IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 19, 2018, accepted June 26, 2018, date of publication July 4, 2018, date of current version July 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2852728

Learning Coexistence Discriminative Features for Multi-Class Object Detection

CHAO YAO[1,2], PENGFEI SUN[1], RUICONG ZHI[2], AND YANFEI SHEN[3,4]
[1]Institute of Sensing Technology and Business, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 10083, China
[3]Sports and Engineering College, Beijing Sport University, Beijing 100084, China
[4]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Corresponding authors: Ruicong Zhi (zhirc@ustb.edu.cn) and Yanfei Shen (syfyhy@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471343 and Grant 61701036 and in part by the Fundamental Research Funds for the Central Universities under Grant 2017RC52.

ABSTRACT Existing methods on object detection have the ability to learn the discriminative features of local regions for object recognition; however, the coexistence relation between the multi-class objects could also benefit recognition. In this paper, we propose to learn the coexistence discriminative features for multi-class object detection. Given an image with multiple class objects, the strong supervision of the region-based annotations are first used as the image-level label to learn the independent discriminative features for each class. Then, the coexistence relation is fused as coexistence feature based on the attention mechanism. By combining the independent discriminative features and coexistence feature, the classification performance of multi-class object proposals can be consistently improved. Experimental results prove that the proposed end-to-end network outperforms the state-of-the-art object detection approaches, and the learned discriminative features can effectively capture the coexistence relations to improve classification performance of multi-class objects in the object detection task.

INDEX TERMS Object detection, faster R-CNN, coexistence relation, multi-class objects, class attention map.

I. INTRODUCTION

In the past few years, deep convolutional neural networks have largely boosted the development of various artificial intelligence applications. Several novel neural network structures, such as VGG [1], GoogLenet [2], ResNet [3] and DenseNet [4], are proposed to improve the learning capability on large-scale datasets. Based on these classical network structures, the performance of object detection has also witnessed significant progress.

Object detection generally aims to localize the instances of object in a given image. In the state-of-the-art object detection framework, the object recognition is formulated as a classification task for the generated bounding box proposals. Bounding box proposals are used to locate the possible objects. The appearance features extracted from each proposal are classified based on their associated class scores independently. For example, R-CNN [5], Fast R-CNN [6], Faster R-CNN [7], R-FCN [8], SSD [9] and YOLO [10] are proposed for object localization and recognition using single region classification. However, these approaches ignore the coexistence relations of multi-class/multi-region in object detection. Therefore, their detection performance is compromised especially in low resolution images, small scale images and heavy occlusion images.

Intuitively, it is believed that the relation of object-to-object coexistence in an image can provide contextual information for challenging object recognition tasks. For example as shown in Fig.1, when one wants to detect a *bottle* in an image, Faster R-CNN [7] usually independently classifies a single proposal region and back-propagates based on ROI proposals. However, the feature of *wine glass* is not sufficient discriminative. As a result, the confidence score of wine glass is low. In fact, objects such as *bottle, table, chair* always simultaneously occur in the same scene. In this case, the coexistence attribute could enhance the discriminative feature of each class. Therefore, incorporating available coexistence object classes could improve the performance of object detection.

In this paper, we propose to develop a coexistence feature map to model multi-class relations for object detection, which can be trained in an end-to-end manner using the

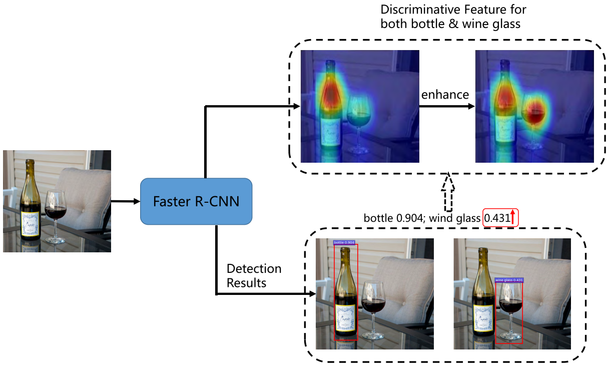


FIGURE 1. **Illustration of our proposed CRN for object detection, where the discriminative features are learned based on attention mechanism. For multi-class objects, Faster R-CNN would weak attention feature for some classes. To enhance the discriminative feature for classification, coexistence features are jointly learned by the independently obtained positive attention features for each class.**

state-of-the-art Faster R-CNN framework. In the task of object detection, the category of each object instance is often recognized individually, without considering a priori knowledge. In our work, firstly, the discriminative features for each class are learned by weakly supervision, which can generate the class activation features for each class. Then, we utilize the class activation feature of all given object categories to learn the multi-class coexistence feature as the prior knowledge of object detection. At last, the captured coexistence features are concatenated into Faster R-CNN, which is used to assist the classification of each bounding box proposals. Experimental results prove the efficiency of the designed network.

In summary, our main contributions are as follows:

1) We adopted a classical image-level classification network to learn the local context with attention mechanism. Moreover, we utilize Coexistence Relation Net (CRN) to learn the coexistence feature of multi-class objects from the attention feature.
2) We concatenated an end-to-end network based on Faster R-CNN for multi-class object detection, which exploits coexistence relations of multi-class objects to enhance the classification feature representation.
3) We comprehensively evaluate the proposed method on the public MS-COCO dataset [11] and PASCAL VOC 2007, 2012 [12]. The experimental results prove that the proposed model has good learning capability and works well on object detection task.

II. RELATED WORK

The task of object detection is to recognize and locate objects of interest in a given image/video. State-of-the-arts object detectors mostly adopt the strong supervision in learning appearance models of object categories, by training the images annotated with bounding boxes and the corresponding category labels. However, these proposal-based classifiers are trained independently for each class, and lack the ability to use the other categories as context.

In fact, contextual information is always believed that it is vital for human to recognize objects [13]. Naturally, for extracting available contextual information in object detection task, the object's local context and the global context have to be considered. Many previous studies have started to exploit contextual information for object detection [14]–[17]. For example, contextual relations can be modeled by learning the discriminative feature in a local region outside of a sliding detection's window [18]. The spatial coexistence of object-to-object is collected by learning inhibitory intra-class and inter-class constraints [19], [20]. Reference [21] proposes to model context relations based on probabilistic models. There are also several methods on discriminative classifiers to explore the context similarity among classes [22]. Recently, the contextual information for object detection are treated as a sequential prediction problem. Reference [23] proposes to sequentially choose detection window to detect objects in an image. Based on Recurrent Neural Networks (RNN) structure, such as [24] and [25], the sequence region proposals are utilized as a cell of contextual information to infer the other possible object categories in a given image. Otherwise, attention mechanism is also introduced to model the context to infer object category. Reference [26] proposed to recurrently generate the attention feature to incorporate the discriminative global context into region-based object detection. However, networks which incorporate context by sequencing the region proposals still not entirely understand the context relations [27].

III. COEXISTENCE RELATION NET FOR OBJECT DETECTION

In this section, we describe the details of the proposed method. The overall framework of our approach is shown in Fig.2. The main structure of network is the Faster R-CNN based on VGG-16. We add a network branch to learn the coexistence feature, where the features from conv 5-3 of Faster R-CNN are feed into the branch and the ROI labels are used as the image-level labels to implement the multi-label classification. Then, the global average pooling (GAP) operation is established to localize the class activation feature [28]. In the following, these feature maps are integrated into a coexistence feature to represent the spatial and coexistence relation in an image by the designed CRN architecture. At last, the feature vectors from CRN and Faster R-CNN are aggregated to calculate the final confidence scores of each proposal. The whole network is trained in an end-to-end manner.

A. ATTENTION FEATURE LEARNING FOR SINGLE CLASS

In case that an image has multi-class objects, the discriminative features for each class are generally learned independently. In our proposed approach, we resize conv 5-3 to a fixed size as input to generate the class activation maps (CAM) [28]. It is noted that, Faster R-CNN can independently learn the discriminative feature for each class, the relations among multi-class are weakened by constraining

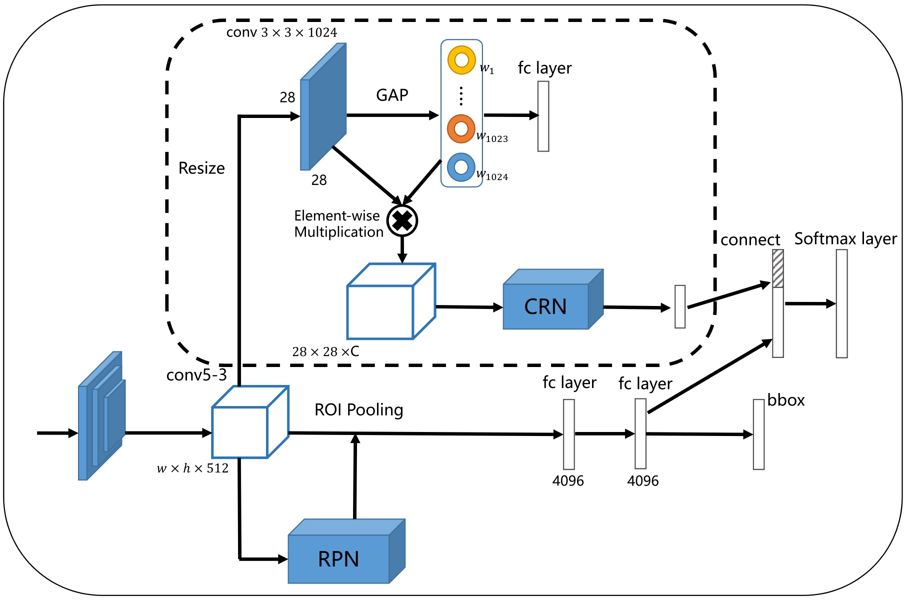


FIGURE 2. Overall framework of our approach, which follows the structure of VGG-16. The whole network consists of two sub-networks: one is Faster R-CNN which independently learns the discriminative feature for each class, another tries to jointly learn the coexistence feature based on the independent attention feature of each class. Input image is fed into Faster R-CNN to produce *conv5-3* feature maps which are fed into two branches, respectively. In the top branch, the feature maps are firstly pooled into a fixed size. Then, the local attention feature maps are generated based on a convolutional layer, GAP and a fc layer. Next, CRN is trained to learn the coexistence feature. In final, the coexistence feature and the ROI pooling feature are connected to be used for object instance classification.

a local region via ROI proposals. But, the convolutional features from conv 5-3 of the pretrained Faster R-CNN network, still carry global contextual information, to some extent. In our work, we believe that the discriminative feature for one class could be semantically related to other classes in a given image. Therefore, our neural network learns to predict such relations for each class with image-level supervision, which follows the setting in CAM [28]. We add a convolutional layer of size 3 × 3, stride 1, pad 1 with 1024 channels, followed by a GAP and a fully connection layer. For a given image, the feature maps from layer conv 5-3 are firstly resized to 28 × 28, after the added convolutional layer, it can be represented as $f^k \in \mathbb{R}^{28\times28\times1024}$, k denotes the channel number ($k = 1, 2, \ldots, 1024$). For the spatial location (x, y), we sum the feature of (x, y) in f^k. By utilizing GAP, the class score can be denoted as

$$S_c = \sum_k w_c^k \sum_{x,y} f^k(x, y) = \sum_{x,y} \sum_k w_c^k f^k(x, y). \tag{1}$$

where w_c^k is the class activation weights of class c corresponding to $k-th$ channel feature map. For each individual class c, the attention feature map can be obtained by using element-wise multiplication operator as

$$M_c(x, y) = \sum_k w_k^c f^k(x, y). \tag{2}$$

Here, attention feature map $M_c(x, y)$ indicates the importance of the activation at spatial (x, y) leading to the classification of an image to class c. Fig.3 shows an example of discriminative feature regions for two classes. It shows that the attention feature supervised by image-level annotation can effectively and independently capture the related visual regions for each class. In addition, since the attention features are learned individually for each category, the attention values for the class-related regions can independently assigned a higher value compared to the other regions. As shown in the above figure, even the cat nears to the cup, the attention feature is also constraint to the local region and the discriminative regions are independent with each other whose strength is enough to be accurately classified.

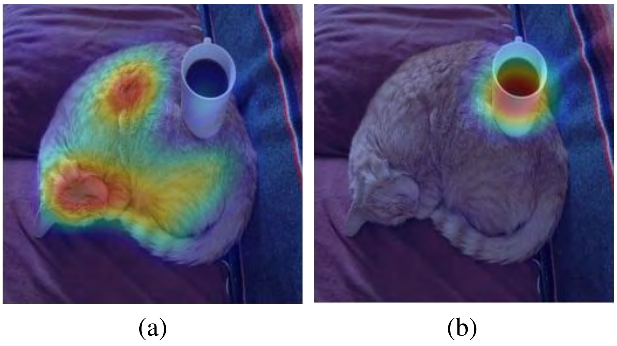

(a)
(b)

FIGURE 3. Discriminative feature for single class by CAM: (a) attention feature for a cat; (b) attention feature for a cup.

B. COEXISTENCE FEATURE LEARNING FOR MULTI-CLASS

Attention feature maps encode rich discriminative information for each class. Meanwhile, the spatial information of attention feature can be localized in each class channel. However, the attention feature for each class are still independent to each other. Therefore, we try to learn the coexistence relation from the weighted attention maps.

Given the attention feature maps $M_c \in \mathbb{R}^{28\times 28\times C}$ (C is the total number of categories), it still carries the global contextual information in different channels. Here, we adopt a network which is similar as that in [29] to learn the global class relation information as coexistence feature (which is equivalent to global contextual information). It should be noted that we apply the similar network to extract the features as that in [29], but the input attention feature maps are different which are explained in Section IV. In our work, our goal is to extract the existence feature for each category and then to construct the relation among different categories as the coexistence feature. Fig.4 illustrates the process to produce the coexistence feature. It can be observed that C-dimensional attention feature maps indeed combine features of all locations, meaning that the spatial relations among multi-class can be accounted. Hence, we add two convolutional layers to capture the spatial relations of multi-class with a $1 \times 1 \times C$ layer and a $1 \times 1 \times 512$ layer. Then, another convolutional layer which has 2048 kernels with size $28 \times 28 \times 1$ is used to learn the semantic relations among multi-class. The intuition is that one class may only semantically relate to a small number of other classes. Thereby, for the third convolution layer, we group 2048 kernels, with each group of 4 kernels corresponding to one feature channel. The 4 kernels in each group convolve the same feature channel independently, and the attention spatial regions of related classes are calculated by different kernels in one group. At last, we can obtain a 2048−dimension feature vector to estimate the label confidences c_{crn} by a fc layer. The objective loss function of the proposed branch is a cross-entropy loss, as

$$J_{loss}(c, c_{crn}) = \sum_{l=1}^{C} c^l \log\sigma(c^l_{crn}) + (1 - c^l_{crn}) \log(1 - \sigma(c^l_{crn})) \tag{3}$$

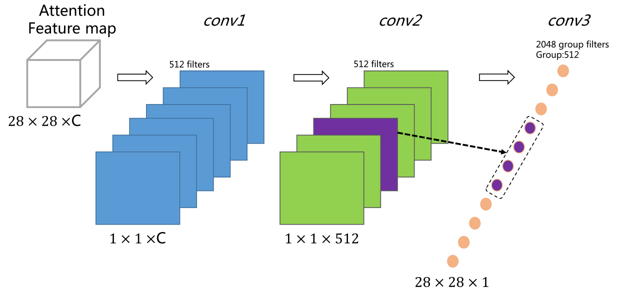


FIGURE 4. Detailed network of CRN. The attention feature maps firstly pass two convolutional layers with a 1 × 1 × C layer and a 1 × 1 × 512 layer, then another convolutional layer which has 2048 kernels with 28 × 28 × 1 is utilized, which is grouped at 4 kernels to 1 channel.

where c^l is the ground-truth label. Here, the trained features denote the coexistence features.

C. OVERALL NETWORK AND TRAINING DETAILS

Based on the two branches, we can obtain two feature vectors: one with 4096 dimensions from ROI pooling and 2 fc layer in Faster R-CNN (denoted as $\mathbf{F}_R$), another with 2048 dimensions from the designed CRN (denoted as $\mathbf{F}_C$). We connect $\mathbf{F}_R$ and $\mathbf{F}_C$ as the final feature representation. Denote $g \in 0, 1, \ldots, C$ as the ground-truth class name, the loss function on each ROI proposal for multiple tasks including classification and bounding box regression is defined as

$$J = J_{cls}([\mathbf{F}_R, \mathbf{F}_C]) + [g \geq 1] J_{reg}(\mathbf{F}_R), \tag{4}$$

where $[\mathbf{F}_R, \mathbf{F}_C]$ indicates connecting two features along the channel axis.

We train the network in multiple steps. Firstly, the Faster R-CNN is fine-tuned based on the pretrained ImageNet model [31]. Secondly, we fix the convolutional layer to train CAM with the classical cross-entropy classification loss. Thirdly, we train CRN by fixing all other layers, and also with the cross-entropy classification loss. Finally, the whole network is jointly fine-tuned with (4).

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

In this part, the proposed framework is evaluated on MS-COCO [11] and PASCAL VOC [12].[1] For MS-COCO, the training set is composed of 82, 783 images, and the contained objects can be categorized into 80 classes, with about 2.9 object labels per image. We use the ‘‘trainval35k’’ set for training and the ‘‘minival’’ set for testing. In the evaluation, toolkits provided are used, and the main metrics (AP and AR) are based on detection average precision/recall. For PASCAL VOC 2007 and 2012, the datasets contain 9, 963 images and 22, 531 images, respectively, which are divided into *train*, *val* and *test* subsets. We train our models on the union of VOC 2007 *trainval* and VOC 2012 *trainval*. The evaluation metrics are *Average Precision* (AP) and *mean of Average Precision* (mAP) complying with PASCAL challenge protocols.

We use TensorFlow to implement our model, which is built on top of the open-source Faster RCNN implementation.[2] During training of our model, each SGD mini-batch consists of 256 randomly sampled object proposals from each randomly chosen image. In each mini-batch, 25% of the object proposals are selected as foreground that have Intersection over Union (IoU) overlap with a ground-truth bounding box of larger than 0.5 and the remaining object proposals that have a maximum IoU with ground-truth in the interval [0.1, 0.5) acting as negative training instances. For data augmentation, images are horizontally flipped with a probability of 0.5.

[1] In fact, PASCAL dataset is not suitable to test context-based object recognition as most of its images contain only a single object class [32], but for the comparison between our framework and the other works, the performance is also evaluated in our work.

[2] https://github.com/endernewton/tf-faster-rcnn

TABLE 1. Performance evaluation on MS-COCO.

Approach	AP	AP-.5	AP-.75	AP-S	AP-M	AP-L	AR-1	AR-10	AR-100	AR-S	AR-M	AR-L
Faster R-CNN [7]	24.2	45.1	23.4	7.4	27.5	38.2	23.6	33.7	34.3	11.7	39.5	54.1
SRN [29]	26.8	47.1	27.3	11.5	30.8	38.4	25.3	37.7	38.4	16.9	43.3	56.2
TF Faster R-CNN [30]	26.5	46.7	27.2	11.8	30.4	37.5	24.9	36.3	37.1	17.3	42.1	52.4
TF Faster R-CNN with ResNet-101 [30]	35.4	55.2	38.2	15.6	40.4	52.2	31.3	46.3	47.4	24.6	53.9	67.5
proposed	26.9	47.3	27.5	12.0	30.9	37.7	25.6	38.6	39.6	19.5	45.2	56.4
proposed with ResNet-101	36.7	56.8	36.4	15.8	41.9	53.6	31.7	47.9	49.5	26.6	56.5	69.3

TABLE 2. Comparison of detection results on COCO between TF faster R-CNN V.S. proposed in different setups.

Approach	Train	Test	stepsize	itersize	AP	AP-.5	AP-.75	AP-S	AP-M	AP-L	AR-1	AR-10	AR-100	AR-S	AR-M	AR-L
TF Faster R-CNN	NMS	NMS	350K	490K	26.5	46.7	27.2	11.8	30.4	37.5	24.9	36.3	37.1	17.3	42.1	52.4
	NMS	TOP	350K	490K	26.9	47.0	27.7	12.0	31.0	38.9	25.3	37.2	38.1	17.6	43.1	54.0
	NMS	NMS	600K	790K	27.9	48.2	29.0	11.8	31.8	40.3	26.0	37.5	38.3	17.6	43.4	55.4
	NMS	TOP	600K	790K	28.3	48.7	29.5	11.8	32.5	41.9	26.2	38.3	39.2	18.0	44.3	56.7
proposed	NMS	NMS	350K	490K	27.6	48.1	28.4	12.0	31.8	38.9	25.9	39.1	40.1	20.1	45.7	56.5
	NMS	TOP	350K	490K	28.0	48.8	28.9	12.2	32.4	39.7	26.3	39.5	40.5	19.8	46.2	57.8
	NMS	NMS	500K	600K	27.7	48.4	28.6	12.1	32.1	39.1	26.2	39.2	40.2	20.3	45.8	56.8
	NMS	TOP	500K	600K	28.2	49.0	29.2	12.3	32.6	40.2	26.4	39.6	40.7	20.2	46.3	58.0

No other data augmentation is used. The learning rate starts with 0.001 and decreases to 0.0001 after 350K iterations with a total iteration number being 490K. The models are trained based on a NVIDIA GeForce Titan X GPU (pascal) and Intel Core i7-4930K CPU @ 3.40 GHz. For training, the average training time for each iteration is about 0.29 second. For testing, on average, the proposed approach processes one image within 0.14 second (excluding object proposal time).

B. PERFORMANCE COMPARISONS ON MS-COCO

In order to verify the effectiveness of the proposed approach, we compare our proposed approach (VGG-16 model and ResNet-101 model) with Faster R-CNN [7], TF Faster R-CNN [30] (VGG-16 model and ResNet-101 model) and SRN [29] (VGG-16 model).[3] As shown in Table.1, we evaluate the object detection performance on MS-COCO based on AP and AR metrics, with similar experimental setup. In this experiment, Faster R-CNN [7] is totally trained for 240K iterations with an initial learning rate of 0.003 and then after 80K iterations with 0.0003. For TF Faster R-CNN and our proposed approach, the experiment setups are set similar as Faster R-CNN. Particularly, we fix 1 image in a batch, and TF Faster-RCNN use *crop-and-resize* instead of ROI pooling, however, we use ROI pooling in our approach. As shown the results in Table.1, our proposed approach achieves the best performance both on AP and AR metrics. Faster R-CNN [30] achieves 26.5% and 35.4% on AP performance with VGG-16 and ResNet-101, respectively. As comparison, our proposed approach further improves the AP performance to 26.9% and 36.7%. Compared to SRN [29], the AP results of both approaches in object detection task are similar, which are 26.8% and 26.9% in case of VGG-16, respectively. However, it should be noted that on AP-S and AR-S performance, the proposed approach achieves better results. As shown in Fig. 5, our attention maps focus on all class-related regions and each attention values are individually assigned, hence, it is easy to find the existence feature for small objects; For the attention maps of SRN, the spatial regularization operator could weaken the class-related feature of small objects.

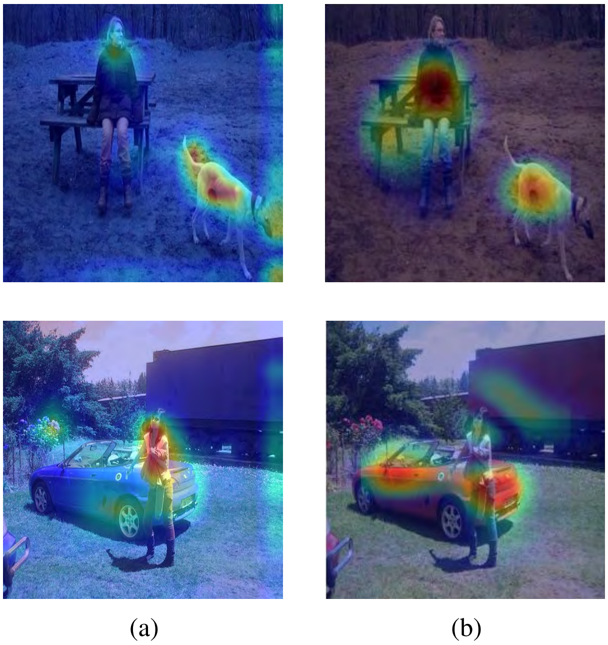


FIGURE 5. Comparison of attention maps with two classes which are obtained by SRN and our proposed approach. (a) SRN; (b) proposed.

In addition, to further verify the performance of the proposed approach, different training and testing setups are employed. Table.2 lists the comparison results between our implemented models and TF-Faster R-CNN models [30] in different experimental setups, which includes different testing modes and iteration numbers, refer to [30]. In particular, the AP performance of our proposed approach is 27.6% compared to 26.5% of TF Faster R-CNN with iteration number 490K, where the learning rate changes after 350K iteration times; for our proposed approach with stepsize 500K and itersize 600K, AP performance is 27.7% which is similar as the AP performance of TF Faster R-CNN with

[3] We adopt the SRN net based on TF Faster-RCNN as the description in [29]. For simplify the training, the backbone is VGG-16.

TABLE 3. Comparison of detection results on PASCAL VOC 2007 between faster R-CNN, ION V.S. proposed.

Approach	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster R-CNN	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
ION [16]	74.6	78.2	79.1	76.8	61.5	54.7	81.9	84.3	88.3	53.1	78.3	71.6	85.9	84.8	81.6	74.3	45.6	75.3	72.1	82.6	81.4
proposed	75.0	76.5	82.1	75.3	59.9	65.6	80.9	87.3	86.7	58.8	79.8	69.8	84.9	84.9	79.8	78.1	46.1	76.8	69.6	81.0	75.2

TABLE 4. Comparison of detection results on PASCAL VOC 2012 between faster R-CNN V.S. proposed.

Approach	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster R-CNN	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
proposed	72.1	83.8	79.2	75.1	55.9	55.9	79.4	80.1	89.0	52.4	76.2	54.3	87.4	79.5	82.4	82.9	45.7	75.9	58.8	80.2	67.4

stepsize $600K$ and itersize $790K$. But for the AR performance, our proposed approach even performs better than TF Faster R-CNN. Compared to 37.5% AR-10 performance of TF Faster R-CNN, that of our proposed approach is 39.2% in case of our iteration time is only $600K$. It means that the proposed approach achieves to enhance the discriminative region for classification. The results in Table.2 can also prove that better AP is able to converge with more iterations.

C. PERFORMANCE COMPARISONS ON PASCAL VOC

As shown in Table.3, we also evaluate our proposed framework compared with Faster R-CNN [7] and ION [16] method which exploits information both inside and outside the region of interest as global context information. The contextual information outside the region of interest is integrated using spatial RNN networks. Applying our approach described above, we obtain a mAP of 75.0%, and the implement details are similar as our previous setting. As comparisons, Faster R-CNN is used as the ablation study which just includes one sub-branch in our framework. The proposed network consists of two sub-network, *i.e.* the state-of-the-art Faster R-CNN network which is used to capture the local discriminative feature, and the CRN network which is utilized to learn the multi-class coexistence contextual feature. the Table.3 shows that 1.8% improvement of mAP can be obtained by incorporating multi-class coexistence feature for detection compared to the Faster R-CNN baseline. We can see that 0.4% improvement in mAP can be observed compared with the general setting of ION work. This validates that the proposed CRN network can provide useful contextual cues for better object detection. The AP performance of some classes obtained by the proposed approach, such as boat, bus and cow etc., are lower than that by Faster R-CNN. However, some classes such as bike, bottle, car and chair etc, achieves a large incasement on AP performance. Moreover, an extra comparison experiment is constructed to verify the performance on PASCAL VOC 2012. Table.4 further shows the evaluation results on PASCAL VOC 2012[4] by using our proposed framework and Faster R-CNN, respectively.

D. VISUALIZATION AND ANALYSIS

The effectiveness of our approach has been quantitatively evaluated in Table.1 and Table.2, we visualize and analyze the learned coexistence feature from our CRN to illustrate the capability. We observe that the learned attention features pass two convolutional layers, the spatial relation and the coexistence relation would randomly distribute

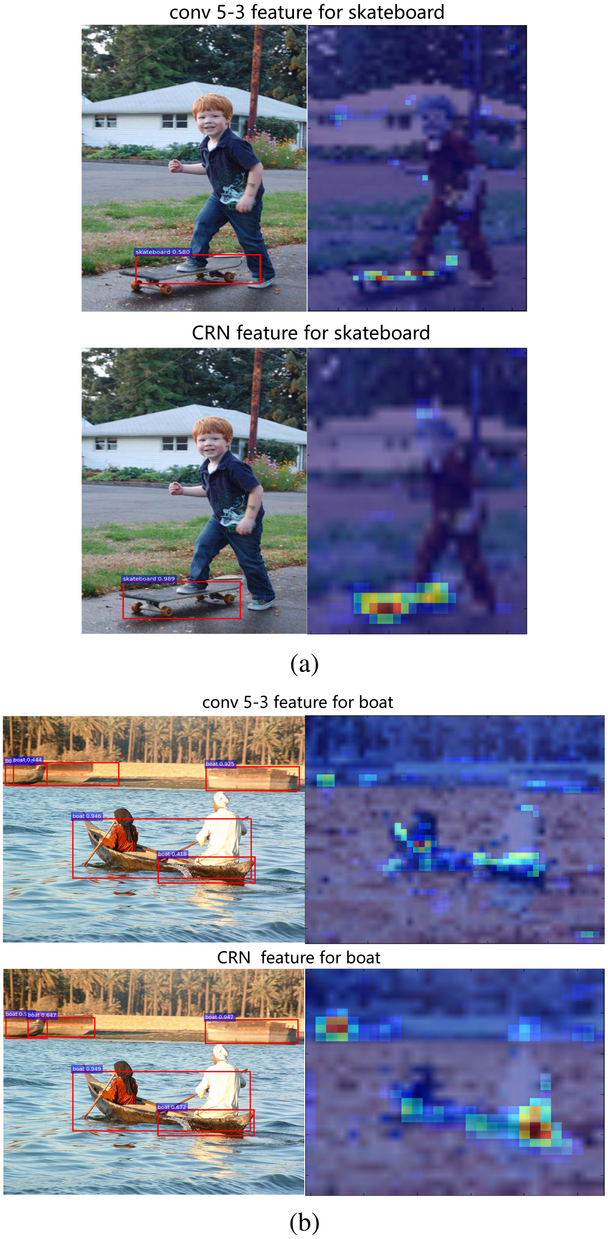
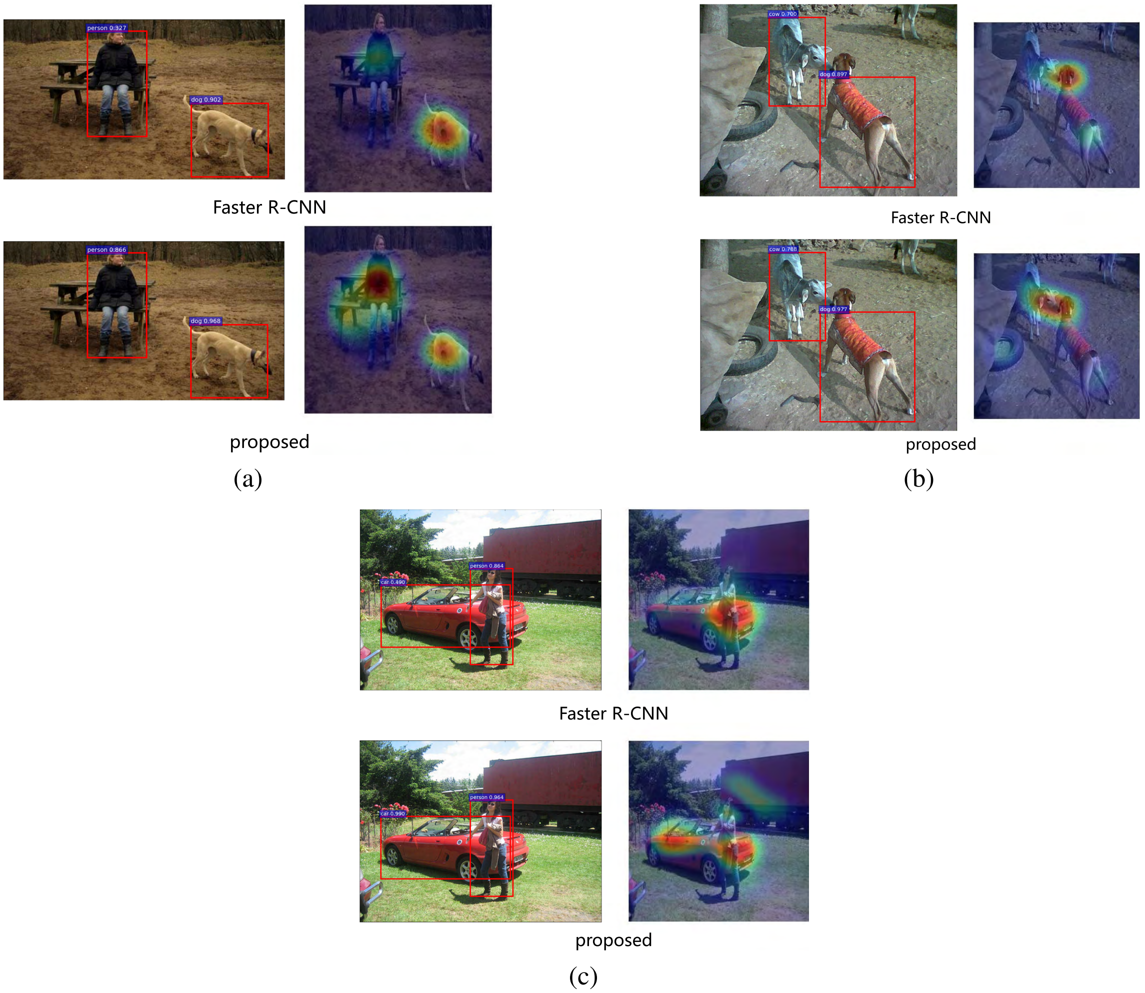


FIGURE 6. Comparison of detection results guided by conv 5-3 feature and coexistence feature. (a) *skateboard* detection results of Faster R-CNN (top) V.S. proposed (bottom); (b) *boat* detection results of Faster R-CNN (top) V.S. proposed (bottom).

[4] http://host.robots.ox.ac.uk:8080/anonymous/LXOLXE.html



Faster R-CNN

proposed

(a)

Faster R-CNN

proposed

(b)

Faster R-CNN

proposed

(c)

FIGURE 7. Comparison of detection results and the CAM features by Faster R-CNN and our proposed approach. (a) Detection results: 32.7% on a person by Faster R-CNN (top) V.S. 86.6% by proposed (bottom), the corresponding features are shown in the right; (b) Detection results: 70.0% on a cow and 89.7% on a dog by Faster R-CNN (top) V.S. 78.8% and 97.7% by proposed (bottom), the corresponding features are shown in the right; (c) Detection results: 49.0% on a car and 86.4% on a person by Faster R-CNN (top) V.S. 99.0% and 96.4% by proposed (bottom), the corresponding features are shown in the right.

different channels. In fact, each channel is corresponding to one specific class. In Fig.6, we provide two such examples. The left sub-figures in Fig.6(a) show the *"skateboard"* detection results from Faster R-CNN (top) and our proposed (bottom), respectively. Since the local discriminative feature of "boat" is weak in conv 5-3, the detection confidence is only 58.0%; however, with the assist of the CRN, the coexistence feature can preserve the attention feature of each class (as shown the right of Fig.6(a)) so that the detection confidence increase to 98.9%. Similar results are also shown in Fig.6(b), the left sub-figures show the *"boat"* detection results from Faster R-CNN (top) and out proposed (bottom), respectively. The class activation maps are also visualized, and the results prove that the enhanced features are helpful to improve the performance of object detection. The final detection scores are largely improved in the given 3 examples, the detailed results are shown in Fig.7.

V. CONCLUSION

In this paper, we propose to produce a coexistence feature to model contextual relations of multi-class for object detection. Firstly, the discriminative feature for each class in an image are learned based on the attention mechanism. Secondly, a CRN network is utilized to integrate the attention feature of each class into coexistence feature vectors for multi-class object detection. At last, the captured contextual information is connected with the feature vectors from Faster R-CNN, and be exploited to assist the classification of each bounding box proposals. Experimental results prove the efficiency of the designed network, and visualization of learned models also shows the proposed approach could effectively capture the coexistence relations of multi-class objects. In the future work, we can implement our proposed CRN branch network based on the other state-of-the-arts networks, such as R-FCN [8], SSD [9] and YOLO [10], and ResNet, DenseNet are also able to be utilized as backbone to further verify the performance of our proposed approach in the deeper network and with longer training time.

REFERENCES

[1] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[2] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[3] 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.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. (2016). "Densely connected convolutional networks." [Online]. Available: https://arxiv.org/abs/1608.06993

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

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

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[8] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 379–387.

[9] W. Liu *et al.* "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[11] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 8693, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014.

[12] 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.

[13] M. Bar, "Visual objects in context," *Nature Rev. Neuro-Sci.*, vol. 5, no. 8, pp. 617–629, 2004.

[14] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2004, pp. 350–362.

[15] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1271–1278.

[16] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[17] Z. Xue, G. Li, and Q. Huang, "Joint multi-view representation and image annotation via optimal predictive subspace learning," *Inf. Sci.*, vol. 451, pp. 180–194, Jul. 2018.

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[19] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 1–12, Oct. 2011.

[20] D. Modolo and V. Ferrari, "Learning semantic part-based models from Google images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1502–1509, Jun. 2018.

[21] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[22] R. G. Cinbis and S. Sclaroff, "Contextual object detection using set-based classification," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 7577, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012.

[23] A. Gonzalez-Garcia, A. Vezhnevets, and V. Ferrari, "An active search strategy for efficient object class detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3022–3031.

[24] X. Chen and A. Gupta. (2017). "Spatial memory for context reasoning in object detection." [Online]. Available: https://arxiv.org/abs/1704.04224

[25] Y. Li, W. Ouyang, X. Wang, and X. Tang. (2017). "ViP-CNN: Visual phrase guided convolutional neural network." [Online]. Available: https://arxiv.org/abs/1702.07191

[26] J. Li *et al.*, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.

[27] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4898–4906.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[29] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5513–5522.

[30] X. Chen and A. Gupta. (2017). "An implementation of faster RCNN with study for region sampling." [Online]. Available: https://arxiv.org/abs/1702.02138

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

[32] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 129–136.

CHAO YAO received the B.S. degree in computer science from Beijing Jiaotong University (BJTU), Beijing, China, in 2009, and the Ph.D. degree from the Institute of Information Science, BJTU, in 2016. From 2014 to 2015, he was a Visiting Ph.D. Student with the École Polytechnique Fédérale de Lausanne, Switzerland. In 2016, he joined the Institute of Sensing Technology and Business, Beijing University of Posts and Telecommunications, Beijing. In 2018, he joined the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing. His current research interests include image and video processing, computer vision, and robot technique.

PENGFEI SUN is currently pursuing the master's degree with the Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include computer vision.

RUICONG ZHI received the Ph.D. degree in signal and information processing from Beijing Jiaotong University in 2010. From 2016 to 2017, she was a Visiting Scholar with the University of South Florida. In 2008, she was a joint Ph.D. Student with the Royal Institute of Technology. She is currently an Associate Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing. She has published over 50 papers. He holds six patents. She was a recipient of over 10 awards, including the National Excellent Doctoral Dissertation Award Nomination. Her research interests include facial and behavior analysis, artificial intelligence, and pattern recognition.

YANFEI SHEN received the M.S. degree in computer science from the Key Laboratory of Multimedia and Network Communication, Wuhan University, China, in 2002, and the Ph.D. degree from the University of Chinese Academy of Sciences in 2014. He is currently an Associate Professor with Beijing Sport University. His research interests include video codec technology, compressed sensing, pattern recognition, machine learning, simultaneous localization, and mapping.