Computer Science and Information Systems 20(2):743–763 https://doi.org/10.2298/CSIS220714015W

Large-scale Image Classification with Multi-perspective Deep Transfer Learning

Bin Wu[1], Tao Zhang[2], and Li Mao[3]

[1] School of Internet of Things Engineering, Jiangnan University,
Wuxi, 214122, China
wubin@jiangnan.edu.cn
[2] China Ship Scientific Research Center,
Wuxi 214122, China
taozhang@jiangnan.edu.cn
[3] School of Artificial Intelligence and Computer Science, Jiangnan University,
Wuxi, 214122, China
wxmaoli@jiangnan.edu.cn

Abstract. Most research efforts on image classification so far have been focused on medium-scale datasets. In addition, there exist other problems, such as difficulty in feature extraction and small sample size. In order to address above difficulties, this paper proposes a multi-perspective convolutional neural network model, which contains channel attention module and spatial attention module. The proposed modules derive attention graphs from channel dimension and spatial dimension respectively, then the input features are selectively learned according to the importance of the features. We explain how the gain in storage can be traded against a loss in accuracy and/or an increase in CPU cost. In addition, we give the interpretability of the model at multiple scales. Quantitative and qualitative experimental results demonstrate that the accuracy of our proposed model can be improved by up to 3.8% and outperforms the state-of-the-art methods.

Keywords: large-scale image classification, channel attention module, spatial attention module, interpretability of the model, multiple scales.

1. Introduction

In computer vision, image classification is a hot subject, and it is also an important foundation in some fields such as object detection [1, 2], face recognition [3], pose estimation [4, 5], population density estimation [6, 7], image segmentation [11, 12] etc. However, due to the large number of image categories and the limitation of computing resources, it is difficult for traditional classification algorithms to achieve a high accuracy. At present, the image classification algorithm mainly relies on the deep neural network model [8, 9, 10, 11]. Different from the traditional image classification method, deep learning scheme does not require complex feature decomposition of the target image. However, the unexplained black box in AI system makes users can only get detection results, and cannot give the reasons [14, 15]. In this regard, the Explainable AI (XAI) is widely developed in recent years [16, 17]. Generally, different decision-makers

focus on different goals in XAI to ensure the high performance and provide a reasonable explainable model synchronously, our work focuses on the robust construction of large-scale image classification model, we also focus on the interpretability of the model at multiple scales.

One main difficult task of image classification is the feature extraction [18, 19]. The so-called feature extraction refers to constructing an algorithm to extract features in the target image, such as the edge feature, the color feature, etc. At present, the best approach is to use deep neural network model to conduct image classification tasks [20, 21]. The experimental result of VGGnet shows that better classification accuracy could be obtained by constructing deeper convolutional neural network model. Inspired by this idea, the deep residual network is constructed by cross-layer connection method, and it achieves higher classification accuracy. GoogLeNet increases the adaptability of the network to different scales, which also obtain better classification accuracy. ResNeXt and Xception added cardinality to the network model, proving that the cardinality can not only reduce the overall parameters of the model, but also has a strong ability of representation. However, considering the fact that the characteristics of different scales are characterized by different feature and highly complementary, it is necessary to consider the complementarity between different feature layers.

The attention mechanism is similar to that of human vision. It depicts a phenomenon, we always focus our attention on the main things to get key information [22]. Attention mechanics have been used in a wide range of scenarios. The neural network could captures more critical information with the help of the attention mechanism. The attention mechanism enables the neural network model to be self-explanatory in the channel dimension and the spatial dimension. In order to give the interpretability of attention mechanism more clearly, the channel attention mechanism (CAM) and the spatial attention mechanism (SAM) are firstly considered in the attention domain. Our attention domain mainly consists of three types: spatial domain, channel domain and mixed domain. In our experiments, it is found that better performance is obtained with using the CAM.

In recent years, many researchers have tried to solve these problems. Wolpert believe that the reason for the evolution of the brain is not for thinking and feeling, but for controlling motion, which is the core idea of the Deep Reinforcement Learning [23, 24]. Curriculum learning and self-paced learning represent the recently proposed learning strategy. Their core idea is to simulate the cognitive mechanism of human beings, they first learn simple and general knowledge structure, then gradually increase the difficulty degree and transition to more complex and professional knowledge. These two methods have great advantages in handling small samples of data and the interpretability of models. However, no one method can solve the problems of small sample size and unrestricted scenes perfectly. In order to clear the aim of this paper, we select some representative works, and provide a comparative analysis of related works in form of table discussing the advantages, disadvantages, techniques, objectives, and limitations etc, as is shown in Table 1.

In view of the existing problems of the field of large scale image classification in complex environments, we construct a novel feature self-selection scheme in the feature layer to get more useful semantic information of the deep network, we try to combine the image loss function and a novel adversarial loss term to optimize our network.

Table 1. Comparative analysis of related works

Algorithms	Advantages	Disadvantages	Techniques	limitations
Detection-based approaches [3-9]	Applicable to simple monitoring environments	The recognition error is high when the background is complex	General feature extraction strategy	Lack of interpretability of constructed model
Attention-based approaches [10-18]	Real-time performance	Work not well under variety of scenes and camera angles	channel attention mechanism, spatial attention mechanism, etc	Complex and unrestricted scenes
Deep learning scheme [19-26]	The detection rate and recognition rate are high and stable	Easy to fall into local optimum, and the decision is short of interpretability	CNN network, RNN network, multi-view deep learning, etc	Lack of interpretability of constructed model
Multi-task learning scheme [28-34]	The model is more stable and can detect multi-scale objects	Training process is complicated and time-consuming, easy to fall into local optimization	learning multiple task classifiers jointly	Complex and unrestricted scenes
Explainable AI [27,35-38]	The interpretable ways of various deep models are given	Theoretical analysis is not enough, Small sample size	Image processing and visual technologies	Complex and unrestricted scenes
Ours	**It combines the advantages of existing algorithms and is suitable for various scenes**	**The real-time performance of the algorithm needs to be improved**	**CNN, Transfer learning, self-paced learning, etc**	**Real-time requirement**

In view of the existing problems of the field of large scale image classification in complex environments, we construct a novel feature self-selection scheme in the feature layer to get more useful semantic information of the deep network, we try to combine the image loss function and a novel adversarial loss term to optimize our network.

Main contributions of this paper are described as follows:

First, our work focuses on the robust construction of image classification model. We propose an multi-channel convolutional neural network model consisting of channel attention module and spatial attention module, which could potentially explain the classification results;

Second, considering the fact that the characteristics of different scales are characterized by different feature and highly complementary, we construct a novel feature self-selection scheme in the feature layer to get more useful semantic information of the deep network;

Third, a nonlinear model based on deep transfer learning is proposed to solve the the classification problem of multiple perspectives, and we construct a novel adversarial loss term to optimize our network;

Last, in order to give the interpretability of the model at multiple scales and reduce the error loss in the constructed model, self-paced learning and transferring learning modules are designed to optimize the constructed model.

In this paper, the traditional deep feature learning is further exploited from the explainable view, and the classification effect is improved by our constructed multi-channel convolutional neural network model consisting of channel attention module and spatial attention module. The main structure of this paper is as follows: Section 2 starts with a brief presentation of necessary related concepts. Section 3 gives our proposed channel and spatial attention module based on explainable machine learning. Designed experiments and discussion are depicted in Section 4. Finally, some conclusions are presented in Section 5.

2. Related Work

Explainable AI (XAI) refers to those Artificial Intelligence techniques aimed at explaining, to a given audience, the details or reasons by which a model produces its output [25]. Consequently, XAI borrows concepts from philosophy, cognitive sciences and social psychology to yield a spectrum of methodological approaches that can provide explainable decisions for users without a strong background on Artificial Intelligence. Therefore, XAI targets at bridging the gap between the complexity of the model to be explained, and the cognitive skills of the audience for which explainability is sought. Interdisciplinary XAI methods have so far embraced assorted elements from multiple disciplines, including signal processing, adversarial learning, visual analytics or cognitive modeling. Although reported XAI advances have risen sharply in recent times [34, 35], there is global consensus around the need for further studies around the explainability of machine learning models. This includes interpretable reasoning of models, neurosymbolic reasoning or systems based on fuzzy rules, etc..

At present, explanations provided by different algorithms are fragmented and independent, which makes it difficult to determine reasonable decisions and explain model structures. Generally, current interpretable classifiers could be classified into the following four kinds: the selection of optimal training set, correlation selection of heat graph, semantic analysis, model visual interpretation [27, 32, 33]. Existing interpretable artifical intelligence models based on the selection of optimal training set can provide the basis behind the classification [34, 37]. However, there is no mechanism for identifying potential misclassification of classifiers in the existing classification model. Warning users about misclassification will help prevent errors from entering the system. One of the reasons for misclassification is the reduction of distance between classes. Some outliers or edge elements of a class can share the common characteristics of adjacent classes. However, there is no mechanism to ensure the number of subclasses of a given class and whether it makes sense to merge closely related subclasses of two adjacent classes into a new class and implement the correct classification. The interpretable artifical intelligence models based on correlation selection of heat graph are another way to understand it. Xu [26] proposes a solution based on database transaction model interpretation, whose explanation is on the basic of logical structure or reasoning. The static structure makes it unsuitable for the deep network classifier. In the constructed model, the system dynamically give appropriate explanations from stored vocabularies, which in turn are generated based on model learning. It provides a consistent view of models and interpretations beyond the scope of existing technology.

Simonyan [29] proposes an interpretable model in which they are interactively validated through visual features and similarity. Moreover, k-means clustering was used to analyze the similar features, so that the average features obtained had greater robustness and relatively low time complexity. In addition, it does not consider the importance or relevance of model, nor does it cluster with respect to output classes.

Some researchers try to use semantic analysis to explain the models. He [28] conducts the explanatory demonstration of the model. By observing the change of the connection mode of the network, the hidden layer is explained visually. However, the feature learned by the network layer are not described in detail. Sengupta [30] proposes an image interpretation and generation method based on semantic analysis. Take an image signature with a fixed length of 8000 to generate a caption. In this model, the correlation of features is firstly determined and the signature generation is carried out on this basis. Since eigenvalues could be of any length, strategies that follow highly correlated features are interpretable. As the power of interpretation becomes more important in intelligent decision-making, AI systems are no longer there to serve as black boxes [31, 37].

Model visual interpretation is also another way to understand the interpretability of the model. Russakovsky [36] proposes a novel model based on the attention mechanism, and it is named the residual attention network. As the network layer deepens, attention modules can extract key information from different layers. Finally, it got 4.8% Top-5 error rate on ImageNet. Hu [38] proposes the SENet 44 network model. In the training process, the model can distinguish the importance of different channels, then enhances useful features and inhibits useless features according to the importance of feature. Finally, it won the ILSVRC2017 classification task championship with a 2.25% Top-5 error rate. Moreover, the text interpretation can't match the characteristics of a certain layer of deep learning network, and it lacks of continuous interpretibility. In general, the text explanation generated for classification comes from training data based on model annotations. Up to now, data labels are set manually and are very subjective, and it doesn't take into account the differences between the different elements. Therefore, it is not possible to determine the relevant region of the image that is most useful for classification. In most cases, experts are encouraged to use their attribute label data as interpretable evidence.

Therefore, this paper proposes a novel image classification framework based on multi-perspective deep transfer learning. Attention mechanism is used to describe varied image characteristics, the self-paced learning strategy is adopted to solve the problems of small number of labeled samples and model interpretation. A nonlinear model based on deep transfer learning is proposed to solve the classification problem of multiple perspectives, also the interpretability of the model is further improved.

3. The Proposed Channel and Spatial Attention Model

In view of the existing problems in the field of image classification in complex environments, this paper considers the solutions based on deep network, and further studies popular algorithms such as deep reinforcement learning, self-paced learning and transfer learning, aiming at the urgent model explanation problems in the framework of

deep network. The overall technical framework containing CAM and SAM is shown in Fig. 1.

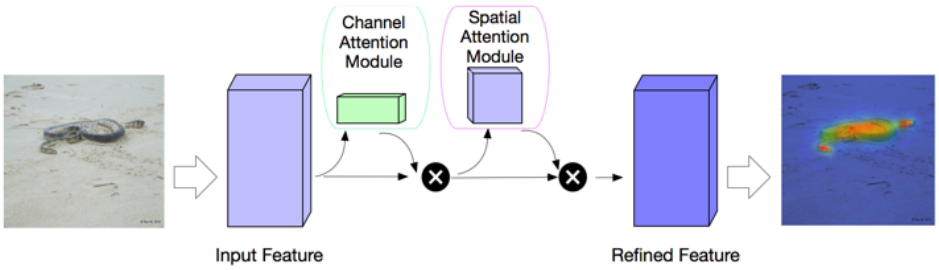

Fig. 1. The overview of model

For the convolutional neural network model, depth, width and attention mechanism are the main factors affecting the accuracy of image classification. At present, attention mechanism contains CAM and SAM. CAM acts on the channel domain, weighting different channel features. For a $C \times H \times W$ feature graph, the weight of channel attention C is different, while the weight of H $H \times W$ is the same. For CAM, the weight of each C on different channel dimensions needs to be learned. To reduce the amount of computation and improve classification accuracy, the pooling layer in the general convolutional neural network directly uses the maximum pooling method or average pooling method to compress the image information. For SAM, only the key information in the spatial features is extracted. We first introduce the general framework of proposed model with CAM and SAM in this section. Finally, we describe how to combine them together.

After convolution operations, an intermediate feature map $F_{\text{in}} \in R^{C \times H \times W}$ is obtained. F_{in} is the input of model. One-dimensional channel map $M_{\text{c}} \in R^{C \times 1 \times 1}$ is inferred after the channel attention module and two-dimensional spatial map $M_{\text{s}} \in R^{1 \times (H \times W) \times (H \times W)}$ is computed, and the corresponding spatial attention module is shown in Fig. 1. The entire attention calculation process could be summarized as:

$$F_1 = M_{\text{c}}(F_{\text{in}}) \otimes F_{\text{in}} \quad (1)$$

$$F_{\text{out}} = M_{\text{s}}(F_1) \otimes F_1 \quad (2)$$

where $\otimes$ denotes element-wise multiplication. F_1 is achieved after F_{in} passing the channel attention module. F_{out} is the final output and the attention value is broadcasted during multiplication process. Attention feature are the multiplication of element levels, which will be propagated automatically, indicating that channel attention broadcasts along spatial dimension, while spatial attention broadcasts along channel dimension.

3.1. Channel Attention Module

The CAM captures the relationship between channel features. It focuses on the key channel information and weakens the influence of the useless channel information [39, 40]. It uses an attention mechanism (similar to the self-attention mechanism, such as query, key, value) to get the similarity between channel graphs, and then use the weight of channel graphs to update. Finally, the matrix of computation attention is obtained, which can enhance the key features. The CAM makes the neural network model pay more attention to the channel features with the key information. On the basis of convolution, we first extrude the feature graph to obtain the global feature of each channel. Then, we use the global feature to get the relationship between different channels and the weight of different channels. Finally, we multiply the weights to get the features on the basis of the original feature graph.

In a convolutional neural network, the convolution operation only performs on image feature space, it is difficult for the convolution module to get the relationship between different feature channels. To get an eigenmatrix, an image needs to go through several convolutional layers and the number of channels represents the number of cores in the convolutional layers. In a normal neural network, the number of convolution kernels is usually as high as 1024 or 2048. Therefore, not each channel is useful for feature extraction. The CAM will help the neural network model select more informative channels. Besides, We encode spatial features using a global average pool that provides feedback for each pixel on the feature map. The following formula shows the global average pool calculation process.

$$F_{\text{avg}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{\text{in}}(i.j) \tag{3}$$

where $F_{\text{avg}} \in R^{C \times 1 \times 1}$ represents the result after implementing the global average pooling on the input feature map F_{in}. In order to capture the relationships between different channels, two conditions need to be met for CAM: firstly, it must be flexible, because it needs to learn the nonlinear relationship between different channels; Secondly, the learning relationship is not mutually exclusive, because it allows for a multichannel feature instead of a hot spot form. We describe the channel attention map M_{c} as follows.

$$M_{\text{c}} = \sigma(W_2 \operatorname{Re} LU(W_1 F_{avg})) \tag{4}$$

where $W_1 \in R^{\frac{C}{\text{r}} \times C}$ and $W_2 \in R^{C \times \frac{C}{\text{r}}}$, σ denotes Sigmoid function. W_1 and W_2 are fully connected layers. To improve the explainable ability of the model, we construct two fully connected layers, namely the bottleneck structure. W_1 denotes the dimensionality reduction layer and the dimensionality reduction factor r is a super parameter. Then the Relu function is used and the W_2 layer restores the dimension to the original number. The process is shown in Fig. 2.

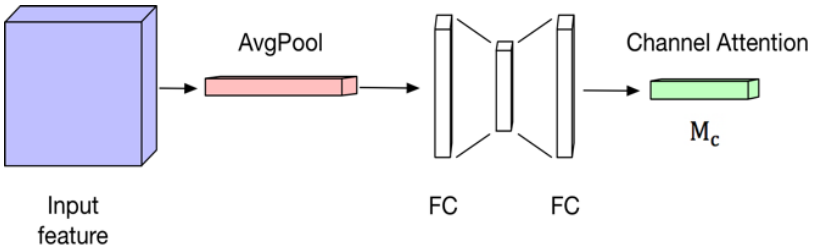

Fig. 2. The details of channel attention module

3.2. Spatial Attention Module

Unlike CAM, SAM only plays the role of distinguishing key information within a single image feature map. First of all, we use average pooling and max pooling to compress the input feature and then we use mean and max operations on the input feature at channel dimension. Finally, we could get two two-dimensional features. Considering different channel size, the two two-dimensional features are combined together to obtain the feature with channel number of 2. And they are convolved to ensure that the resulting features are consistent with the input feature in spatial dimension.

Spatial attention map is generated by adding the internal relationship, as shown in Fig. 3. $A \in R^{C\times H\times W}$ represents the input of the SAM. After conducting convolutional layers, feature maps B, C and D ($(B,C,D) \in R^{C\times H\times W}$) are generated. We reshape B and C to $R^{C\times H\times W}$ and $H\times W$ represents the resolution in spatial module.

Then we get $R^{(H\times W)\times(H\times W)}$ by operating a matrix multiplication between the transpose of B and C. The spatial attention map M_c could be obtained when a softmax layer is applied. M_c is computed as follows:

$$M_{sij} = \frac{\exp(B_i \times C_j)}{\sum_{i=1}^{H\times W} \exp(B_i \times C_j)} \tag{5}$$

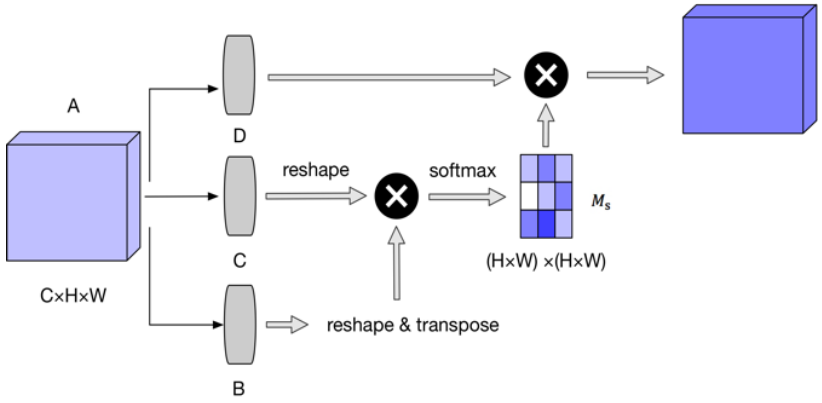

Fig. 3. The details of spatial attention module

3.3. Explainable Object Detection Model of Combing Self-paced Learning and Deep Reinforcement Learning

The deep network model is designed for learning across model characteristics. This paper proposes a new application based on the deep network framework, using deep network to learn multi-mode. In particular, this paper demonstrates that cross-modal feature learning could effectively relate different features to obtain a discriminant feature. In addition, the paper designed how to learn a shared feature between multiple modes and evaluate it on a particular task. Multi-mode explainable deep network model can be seen in Fig. 4. This model consists of three streams, video information, text information and audio information. The structure of the three streams is identical, each consisting of eight layers (including the input layer).

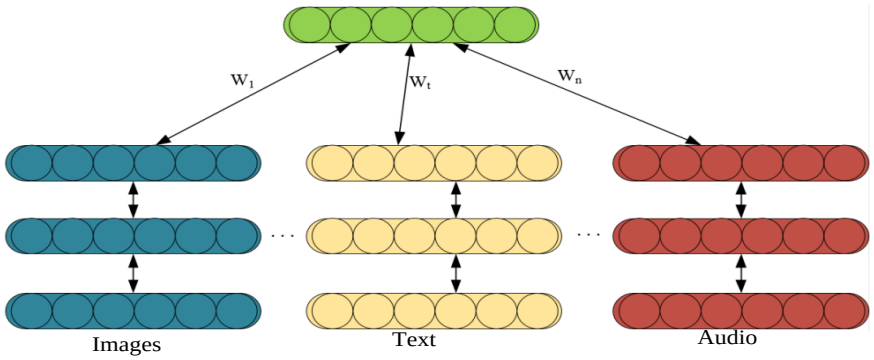

Fig. 4. Proposed multi-mode explainable deep network model

There are two problems with the traditional multimodal model. First, there is no clear goal for the model to find correlations across modes. Some hidden layer units adjust the parameters only for one stream, and others adjust the parameters only for other stream, so that it is possible for the model to find the desired feature. Second, there is only one mode for supervised training and testing in the cross-modal learning arrangement, which makes the model unexplainable. If there is only one modal representation, it is necessary to integrate observable variables that are not observed. Therefore, this paper proposes a deep self-coding model to solve the above problems. Inspired by the noise-reducing self-coding model [41], this paper proposes a training multi-mode deep self-coding model (Fig. 4), which uses an extended (extended single-mode input) but noisy data set. In fact, the model is still required to reconstruct the three modes when one mode uses zero as input and the other uses the original value as input when expanding. Therefore, one-third of the training data is input only by video, one-third of the training data is input only by text, and the last third is input only by voice. This model can be viewed as an example of multitasking learning.

Algorithm 1: Training process of our proposed network

Input: training data set $D\{(x_1,l_1),(x_2,l_2),\Lambda,(x_t,l_T)\}$

Weight K

Initialize the experience pool *M*

Randomly initialize network parameters θ

for k=1 to K do :

Scramble training data sets D

Initialize the initial state s_1 to x_1

for t=1 to T do :

Choose an action based on the ε-greed strategy

$a_t = \pi_\theta(s_t)$

$r_t, \text{terminal}_t = \text{Reward}(a_t, l_t)$

Set the state s_{t+1} to x_{t+1}

Store experience data into the experience pool M

from M

Set y_j as:

If $\text{terminal} = \text{True}$, set as r_j ;

If $\text{terminal} = \text{False}$, set as $r_j + \gamma \max_a Q(s, a_j, \theta_{k-1})$

Gradient descent

$L_\theta = (y_i - Q(s_j, a_j; \theta_{k-1}))^2$

If $\text{terminal}_t = \text{True}$,

break

When designing the intensification strategy, this paper uses the Q network to interact with its environment during the data generation phase. The system looks at the current scene, which consists of multi-mode data, and takes actions using the ε-greedy strategy. This environment in turn provides scalar rewards. Interaction experiences are stored in replay memory M. Replaying M preserves N recent experiences, which are then used to update the network parameters during the training phase. In the training stage, the network structure will use the data stored in replay memory M to train the network. Assume that the superparameter n represents the number of experiences replay, and for

each experience replay, a mini-cache B containing several interactions is randomly sampled from the finite size replay memory M. The model will be trained by sampling from cache B, and the parameters of the network will be updated iteratively in the direction of The Behrman target. The algorithm is divided into two phases to avoid latency. Therefore, this paper divides the algorithm into two stages: in the first stage, the robot collects data through limited time interaction with human beings; In the second stage, the training phase is activated to train the multimodal depth Q network.

In order to avoid non-convex optimization problems falling into poor local solutions, the proposed network optimization method adopts multiple random initializations to train the model, and then chooses the initialization network with the best performance to construct the model. However, this method is too adhoc and the calculation cost is too high. Self-learning is the best solution to non-convex optimization problems. The aim of curriculum learning is to simulate the cognitive mechanism of human beings by first learning simple and universal knowledge structure and then gradually increasing the difficulty to learn more complex and specialized knowledge. However, self-paced learning has been improved in curriculum learning. Instead of assigning prior knowledge to sample learning sequence in advance, the learning algorithm itself determines the next learning sample in each iteration. Detailed algorithm is described as Algorithm 1.

As mentioned in the introduction part, the traditional method uses L2-based regression to train a multi-scale network and finally forms a fused prediction density map. It also mentions why it is only based on L2 regression. In order to solve such problems and make the final density clearer, we choose to use adversarial loss. The adversarial generative network involves the model generator G and model discriminator D. The two are like playing a minmax game: the image generated by the training generator G deceives model D, while the aim of training D is to distinguish the synthesized image from the actual image, if it is inconsistent, it is Fake, and if it is consistent, it is True [42]. In our method, this adversarial loss is defined as follows:

$$L_A(G,D) = E_{x,y\sim P_{data}(x,y)}[LogD(x,y)] + E_{x\sim P_{data}(x)}[Log(1-D(x,G(x)))]. \quad (6)$$

Where x denotes the training patch and y denotes the corresponding ground heat map. The aim of G is to minimize this goal, and D tries to maximize it.

The traditional pixel-by-pixel Euclidean loss is based on large deviations between pixels, it will make the feature map fuzzy when facing sharp edges or outliers, and thus the generated density map will become fuzzy. But the adversarial loss discards the large deviation between the existing pixels. It is a binary judgment for each pixel, either true or false.

Because of the lack of punishment constraint on the ground real image, only using adversarial losses can sometimes lead to abnormal spatial structure. As suggested in previous work, we also use two common losses to smooth the solution. Details are as follows:

• Euclidean loss: In our model, the L2 loss is also adopted to change the estimated density map G into the discriminator model D to approximate the basic facts in the sense of L2. Assuming that W×H resolution image with c channels is constant, we design the following rule to depict the pixel-by-pixel loss:

$$L_E(G) = \frac{1}{C}\sum_{c=1}^{c} \| P^G(c) - P^{GT}(c) \|_2^2 . \quad (7)$$

Where $P^G(c)$ represents the pixel generating the heat map and $P^{GT}(c)$ indicates the pixel of the ground truth map, here we set C = 3.

• Perceptual loss: This kind of loss function was originally added into the image task by Johnson [43]. For image conversion and super-resolution tasks, it compares the features obtained by convolution of the real picture with the features obtained by the convolution of the generated picture, making the high-level information (content and global structure) more similar, whose aim is to minimize the perceived difference between the two. The definition of perceived loss is as follows:

$$L_p(G) = \frac{1}{c}\sum_{c=1}^{c} \| f^G(c) - f^{GT}(c) \|_2^2 . \quad (8)$$

Where $f^G(c)$ represents the pixels in the advanced receptiveness features of the previous heat map and $f^{GT}(c)$ denotes the pixels in the following receptiveness features of the ground truth, it should be noted that C is set to 128.

So the overall first-stage loss could be defined as follows:

$$L_I = ArgMin_G Max_D L_A(G,D) + \lambda_e L_E(G) + \lambda_p L_p(G). \quad (9)$$

Among them, λ_e and λ_p are the weights that measure Euclidean loss and perceptual loss. In this experiment, we set $\lambda_e = \lambda_p = 150$.

During our model training process, the patches are sent to G-large and G-small, respectively. To obtain the estimated heat map P-parent and generated by G-small at the end of the subnetwork, the four sub-pictures P-child are spliced to form P-concat. The constraint of cross-scale consistency is imposed on P-concat and P-parent. In general, given the W×H density map with c channels, the cross-scale consistency constraint based on L2 loss is defined as follows:

$$L_c(G) = \frac{1}{c}\sum_{c=1}^{c} \| P^{prt}(c) - P^{cnt}(c) \|_2^2 . \quad (10)$$

Where $P^{prt}(c)$ denotes the pixels in the parent block heat map, and $P^{cnt}(c)$ denotes corresponding pixels after the child block density map is spliced, C is set to 3.

Algorithm 2: Reward scheme

Deifne Reward function $(a_t \in A, l_t \in L)$:
  Initialize Terminal as False
  If $s_t \in D_p$,then
    If $a_t = l_t$, then
      Set r_t as 1
    else
      Set r_t as -1
      $terminal_t = True$,
  else
    if $a_t = l_t$, then
      Set r_t as λ
    else
      Set r_t as $-\lambda$
  return r_t , $terminal_t$

The final goal is to combine the four loss functions mentioned above to achieve the final loss function:

$$L_{II} = L_1 + \lambda_c L_c(G). \quad (11)$$

Among them, λ_c is a predefined weight to achieve cross-scale consistency with respect to constructed loss function. It's worth noting that if λ_c is set to 0, the two normal forms will be generated independently. In addition, we define the corresponding reward scheme, as is shown in Algorithm 2.

4. Experimental Results and Analysis

4.1. Datasets and Experiments Settings

The dataset is used in the experiment including ImageNet- 1K and Cifar-10050 [29, 31, 44, 45, 46]. The Cifar-100 contains 100 classes and each class has 600 color images. But each images is only a size of 32*32. Five hundred images in each class serve as the training set and the rest as test set. For each image, it has two labels, fine-labels and coarse-labels, which represent the fine-grained and coarse-grained labels of the image respectively, and Cifar-100 is hierarchical. In Fig. 5, we extracted the images from Cifar-100 as the visual example.

ImageNet-1K is an image dataset and each concept image is quality-controlled and manually tagged. At present, ImageNet-1K consists of 1,4197,122 images. The major categories include: animal, bird, fish, flower etc. In Fig. 6 shows the visual examples of the ImageNet-1K.

In experiment, we compare the effectiveness between the CAM and the SAM. We compared 4 different network models: baseline network, baseline network with CAM, baseline network with SAM, baseline network with CSM. For the unbalanced data classification model based on our proposed model, we uses the ε-greedy strategy to determine the parameters of models. The probability of exploration decays linearly from 1.0 to 0.01. The size of the experience pool is set to 50,000, and the interaction between intelligent agent and the environment is about 120,000 steps. The discount factor for instant rewards is set to 0.1. The gradient of cost function to parameters is calculated by using the data of the whole training set. Adadelta algorithm [25] is used to update network parameters (aimed at solving the problem of plummeting learning rates), and the learning rate is 0.002. For other algorithms, the optimizer is Adam [44], the learning rate is 0.0005, and the batch size is 64. Training rounds are 100.

In this comparative experiment with the state the art, resNET-50 and RESNET-200 models were used to validate this strategy on CIFAR-10 and ImageNet-1Kdataset. Epoch was set to 270, batch size was set to 4096, and learning rate was set to 0.01. The comparison was done based on the same datasets and rules.

Finally, we make some comparisons between our method and existing methods, and we draw the following conclusions:

1. Data enhancement is relatively complete, and the scale variation of multi-scale training is relatively small;

2, the learning rate is directly default, without warm up and decay;

3. The weighted processing of loss is similar in the four datasets, and some are not even set;

4. Anchor matching strategy is also similar after verification, for example. each YOLO layer matches the largest IOU.

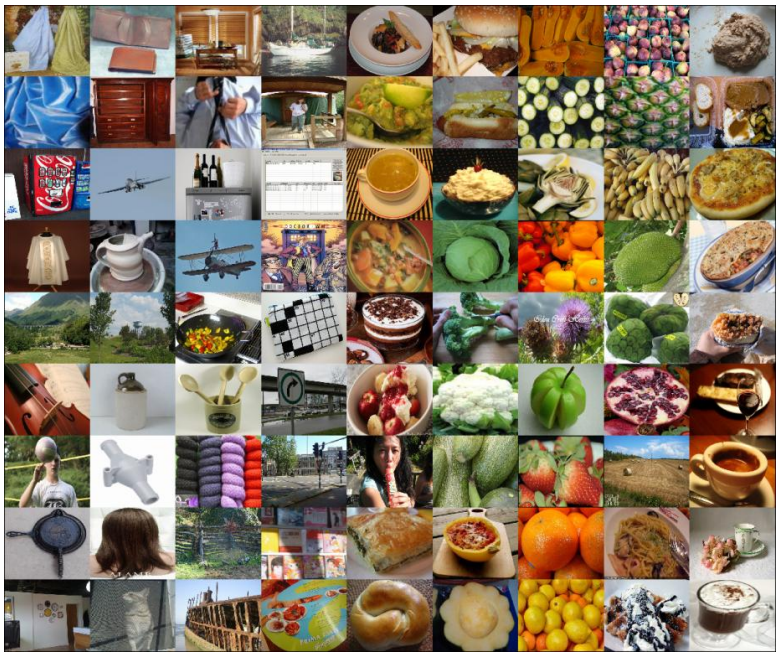

Fig. 5. Visual Examples of the Cifar-100 [31]

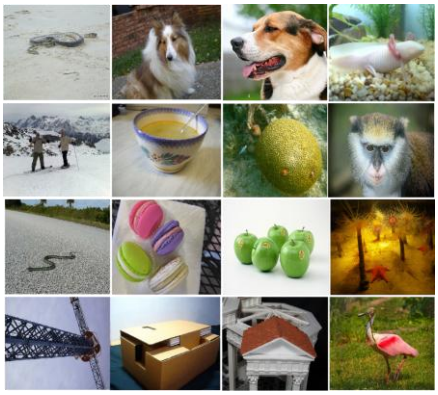

Fig. 6. Visual examples of the ImageNet-1K

4.2. Image Classification on ImageNet-1K

In this part, we use ResNet and WideResNet as the baseline model. On this basis, we add attention mechanism for comparison. The extensive image classification experiments are conducted based on the ImageNet-1K. The structure of ResNet with adding SE module is shown in Figure 7 and the results of the experiment is shown in Table **2**. The experiment still prove that networks with CSM performs better than the baseline module, indicating that attention mechanism can be well used in the various network models. Besides, the depth and width of the neural network also greatly affect image classification accuracy. SENet won the ILSVRC2017 classification task championship. But CSM fuses channel features with spatial features for better representation capabilities, and CSM performs better than SENet. On the other hand, it also proves that the quality of attention model determines the final classification accuracy, which is related to the interpretability of the model. The obtained prediction density map after fusing feature self-learning idea is nearer to the ground-truth map. This phenomenon can also be observed in Fig.6, the feature self-learning loss module plays an important role in the corresponding changes of MAE accuracy.

Table 2. Results of the experiment on the ImageNet-1K dataset

Architecture	Top-1 error (%)	Top-5 error (%)
ResNet-18	29.62	10.56
ResNet-18+SE	29.43	10.24
ResNet-18+CSM	29.29	10.12
ResNet-34	26.71	8.62
ResNet-34+SE	26.16	8.37
ResNet-34+CSM	26.03	8.28
ResNet-50	24.55	7.52
ResNet-50+SE	23.21	6.74
ResNet-50+CSM	22.78	6.57
WideResNet18(widen=1.5)	26.86	8.91
WideResNet18(widen=1.5) +SE	26.23	8.51
WideResNet18(widen=1.5) +CSM	26.14	8.49

4.3. Image Classification on CIFAR-100

Based on Cifar-100, we conduct image classification experiment to verify the effectiveness of the CSM. ResNet and WideResNet are also used as baseline model. Table 3 shows the experimental results. The experimental results prove that the combination of CAM and SAM can improve classification accuracy. Besides, the depth and width of the neural network also greatly affect image classification accuracy. As we can see, our proposed method has the minimum absolute error and the minimum mean square error on CIFAR-100 dataset. Compared with other algorithm, The accuracy of the model with added attention and transfer learning strategies is greatly improved,

which reflects the best performance of the our proposed algorithm. The model is also more robust to moderately dense data sets.

Table 3. Results of the experiment on the ImageNet-1K dataset

Architecture	Accuracy (%)
ResNet-18	91.7
ResNet-18+CSM	93.1
ResNet-34	92.4
ResNet-34+CSM	93.8
ResNet-50	92.9
ResNet-50+CSM	94.3
WideResNet18(widen=1.5)	92.8
WideResNet18(widen=1.5) +CSM	94.2

As can be seen from Table **3**, the prediction density map that we got after adding feature self-learning is closer to the true value, so we can draw a conclusion that we can more accurately describe the feature through adding feature self-learning Loss reduction and corresponding changes in MAE accuracy (see Fig.7 and Fig.8), we also give the performance of net similarity.

The proposed model is evaluated with several latest models on our constructed dataset and two benchmarks, and the results are given in Tables 4 and 5. From all the tables, we notice that our method is always much better than the previous method. Table 4 gives a comparison of the ImageNet-1K, and their images are closer to the real monitoring scene than other data sets. Our proposed model has achieved considerable improvement compared to the existing technology. Table 5 indicates that proposed method obtained the best MAE and competitive MSE among the five latest methods in the CIFAR-100 dataset. As we can see, our model performs best in the most compared algorithms both in terms of MAE and MSE. The spatial complexity of the network is only related to the size of the convolution kernel, the number of channels and the depth of the network. However, it has nothing to do with the size of input data. Since the network model designed by us adopts optimization and transfer learning strategies, compared with other existing popular algorithms, the complexity of our designed model is lower.

The result of experiment is shown in Figure 7. As is shown in Figure 7, it is easily concluded that the ResNet-50 model with CSM has achieved higher accuracy. We can observe that CAM perform better than SAM. Besides, the combination of two attention modules can bring a better performance. The experiment shows that it is effective to conduct CAM and SAM at the same time. The experimental results are shown in Table 6. Based on the above analysis of the experiment, we find that it is effective to conduct CAM and SAM at the same time for improving the interpretability ability of neural networks. So, in the ablation experiment, we compare the effect of CAM and SAM in different order of use. The baseline network with CAM and SAM, baseline network with SAM and CAM are conducted respectively. Figure 7 shows the result of experiment.

Table 4. Comparison in the ImageNet-1K data set

Methods	MAE	MSE
Optimization Algorithm [44]	419	541
XAI [27]	467	498
Software-Defined [13]	377	509
Cost-Driven [8]	322	**341**
Aquila Optimizer[45]	318	439
ours	**280**	379

Table 5. Comparison in the ImageNet-1K data set

Methods	MAE	MSE
Optimization Algorithm [44]	181	277
XAI [27]	154	229
Software-Defined [13]	110	173
Cost-Driven [8]	101	152
Aquila Optimizer[45]	97	145
ours	**91**	**137**

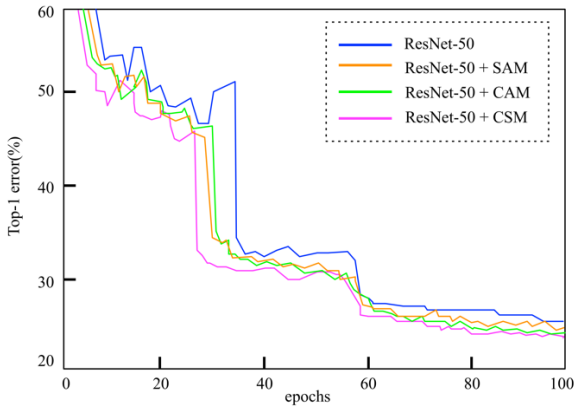

Fig. 7. Comparison of different network models

Similar to the above experiment, adding the attention module still bring improvement on image classification accuracy (as is shown in Figure 8). We can observe that the CAM-first combination method achieves better performance. The order in which models are executed can affect the accuracy, which is a supplement to the interpretability of the model.

Table 6. Comparison of different network models

Architecture	**Top-1 error (%)**
ResNet-50	24.55
ResNet-50+SAM	23.47
ResNet-50+CAM	23.21
ResNet-50+CSM	22.78

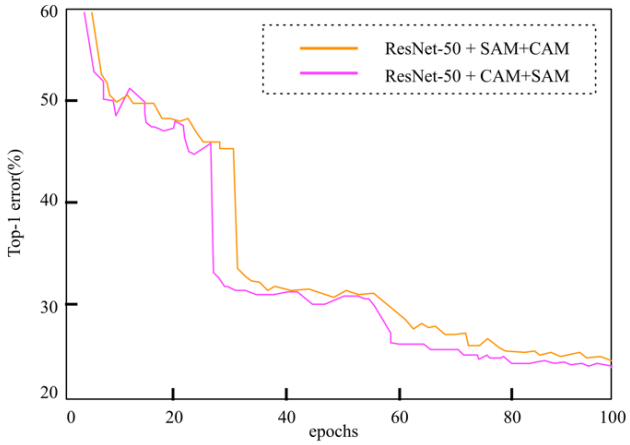

Fig. 8. Comparison of different network models

5. Conclusions

Different from the previous works on image classification, we propose proposes an explainable multi-channel convolutional neural network model consisting of channel attention module and spatial attention module. This module derives the attention map by CAM and SAM respectively. The CAM selectively enhances some feature channels and suppresses some feature channels by learning the relational mapping. The SAM aggregates features by weighting features at spatial dimension. We conducted a lot of image classification experiments for comparison on the ImageNet-1K and Cifar-100 dataset. In fact, the attention module can be well embedded in different deep neural networks and it improves the explainable deep neural network's ability of expression. Besides, the width and depth of the neural networks are also researched through self-paced learning and transferring learning scheme.

In the future research, we will pay more attention to the improvement of real-time algorithm and effectiveness of explainable deep learning, also be able to obtain more accurate data collection to develop an intelligent prediction machines for image classification based on more effective machine learning approaches.

Acknowledgment. This work is supported by the 111 Project (B12018); Innovative Research Foundation of Ship General Performance (14422102).

References

1. Diba, A., Sharma, V.: Weakly super-vised cascaded convolutional networks. In: IEEE conference on computer vision and pattern recognition (CVPR). New York, NY, 5131-5139. (2017)

2. Ren, S. H., He, K. M., Girshick, R.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Analysis And Machine Intelligence, Vol. 39, 1137–1149. (2017)
3. Ross, G., Donahue, J., Darrell, T., Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR), Columbus, Ohio, 580–587. (2014)
4. Chen, X., Li, M., Hsu, C. H.: DNNOff: Offloading DNN-based Intelligent IoT Applications in Mobile Edge Computing. IEEE Transactions on Industrial Informatics, 612-623. (2021)
5. Chen, X., Chen, S. H., Ma, Y., Huang, G.: An Adaptive Offloading Framework for Android Applications in Mobile Edge Computing. SCIENCE CHINA Information Sciences, Vol. 62, No. 8, 821-832. (2019)
6. Zhang, Y., Huang, G., Liu, X. Z., Zhang, W., Mei, H.: Refactoring android Java code for on-demand computation offloading. ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, 980–987. (2012)
7. Chen, X., Hu, J., Chen, Z. Y., Lin, B., Min, G. Y.: A Reinforcement Learning Empowered Feedback Control System for Industrial Internet of Things. IEEE Transactions on Industrial Informatics, 1109-1124. (2021)
8. Lin, B., Huang, Y., Zhang, J. S., Hu, J., Li, J.: Cost-Driven Offloading for DNN-based Applications over Cloud, Edge and End Devices. IEEE Transactions on Industrial Informatics, Vol. 16, No. 8, 5456-5466. (2020)
9. Lin, B., Zhu, F., Zhang, J. S., Chen, J., Mauri, J. L.: A Time-driven Data Placement Strategy for a Scientific Workflow Combining Edge Computing and Cloud Computing. IEEE Transactions on Industrial Informatics, Vol. 15, No. 7, 4254-4265. (2019)
10. Gong, C., Ding, Y., Han, B., Yang, J., Tao, D. C.: Class-Wise Denoising for Robust Learning under Label Noise. IEEE Transactions on Pattern Analysis and Machine Intelligence,12-28. (2022)
11. Gong, C., Wang, Q. Z., Liu, T. L., Hu, J., Yang, J., Tao, D. C.: Instance-Dependent Positive and Unlabeled Learning with Labeling Bias Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence,189-199. (2021)
12. Chen, X., Zhu, F. N., Zhang, J. S., Chen, Z. Y., Min, J. Y.: Resource Allocation for Cloud-based Software Services Using Prediction-Enabled Feedback Control with Reinforcement Learning. IEEE Transactions on Cloud Computing, 111-125. (2022)
13. Huang, G., Luo, C. R., Wu, K. D., Ma, Y., Zhang, Y.: Software-Defined Infrastructure for Decentralized Data Lifecycle Governance: Principled Design and Open Challenges. IEEE International Conference on Distributed Computing Systems, 236-252. (2019)
14. Florian, S., Kalenichenko, D., Philbin, J. S.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: IEEE conference on computer vision and pattern recognition (CVPR), 815–823. (2015)
15. Leung, Y., Ji, N. N., Ma, J.: An integrated information fusion approach based on the theory of evidence and group decision-making, Information Fusion,Vol. 14, No. 4, 410-422. (2013)
16. Parkhi, O. M., Andrea, V., Andrew, Z.: Deep Face Recognition. In: British machine vision conference (BMVC), 110-119. (2015)
17. Jamie, S., Andrew, F., Mat, C., Alex, K., Blake, N.: Real-Time Human Pose Recognition in Parts from Single Depth Images. In: IEEE conference on computer vision and pattern recognition (CVPR), 1297–1304. (2011)
18. Newell, A., Yang, K. Y., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In: Computer Vision-ECCV, 483–99. (2016)
19. Alex, K., Sutskever, I., Geoffrey, E., Hinton, J.: ImageNet Classification with Deep Convolutional Neural Networks. Communications of The ACM. Vol. 60, 84–90. (2017)
20. Cires, A., Dan, U., Juergen, S.: MultiColumn Deep Neural Networks for Image Classification. In: IEEE conference on computer vision and pattern recognition (CVPR), 3642–3649. (2012)

21. Chen, S., Gong, C., Yang, J., Niu, G., Sugiyama, M.: Learning Contrastive Embedding in Low-Dimensional Space. Annual Conference on Neural Information Processing Systems. Berlin Heidelberg New York. (2022)
22. Pentina, A., Sharmanska, V., Lampert, C. H.: Curriculum learning of multiple tasks. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, 2547–2554. (2015)
23. Lin, L., Wang, K., Meng, D., Hu, J., Li, J.: Active Self-Paced Learning for Cost-Effective and Progressive Face Identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 7-19. (2017)
24. Holzinger, A., Biemann, C., Constantinos, S. P., Douglas, B.: What do we need to build explainable ai systems for the medical domain?, arXiv:1411.1784 (2017)
25. Arrieta, A. B., Rodriguez, N. D., Ser, J. D., Bennetot, A., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, vol. 58, 82-115, June (2020)
26. Xu, D., Yang, W., Xavier, A. P., Wang, X. J., Sebe, N.: Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In NIPS, 3961-3970. (2017)
27. Sachan, S., Yang, J., Xu, D. L., Benavides, D., Li, Y.: An explainable AI decision-support-system to automate loan underwriting. Expert Systems with Applications, Vol. 144, 113-130. (2020)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), 28-34. (2016)
29. Simonyan, K., Zisserman, A., Zhang, J. S., Hu, J., Li, J.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv preprint arXiv: 1409.1556 (2020)
30. Abhronil, S., Ye, Y., Wang, R., Liu, C., Roy, K.: Going Deeper in Spiking Neural Networks: VGG and Residual Architectures. Neuroscience. Vol. 13, 95-95. (2009)
31. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv: 1611.05431 (2016)
32. Fei, W., Jiang, M., Qian, C., Yang, S., Li, C.: Residual Attention Network for Image Classification. In IEEE conference on computer vision and pattern recognition (CVPR), 6450–6458. (2017)
33. Yang, K. Y., Qinami, K., Li, F. F., Deng, J., Olga, J.: Towards Fairer Datasets: Filtering and Balancing teh Distribution of the People Subtree in the ImageNet Hierarchy. In: Conference on fairness, accountabiility and transparency (FAT) (2020)
34. Olga, R., Deng, J., Su, H., Krause, J., Ma, J.: ImageNet Large Scale Visual Recognition Challenge. International journal of computer vision (IJCV), 1021-1034. (2015)
35. Deng, J., Berg, A., Li, K., Hu, J., Li, F. F.: What does classifying more than 10,000 image categories tell us? In: Computer Vision-ECCV (2010)
36. Russakovsky, O., Li, F. F.: Attribute Learning in Largescale Datasets. In: Computer Vision-ECCV (2010)
37. Deng, J., Dong, W., Socher, R., Li, F. F.: ImageNet: A Large-Scale Hierarchical Image Database. In: IEEE international conference on computer vision and pattern recognition (CVPR), 1354-1361. (2009)
38. Hu, J., Shen, L., Sun, J.: Squeeze-and-excitation networks. In: IEEE conference on computer vision and pattern recognition (CVPR), 7132-7141. New York, USA. NY: IEEE (2019)
39. Zhang, Y. L., Li, K. P., Li, K., Wang, L., Zhong, B.: Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In: Computer Vision ECCV. 294–310. (2010)
40. Woo, S., Park, J., Lee, J. Y., Kweon, J.: CBAM: Convolutional Block Attention Module. In: Computer Vision-ECCV. 3-19. (2010)
41. Chen, L., Zhang, H. W., Xiao, J., Nie, L. H., Liu, W.: SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In: IEEE international conference on computer vision and pattern recognition (CVPR). 6298–6306. (2017)

42. Zhao, H. S., Zhang, Y., Shi, J. S., Lin, J., Jia, J.: PSANet: Point-Wise Spatial Attention Network for Scene Parsing. In: Computer Vision-ECCV. 267–283. (2018)
43. Meng, X., Huang, Y., Zhang, J. S., Hu, J., Li, J.: Feature selection and enhanced krill herd algorithm for text document clustering. Computing reviews, Vol. 60, No. 8, 318-328. (2019)
44. Abualigah, L., Diabat, A., Mirjalili, S.: The Arithmetic Optimization Algorithm. Computer Methods in Applied Mechanics and Engineering, 376-389. (2021)
45. Laith, A., Dalia, Y., Mohamed, A. E., Amir, H. J., Li, J.: Aquila Optimizer: A novel meta-heuristic optimization algorithm, Computers & Industrial Engineering, Vol. 157, 28-41. (2021)
46. Johnson, J., Alahi, A., Li, F. F.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. Springer, 120-138. (2016)

Bin Wu received the B.Sc. degree from Jiangnan University, Wuxi, China, in 1996, and the M.S. degree from Jiangnan University, Wuxi, China, in 2005. He is currently a lecturer with the School of Internet of Things Engineering, Jiangnan University. His major research interests include visual surveillance, object detection, integrated circuit design and application of embedded system.

Tao Zhang received the bachelor's degree from Henan Polytechnic University, Jiaozuo, China, in 2008, and the Ph.D degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2016. He is currently an Associate Professor with the Jiangsu Provincial Engineering Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, China. He has led many research projects (e.g., the National Science Foundation and the National Joint Fund), He has authored over thirty quality journal articles and conference papers. His current research interests include data mining, information systems, wireless network, artificial intelligence, IoT and security, medical data analysis, visual surveillance, scene understanding, behavior analysis, object detection, and pattern analysis.visual monitoring, scene understanding, behavior analysis, target detection and pattern analysis.

Li Mao received the B.Sc. degree from Southeast University (Nanjing, China) in 1990 and M.S. degree from Donghua University (Shanghai, China) in 2003. He is currently an associate professor at the department of computer science, Jiangnan University. His major research interests include visual surveillance, object detection, and data mining.

Received: July 14, 2022; Accepted: December 18, 2022.