MHA-CNN: AIRCRAFT FINE-GRAINED RECOGNITION OF REMOTE SENSING IMAGE BASED ON MULTIPLE HIERARCHIES ATTENTION

Yonghao Yi[1], *Yanan You*[1*], *Wenli Zhou*[1], *Gang Meng*[2]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2] Beijing Institute of Remote Sensing Information
Beijing, China

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884219

ABSTRACT

Aircraft fine-grained recognition of remote sensing image is widely exploited in both military and civilian fields, which the similar physical structure and the changeable attitude between variable aircraft types makes this task challenge. Great progress has been made by proposal models based on convolutional neural network. However, previous works did not focus on local and multiple hierarchy features. In this paper, we propose a framework based on multiple hierarchy network and attention module for aircraft recognition. Compared to the existing methods, the extraction and enhancement of features proposed by us has been greatly improved. Remarkable results have been achieved on our dataset Aircraft-16 and the MTARSI dataset.

Index Terms— Fine-grained recognition, remote sensing, deep learning, aircraft target.

1. INTRODUCTION

Fine-grained image recognition of visible spectral remote sensing image has become more and more significant in recent years. Remote sensing image target recognition advances from the traditional method of manual features to the deep learning algorithm. Compared with other targets (ship, car, etc.), recognize the category of aircraft has been a typical task in remote sensing image due to its high value in application.

In fact, existing following problem makes this task challenge: The similar physical structure and the changing attitude between variable aircraft types and inevitable shadow changing, cloud occlusion in visible remote sensing image. Previous fine-grained image recognition method [1] generally leverages the extra annotations of bounding box as well as parts. As a result of the deep learning, most of works utilize the convolutional neural networks to distill higher semantic information. Further research [2] shows that the convolutional neural network (CNN) is a feasible approach for the aircraft recognition. The existing methods mainly focus on the single global features of the aircraft target. Zhao's proposal [3], an framework based on landmark detection, can not effectively deal with the problem of incomplete aircraft structure. Another work [4] emphasizes the correlation between channels.

Although worthwhile works have been proposed, the network missing attention local features and multi-scale features performances insufficient in more complex tasks. For visible spectral remote sensing images, (1) it is difficult to ensure that the target is sufficiently clear and complete, (2) and it is impossible to eliminate the noise interference caused by imaging angle, day and night, and cloud or fog. To solve the difficulties above, we utilize the hierarchical network to leverage the features of targets in multiple scales. In addition, in order to improve the significance of features and reduce the impact of interference, we propose to adopt the attention module to filter the disturbed responses that may affect the recognition results. In fact, the network only need to effectively extract the most significant features supporting aircraft target recognition, avoiding the interference of weak and fuzzy features.

In this work, we implement the multiple pyramid hierarchy on ResNet-50 [5] with multiple attention to localize and extract efficiently the significant region feature. Remarkable results have been achieved on our dataset Aircraft-16. We also evaluate the performance of the model in the MTARSI dataset [6], and reach the state-of-the-art.

2. PROPOSAL METHODS

In this section, we will introduce the details of our proposed aircraft recognition without any bounding-box annotations. The whole network is divided into five steps as illustrated in Fig. 1.

First, the remote sensing image (slice) containing the aircraft target feeds into several layers in the feature extractor (e.g., ResNet-50 [5]) to generate output feature maps. Multiple Hierarchy Network as illustrated in Fig. 1 (a) is constructed based on the output feature maps. Next, these multiple hierarchy features pass through attention modules in (b) to generate the attention feature maps in (c). Furthermore, the

*Corresponding to youyanan@bupt.edu.cn.

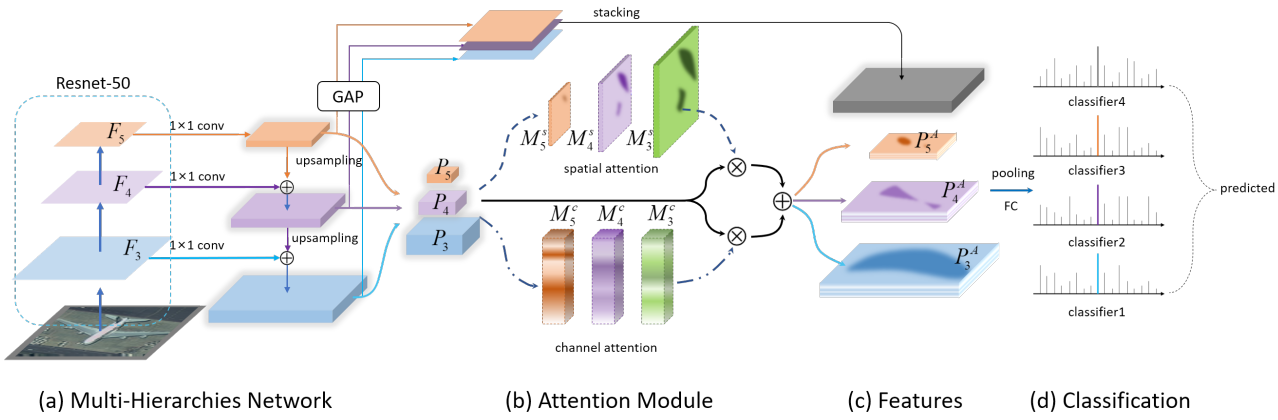

Fig. 1. Overview of the multiple hierarchies attention convolutional neural network (MHA-CNN) for aircraft recognition.

stacking features block in (c) entries the recognition network in (d).

2.1. Multiple Hierarchies Network

With the help of Feature Pyramids Network (FPN for short) [7], [8], we construct a structure with lateral connection to maintain the current scale information and a up-to-down structure to transfer the high-level semantic information to the lower level. Thus, feature maps from top to down contain the highest-level semantic information.

We denote the output feature maps from Layer i of the ResNet-50 as F_i. F_3, F_4 and F_5 are extracted to construct the pyramid structure. For the lateral connection, 1×1 convolution layer is used to unify the number of channels of different features F_i (we introduce the super parameter κ to represent the unified number of channel). For up-to-down pyramid, upsampling each channel of the higher layer feature map F_{i+1} to unify the size of the feature map of the previous and the current layer. Then, the fused features are obtained by adding the two adjacent features of element positions, which is defined as P_i:

$$P_i = upsampling(P_{i+1}) + conv(F_i) \tag{1}$$

where $P_i (i = 3, 4, 5)$ denotes the i^{th} output feature map of multiple hierarchies network. $upsampling(\cdot)$ is the upsampling function and $conv$ is the 1×1 convolution layer.

The feature stacking network enables the network to capture features at different resolutions.

2.2. Attention Module

The attention mechanism such as SE-Net [9] in Computer Vision not only enable us to discard key points and regional bounding boxes except extra annotations, but more importantly, filter out noise interference in features.

(1) Channel Attention Module: For each layer of the channel dimension of the feature map P_i, it can be regarded as the response intensity map of the network to different modes in the output image. However, neither all channels are helpful for recognition, nor each channel responds to diverse patterns. At this time, we need to use the channel attention module (CAM) to screen out the channels that are beneficial to aircraft recognition. The CAM acts like a filter here.

Specifically, according to feature map P, We use convolution layer [9] instead of fully-connected layer to obtain channel attention map M_i^c:

$$M_i^c = \varsigma\{W_2 * [\rho(W_1 * GAP(P_i) + b_1)] + b_2\} \tag{2}$$

where GAP is the global average pooling (output size set to 1), and $*$ represent the convolution. W_1, b_1 and W_2, b_2 represent the parameters of convolution layers. ς, ρ are the $Sigmoid$, $ReLU$ activation function. We denote the i^{th} channel attention map as M_i^c.

(2) Spatial Attention Module: [10] mentions that after the channel attention squeezes and excites the features, a spatial attention module (SAM) could be used to select the "focus" position to highlight the features. More experiments show that the output of a convolution layer that can "close the field of vision" is used as a SAM, which is connected in parallel with the CAM and achieves the best effect.

$$M_i^s = \varsigma(W_3 * P_i + b_3) \tag{3}$$

where M_i^s is the i^{th} attention map of SAM, w_3, b_3 is the weight and bias of the convolution layer.

Based on the generated attention maps, we further calculate the final attention pooling features which we denote as P_i^A:

$$P_i^A = P_i \circ M_i^c + P_i \circ M_i^s \tag{4}$$

where $\circ$ represents the Hadamard product.

Fig. 2. Some examples of Aircraft-16.

2.3. Multiple Hierarchies Attention Features for Recognition

Multiple Hierarchies Attention CNN (MHA-CNN for simplicity) which we propose can obtain attention feature maps in different scales. Multiple attention feature maps P_i^A $(i = 3, 4, 5)$ after pooling the attention map of aircraft targets are obtained by forward propagation. In addition, there are global feature vectors which are expanded and spliced from the global feature map P_i $(i = 3, 4, 5)$ after global average pooling to unify the output size to $batchsize \times \kappa \times 1 \times 1$.All these features are sent to the recognition network with max-pooling layer and fully connected layer at the end respectively. The weights of these recognition networks are equal. We use CrossEntropy Loss to optimize network parameters.

3. EXPERIMENTS

In experiments, the batch size is 16, κ set to 256, and the learning rate is fixed to $5e^{-4}$.We conduct the experiments using Pytorch on the Ubuntu platform, and 2 NVIDIA RTX 3090 GPUs are used.

3.1. Datasets

In order to evaluate the role of our proposed model, MHA-CNN, in fine-grained recognition of aircraft targets in remote sensing images, we have done experiments on two remote sensing image datasets. All data of contains only the supervision information of the category to which the target belongs. Table 1 lists the details of our experimental dataset:

3.1.1. Aircraft-16

At present, there are problems such as few available target categories, few target scenes, and low target resolution in the public aircraft dataset. We cut out the target aircraft from the original image from the available remote sensing image aircraft target detection dataset, by considering factors such as quantity, geographic location, natural weather conditions. In this way, a fine-grained recognition dataset containing 16 types of remote sensing image aircraft targets (we denote it as Aircraft-16 simplicity) is produced. The images and categories of some examples are shown in the Fig. 2.

Table 1. Statistics of the experimental datasets.

Dataset	#Class	#Train	#Test
Aircraft-16	16	6335	2133
MTARSI	20	7500	1885

Table 2. Ablation analysis of our proposal on Aircraft-16. "+ SA" stands for the spatial attention module connected after the attention. "Time" denotes the total inference time.

Model	Accuracy (%)	Time (s)
ResNet-50	85.00	10.44
Multiple Hierarchies Network	89.03	12.19
Attention Module	89.55	12.38
MHA-CNN + SA	89.64	12.53
MHA-CNN	**89.73**	12.27

3.1.2. MTARSI [6]

Multi-type aircraft remote sensing images (MTARSI) [6] are the only publicly available fine-grained recognition datasets of remote sensing image targets.

3.2. Ablation Study

To fully inspect our method, Table 2 and Table 3 provide the detailed ablation analysis.

(1) Effectiveness of attention module and multiple hierarchy structure: Table 2 shows the impact of different module. With only Multiple Hierarchy Network, the network at all hierarchies can not focus the detail, i.e., all layers focus on the same pattern (such as tail). The accuracy is 0.7% lower than the best. Without multi-level network, the decline of accuracy indicate that the multi-level information can not be obtained. After adding the serial attention mechanism, the network efficiency and performance are reduced.

Moreover, note that the performance saturates after extending the multiple hierarchies network to three layers because the output features in the first layer of ResNet-50 are too redundancy to be utilized.

(2) Comparison with the Proposal Models: Table 3 lists the test results of our proposal and other comparison models on MTARSI [6] and Aircraft-16. Light-CNN [4] uses the depth separable convolution to replace all convolutional layers, and uses the special loss function containing the channel attention structure to optimize all parameters.Since there is not much open source code for comparison in the field of fine-grained recognition of aircraft targets in remote sensing images, only some models are performed. Reproduce and test on our dataset. At present, the accuracy of our model has reached SOTA.

Table 3. Experimental results on MTARSI [6] and Aircraft-16. All of the values denotes the accuracy in percentage.

Model	MTARSI	Aircraft-16
ResNet-18	88.10 [4]	83.36
ResNet-50	89.61 [6]	85.00
LCNN [4]	89.50 [4]	83.03
MHA-CNN	**99.89**	**89.73**

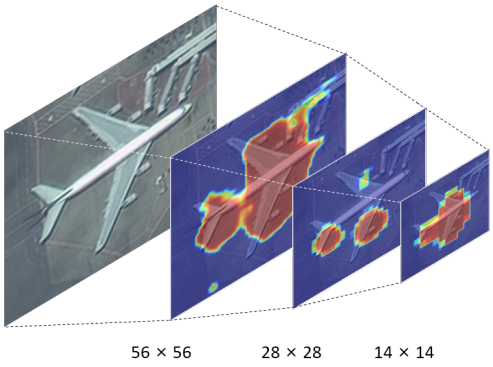

Fig. 3. Visualization of the attention regions. The numbers represent the size, $w \times h$, of the different scale feature maps

3.3. Visualization

We visualize the attention map generated by attention module. The heatmap indicates the interests of multiple layer in attention module. As shown in Fig. 3, the range of multi-scale features that each layer focuses on is slightly different. Similarly, the humans may rely on these features to recognize.

4. CONCLUSION

In this paper, we propose a convolutional neural network with multiple hierarchies attention (MHA-CNN) for remote sensing aircraft target recognition, which can utilize the multi-scale features to localize the significant region and enhance the features through the attention modules. A large number of results demonstrate the superior performance on both our Aircraft-16 and the public MTARSI. Our model is steady to identify aircraft targets in remote sensing images which are incomplete and greatly influenced by light, shadow, and clouds. Therefore, we will continue to promote this research to improve the accuracy of the network attention map, so that the network can acquire key features that are beneficial to recognition more efficiently.

5. ACKNOWLEDGEMENT

This research was supported by National Natural Science Foundation of China (62101060), and Beijing Natural Science Foundation, China (Grant No. 4214058).

6. REFERENCES

[1] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] A. Zhao, K. Fu, S. Wang, J. Zuo, Y. Zhang, Y. Hu, and H. Wang, "Aircraft recognition based on landmark detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1413–1417, 2017.

[4] Y. Pan, L. Tang, and B. Zhao, "Lightweight fine-grained recognition method based on multilevel feature weighted fusion," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 4767–4770.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] Z.-Z. Wu, S.-H. Wan, X.-F. Wang, M. Tan, L. Zou, X.-L. Li, and Y. Chen, "A benchmark data set for aircraft type recognition from remote sensing images," *Applied Soft Computing*, vol. 89, p. 106132, 2020.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[8] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "Augfpn: Improving multi-scale feature learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 595–12 604.

[9] 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.

[10] 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.