Few-Shot Scene Classification with attention mechanism in Remote Sensing

LI Xuanye1, LI Hongguang2,*, YU Ruonan1, Wang Fei1

1 School of Electrical and Information Engineering, Beihang University, Beijing, China
2 Unmanned System Research Institute, Beihang University, Beijing, China
*Corresponding author’s e-mail: lihongguang@buaa.edu.cn

Abstract: Remote sensing scene classification is a hot research topic in computer vision and it is of great significance to the semantic understanding of remote sensing images. At present, remote sensing scene classification methods based on deep learning occupy a dominant position in this field. However, it suffers from the lack of samples and poor model generalization ability in actual application scenarios. Therefore, this paper proposes a few-shot remote scene classification method based on attention mechanism, and designs a structure of dual-branches similarity measurement. This method is based on the meta-learning training strategy to divide the dataset into tasks. At the meantime, the input images are divided into blocks to preserve the feature distribution in the remote sensing image. Then the lightweight attention module is introduced into the feature extraction network to reduce the risk of overfitting and ensure the acquisition of discriminative features. Finally we add a dual-branches similarity measurement module on the basis of Earth Mover's Distance to improve the discriminative ability of the classifier. The results show that compared with the classic small-sample learning method, the few-shot remote scene classification method proposed in this paper can significantly improve the classification performance.

1. Introduction
In recent years, aerial remote sensing and UAV technology has developed rapidly. As an effective information carrier with rich shape, texture and scene semantic information, aerial remote sensing images are widely used in intelligence reconnaissance, environmental monitoring, natural disaster prevention, and water resource protection applications. However, the massive amount of high-definition remote sensing image data generated by aerial photography has exceeded the ability of manual real-time interpretation and understanding, making it difficult to obtain relevant information in a timely and accurate manner. As one of the research hotspots in the field of computer vision, remote sensing image scene classification [1]-[2] is of great significance to the understanding of the scene and semantic information contained in the data, and is an important basis for subsequent projects and military operations.

Compared with the traditional image classification methods, the remote sensing image scene classification task shows the characteristics of complex scene information and difficult data annotation. At the same time, due to the characteristics of unmanned aerial vehicles and the variability of their mission scenes, some application scenes do not have the objective conditions for long-term collection of remote sensing image data, which leads to the problems of less sample data and poor generalization ability of models in the practical application of remote sensing image scene classification. The
implementation of remote sensing image scene classification methods based on deep learning is more difficult and the performance is significantly reduced. Therefore, it is required to construct a remote sensing image scene classification models with strong generalization ability based on a small number of data samples to solve these practical problems. At the same time, few-shot scene classification in remote sensing faces more challenges than that in natural scenes. The special characteristics of remote sensing images are mainly manifested in:

- Interference from irrelevant objects: remote sensing images are top-down views and inevitably contain numerous objects that are irrelevant to the semantic class of the scene.
- Large intra-class variation: remotely sensed features are complex and diverse, and remotely sensed images can be affected by a variety of imaging conditions [11].
- Inter-class similarity: different scene images may contain similar feature types.

In recent years, there has been a proliferation of few-shot learning methods. Currently, combining meta-learning [3]-[4] and metric learning [5]-[6] is the mainstream development direction. Among them, this kind of method firstly constructs tasks of small samples based on meta-learning, which enhances the generalization performance of the model for new tasks. Secondly, it makes predictions by measuring the similarity between the features of images. Related algorithms include the classical Matching network [7], Relation network [8], Prototypical network [9], etc. In 2020, Chi Zhang [10] et al. proposed a method based on image region block matching, using EMD (Earth Mover’s Distance) distance to calculate the correlation between image representations.

Based on the EMD matching block, this paper proposes an algorithm of few-shot scene classification in remote sensing based on attention mechanism, and designs the dual-branches structure to measure the similarity of images. (1) The method is based on a meta-learning training strategy for task-based partitioning of the dataset; (2) to preserve the distribution of remote sensing image features, the input images are overlapped into blocks; (3) a lightweight attention module is introduced into the feature extraction network to reduce the risk of overfitting while ensuring the acquisition of discriminative features; (4) the dual-branches similarity measurement module is added to the EMD distance matching block to improve the discriminative ability of the classifier. Experiments on two popular remote sensing scene datasets, UCMerced_LandUse [12] and NWPU-RESISC45 [13], show that the method in this paper achieves significant improvement in the accuracy of remote sensing scene classification compared with the classical few-shot image classification and deepEMD [10] methods.

2. Methodology

This paper proposes a few-shot scene classification algorithm based on attention mechanism in remote sensing whose network model structure is shown in Fig. 1. Firstly, the algorithm divides the data set into tasks based on meta-learning strategy, and feeds the overlapped and blocked support set and query set images into a feature extraction network with shared weights.
2.1. Meta-learning task division

Deep neural networks are usually trained iteratively based on a specific large dataset to obtain the model parameters appropriate for the task. However, there are differences between this learning strategy and the human model of acquiring new knowledge. Humans have the ability to learn by example and use their previous learning experience to complete new tasks quickly.

Therefore, this paper adopts the meta-learning training strategy to divide the data set into different image categories in the train set and the test set for meeting the requirements of the model for the generalization ability of the new task. For example, in the N-way M-shot, N categories are randomly selected in the training set, and M samples are randomly selected from each of these N categories, making a total of N*M samples to form the support set in this meta-task, assuming that the labels of the support set are known. The extraction rules for the test task are the same as above. Meta-learning training based on the divided tasks allows the model to acquire prior knowledge that enables it to generalise better in tasks containing new categories.

2.2. Image overlap chunking

Remote sensing images have a unique perspective, wide field of view, large image size and rich and diverse features. In contrast, most remote sensing images do not contain iconic features or iconic features are not located in the centre of that image. Therefore, if the whole image is fed into the feature extraction network and pooled through several layers of downsampling, it will result in a large number of detailed features in the original image being lost.

To address the above issues, the image is overlapped and partitioned before feature extraction. In this paper, the original image is cut according to the positions of top left, bottom left, top right, bottom right and centre. The discriminative features in the remote sensing image are retained to a great extent. It also serves as a data augmentation to reduce the risk of overfitting.

2.3. Lightweight spatial-channel attention-based feature extraction

To obtain the high-dimensional abstract features of an image, a feature extraction model needs to be constructed to learn a practical feature embedding space in which to measure the degree of similarity...
between images before performing the similarity metric. In this paper, ResNet-12 [14] is used as the backbone network for feature extraction.

![ResNet-12](image)

**Fig.2** (a) Residual block structure; (b) Residual block structure with channel-spatial attention module

ResNet consists of a series of residual blocks, and the residual block structure is shown in Fig 2(a). Each residual block establishes a cross-layer connection structure between its input and output which allows a better transfer of the reverse gradient to the shallow network and overcomes the problem of gradient disappearance in deep neural networks. The mathematical representation of the residual structure is shown below:

\[ x_{t+1} = x_t + F(x_t, W_t) \]  

where \( x_t \) denotes the direct mapping of the residual block input and \( F(x_t, W_t) \) consists of a number of convolutional layers.

Remote sensing images are rich in spatial information. At the same time, the convolution operation maps the features to a high-dimensional space, and the data channels are rich in detailed features. To overcome the "inter-class similarity" and "intra-class variability" of remote sensing images, the feature extraction network needs to be able to acquire discriminative features.

Attention mechanisms enable neural networks to focus accurately on effective elements and have become an important component in improving the performance of deep neural networks [16][17]. Attention mechanisms are mainly divided into spatial and channel attention, capable of capturing pixel-level and data channel dependencies, respectively. In computer vision research, channel attention and spatial attention in combination often result in performance gains. At the same time, feature extraction networks in few-shot learning have higher restrictions on model complexity, with higher model complexity exposing higher risk of overfitting. In order to balance the needs of network feature extraction capability and model complexity, this paper introduces a lightweight channel-space attention module based on Shuffle units in the residual structure, as shown in Fig 2(b).

The process of implementing lightweight channel-space attention based on Shuffle units [18] consists of the following steps: feature grouping, channel attention, space attention, and sub-feature aggregation.
1) Feature grouping
For a feature map $X$ with scale $C \times H \times W$, the feature map is first divided into $G$ groups along the number of channels, i.e. $X = [X_1, ..., X_G]$, $X_k \in R^{C/G \times H \times W}$. By dividing $X_k$ into two groups $X_{k1}$ and $X_{k2}$ again according to channels, the number of channels in each group becomes $C/2G$, and the two branches are processed for channel and spatial attention respectively.

2) Channel attention
Integrating global information about the channel based on global average pooling generates a channel statistics vector with a scale of $C/2G \times 1 \times 1$. The global average pooling is calculated as follows:

$$c = f_{\text{channel}}(X_{k1}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{k1}(i,j)$$

(2)

The output feature $X'_{k1}$ of the channel attention module is obtained by weighting the $X_{k1}$ feature map based on the global statistics of the channel.

$$X'_{k1} = \text{sigmoid}((w_1 c + b_1) \cdot X_{k1})$$

(3)

3) Spatial attention
Image spatial statistics are obtained based on group normalization. The final spatial attention output $X'_{k2}$ is obtained by augmenting the feature representation with a fully connected layer.

$$X'_{k2} = \text{sigmoid}((w_2 \cdot \text{GN}(X_{k2}) + b_2) \cdot X_{k2})$$

(4)

4) Sub-feature aggregation
The feature vectors obtained from the two branches are spliced and the number of channels becomes $C/G$. The cross-group feature exchange is achieved along the channel dimension based on channel shuffle [15].

The SA-module has a number of channels per branch of $C/2G$ and a total number of parameters of $3C/G$, with $G$ taken as 8, 16, 32, etc. The SA-module is therefore a lightweight network structure, which greatly reduces the complexity of the model with the addition of an attention mechanism.

2.4. Dual-branch discriminative distance metric block based on channel attention

2.4.1. Earth Mover's Distance. Earth Mover's Distance [19] is commonly used to solve the distribution comparison problem in image retrieval, which measures the minimum cost required to transform
between different data distributions as an indication of how similar two distributions are. Assume that the distributions of the two images to be tested are \( S = \{ s_i | i = 1, 2, ..., m \} \) and \( D = \{ d_j | j = 1, 2, ..., k \} \), and the cost of the transformation from \( s_i \) to \( d_j \) is \( c_{ij} \), and \( C \) is an \( m \times k \) matrix. A matrix \( F \) needs to be found that satisfies:

\[
\minimize \sum_{i=1}^{m} \sum_{j=1}^{k} c_{ij} f_{ij}
\]

subject to \( f_{ij} \geq 0, i = 1, ..., m, j = 1, ..., k \)  \( (5) \)

\[
\sum_{j=1}^{k} f_{ij} = s_i, i = 1, ..., m
\]

\[
\sum_{i=1}^{m} f_{ij} = d_i, i = 1, ..., k
\]

During the calculation, if a matching result is found that minimises the overall matching cost, this minimum cost is the Earth Mover's Distance.

Because there are no instance-level targets in remote sensing images in most cases, distribution information is more important than single category representation. Therefore, the remote sensing images should be overlapped and segmented before being input into the network, and after feature extraction a set of local feature vectors of the input image is obtained, which then forms the distribution representation of the image. Based on this, the optimal matching cost between the distributions is calculated. The smaller the matching cost is, the more similar the two images are.

2.4.2. Structure of the dual-branches similarity measurement. Feature extraction yields a local feature map for each segmented image block as \( h_i (i = 1, ..., 5) \). It is necessary to globally pool feature maps to obtain local feature vectors, in which global pooling is often characterized by global maximum pooling and global average pooling. Maximum pooling focuses on the largest and fastest responsive part of the feature, while average pooling focuses on the overall feature. This paper uses global maximum pooling and global average pooling in parallel to obtain the local feature vector, and designs two branches to obtain the significant local feature vector and the comprehensive local feature vector of the image respectively to consider both maximum response features and global features.

In order to improve the learnability and feature characterization ability of the two branches, and to enhance the discriminatory ability of the subsequent classifier for "intra-class variability" and "inter-class similarity" between distributed features, we consider adding a channel attention-based SE module[20] to each of the two branches. The principle of the SE module is the same as that of the channel attention section in the previous subsection, and the structure of the two-branch discriminative network is shown in Fig 4.
The image distribution features are calculated as shown below, first obtaining the local feature vector:

\[ r_{\text{max}} = \text{GMP}(SE_{\text{max}}(h_i)) \]
\[ r_{\text{avg}} = \text{GAP}(SE_{\text{avg}}(h_i)) \]  

(6)

where \( r_{\text{max}} \) and \( r_{\text{avg}} \) are the saliency distribution features of the support set and query set images respectively, and \( SE_{\text{max}} \) and \( SE_{\text{avg}} \) denote the saliency features of the maximum and average respectively.

The number of overlapping chunks is 5, i.e., \( i = 1, \ldots, 5 \). So the overall distribution characteristics of the image are expressed as:

\[ R_{\text{max}} = \{ r_{\text{max}} \}_{i = 1, \ldots, 5} \]
\[ R_{\text{avg}} = \{ r_{\text{avg}} \}_{i = 1, \ldots, 5} \]  

(7)

The resulting similarity of the query set image to the support set image is expressed as:

\[ D = \frac{\text{EMD}(R_{\text{max},s},R_{\text{max},q}) + \text{EMD}(R_{\text{avg},s},R_{\text{avg},q})}{2} \]  

(8)

where \( R_{\text{max},s} \) and \( R_{\text{max},q} \) denote the saliency distribution features of the support set and query set images respectively, and \( R_{\text{avg},s} \) and \( R_{\text{avg},q} \) denote the comprehensive distribution features of the support set and query set features respectively.

3. Experiment

3.1. Dataset

3.1.1. UCMerced LandUse. The UCMerced_LandUse dataset consists of remote sensing images of 21 categories of ground scenes such as rivers, buildings, and car parks, each category contains 100 remote sensing images of varying angles, with 256*256 pixels and a spatial resolution of 0.3m/pixel, as shown in Fig 5.
3.1.2. NWPU-RESISC45. The NWPU-RESISC45 dataset consists of remote sensing images of 45 ground scenes such as deserts, forests and lakes. Each category contains 700 remote sensing images with 256*256 pixels and a spatial resolution of 0.2-30m/pixel, as shown in Fig 6.

3.2. Metric
The metric used to evaluate the performance of the algorithm is the average classification accuracy. The average classification accuracy is calculated as shown below:

\[
\text{Acc} = \frac{1}{M} \sum_{i=1}^{M} \frac{N_{\text{correct}}^i}{N_{\text{total}}}
\]

(9)

where M denotes the number of small sample classification tasks, \(N_{\text{correct}}^i\) denotes the number of correctly classified samples in each task, and \(N_{\text{total}}\) denotes the total number of samples to be classified in the task.

3.3. Training Details
The algorithmic model implementation in this paper is based on the pytorch deep learning framework with an Nvidia Titan X (Pascal) graphics card with 12GB of video memory.

During training and testing, the meta-tasks were generated by N-way M-shot division rules, in which each image classification is unique. Considering the limitation of computational resources, this paper sets the tasks as 5-way 1-shot, i.e. each task contains a total of 5 samples from the support set and 75 samples from the query set. The specific training parameters are set as follows:

Using SGD as the neural network optimizer, the initial learning rate was 0.0005, and the learning rate decayed to 0.5 times of the previous stage every 10 epochs, with a total of 50 epochs trained. The input data dimension of the network was 84*84*3, and the batch size was \(N_{\text{batch size}} = N_{\text{ways}} \times (N_{\text{supports}} + N_{\text{queries}}) \times N_{\text{patches}}\), where \(N_{\text{ways}}\) represents the number of categories to be
distinguished in each meta-task, $N_{\text{support}}$ refers to the number of support set images in each category, $N_{\text{query}}$ represents the number of query set images in each category, and $N_{\text{patches}}$ refers to the number of overlapping chunks done on the images before input.

4. Results

4.1. Evaluation of Model performance
The grouping number $G$ shall be an indispensable hyperparameter setting given that SA-module requires grouping of feature maps. In order to obtain the optimal grouping parameter $G$, the grouping number $G$ is set to 8, 16 and 32 under the two-branch discriminant structure with UCMerced_LandUse as the dataset to test the influence of different grouping numbers on the performance of the algorithm.

![Fig.7 The influence of G](image)

As shown in Fig. 7, adding SA-module to the feature extraction network structure can significantly improve the model performance. With the increase of $G$ value and the number of groupings, the size of SA-module becomes smaller. It is noticed that the number of parameters brought by this module to the feature extraction network decreases, which reduces the complexity of the model and minimizes the overfitting problem of few-shot learning. Therefore, the performance of the model is improved when the groupings $G$ increases. In the following experiments the number of groupings $G$ for the fixed SA-module is 32.

To verify the validity and reasonableness of the algorithmic model in this paper, the ablation experiments of several network structures are shown below:

1) Structure without SA-module and dual-branches similarity measurement (origin)
2) Structure with dual-branches similarity measurement which has no SE-module (Bi_withoutse)
3) Structure with dual-branches similarity measurement which has SE-module (Bi_withse)
4) SA-module only, without dual-branches similarity measurement (SA-module)
5) SA-module and dual-branches similarity measurement without SE-module (SA-module+Bi_withoutse)
6) SA-module and dual-branches similarity measurement with SE-module (SA-module+Bi_withse)

Table 1 lists the performance comparisons for each of the above experimental setups.

| Structure                               | UCMerced_LandUse | NWPU-RESISC45 |
|-----------------------------------------|-----------------|---------------|
| Origin                                  | 65.47%          | 58.59%        |
| Bi_withoutse                            | 66.93%          | 60.78%        |
| Bi_withse                               | 67.34%          | 60.91%        |
As can be seen from Table 1, the proposed method generally improves the model performance further on the UCMerced_LandUse and NWPU-RESISC45 datasets, which shows the effectiveness and reasonableness of the network structure combining the SA-module and the two-branch discriminant module in the few-shot remote sensing image classification problem. On the UCMerced_LandUse dataset, the optimal performance was obtained for the SA-module and the dual-branches similarity measurement without the SE-module; however, on the NWPU-RESISC45 dataset, the optimal performance was obtained for the SA-module and the dual-branches similarity measurement with the SE-module.

By comparing the training and validation accuracy of the two datasets in Fig. 8, it can be found that the validation accuracy of UCMerced_LandUse is better than NWPU-RESISC45. But there is a big difference in the training and validation accuracy of the UCMerced_LandUse, showing a more obvious over-fitting tendency. On the contrary, there is little difference between the training accuracy and validation accuracy of the training process on NWPU-RESISC45. Based on this inference, the algorithm has a strong generalization ability for NWPU-RESISC45. The research results are consistent with the trend of network complexity shown in Table 1. Based on the feature extraction backbone network with SA module, compared with the dual-branches similarity measurement without SE module, the performance of the dual-branches similarity measurement with SE module on the UCMerced_LandUse is slightly reduced, while the performance on the NWPU-RESISC45 is slightly improved. The difference between the two datasets is that NWPU-RESISC45 contains more scene categories, so it can be inferred that the more categories included in the training process, the stronger the model's generalization ability to new categories.

![Fig.8 Train accuracy and validation accuracy of UCMerced_LandUse (the left) and NWPU-RESISC45 (the right)](image)

### 4.2. Comparisons with Other few-shot learning Methods

| Methods       | UCMerced_LandUse | NWPU-RESISC45 |
|---------------|------------------|---------------|
| MAML          | 51.90%           | 48.40%        |
| Prototypical Network | 50.45%       | 54.78%        |
| Relation Network  | 57.70%            | 60.63%        |
| DeepEMD      | 65.47%           | 58.59%        |
| Ours         | 67.87%           | 62.74%        |

In order to verify the effectiveness of this algorithm, the performance of this algorithm model was compared with that of classical few-shot learning algorithms.
The classical few-shot learning algorithms are as follows.

1) MAML

The aim of MAML [21] is to obtain a good set of initialization parameters through training. The requirement to quickly adapt to a new task using a small number of samples is achieved by performing several steps of parameter tuning based on the good initialization parameters.

2) Prototypical Network

Prototypical Network maps the support set and query set samples into the feature space to calculate the centre of each category, and determine the category of the query set samples by comparing the Euclidean distance from the query set features to the category centre.

3) Relation Network

The Relation Network maps the samples into the feature space and constructs a neural network to calculate the distance between the samples and analyse the degree of similarity accordingly.

4) DeepEMD

DeepEMD maps the samples into the feature space through a neural network and uses the EMD distance as a measure of similarity.

As can be seen from Table 2, the algorithm proposed in this paper has greater advantages in performance than the algorithm mentioned above, which has achieved better performance on two datasets of remote sensing, thus proving the effectiveness of the algorithm in classifying remote sensing images with a small number of shots.

5. Conclusion

For addressing the problems of small sample data and poor generalization ability of the model in the application scenario, this paper proposes an algorithm of few-shot scene classification in remote sensing based on attention mechanism and the meta-learning framework, then designs a dual-branches structure for similarity measurement.

Aiming at the problems that the distribution characteristics of remote sensing images are not obvious and the key role is prominent, the input images are overlapped and partitioned. In order to solve the difficulties of "inter-class similarity" and "intra-class difference" in remote sensing images, a spatial and channel-based attention module is added to the feature extraction network, and a Shuffle unit is embedded in this attention module to avoid the increase of model complexity caused by attention, which leads to the increase of overfitting in few-shot learning tasks. Eventually, the dual-branches module was added to the similarity measurement block in order to enhance the characterisation of discriminative features. The effectiveness of this scheme was verified on two benchmark remote sensing image classification datasets.

However, the current mainstream work mainly focuses on improving performance. The ability to deploy hardware for few-shot scene classification in remote sensing is the key to practical application. Therefore, it is necessary to conduct in-depth research on the lightweight design of the model to achieve the above results in the future.

Acknowledgments

This subject is supported by National Natural Science Foundation of China (NSFC Grant No. 62076019).

References

[1] C. Tao, W. Lu, J. Qi, and H. Wang, “Spatial information considered network for scene classification,” IEEE Geosci. Remote Sens. Lett., 2020.
[2] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, “Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification,” ISPRS J. Photogramm. Remote Sens., vol. 138, pp. 74–85, 2018.
[3] Y. Bengio, S. Bengio, and J. Cloutier, Learning a synaptic learning rule. Citeseer, 1990.
[4] S. Thrun and L. Pratt, Learning to learn. Springer Science & Business Media, 2012.
[5] HU, Junlin; LU, Jiwen; TAN, Yap-Peng. Discriminative deep metric learning for face verification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 1875-1882.

[6] SOHN, Kihyuk. Improved deep metric learning with multi-class n-pair loss objective. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016. p. 1857-1865.

[7] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra et al., “Matching networks for one shot learning,” in Advances in Neural Information Processing Systems, 2016, pp. 3630–3638.

[8] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.

[9] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in Advances in Neural Information Processing Systems, 2017, pp. 4077–4087.

[10] C. Zhang, Y. Cai, G. Lin, and C. Shen, “DeepEMD: Few-shot image classification with differentiable earth mover’s distance and structured classifiers,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 203–12 213.

[11] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, “When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 5, pp. 2811–2821, 2018.

[12] 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.

[13] 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.

[14] 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.

[15] ZHANG, Xiangyu, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 6848-6856.

[16] Y. Cao, Jiaru Li, Stephen Lin, Fangyun Wei, and Han Hu, “Genet: Non-local networks meet squeeze-excitation networks and beyond,” CoRR, vol. abs/1904.11492, 2019.

[17] Sanghyun Woo, Jongchan Park, Joon-young Lee, and In So Kweon, “CBAM: convolutional block attention module,” in Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, 2018, pp. 3–19.

[18] YANG, Qing-Long Zhang Yu-Bin. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. arXiv preprint arXiv:2102.00240, 2021.

[19] RUBNER, Yossi; TOMASI, Carlo; GUIBAS, Leonidas J. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 2000, 40.2: 99-121.

[20] HU, Jie; SHEN, Li; SUN, Gang. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7132-7141.

[21] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” preprint 2017, arXiv:1703.03400.