IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 19, 2019, accepted July 16, 2019, date of publication July 18, 2019, date of current version August 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929512

Multi-Label Image Classification by Feature Attention Network

ZHENG YAN[1], **WEIWEI LIU**[1,2], **SHIPING WEN**[2,3], **AND YIN YANG**[4]

[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]Research Institute in Shenzhen, Huazhong University of Science and Technology
[4]College of Science, Engineering and Technology, Hamad bin Khalifa University, Doha 5855, Qatar

Corresponding author: Shiping Wen (wenshiping@uestc.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61673187, in part by the Shenzhen Innovative, Research Institute in Shenzhen, under Grant JCYJ20170307160806070, in part by the Huazhong University of Science and Technology, and in part by the NPRP through the Qatar National Research Fund under Grant NPRP 9-466-1-103.

ABSTRACT Learning the correlation among labels is a standing-problem in the multi-label image recognition task. The label correlation is the key to solve the multi-label classification but it is too abstract to model. Most solutions try to learn image label dependencies to improve multi-label classification performance. However, they have ignored two more realistic problems: object scale inconsistent and label tail (category imbalance). These two problems will impact the bad influence on the classification model. To tackle these two problems and learn the label correlations, we propose feature attention network (FAN) which contains feature refinement network and correlation learning network. FAN builds top-down feature fusion mechanism to refine more important features and learn the correlations among convolutional features from FAN to indirect learn the label dependencies. Following our proposed solution, we achieve performed classification accuracy on MSCOCO 2014 and VOC 2007 dataset.

INDEX TERMS Deep neural network, multi-label recognition, label correlation, attention.

I. INTRODUCTION

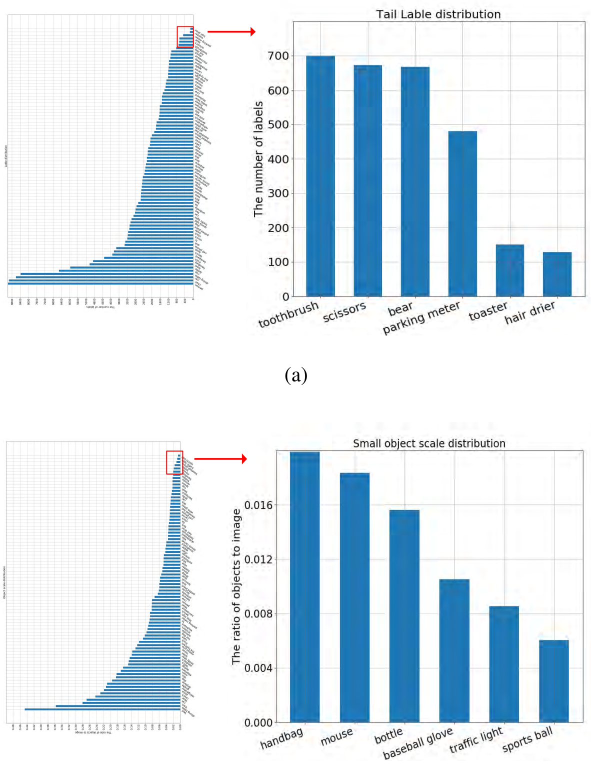
Multi-label image classification aims to recognize the different objects or attributes in images. Compared with the single label image classification, which predicts only one label to each image, multi-label classification is more complicated. The labels of each image are different and the number of labels in per image is not fixed. Actually One can surmise whether other labels exist in this image according to predicted labels due to the label correlation. The key to solve the multi-label problem is to exploit the label correlation to precisely predict labels in images. The label correlation learning is long-standing problem as it is abstract and difficult to model directly.

With the development of machine learning and deep learning technologies, a lot of solutions [2], [3], [8], [34], [37], [53] are proposed to learn the label correlation and have achieved promising performance on different benchmarks. However, they all ignored two realistic problems in multi-label classification: object scale inconsistent and label tail (category imbalance) as shown in Figure 1. Object scale inconsistent: In the actual applications, the proportion of different objects in images is different such as person and tennis ball. Small objects in images are more difficult to identify than big objects. Label tail: label tail can also be viewed as category imbalance which manifests itself as a long tail distribution of labels. It is difficult for the algorithm to learn the informative features of tail objects and accurately identify the tail labels, because tail labels appear in dataset with very few times. Actually, frequently occurring categories are more easily identified.

Both object scale inconsistent and label tail are common phenomenons in realistic datasets. In a deep neural network, the features of the last few layers have a larger receptive field. If the receptive field is much larger than the object size, the feature of small objects is easily overlooked. Further, deep networks will pay attention on easily identifiable categories such as *person* and *car*. Meanwhile, for the tail objects, they only appear a few times in dataset so that neural network cannot learn generic distinguishable features from limited data. Small objects and tail labels usually have low recognition performance than other categories. This will affect the overall performance and versatility of the algorithm. However, there exists a same key point between object scale inconsistent

The associate editor coordinating the review of this manuscript and approving it for publication was Chenguang Yang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/



(a)

(b)

FIGURE 1. The illustration of COCO2014 dataset (a) label number distribution (b) object scale distribution. We zoom in to show the data of the red box. Tail labels only appear a few times in the COCO2014 dataset, and the small objects has a small proportion in images, which will bring difficulties for image classification.

and tail label, which is the lack of informative and representive features to classify these categories.

To challenge the multi-label image classification task, we proposed Feature Attention Network to mine more representative features and learn label correlation based on self-attention mechanism.

Our Feature Attention Network contains two sub networks named Feature Refinement Network and Correlation Learning Network. Feature Refinement Network aims to solve the object scale inconsistent and the label tail problem by mining informative features, and the Correlation Learning Network for learning the label correlation indirectly by learning semantic and spatial dependencies among features.

In order to recognize multi-scale objects, the multi-scale feature and the context information are important and useful. We extract multi-scale feature for recognition. Smaller objects usually are obvious in low-level features (spatial feature) and disappear in high-level features (semantic feature). Therefore, it is necessary to reasonably exploit the multi-scale feature. However, not all features are informative, we should highlight important features and underrate the less importance ones. Therefore, we proposed Feature Refinement Block to select the useful and outstanding features, inspired by SEnet [15].

Correlation Learning Network aims to learn the label correlation by model convolutional feature dependencies. Label correlation is long-standing but key problem in multi-label classification. A lot of methods [3], [34], [37] try to model the label dependencies indirectly due to its abstract nature. In our proposed solution, we learn the feature correlation by self-attention [33] method. Convolutional feature contains pixel intensity information and spatial distribution information. Correlation Learning Network integrates the multi-scale features from Feature Refinement Network. It can explicitly exploit the feature intensity and spatial information to get the new feature which considers label correlation and further solves the object scale inconsistent and label tail problem.

In this paper, we reconsider the large-scale multi-label image classification task. We point out ignored problems in multi-label image classification: object scale inconsistent and label tail problem. Then, we propose Feature Attention Network, which not only solves the above two problems, but also learns the label relationship. Our experiment results on MSCOCO 2014 and PASCAL VOC 2007 demonstrate the effectiveness of our solution.

II. RELATED WORKS

A. MULTI-LABEL CLASSIFICATION

Instead of transforming the multi-label problem [8], [26], WARP [11] proposed to exploit the advantage of convolution features to multi-label annotation and analyze key components that improve performance. Hypotheses-CNN-Pooling (HCP) [39] proposed to use the max pooling to aggregate different results of each specific object hypotheses. CNN-RNN [34] built a joint CNN-RNN network to learn joint image-label embedding in which semantic label relevance is considered. Other works like [2], [3], [37] used RNN to reason or find the corresponding attention regions in terms of multi-label classification. Those solutions can only predict the top-k labels not unfixed label. SRN [53] learned the class-wised attention maps and captured the potential correlation between them by doing spatial regularization on feature maps. However, despite the better performance of these methods, these methods ignore the object scale inconsistent and label tail issues.

B. ATTENTION

Attention plays an important role in both computer vision and neural language processing field. Some models introduce supervise information to capture the context information or long-range dependencies among features in action recognition [4], [10], [25]. Apart from this, SEnet proposed Squeeze and Excitation Module to adaptively recalibrate channel-wise features without extra supervised information. Meanwhile, work [33] proposed self-attention mechanism to draw global dependencies between input and output and achieved great success in machine learning. Further, non-local operation [36] was introduced to relate the response of a position to the features of all positions. Non-local has improvements on many computer vision tasks. Work [14] improved the object detection performance by well-design relevance learning network. Attention mechanism has been proven to be effective in learning label dependencies.

C. MULTI-SCALE

object scale inconsistent is more realistic and long-standing problem. Multi-scale features have been used to improve the object detection performance [20], [23], [27]. The top-level features of deep neural networks have rich semantic information and have small size but larger receptive field that is useful to recognize bigger objects. The features of first few layers contain rich spatial information which represent the simple understanding of images by neural networks, and has bigger size but smaller receptive field. Therefore, the larger scale feature is useful to find small objects. However, not all features are useful. It is necessary to select the informative features forward to the output.

D. LABEL TAIL

Work [38] pointed out that tail labels have less impact in terms of Top-k precision and nDGG@k metric. Therefore, they develop a low-complexity multi-label algorithm by trimming tail labels adaptively. However, a good multi-label classification model should not be limited by unfixed number of labels. In object detection field, object detection methods [9], [12], [29] use the online hard example mining to balance the positive and negative sample ratio. However, you have no idea which is positive examples before get the results. Focal loss [21] put more attention on hard misclassified examples by changing the loss function. Our model solves the problem of low classification accuracy of tail labels by finding more fine-grained and discriminative features.

III. PROPOSED SOLUTION

In this section, we detail our proposed solution. We firstly analysis the problem in multi-label image classification. Then, we point out how we solve object scale inconsistent and label tail problem using Feature Refinement Network. Finally, we detail our Correlation Learning Network for learning label relevance.

In this section, we detail our proposed Feature Attention Network (FAN). Feature Attention Network consists of backbone network, Feature Refinement Network (FRN) and Correlation Learning Network (CLN). Backbone network can be classic network such as VGG, Resnet. FRN and CLN are introduced as followed.

A. FEATURE REFINEMENT NETWORK

The recognition accuracy of small objects and tail label is usually lower than other labels that are easier to recognition.

Features of small objects are easily ignored by deep neural networks due to the convolutions with stride and pooling operation. On the other hand, the tail label objects appear a few times in dataset. Lack of training data leads to underrepresentation of tail label by neural networks. Therefore, the same problem exists between the object scale inconsistency and the label tail problem: the lack of informative and fine-grained features. In our solution, we build feature recalibrate mechanism to mine the useful features, which exploit the global context information and multi-scale features reasonably.

How to use multi-scale features has been widely studied in object detection task [18], [20], [21], [23], [28], [35], [41], [44], [45] and segmentation task [1], [6], [19], [30], [42], [43], [46], [48], [49], [51]. It is useful for multi-scale objects recognition that using multi-scale feature. In our solution, we also use multi-scale features. More importantly, we recalibrate learned multi-scale. Recalibrate features can help our model to locate more representative and informative features. Not only that, we also use high-level features to guide the refinement of low-level features. Specifically, we build top-down feature propagation mechanism like FPN [20]. However, feature is transformed and recalibrated by Residual Transform Block(RTB) and Feature Refinement Block (FRB) respectively, before feature is passed to the next stage.

1) RTB

Residual Transform Block is used to transform features of different resnet stages to same level space, in Figure 3, which is benefit to followed feature fusion and recalibrate operation. RTB contains a convolution layer and a residual block. RTB acts as a buffer between the backbone network and the Feature Refinement Network. RTB is similar with common residual block in Resnet [13]. In RTB, we firstly use a convolution to reduce the dimension of inputs to K. In our model, we set K as 512. Then, a residual block is followed to transform the feature. And finally, we use the pooling operation to halve the size of the feature maps. The average pooling with 2 × 2 kernel and stride 2 is used in our solution. However, for the bigger resolution feature maps like $Block2$ stage, we use more RTBs and average pooling to get uniform size feature maps.

2) FRB

Feature Refinement Block is used to fusion and recalibrate features of different convolution stage. It highlights the informative and discriminative features and pay less attention on unimportant features. It is achieved by self-attention mechanism. FRB will learn a weighted vector from different stage features. The weighted vector will serve as a $attentionvector$ to recalibrate feature. This can highlight features that are useful for small objects and tail label recognition. Specially, we concatenate high-level features x_h and low-level features x_l in channel dimension to get new features $x_c, x_c \in R^{C*H*W}$. High-level features x_h have rich semantic information but less spatial information. It can be viewed as semantic supervised information to guide the recalibrate of low-level features x_l. Then, we use the global max pooling to capture the global context information.

$$z_c = F_{sq}\left(x_c\right) = \underset{x_c(i,j)}{\arg\max} f\left(x_c\left(i,j\right)\right) \tag{1}$$

where F_{sq} denotes the global max pooling. It takes x_c as input to calculate the vector z_c, $z_c \in R^{C*1*1}$.

$$\tilde{z}_c = \sigma\left(F_{tr}\left(z_c\right)\right), \quad \tilde{z}_c \in R^{C*1*1} \tag{2}$$

where F_{tr} refers to the convolution layers with relu activation followed. σ is sigmoid function. $\tilde{z}_c$ is learned weighted vector which go from 0 to 1 because of sigmoid function. Notice that $\tilde{z}_c$ is learned from x_c but works on x_l.

$$\tilde{x}_l = \tilde{z}_c * x_l, \quad \tilde{x}_l \in R^{C*H*W} \tag{3}$$

where $\tilde{x}_l$ is refined features which will guide the next feature refinement iteration. We expand the dimension of $\tilde{z}_c$ to R^{C*H*W} before channel-wise multiplication. We iteratively use FRB to recalibrate features from top to down.

3) GMP

Note that we use global max pooling in FRB. Other works had verified the global average pooling (GAP) is effective in image classification [13], [37] and semantic image segmentation [24] respectively. However, in our feature refinement and fusion process, we wish our model pays more attention to representative and discriminative features. Especially global average pooling will miss responses from small objects, when the feature maps have larger resolution. However, Global Max Pooling (GMP) will select the max response point as the global representation in terms of responding feature map. It will not ignore the responses from small objects or tail label objects. Therefore, we use global max pooling to capture global context information. Our ablation experiments demonstrate GMP is better than GAP.

In conclusion, we build Feature Refinement Network to fusion multi-scale features and mine representative and discriminative features. And global max pooling is used in Feature Refinement Block to capture context information. These is benefit to recognize small objects and tail label.

B. CORRELATION LEARNING NETWORK

This section introduces our Correlation Learning Network which learn feature spatial dependencies and semantic relevance by self-attention.

Some works use LSTM to locate attentional and informative regions that related to different semantic objects, and further predict semantic labeling scores on the located regions. LSTM can capture the global dependencies of located regions. However, SRN learns attention map for each label and further performs spatial regularity on learned features maps. We design Correlation Learning Network (CLN) to learn semantic dependencies and spatial relevance of features simultaneously by self-attention mechanism. Specifically, CLN learns attention responses based on relationships between different positions of feature. The response of any location to attention feature is related to the feature of other locations. The formula is as followed.

$$\tilde{f}(\ \ , x_j) = \sum_{\forall j} f\left(x_i, x_j\right) * g\left(x_i, x_j\right) \tag{4}$$

where $\tilde{f}(x_i)$ is attention feature scalar, where i is the index of output feature position in space, and x is an input signal. The response of $\tilde{f}(x_i)$ is related to all positions ($\forall j$) of feature $f(x_j)$. Here $g(x_i, x_j)$ is a binary function. It computes an attention matrix for regularizing feature f. We consider $g(x)$ as a linear embedding operation. In our solution, g is defined as a dot product function as followed.

$$g\left(x_i, x_j\right) = \frac{1}{C\left(x\right)} \theta\left(x_i, x_j\right) \bullet \phi\left(x_i, x_j\right) \tag{5}$$

where θ and ϕ are different image features. Here $C(x)$ is normalized function. We use softmax function in our network. $g(x)$ is used to compute attentional weight matrix with semantic relevance of features considered. In our solution, θ and ϕ are refined feature P_3 and P_2 respectively from Feature Refinement Network. Compared with feature P_2 and P_3, P_4 has rich semantic information. Therefore, we use learned attention matrix to regularize P_4.

C. LOSS FUNCTION

Previous works like [3], [11], [34], [37], [39], [52] can only predict the top-k predictions. However, the number of label of each image is unfixed. Work [17] can predict unfixed number of labels through adaptive thresholds learned by designed label decision module. Our model output prediction scores with dimension R^C, where C is the number of classes each dataset. The correlation information is considered before predictions are output. We can view multi-label outputs as a collection of multiple two-label output. Therefore, 0 is threshold for screening prediction scores. It is benefit to predict unfixed numbers of label. In our model, we use the multi-label soft margin loss to optimizer our model. For each minibatch, we can calculate loss using the following formula:

$$\begin{aligned} Loss(\hat{y}, y) = -\frac{1}{n}\sum_{i}^{n} y_i * log(\frac{1}{1+exp(-\hat{y}_i)}) \\ +(1-y_i) * log(\frac{exp(-\hat{y}_i)}{1+exp(-\hat{y}_i)}) \end{aligned} \tag{6}$$

where $\hat{y}$ is the predicted label and y is the ground truth one-hot label.

IV. EXPERIMENTS

To valid the effectiveness of our proposed solution, we carry out experiments on MSCOCO2014 and VOC2007 dataset. The results of both datasets demonstrate that our solution has out-standing performance. In this section, we firstly introduce the datasets we used and our implementation details are followed. In the following, we compare with other best multi-label classification methods, and perform some ablation experiments to evaluate each module of our model. Finally, we report our results on both dataset, and analysis it in detail.

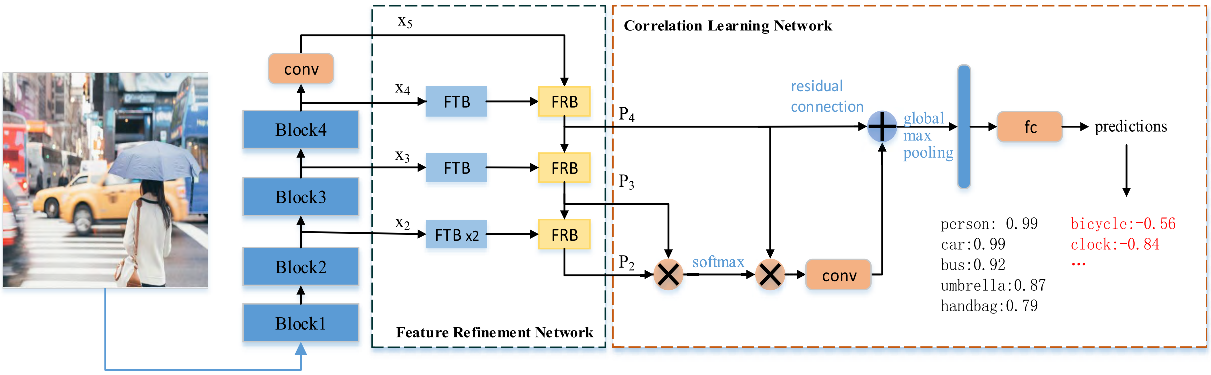


FIGURE 2. The illustration of our deep framework. Our proposed model contains three parts: Backbone, Feature Refinement Network and Correlation Learning Network. Block1-4 denote different convolutional stages. *conv* is single convolution layer. *fc* is the fully connection layer. FTB and FRB is feature transform block and Feature Refinement Block respectively, where *x* and *P* denote features of different stages, respectively. The blue fronts denote corresponding math operation.

A. DATASET

MSCOCO2014: MSCOCO2014 is a good object recognition dataset, which contains 82783 images for training and 40504 for validation of 80 different object categories. This dataset is primarily built for object recognition task in the context of scene understanding.

PASCAL VOC2007: VOC2007 is another well-known object recognition dataset, which contains 5011 images for train and 4952 image for validation of 20 object categories. Compared with COCO dataset, VOC have less training data. The annotated objects in VOC usually have larger scale and no serious tail label problem.

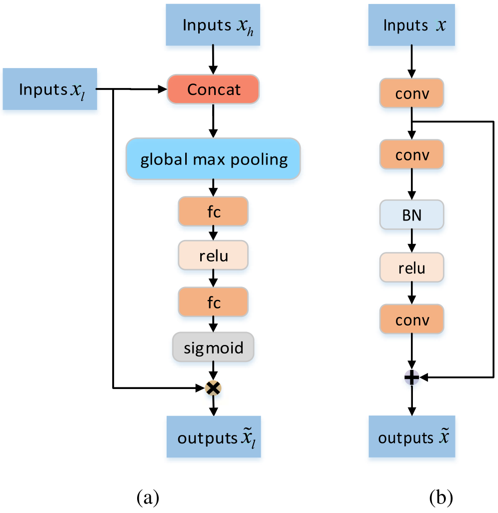
B. IMPLEMENTATION DETAILS

Our deep neural model contains two parts: Resnet-101 [13] and Feature Attention Network. Resnet is used to extract the image feature for Feature Attention Network. To fairly compare with other methods, we also use VGG [32] as backbone to demonstrate the effectiveness of our solution.

1) NETWORK IMPLEMENTATION

Our deep framework is shown in Figure 2, which contains backbone network, Feature Refinement Network and Correlation Learning Network. In addition, Feature Refinement Network consists of two components: Feature Transform Block (FTB) and Feature Refinement Block (FRB), schematically depicted in Figure 3. FTB is used to transform features into low-dimension space, and get compact and information representations. Firstly, we use a convolution layer with 1*1 size kernel to reduce the dimension of inputs to 512. Then, two another convolution layers with 3*3 kernel size are followed. Batchnorm [16] and Relu activation function are following the first convolution. Finally, we can get transformed results by residual connection. For feature x_2, we use two FTBs to map features. We use average pooling with kernel 2 and stride 2. FTB acts as a buffer between backbone and feature refinement block.

In FRB, global max pooling is used to get compressed feature vector. Fully connected layer learns channel-wised weights θ, which has the same channel number as low-level feature.



(a)

(b)

FIGURE 3. The illustration of our proposed feature transform block and feature refinement block. (a) Feature Refinement Block. (b) Feature Transform Block.

$$w = f\left(x_l, x_h; w_f\right) \tag{7}$$

$$\theta = \sigma\left(w\right) = \frac{exp\left(w_j\right)}{\sum_{i=1}^{C} exp\left(w_i\right)} \tag{8}$$

where f is fully connection layers with weights w_f. w is the results of fully connection layers. We calculate the weight θ by the sigmoid function σ. θ ranges from 0 to 1. The value of θ closed to 1 indicates that corresponding channel is important than other channels and vice versa. Finally, the final refined feature is computed by point-wised multiplication.

$$f_r = x_l \odot \theta \tag{9}$$

where $\odot$ denotes channel-wised multiplication.

TABLE 1. Comparison results of average precision and mAP of other methods and our method on the MSCOCO dataset. The bold front is used to mark the best results.

	ALL							TOP-3					
Methods	mAP	F1-C	P-C	R-C	F1-O	P-O	R-O	F1-C	P-C	R-C	F1-O	P-O	R-O
WARP [11]	-	-	-	-	-	-	-	55.7	59.3	52.5	60.7	59.8	61.4
CNN-RNN [34]	-	-	-	-	-	-	-	60.4	66.0	55.6	67.8	69.2	**66.4**
RDAR [37]	-	-	-	-	-	-	-	67.4	79.1	58.7	72.0	84.0	63.0
RARL [3]	-	-	-	-	-	-	-	66.2	78.8	57.2	71.1	84.0	61.6
VGG	67.8	63.3	72.0	56.4	68.9	76.8	62.4	60.4	75.1	50.5	66.4	81.5	66.0
Ours(VGG-FAN)	73.7	68.5	80.0	59.8	73.0	83.1	65.2	67.6	81.6	57.7	72.8	86.8	62.6
ResNet101 [53]	75.2	69.5	80.8	63.4	74.4	82.2	68.0	65.9	84.3	57.4	71.7	86.5	61.3
ResNet-SRN [53]	77.1	71.2	81.6	65.4	75.8	82.7	69.9	67.4	85.2	58.8	72.9	87.4	62.5
Ours(Resnet-FAN)	**81.8**	**77.1**	**84.6**	**70.8**	**79.5**	**85.2**	**74.5**	**73.6**	**88.4**	**63.0**	**75.9**	**89.7**	65.7

TABLE 2. Comparison of average precision and mAP of other methods and our method on VOC dataset. The best evaluation value is highlighted in bold front.

Methods	Aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	MAP
CNN-SVM [52]	88.5	81.0	83.5	82.0	42.0	72.5	85.3	81.6	59.9	58.5	66.5	77.8	81.8	78.8	90.2	54.8	71.1	62.6	87.2	71.8	73.9
CNN-RNN [34]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
VeryDeep [32]	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	87.8	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
RLSD [52]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
HCP [40]	98.6	97.1	**98.0**	95.6	75.3	**94.7**	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
FeV+LV [50]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	**97.7**	**95.9**	**98.6**	77.6	88.7	78.0	98.3	89.0	90.6
RDAR [37]	98.6	**97.4**	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	**87.7**	96.8	97.5	93.8	98.5	**81.6**	93.7	82.8	98.6	89.3	91.9
RARL [3]	98.6	97.1	97.1	95.5	75.6	92.8	**96.8**	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	**83.2**	98.5	89.3	92.0
Ours(Resnet-FAN)	**99.5**	96.6	**98.0**	**97.8**	**76.7**	92.7	96.0	**98.3**	**81.9**	**94.9**	81.8	**98.1**	**97.7**	94.6	**98.6**	81.5	**96.0**	80.1	98.8	**92.6**	**92.6**

2) TRAINING DETAIL

We use Resnet-101 or VGG as backbone of our model and load the weights pre-trained on ImageNet dataset [5]. We train our deep neural network in end-to-end way, using mini-batch stochastic gradient descent (SGD) with momentum factor 0.9 and weight decay 1e-4. We set batch size as 16. We use random crop and random horizontal flip in training for both datasets. In training process, we assign different learning rates to different network layers. Specifically, in the early stage of training process, we set the learning rate of Feature Attention Network as 0.1, Resnet-101 for 0.01. This will increase the speed of training. The learning rate is multiplied by 0.1, when the test accuracy is basically not increasing. The input image size was set as 448*448.

3) METRICS

We use the same seven metrics as [34], [47], [53] used to evaluate our proposed solution. The metrics used include macro/micro precision (P-C/P-O), macro/micro recall (R-C/R-O), macro/micro F measure (F-C/F-O) and Mean Average Precision (*MAP*). Precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. Precision and recall do not make scene in the isolation from each other.

$$PO = \frac{\sum_{i}^{C} TP_i}{\sum_{i}^{C} (TP_i + FP_i)}, \quad RO = \frac{\sum_{i}^{C} TP_i}{\sum_{i}^{C} (TP_i + FN_i)} \tag{10}$$

$$PC = \frac{1}{C} \sum_{i}^{C} \frac{TP_i^C}{TP_i^C + FP_i^C}, \quad RC = \frac{1}{C} \sum_{i}^{C} \frac{TP_i^C}{TP_i^C + FN_i^C} \tag{11}$$

where *TP*, *FP*, *FN* denote true positive, false positive and false negative respectively. *F* measure is a balanced metric considering precision and recall simultaneously. In our paper, we use $F1$ measure. Mean Average Precision is the mean value of class-wised average precision. Therefore, *F* measure and *MAP* are more important metrics.

$$F1O = \frac{2 * (PO * RO)}{PO + RO}, F1C = \frac{2 * (PC * RC)}{PC + RC} \tag{12}$$

C. EVALUATION

We compare our proposed solution against previous best multi-label image classification methods on MSCOCO 2014 [22] dataset and PASCAL VOC 2007 dataset [7]. The evaluation results are shown in Table 1 (for COCO) and Table 2 (for VOC). Some methods only provide top-k predictions. To compare with them fairly, we also compute top-k metrics based on our top-k prediction. In other best methods and our method, the hyper-parameter k is 3. Clearly, our proposed approach outperforms baseline and SRN [53] greatly, and improves the *mAP* performance from 77.1% to 81.8%. For the balanced metrics $F1 - C$, $F1 - O$, we all get state-of-art performance. Compared with other methods [3], [34], [37] which use RNN to learning label correlation information, our results has significant improvement than them.

1) RESULTS FOR TAIL LABEL AND SMALL OBJECT

To illustrate how our solution improves the classification accuracy of tail labels and small object categories, we select six labels with the fewest occurrences and six labels with the smallest percentage of the images in COCO dataset, and show the AP improvement of them in Table 3. We can easily know from Table 3 that our solution can greatly improve AP value

TABLE 3. The increase in the average precision (AP) of tail labels in the coco dataset.

	tooth brush	scissors	bear	parking meter	toaster	hair direr
VGG	42.9	35.3	91.4	54.4	6.0	5.3
VGG+FAN	50.8	42.7	91.8	76.6	11.5	6.0
Resnet101	60.0	51.8	91.4	58.4	9.5	3.8
Resnet101+FAN	72.2	64.6	96.8	66.3	16.0	30.1

TABLE 4. The increase in the average precision (AP) of small object labels in the coco dataset.

	handbag	mouse	bottle	baseball glove	traffic light	sport ball
VGG	39.6	74.7	53.4	89.7	70.4	69.1
VGG+FAN	44.5	80.1	64.7	93.3	81.2	81.0
Resnet101	47.6	73.2	60.1	89.6	79.0	77.7
Resnet101+FAN	51.9	87.6	73.0	95.2	86.2	87.1

TABLE 5. Detailed results of each component of our proposed solution on COCO dataset. FRN: feature refinement network. CLN: correlation learning network.

Method				Metrics						
Vgg16	Resnet101	FRN	CLN	MAP	P-O	R-O	F1-O	P-C	R-C	F1-C
✓				67.8	76.8	62.4	68.9	72.0	56.4	63.3
✓		✓		71.8	77.1	67.0	71.7	74.8	60.8	67.1
✓			✓	68.6	78.5	62.3	69.5	75.5	54.8	63.5
✓		✓	✓	73.7	83.1	65.2	73.0	80.0	59.8	68.5
	✓			75.2	82.2	68.0	74.4	80.8	63.4	69.5
	✓	✓		80.6	**85.3**	72.8	78.5	82.1	68.7	74.8
	✓		✓	80.2	83.1	72.7	77.6	80.7	68.9	74.3
	✓	✓	✓	**81.8**	85.2	**74.5**	**79.5**	**84.6**	**70.8**	**77.1**

of tail label, especially Resnet is used as feature extractor. We also show the effectness of our approach on small object categories classification, in Table 4. The classification accuracy of VGG with FAN can exceed the classification accuracy of Resnet101 baseline, which can demonstrate the effectiveness of our Feature Attention Network on multi-label image classification.

D. ABLATION STUDY

To evaluate our design modules, we decompose our approach and reveal the effect of each component in COCO [22] and VOC dataset [7]. COCO dataset is more complicated and realistic in image scene than VOC dataset.

1) ABLATION FOR FEATURE REFINEMENT NETWORK

Feature Refinement Network aims to learn informative and discriminative features, which is benefit to classify the small scale objects and tail label objects. We use Resnet-101 as our backbone and also make compared experiments with Resnet-101. The ablation results are shown in Table 5. We can easily know that our feature refinement network improve classification performance greatly, especially in recall rate RC, RO. The increase of recall rate means that the increase of the number of predicted positive labels. This indicates feature refinement network can predict more positive labels compared with baseline. Actually, more features are benefit to find negligible object. When correlation learning network is not used, we use joint predictions from P_2, P_3 and P_4 to get final predicted scores.

2) ABLATION FOR CORRELATION LEARNING NETWORK

Correlation Learning Network is responsible for learning label dependencies. Label dependencies play important role in image recognition and scene understanding. We can exploit the predicted labels and label dependencies information to reason possible positive labels in image. Our ablation results are shown in Table 5. Compared with Resnet-101 results, Resnet-101 with Correlation Learning Network improves F score and MAP a lot due the increase of recall score. From results of Resnet101 + CLN, CLN has less impact on precision score PC. The increase of recall score means the increase of number of predicted false negative labels. When feature refinement network is not used, x_2, x_3 and x_4 is equal to P_2, P_3 and P_4. That means feature attention matrix is computed by x_2 and x_3. Our final results from Resnet-101 with joint FRN and CLN can demonstrate the improvement of CLN.

TABLE 6. Compared results of global average pooling and global max pooling on COCO2014 dataset. GAP: global average pooling. GMP: global max pooling. Our approach with GMP has better performance.

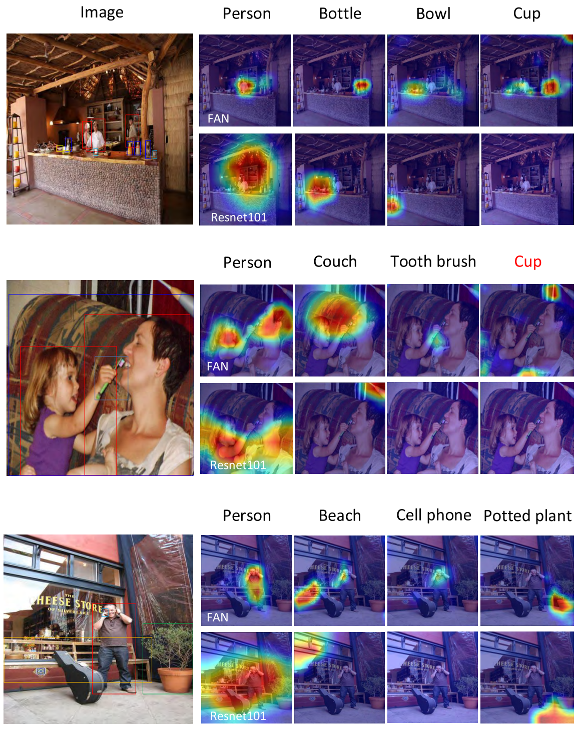
Metrics	FAN(GAP)	FAN(GMP)
MAP	81.2	**81.8**
P-O	85.7	85.1
R-O	73.4	**74.5**
F1-O	79.1	**79.5**
P-C	82.9	**84.6**
R-C	69.2	**70.8**
F1-C	75.4	**77.1**

3) ABLATION FOR GLOBAL MAX POOLING

As described in the section III-A. We use global max pooling instead of global average pooling to capture global context information. Global max pooling is sensitive to obvious responses and will not miss features of small objects. We made ablation experiments to valid the function of global max pooling. Its results are shown in Table 6. Global max pooling improve the performance of PC and F score in our solution.

TABLE 7. Detail results of our proposed approach and baseline on VOC2007 dataset.

Metrics	VGG16	Resnet101	FAN(vgg)	FAN(resnet)
MAP	86.5	87.8	91.1	**92.6**
P-O	82.9	88.8	**91.2**	90.3
R-O	81.1	78.6	82.6	**86.8**
F1-O	82.0	83.6	86.7	**88.5**
P-C	82.9	86.8	**89.7**	88.1
R-C	76.3	76.9	79.1	**85.6**
F1-C	79.5	81.6	84.1	**87.0**



(a)

FIGURE 4. Visualized feature maps from COCO dataset. We make compare with Rsenet101 baseline in locating multi-scale objects on the image. FAN is our method with Resnet101 as backbone network. Label in black are ground truth labels and red one are false labels. It suggests that FAN can more accurately locate the object corresponding to ground truth labels in the image.

E. VISUALIZATION

To further illustrate the effect of our FAN on solving the tail label and object scale inconsistent problems, we visualize learned feature maps using CAM method [31] in Figure 4. The visualized results show that FAN with Resnet101 as backbone can locate negligible objects more accurately than Resnet101. It suggests that our network is trained to capture semantic and spatial dependencies of objects in the image.

V. CONCLUSION

In this paper, we proposed Feature Attention Network for large-scale multi-label image classification. On one hand, we proposed the recalibrated feature to make our deep model pay more attention on small objects and tail label objects. On the other hand, we designed correlation learning module to learn semantic and spatial dependencies of objects based on the attention mechanism. Our ablation experiments also demonstrated the effectiveness of each component of our model. We also validated the role of global max pooling in capture context information. Extensive evaluations on MSCOCO2014 and VOC2007 datasets confirm that our proposed Feature Attention Network outperforms other multi-label image classification methods. Visualization results show that FAN can accurately locate the objects in the images, which is benefit to small objects and tail label recognition.

REFERENCES

[1] Y. Cao, Y. Cao, S. Wen, Z. Zeng, and T. Huang, "Passivity analysis of delayed reaction–diffusion memristor-based neural networks," *Neural Netw.*, vol. 109, pp. 159–167, Jan. 2019.

[2] S.-F. Chen, Y.-C. Chen, C.-K. Yeh, and Y.-C. F. Wang, "Order-free rnn with visual attention for multi-label classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[3] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[4] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3218–3226.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[6] M. Dong, S. Wen, Z. Zeng, Z. Yan, and T. Huang, "Sparse fully convolutional network for face labeling," *Neurocomputing*, vol. 331, pp. 465–472, Feb. 2019.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[8] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, vol. 73, no. 2, pp. 133–153, 2008.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[10] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1080–1088.

[11] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," Dec. 2013, *arXiv:1312.4894*. [Online]. Available: https://arxiv.org/abs/1312.4894

[12] K. He, G. Gkioxari, and P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Mar. 2017, pp. 2961–2969.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Mar. 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[17] Y. Li, Y. Song, and J. Luo, "Improving pairwise ranking for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1837–1845.

[18] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Detnet: A backbone network for object detection," Apr. 2018, *arXiv:1804.06215*. [Online]. Available: https://arxiv.org/abs/1804.06215

[19] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, "CLU-CNNs: Object detection for medical images," *Neurocomputing*, vol. 350, pp. 53–59, Jul. 2019.

[20] T.-Y. Lin and P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 936–944.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, and P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[24] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: https://arxiv.org/abs/1506.04579

[25] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 414–428.

[26] J. Read, L. Martino, and D. Luengo, "Efficient Monte Carlo methods for multi-dimensional learning with classifier chains," *Pattern Recognit.*, vol. 47, no. 3, pp. 1535–1546, 2014.

[27] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[28] G. Ren, Y. Cao, S. Wen, Z. Zeng, and T. Huang, "A modified Elman neural network with a new learning rate scheme," *Neurocomputing*, vol. 286, pp. 11–18, Apr. 2018.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2015, pp. 91–99.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.

[34] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2285–2294.

[35] S. Wang, Y. Cao, T. Huang, and S. Wen, "Passivity and passification of memristive neural networks with leakage term and time-varying delays," *Appl. Math. Comput.*, vol. 361, pp. 294–310, Nov. 2019.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[37] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 464–472.

[38] T. Wei and Y.-F. Li, "Does tail label help for large-scale multi-label learning," in *Proc. IJCAI*, Jul. 2018, pp. 2847–2853.

[39] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "CNN: Single-label to Multi-label," Jul. 2014, *arXiv:1406.5726*. [Online]. Available: https://arxiv.org/abs/1406.5726

[40] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Jun. 2015.

[41] S. Wen, Z. Q. M. Chen, X. Yu, Z. Zeng, and T. Huang, "Fuzzy control for uncertain vehicle active suspension systems via dynamic sliding-mode approach," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 1, pp. 24–32, Jan. 2017.

[42] S. Wen, R. Hu, Y. Yang, Z. Zeng, T. Huang, and Y.-D. Song, "Memristor-based echo state network with online least mean square," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[43] S. Wen, T. Huang, X. Yu, Z. Q. M. Chen, and Z. Zeng, "Aperiodic sampled-data sliding-mode control of fuzzy systems with communication delays via the event-triggered method," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 5, pp. 1048–1057, Oct. 2016.

[44] S. Wen, W. Liu, Y. Yang, Z. Zeng, and T. Huang, "Generating realistic videos from keyframes with concatenated GANs," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[45] S. Wen, S. Xiao, Y. Yang, Z. Yan, Z. Zeng, and T. Huang, "Adjusting learning rate of memristor-based multilayer neural networks via fuzzy method," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 38, no. 6, pp. 1084–1094, Jun. 2019.

[46] S. Wen, X. Xie, Z. Yan, T. Huang, and Z. Zeng, "General memristor with applications in multilayer neural networks," *Neural Netw.*, vol. 103, pp. 142–149, Jul. 2018.

[47] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3780–3788.

[48] X. Xie, S. Wen, Z. Zeng, and T. Huang, "Memristor-based circuit implementation of pulse-coupled neural network with dynamical threshold generators," *Neurocomputing*, vol. 284, pp. 10–16, Apr. 2018.

[49] X. Xie, L. Zou, S. Wen, T. Huang, and Z. Zeng, "A flux-controlled logarithmic memristor model and emulator," *Circuits, Syst., Signal Process.*, vol. 38, no. 4, pp. 1452–1465, 2019.

[50] H. Yang, J. T. Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 280–288.

[51] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," Apr. 2018, *arXiv:1804.09337*. [Online]. Available: https://arxiv.org/abs/1804.09337

[52] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu, "Multilabel image classification with regional latent semantic dependencies," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2801–2813, Oct. 2018.

[53] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," Mar. 2017, *arXiv:1702.05891*. [Online]. Available: https://arxiv.org/abs/1702.05891