more powerful theory of explainable causal effect in future work.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgement

This work was supported without any funding.

References

[1] Zhang, X., Lv, Y., Yao, L., Xiong, W., & Fu, C. (2020). A New Benchmark and an Attribute-Guided Multilevel Feature Representation Network for Fine-Grained Ship Classification in Optical Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1271-1285. [CrossRef]

[2] Li, D., Liu, R., Tang, Y., & Liu, Y. (2024). PSCLI-TF: Position-Sensitive Cross-Layer Interactive Transformer Model for Remote Sensing Image Scene Classification. *IEEE Geoscience and Remote Sensing Letters*, 21, 1-5. [CrossRef]

[3] Lan, J., & Wan, L. (2009). Automatic ship target classification based on aerial images. In *2008 International Conference on Optical Instruments and Technology: Optical Systems and Optoelectronic Instruments* (Vol. 7156, p. 715612). SPIE. [CrossRef]

[4] Xiong, W., Xiong, Z., Cui, Y., & Lv, Y. (2020). A Discriminative Distillation Network for Cross-Source Remote Sensing Image Retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1234-1247. [CrossRef]

[5] Chen, J., Chen, K., Chen, H., Li, W., Zou, Z., & Shi, Z. (2022). Contrastive Learning for Fine-Grained Ship Classification in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-16. [CrossRef]

[6] Chen, Y., Zhang, Z., Chen, Z., Zhang, Y., & Wang, J. (2022). Fine-Grained Classification of Optical Remote Sensing Ship Images Based on Deep Convolution Neural Network. *Remote Sensing*, 14(18), Article 18. [CrossRef]

[7] Xiong, W., Xiong, Z., & Cui, Y. (2022). An Explainable Attention Network for Fine-Grained Ship Classification Using Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14. [CrossRef]

[8] Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 5219-5227. [CrossRef]

[9] Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. *arxiv preprint arxiv:1609.02907*. [CrossRef]

[10] Guo, Y., Bo, D., Yang, C., Lu, Z., Zhang, Z., Liu, J., Peng, Y., & Shi, C. (2023). Data-centric Graph Learning: A Survey. *arxiv preprint arxiv:2310.04987*. [CrossRef]

[11] Yang, Y., Tang, X., Cheung, Y.-M., Zhang, X., & Jiao, L. (2023). SAGN: Semantic-Aware Graph Network for Remote Sensing Scene Classification. *IEEE Transactions on Image Processing*, 32, 1011-1025. [CrossRef]

[12] Ge, Y., Xiao, Y., Xu, Z., Zheng, M., Karanam, S., Chen, T., Itti, L., & Wu, Z. (2021). A Peek Into the Reasoning of Neural Networks: Interpreting with Structural Visual Concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2195-2204. [CrossRef]

[13] Hu, H., Yao, M., He, F., & Zhang, F. (2022). Graph Neural Network via Edge Convolution for Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters, 19, 1-5. *IEEE Geoscience and Remote Sensing Letters*. [CrossRef]

[14] Di, Y., Jiang, Z., & Zhang, H. (2021). A Public Dataset for Fine-Grained Ship Classification in Optical Remote Sensing Images. *Remote Sensing*, 13(4), Article 4. [CrossRef]

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep

Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. [CrossRef]

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818-2826. [CrossRef]

[17] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261-2269. [CrossRef]

[18] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arxiv preprint arxiv:1704.04861*. [CrossRef]

[19] Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800-1807. [CrossRef]

[20] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arxiv preprint arxiv:1409.1556*. [CrossRef]

[21] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500). [CrossRef]

[22] Lin, T.-Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN Models for Fine-Grained Visual Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1449-1457. [CrossRef]

[23] Fu, J., Zheng, H., & Mei, T. (2017). Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4476-4484. [CrossRef]

[24] Chen, Y., Bai, Y., Zhang, W., & Mei, T. (2019). Destruction and Construction Learning for Fine-Grained Image Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152-5161. [CrossRef]

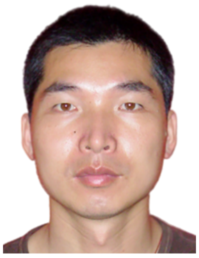
[25] Zheng, H., Fu, J., Zha, Z.-J., & Luo, J. (2019). Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5007-5016. [CrossRef]

[26] Sun, H., He, X., & Peng, Y. (2022). SIM-Trans: Structure Information Modeling Transformer for Fine-grained Visual Categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5853-5861. [CrossRef]

[27] Shi, Q., Li, W., & Tao, R. (2018). 2D-DFrFT Based Deep Network for Ship Classification in Remote Sensing Imagery. In *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, 1-5. [CrossRef]

[28] Shi, Q., Li, W., Tao, R., Sun, X., & Gao, L. (2019). Ship Classification Based on Multifeature Ensemble with Convolutional Neural Network. *Remote Sensing*, 11(4), Article 4. [CrossRef]

Haoran Li received the B.S. and M.S. degrees from Naval Aviation University, Yantai, China, in 2020 and 2022 respectively, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include information fusion, and deep learning with their applications in remote sensing. (E-mail: rizhaolihaoran@163.com)



Wei Xiong received the B.S., M.S., and Ph.D. degrees from Naval Aviation University, Yantai, China, in 1998, 2001 and 2005, respectively. From 2007 to 2009, he was a Post-Doctoral Researcher with the Department of Electronic Information Engineering, Tsinghua University, Beijing. He is currently a Full Professor with the Naval Aviation University. He is the Member and Director General of Information Fusion Branch of Chinese Society of Aeronautics and Astronautics. His research interests include pattern recognition, remote sensing and multi-sensor information fusion. (E-mail: xiongwei@csif.org.cn)

Yaqi Cui received the B.S., M.S. and Ph.D. degrees from Naval Aviation University, Yantai, China, in 2008, 2011 and 2014, respectively. He is an associate professor with Naval Aviation University. His research interests include information fusion, machine learning, and deep learning with their applications in information fusion. (E-mail: cui_yaqi@126.com)

Zhenyu Xiong received the B.S. and M.S. degrees from Naval Aviation University, Yantai, China, in 2018 and 2020 respectively, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include information fusion, and deep learning with their applications in remote sensing. (E-mail: x_zhen_yu@163.com)