# Multidimensional Attention Learning for VHR Remote Sensing Imagery Recognition

Jie Fang, Xiaoqian Cao, Pengfei Han, and Dianwei Wang

***Abstract*— This work presents a multidimensional attention learning-based lightweight convolutional neural network for very-high-resolution (VHR) remote sensing imagery recognition, which incorporates channel attention, spatial attention, and saliency sampler attention into the backbone to improve its recognition performance. Specifically, channel attention can alleviate the interference of original feature cube by adaptively giving different coefficient weights to different feature channels. Spatial attention can emphasize the discriminative regions by using the sum of activation value in different locations of the image to weight the original feature cube. Saliency sampler attention can increase the influence of interesting regions on the final representation according to the saliency priors. In general, different from the existing methods, this work utilizes the constraints rather than the model scale to improve the recognition performances of the network. In addition, the novel flooding loss is used to optimize the network, which can improve the performance of the framework by alleviating the severe overfitting problem.**

***Index Terms*— Flooding loss, multidimensional attention, very-high-resolution (VHR) imagery recognition.**

## I. Introduction

VERY-HIGH-RESOLUTION (VHR) remote sensing imagery recognition is a basal technique in the computer vision field, which aims to attach the specific semantic label to each image according to its apparent properties and content information. During the last decade, this task has become a hot topic since its important scientific research and practical application values in the Earth observation field. Compared to the natural image, the geometry and spatial structures of VHR remote sensing imagery are more complex. Specifically, VHR remote sensing imageries are photographed from high altitude and, hence, accompanied by variable angles and scales. Consequently, VHR remote sensing imageries cannot be easily modeled through traditional statistical handcrafted descriptors, such as SIFT [1] and LBP [2].

With the rapid progress of machine learning and artificial intelligence, deep neural network, especially deep convolutional neural network-based methods, has achieved competitive performances on many computer vision tasks, such as semantic segmentation [3], [4], salient object detection [5], [6], road crack detection [7], [8], and image quality prediction [9], [10]. As for VHR remote sensing imagery recognition, Fang *et al.* [11] proposed to incorporate statistical frequency domain information into the traditional space-domain feature representation framework to improve the rotation robustness of the model. Yuan *et al.* [12] exploited discriminative feature representations by emphasizing the location relationships of different local regions in the image. Huo *et al.* [13] proposed a multiscale-variational model to address the variable scale problem of the VHR remote sensing imagery.

Comparatively, the deep neural network-based methods have achieved relative performances on VHR remote sensing imagery recognition tasks, there still exist several disadvantages. First of all, the scales of the most existing models are relatively large, and they cannot be transformed into mobile devices conveniently, which limits their practical generalization and application. In addition, the existing methods often take the whole image as the input of the network to finalize recognition, which may ignore some important discriminative local objects since lack of effective guidance. Consequently, the recognition performances of these methods are not satisfactory as we expected. In these cases, Zhang *et al.* [14] proposed gradient boosting random convolutional network (GBRCN), which only contains two convolutional layers, and hence, its model scale is relatively small. Yang *et al.* [15] proposed object-guided classification methods based on joint use of deep-learning classifier and detector (OGJCD), which first searches the interesting objects in the image through a pretrained object network and then constructs discriminative features according to the detection results.

Even though the aforementioned methods can address the too large model scale and discriminant deficiency problems, they brought to new limitations. Specifically, GBRCN cannot work well when meeting large-scale and complex-geometry data sets since its insufficient nonlinear representation capability. OGJCD needs high computational load and the similar data distributions between the presented recognition set and the training set of the object detection model.

To address the aforementioned problems, we propose a multidimensional attention learning mechanism for VHR remote sensing imagery recognition. Specifically, we utilize the lightweight DenseRes50-Net (ResNet50 equipped with dense connections) as the backbone of the recognition framework, which can effectively control the network scale.

Manuscript received October 27, 2020; revised November 27, 2020; accepted December 13, 2020. Date of publication December 24, 2020; date of current version December 29, 2021. This work was supported in part by the Special Project of Strengthening Police With Science and Technology of Ministry of Public Security under Grant 2019GABJC42 and in part by the Fund of Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data, Yan'an University, under Grant IPBED6. *(Corresponding author: Jie Fang.)*

Jie Fang, Pengfei Han, and Dianwei Wang are with the Center for Imagery and Information Processing (CIIP), School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: 2443952262@qq.com).

Xiaoqian Cao is with the School of Electric and Control Engineering, Shaanxi University of Science and Technology, Xi'an 710021, China.

Digital Object Identifier 10.1109/LGRS.2020.3045496

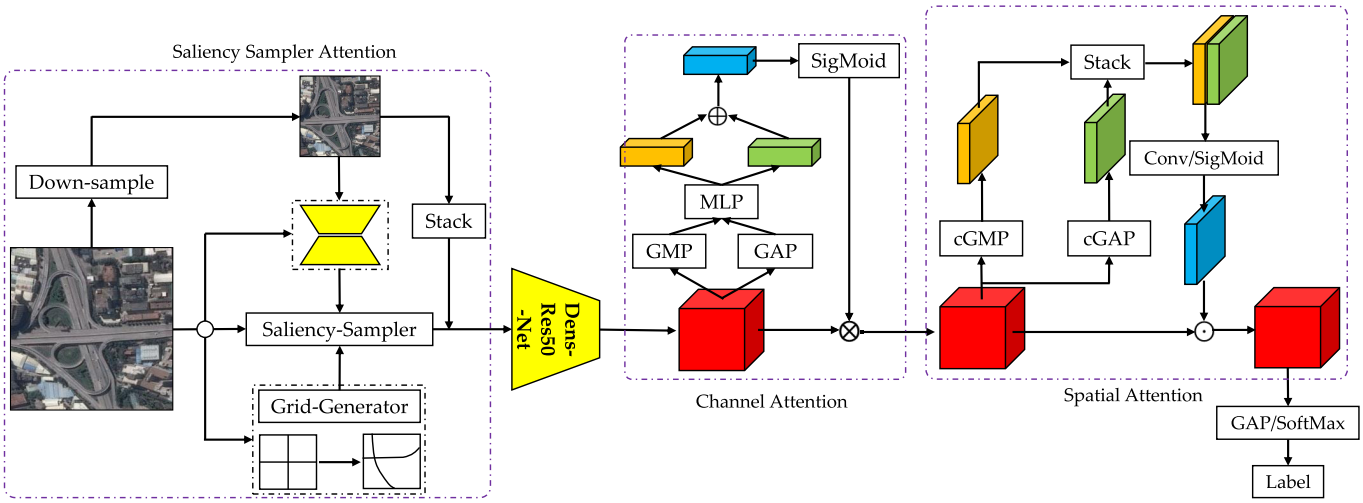


Fig. 1. Architecture of the proposed method multidimensional attention learning mechanism-based lightweight convolutional neural network, which mainly contains a saliency detection network, a conventional classification network, a channel attention module, and a spatial attention module.

In addition, we incorporate channel attention, spatial attention, and saliency sampler attention mechanisms into the framework as the constraints to improve its feature representation capability. In some detail, channel attention mechanism can filter out noisy and redundant feature channels to enhance the feature. Spatial attention mechanism attention can automatically emphasize the interesting object regions and improve the discrimination of the final feature vector. Combined with spatial attention mechanism, saliency sampler attention mechanism can enlarge the discriminative regions and compress the background regions and strengthen the recognition performance of the network furthermore. Finally, we use the flooding loss to optimize the proposed network, which can address the overfitting problem under small-scale training set. In summary, the contributions of this letter can be summarized as follows.

1) We propose a lightweight convolutional neural network for VHR imagery recognition, which can be deployed to mobile terminal conveniently since its small model scale and low computational load.
2) We incorporate channel attention, spatial attention, and saliency sampler mechanisms into our framework, which can effectively enhance the representation ability of the model through latent constraints.
3) We utilize the novel flooding loss to optimize the network, which can alleviate the overfitting problem especially when the scale of training set is not large enough.

## II. Proposed Method

### A. Overview

This section introduces the overall framework of the proposed multidimensional learning mechanism, and its diagram is shown in Fig. 1. Specifically, the original image is downsampled with a mean value window and a saliency-based sampler module to obtain two different sampled images. The former is to ensure the structure information of the image, whereas the latter is to emphasize the discriminative objects. Then, the sampled images are fed into a DenseRes50Net. In addition, outputs of two networks are concatenated and fused through a channel–spatial attention (CSA) module to finalize the recognition. Specifically, the CSA module contains a channel attention module and a spatial attention module, which can enhance the representation capability of the framework.

### B. Channel Attention

The procedures of channel attention can be described as follows. First of all, the original feature cube $\mathbf{X} \in \mathbb{R}^{W\times H\times C}$ is fed into a global average pooling (GAP) layer and a global max pooling (GMP) layer to obtain two vectors $\mathbf{v}^A \in \mathbb{R}^C$ and $\mathbf{v}^M \in \mathbb{R}^C$, which are defined in the following equation:

$$\mathbf{v}^A = \mathrm{GAP}(\mathbf{X}), \quad \mathbf{v}^M = \mathrm{GMP}(\mathbf{X}) \tag{1}$$

where $W$, $H$, and $C$, respectively, denote the width, height, and number of channels of $\mathbf{X}$. The $i$th element of $\mathbf{v}^M$ and $\mathbf{v}^A$ can be calculated by the following equation:

$$\mathbf{v}_i^A = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\mathbf{X}_i^{(w,h)}, \quad \mathbf{v}_i^M = \max_{1\le w\le W}\max_{1\le h\le H}\mathbf{X}_i^{(w,h)} \tag{2}$$

where $\mathbf{X}_i^{(w,h)}$ denotes the $(w,h)$th element in $i$th channel $\mathbf{X}$.

Second, these two vectors are encoded with two fully connected networks AEDNet and MEDNet to obtain two original activation weight vectors $\omega^A$ and $\omega^M$, which are defined in the following equation:

$$\omega^A = \mathrm{MADNet}\left(\mathbf{v}^A\right), \quad \omega^M = \mathrm{MEDNet}\left(\mathbf{v}^M\right). \tag{3}$$

Then, $\omega^A$ and $\omega^M$ are added together and then activated by a sigmoid layer to obtain the final activation weight vector $\omega$, which is defined in the following equation:

$$\omega = \mathrm{sigmoid}\left(\omega^M + \omega^A\right). \tag{4}$$

Finally, the multiplication of the original feature cube and the coefficient vector becomes the consequent output of this module, which is defined in the following equation:

$$\mathbf{O}^{ch} = \omega \bigotimes \mathbf{X} \tag{5}$$

where $\bigotimes$ denotes the expanded multiplication operator, and the $i$th channel of $\mathbf{O}^{ch}$ can be calculated with the following equation:

$$\mathbf{O}_i^{ch} = \omega_i \cdot \mathbf{X}_i \tag{6}$$

where $\omega_i$ denotes the $i$th element of $\omega$ and $\mathbf{X}_i$ represents the $i$th channel of the input feature cube $\mathbf{X}$.

### C. Spatial Attention

This section details the inner architecture of the spatial attention module. First of all, the original feature cube is fed into a channel-based GAP (cGAP) and a channel-based GMP (cGMP) layer to obtain two matrices $\mathbf{M}^A$ and $\mathbf{M}^M$, which are defined in the following equation:

$$\mathbf{M}^A = \text{cGAP}(\mathbf{X}), \quad \mathbf{M}^M = \text{cGMP}(\mathbf{X}) \tag{7}$$

where $\mathbf{X}$ is the input feature cube, and the $(w, h)$th element of $\mathbf{M}^A$ and $\mathbf{M}^M$ is calculated by the following equation:

$$\mathbf{M}_{w,h}^A = \frac{1}{C}\sum_{c=1}^{C} \mathbf{X}_c^{(w,h)} \quad \mathbf{M}_{w,h}^M = \max_{1\le c\le C} \mathbf{X}_c^{(w,h)} \tag{8}$$

where $C$ denotes the channel number of the input feature cube. Then, $\mathbf{M}^A$ and $\mathbf{M}^M$ are concatenated as $\mathbf{M}^{AM}$, which is defined in the following equation:

$$\mathbf{M}^{AM} = \left[\mathbf{M}^A \quad \mathbf{M}^M\right]. \tag{9}$$

In addition, a $1\times1$ convolutional layer and a sigmoid activation layer are applied to $\mathbf{M}^{AM}$ to obtain the spatial attention map $\mathbf{M}^{\omega}$, which is defined in the following equation:

$$\mathbf{M}^{\omega} = \text{sigmoid}\left(\text{Conv}^{1\times1\times1}\left(\mathbf{M}^{AM}\right)\right). \tag{10}$$

Finally, we can obtain the final output by multiplying the input feature cube $\mathbf{M}^{\omega}$ and spatial attention map, which is defined in the following equation:

$$\mathbf{O}^{sp} = \mathbf{M}^{\omega} \odot \mathbf{X} \tag{11}$$

where $\odot$ denotes the channel-wise multiplication operator, and the $i$th channel of $\mathbf{O}^{sp}$ can be calculated with the following equation:

$$\mathbf{O}_i^{sp} = \mathbf{M}^{\omega} \cdot \mathbf{X}_i \tag{12}$$

where $\cdot$ denotes the dot product.

### D. Saliency Sampler Attention

The saliency sampler attention framework can be found in [16], here we describe it in brief. First of all, the original high-resolution image $I$ is downsampled with ratio $r$ to obtain the corresponding low-resolution image $I_l$. Then, $I_l$ is fed into a saliency network to obtain the saliency map $S$, which can be formulated with the following equation:

$$S = f_s(I_l) \tag{13}$$

where $f_s(\cdot)$ denotes the saliency network. Finally, the $S$-based grid generator is used to downsample the original high-resolution image $I$, which can be formulated with the following equation:

$$J = g(I, S) \tag{14}$$

where $J$ denotes the final output of this module and $g(\cdot)$ represents the sample strategy described in [16], which can effectively enlarge the interesting regions while compressing the redundant regions in the image and further improve the consequent feature representation ability of the system.

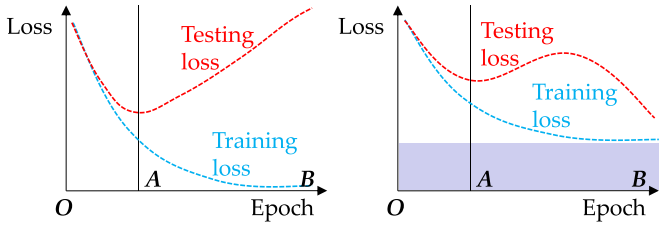


Fig. 2. Superiority of the flooding loss. The left figure shows the training loss and testing loss of conventional cost function, whereas the right one demonstrates the tendency of flooding-based cost function.

### E. Cost Function

The conventional cross-entropy loss is defined as follows:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} \mathrm{I}\{l^n = c\} \log\frac{e^{\theta_c^T x^n}}{\sum_{t=1}^{C} e^{\theta_c^T x^n}} \tag{15}$$

where $N$ denotes the sample number of the training set, $C$ represents the category number of the data set, $\theta$ denotes the parameters of the network, and $\mathrm{I}\{\cdot\}$ represents the indictor function, which equals 1 when the condition satisfies while 0 otherwise. In addition, inspired by the idea proposed in [17], we used flooding loss to alleviate the overfitting problem, which is defined in the following equation:

$$\tilde{\mathcal{L}}(\theta) = |\mathcal{L}(\theta) - b| + b \tag{16}$$

where $b$ is a hyperparameter, which is used to control the training degree. Specifically, the smaller $b$ means the heavier training degree. In addition, the superiority of the flooding loss can be reflected directly in Fig. 2, from which we can see that the incorporation of perturbation $b$ can contribute to the secondary decline of testing loss and improve the recognition performance further. Specifically, when $\mathcal{L}(\theta) \ge b$, the flooding cost function is optimized in decreased detection and, hence, ascend the fitting and discriminative capability of the model for training set progressively. Otherwise, when $\mathcal{L}(\theta) < b$, the flooding cost function is optimized in increased direction, which can alleviate the overfitting of the model for training set and, hence, enhance the generalization and robustness of the model.

## III. Experiments

### A. Data Sets

In order to demonstrate the effectiveness and superiority of the proposed method, we test our method on three public and challenging data sets, including AID [18], NWPU-RESISC45 [19], and OPTIMAL31 [20]. Specifically, the AID

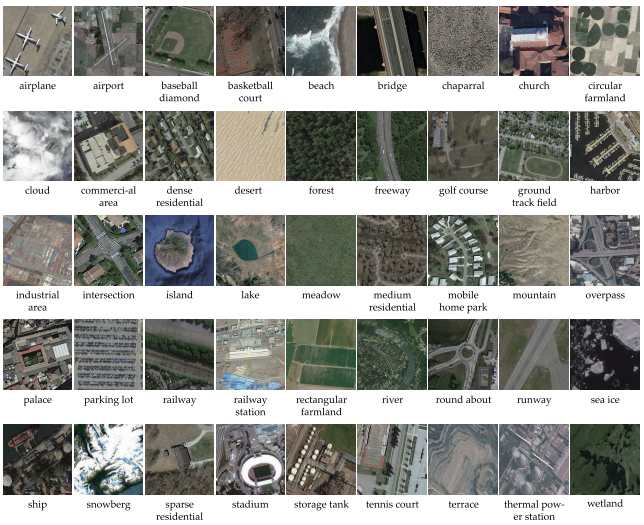


Fig. 3. Some samples of the NWPU-RESISC45 data set.

data set contains 10 000 images within 30 classes, and the sample of each class varies from 220 to 420. The NWPU-RESISC45 data set contains 31 500 images within 45 classes, and each class contains 700 images; some examples are shown in Fig. 3. The OPTIMAL31 data set contains 1860 images within 31 classes, and each class contains 60 images.

### *B. Experimental Settings*

The algorithm is implemented with Pytorch, and the testing platform is X99UD4 of GIGABBYTE, GPU (8G×8) of Titan X. The proposed model is trained by mini-batch Adam strategy with batch size 32. The weight decay is set to $5\times10^{-5}$. The initial learning rate is set to $10^{-4}$, the momentum is set to 0.9, and the model is trained for 100 epochs.

### *C. Evaluation Metrics*

We use the overall accuracy (OA) to test the performances of different methods, and its definition is shown in the following equation:

$$OA = \frac{\sum_{i=1}^{N_c} p_{ii}}{\sum_{i=1}^{C}\sum_{j=1}^{C} p_{ij}} \tag{17}$$

where $N_C$ represents the class number of the data set and $p_{ij}$ represents the number of samples with true label $m$ and predicted label $n$.

### *D. Contrasting Methods*

In order to validate the superiority of the proposed method, we compare it with eight existing state-of-the-art methods, including BoCF [21], Fine-tuned Alex-Net [19], Fine-tuned GoogLeNet [19], Fine-tuned VGG16-Net [19], D-CNN [22], CAD [23], GBRCN [14], and OGJCD [15].

### *E. Experimental Results*

*1) Ablation Experimental Results:* The ablation experimental results on the AID data set are shown in Table I. Specifically, we randomly choose 10% in the data set as the training set and the left as the testing set. In addition, Cha-Att, Spa-Att, Sam-Att, and Flo-Loss indicate the channel attention module, spatial attention, saliency sampler attention module, and flooding loss, respectively. From Table I, we can see that all the incorporated attention modules can attribute to the recognition performance to a certain extent. Specifically, when the channel attention is incorporated into the backbone network, it achieves a 3.67% increment in terms of OA. When the channel attention module, spatial attention module, and saliency sampler attention module are simultaneously incorporated into the network, it achieves 8.71% OA increment, compared to the original backbone network. In addition, when we use to replace the conventional cross-entropy loss with the flooding loss, it achieves 1.85 OA increment further. The reasons mainly contain three aspects, which are described as follows.

TABLE I
ABLATION EXPERIMENTAL RESULTS ON THE AID DATA SET (RATIO = 0.1)

| Cha-Att | Spa-Att | Sam-Att | Flo-Loss | *OA* (%) |
|---|---|---|---|---|
| – | – | – | – | 80.91 |
| √ | – | – | – | 84.58 |
| √ | √ | – | – | 86.73 |
| √ | √ | √ | – | 89.62 |
| √ | √ | √ | √ | 91.47 |

1) The channel attention can alleviate the interferences of redundant and noisy channels of the feature cube. Besides, the spatial attention can enhance the discriminative regions while suppressing other background and irrelevant regions.
2) The saliency sampler attention module can enlarge the interesting districts while compressing other districts in the original input image and improving sufficient useful information for the consequent processes.
3) The flooding loss can alleviate the overfitting problem by lightening the severe constraints of the model for the training set.

*2) Contrasting Experimental Results:* The contrasting experimental results are shown in Table II, from which we can see that three fine-tuned CNN-based methods achieve better performances than BoCF [21]. The reason is that end-to-end backward learning style optimizes the feature extractor and classifier simultaneously, and it can reinforce each other and further dig out more discriminative semantic information of the image. Besides, D-CNN [22] achieves more comparative performance than conventional CNNs because the metric learning-based optimization strategy can distinguish the diversity among different category pairs, which can improve the discrimination capability of the model. In addition, OGJCD [15] surpasses D-CNN because the object-guided strategy can exploit more discriminative characteristics of the image through filtering out complex background information. Finally, our method achieves better recognition performance than OGJCD [15], and the reasons can be concluded as the following three points.

TABLE II
CONTRASTING EXPERIMENTAL RESULTS ON THREE DATA SETS (RATIO = 0.4)

| Dataset | Method | *OA*(%) | *Std* |
|---|---|---|---|
| NWPU-45 | BoCF [21] | 84.42 | 0.18 |
| | Fine-tuned AlexNet [19] | 85.04 | 0.25 |
| | Fine-tuned GoogLeNet [19] | 89.75 | 0.19 |
| | Fine-tuned VGG16-Net [19] | 86.33 | 0.20 |
| | D-CNN [22] | 91.84 | 0.25 |
| | CAD [23] | 94.40 | 0.34 |
| | GBRCN [14] | 87.91 | 0.28 |
| | OGJCD [15] | 94.62 | 0.22 |
| | MAL (ours) | 95.94 | 0.23 |
| OPTMAL31 | BoCF [21] | 79.78 | 0.20 |
| | Fine-tuned AlexNet [19] | 81.34 | 0.18 |
| | Fine-tuned GoogLeNet [19] | 83.61 | 0.27 |
| | Fine-tuned VGG16-Net [19] | 87.91 | 0.32 |
| | D-CNN [22] | 92.08 | 0.35 |
| | CAD [23] | 93.12 | 0.24 |
| | GBRCN [14] | 88.25 | 0.30 |
| | OGJCD [15] | 93.91 | 0.28 |
| | MAL (ours) | 94.90 | 0.31 |
| AID | BoCF [21] | 81.58 | 0.29 |
| | Fine-tuned AlexNet [19] | 82.43 | 0.31 |
| | Fine-tuned GoogLeNet [19] | 83.91 | 0.24 |
| | Fine-tuned VGG16-Net [19] | 86.25 | 0.27 |
| | D-CNN [22] | 91.16 | 0.20 |
| | CAD [23] | 94.87 | 0.29 |
| | GBRCN [14] | 88.28 | 0.31 |
| | OGJCD [15] | 96.55 | 0.24 |
| | MAL (ours) | 97.06 | 0.25 |

1) The channel attention, spatial attention, and saliency sampler attention modules are incorporated into the baseline framework, which can alleviate the redundant feature channels and enlarge and emphasize the discriminative regions of the image to enhance the discrimination of the final feature.
2) Different from the transfer learning mechanism used in OGJCD, our model is completely trained with the only data set specific to the very task and, hence, avoids the data gap problem.
3) The flooding loss can effectively avoid the overfitting problem, which can further strengthen the generalization of the method. Generally speaking, the proposed method achieves a relatively satisfactory performance.

## IV. CONCLUSION

In this letter, we propose a multidimensional attention learning-based lightweight convolutional neural network for VHR remote sensing imagery recognition. Specifically, we incorporate channel attention, spatial attention, and saliency sampler attention modules into the backbone network to enhance its representation capability. In other words, we utilize the constraints coding rather than the data driven under large-scale labeled samples to improve its performance, which can reduce the computational load of the model in practical applications. In addition, we use the flooding loss to optimize the network, which can alleviate the overfitting problem in the training phase and improve its generalization further.

## REFERENCES


[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[2] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[3] J. Fang and X. Cao, "GAN and DCN based multi-step supervised learning for image semantic segmentation," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Springer, 2018, pp. 28–40.
[4] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Spatial structure preserving feature pyramid network for semantic image segmentation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3, pp. 1–19, 2019.
[5] X. Li, D. Song, and Y. Dong, "Hierarchical feature fusion network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 9165–9175, 2020.
[6] L. Han, X. Li, and Y. Dong, "Convolutional edge constraint-based U-net for salient object detection," *IEEE Access*, vol. 7, pp. 48890–48900, 2019.
[7] J. Fang, B. Qu, and Y. Yuan, "Distribution equalization learning mechanism for road crack detection," *Neurocomputing*, 2019.
[8] S. Wu, J. Fang, X. Zheng, and X. Li, "Sample and structure-guided network for road crack detection," *IEEE Access*, vol. 7, pp. 130032–130043, 2019.
[9] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432–442, Sep. 2018.
[10] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Trans. Image Process.*, vol. 29, pp. 7414–7426, 2020.
[11] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space–frequency joint representation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7492–7502, Oct. 2019.
[12] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, Mar. 2019.
[13] L.-J. Huo, H. E. Bin, and D.-B. Zhou, "A destriping method with multi-scale variational model for remote sensing images," *Opt. Precis. Eng.*, vol. 25, no. 1, pp. 198–207, 2017.
[14] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.
[15] X. Yang, W. Yan, W. Ni, X. Pu, H. Zhang, and M. Zhang, "Object-guided remote sensing image scene classification based on joint use of deep-learning classifier and detector," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2673–2684, 2020.
[16] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Learning to zoom: A saliency-based sampling layer for neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 51–66.
[17] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, "Do we need zero training loss after achieving zero training error?" 2020, *arXiv:2002.08709*. [Online]. Available: http://arxiv.org/abs/2002.08709
[18] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[19] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[20] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
[21] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.
[22] 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, May 2018.
[23] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based densenet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, Jul. 2020.