Attention CoupleNet: Fully Convolutional Attention Coupling Network for Object Detection

Yousong Zhu, Chaoyang Zhao, Haiyun Guo, Jinqiao Wang, *Member, IEEE*, Xu Zhao, and Hanqing Lu, *Senior Member, IEEE*

***Abstract*— The field of object detection has made great progress in recent years. Most of these improvements are derived from using a more sophisticated convolutional neural network. However, in the case of humans, the attention mechanism, global structure information, and local details of objects all play an important role for detecting an object. In this paper, we propose a novel fully convolutional network, named as Attention CoupleNet, to incorporate the attention-related information and global and local information of objects to improve the detection performance. Specifically, we first design a cascade attention structure to perceive the global scene of the image and generate class-agnostic attention maps. Then the attention maps are encoded into the network to acquire object-aware features. Next, we propose a unique fully convolutional coupling structure to couple global structure and local parts of the object to further formulate a discriminative feature representation. To fully explore the global and local properties, we also design different coupling strategies and normalization ways to make full use of the complementary advantages between the global and local information. Extensive experiments demonstrate the effectiveness of our approach. We achieve state-of-the-art results on all three challenging data sets, i.e., a mAP of 85.7% on VOC07, 84.3% on VOC12, and 35.4% on COCO. Codes are publicly available at https://github.com/tshizys/CoupleNet.**

***Index Terms*— Object detection, cascade attention, global structure, local parts.**

I. Introduction

OBJECT detection aims to locate and classify all targets in the image or video. Compared to specific object detection, such as face, pedestrian and vehicle detection, general object detection often faces more challenges due to the complex inter-class appearance variations, like non-grid deformations, truncations, occlusions and inter-class interference. However, no matter how complicated the objects are, when humans identify a target, we first focus on where it is, and then the recognition of object categories is subserved by both a global process that retrieves structural information and a local process that is sensitive to individual parts. This motivates us to build a detection model which considers attention mechanism, global and local information of objects.

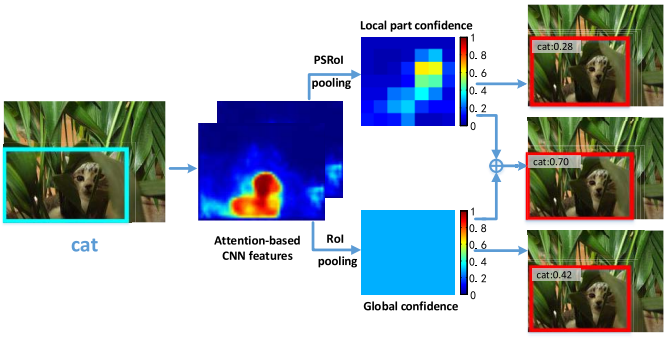

Fig. 1. A toy example of object detection by applying our ACoupleNet. The attention-based CNN features are generated firstly to provide a global location perception. Subsequently, by coupling the local part information and global structure of the object, we can detect the cat accurately with a confidence score of 0.70. Best viewed in color.

Thanks to the development of Convolutional Neural Networks (CNN) [1], recent CNN-based object detection algorithms [2]–[5] have been proposed consecutively and achieved significant progress in generic benchmarks, *e.g.* PASCAL VOC [6] and MS COCO [7]. However, as far as we know, none of these methods have applied the attention mechanism to the image feature extraction network to achieve state-of-the-art results in object detection task. Actually, the attention mechanism has been widely investigated in machine translation [8], speech recognition [9] and image captioning [10] fields. In these fields, the attention model is usually applied to weight the importance of specific parts from a source input, such as words in machine translation, fragment sequences in speech recognition and regions in image captioning. It is quite obvious that the attention mechanism can help to extract the contents we are interested in within a given input. Therefore, it's reasonable to exploit the attention mechanism to estimate the exact locations of objects from the complex scenarios and provide preliminaries and heuristics for accurate object detection. As shown in Figure 1, we first highlight the target regions and suppress the unrelated backgrounds by generating attention-aware convolutional features. The attention-aware

Manuscript received January 10, 2018; revised June 15, 2018 and July 24, 2018; accepted August 9, 2018. Date of publication August 13, 2018; date of current version September 19, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61772527. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Julian Fierrez. *(Corresponding author: Haiyun Guo.)*

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yousong.zhu@nlpr.ia.ac.cn; chaoyang.zhao@nlpr.ia.ac.cn; haiyun.guo@nlpr.ia.ac.cn; jqwang@nlpr.ia.ac.cn; xu.zhao@nlpr.ia.ac.cn; luhq@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2865280

features are then fed into the subsequent detection subnetwork to better extract object-related features.

Based on the aforementioned object-aware features, we then focus more on better classifying the object categories. As two representative region-based CNN detection approaches, Fast/Faster R-CNN [2], [5] extracts the global structure representation of the object via the RoI pooling layer to predict the category of each region proposal while R-FCN [4] conducts the inference with the position-sensitive score maps. Through removing the RoI-wise subnetwork, R-FCN has achieved higher detection speed while keeping the detection performance. However, the global structure information is ignored by the PSRoI pooling. As shown in Figure 1, using PSRoI pooling to extract local part information for final object category prediction, R-FCN leads to a low confidence score of 0.28 for the cat detection since most of the local responses are disturbed by the noisy background. Conversely, the global structure of cat could be extracted by the RoI pooling, but the confidence score is 0.42, which is also very low for the incomplete structure of cat. By coupling the global confidence with the local part confidence together, we can obtain a more reliable prediction with the confidence score of 0.70.

In fact, the idea of fusing global and local information together is widely used in lots of visual tasks. In fingerprint recognition, Gu *et al.* [11] combined the global orientation field and local minutiae cue to largely improve the performance. In clique-graph matching, Nie *et al.* [12] proposed a clique-graph matching method by preserving global clique-to-clique correspondence and local unary and pairwise correspondences. In scene parsing, Zhao *et al.* [13] designed a pyramid pooling module to effectively extract hierarchical global contextual prior, and then concatenated it with the local FCN feature to improve the performance. In traditional object detection, Felzenszwalb *et al.* [14] incorporated a global root model and several finer local part models to represent highly variable objects. All of which show that effective combination of the global structural properties and local fine-grained details can achieve complementary advantages. Thus the keypoint lies in how to design a structure to explicitly collect these two messages and effectively couple them together for object detection.

Therefore, in this paper, we first propose a novel cascade attention structure to gradually locate the objects from the image for object detection. In contrast to some common unsupervised methods, our attention structure uses pixel-level supervision and can be regarded as a spatial global-to-local glimpse against image. Second, to fully explore the global and local clues of objects, we propose a unique fully convolutional coupling structure, which couples the global structure and local parts of object to boost the detection accuracy. Our network is fully convolutional and denoted as Attention CoupleNet (ACoupleNet).

Specifically, the cascade attention structure is naturally encoded into the backbone to generate attention-aware convolutional features which used for the subsequent refining operation. Then the object proposals obtained by the RPN are fed into the coupling module which consists of two branches. One branch adopts the PSRoI pooling to capture the local part information of the object, while the other employs the RoI pooling to encode the global and context information. Moreover, we design different coupling strategies and normalization ways to make full use of the complementary advantages between the global and local branches. With the coupling structure, our network can jointly learn the local, global and context expression of the objects, which makes the model to have a more powerful feature representation capacity and generalization ability. Extensive experiments show competitive results on PASCAL VOC 07/12 and MS COCO compared to other state-of-the-art detectors, even with model ensemble approaches. In summary, our main contributions are as follows:

1. A novel cascade attention structure is designed to automatically refine the regions of object and is easy to incorporate with the existing deep networks in a learnable fashion.
2. A fully convolutional coupling structure is proposed to jointly learn the local, global and context information of the object. Different normalization methods and coupling strategies are evaluated to mine the compatibility and complementarity between the global and local information.
3. We achieve the state-of-the-art results on all three challenging datasets, *i.e.* a mAP of 85.7% on VOC07, 84.3% on VOC12, and 35.4% on MS COCO.

A preliminary conference version of this paper can be referred to [15]. Compared to [15], this study contains: (a) a fully convolutional network, named as Attention CoupleNet, is proposed to mine the attention, local and global information for robust object detection. (b) we propose a novel cascade attention structure to gradually highlight the target regions and guide the feature learning by predicting a series of class-agnostic attention maps. As a new algorithmic enhancement, the cascade attention module further perfects the detection stage and improves the detection performance; (c) we perform extensive experiments on MS COCO and PASCAL VOC to validate the effectiveness of the attention module. More extra experiments and theoretical analyses are also added to make the detection framework more solid.

II. Related Work

A. Hand-Crafted Detectors

Before the arrival of CNN, object detection has been dominated by traditional paradigms [14], [16]–[19], which apply the hand-crafted features and classifiers on dense image windows to find the objects. As one of an outstanding framework, DPM [14] described the object system using mixtures of multi-scale deformable part models, including a coarse global root model and several finer local part models. The root model extracts structural information of the objects, while the part models capture local appearance properties of an object. The sum of root response and weighted average response of each part is used as the final confidence of an object. Although DPM provides an elegant framework for object detection, the hand-crafted features, *i.e.* improved HOG [20],

are not discriminative enough to express the diversity of object categories. This is also the main reason that CNN completely surpassed the traditional methods in a short period time.

B. CNN-Based Detectors

Benefited from the tremendous development of deep neural networks [1], [21]–[23], considerable object detection methods based on deep learning have been proposed [3], [24]–[28]. Based on whether generating region proposals, these methods can be roughly divided into two categories, *i.e.*, one-stage methods such as YOLO [26], DenseBox [29], SSD [25], RetinaNet [30], and two-stage methods such as Fast/Faster R-CNN [2], [5], R-FCN [4], Mask R-CNN [31]. Although the one-stage methods has a superiority in speed, the two-stage methods still dominates the detection accuracy on generic benchmarks [6], [7]. The advantages of two-stage methods lies in: First, by exploiting a divide-and-conquer strategy, the two-stage framework is more stable and easier to converge. Second, without the complicated data augmentation and training skills, you can still easily achieve state-of-the-art performance. The main reason for these advantages is that there is a certain structure to encode translation variance features for each proposal, since in deep networks, higher-layers contain more semantic meaning and less location information. As a consequence, a RoI-wise subnetwork [2], [5] or a position-sensitive RoI pooling layer [4] is used to achieve the translation variance in two-stage systems. However, all the existing two-stage methods utilize either the region-level or part-level features to learn the variations, where each one alone is not representative enough for a variety of challenging situations. Therefore, this motivates us to design a certain structure to take advantages of both the global and local features of the object.

C. Context

In addition, context [32] is known to play an important role in visual recognition. Considerable works have been proposed for exploting context in object detection. Bell *et al.* [33] explored the use of recurrent neural networks to model the contextual information. Gidaris and Komodakis [34] proposed to utilize multiple contextual regions around the object. Cai *et al.* [35] collected the context by padding the proposals for pedestrian and car detection. Similar to these works, we also absorb the context prior to enhance the global feature representation.

D. Attention

The attention mechanism has been successfully applied in machine translation [8], speech recognition [9] and image captioning [10]. There are also several works introducing attention-based models for visual recognition tasks. Wang *et al.* [36] designed a residual attention network by stacking attention modules which generates attention-aware features for image classification. Sharma *et al.* [37] predicted visual attention in each frame using a recurrent LSTM model for action recognition. Chen *et al.* [38] proposed an attention model that learns to softly weight the multi-scale features for semantic segmentation. Liu *et al.* [39] introduced a reinforcement learning-based fully convolutional attention localization network to adaptively select multiple attention regions for fine-grained recognition. All of these works show that the attention mechanism has the ability to focus on some salient regions. Therefore, motivated by these methods, in this paper, we extend the attention mechanism to deal with the task of object detection. Instead of stacking attention module naively without any supervision, the attention block we propose is embedded into the backbone regularly with the instance segmentation annotations as the supervised information.

III. Attention CoupleNet

In this section, we first introduce the architecture of the proposed ACoupleNet for object detection. Then we explain in detail how the cascade attention works and how we incorporate local representations, global appearance and contextual information for robust object detection.

A. Network Architecture

The architecture of our proposed ACoupleNet is illustrated in Figure 2. Our ACoupleNet includes three different branches: a) a cascade attention fully convolutional network to highlight the regions of interests and to weaken the interferences of background area, denoted as cascade attention FCN. b) a local part-sensitive fully convolutional network to learn the object-specific parts, denoted as local FCN; c) a global region-sensitive fully convolutional network to encode the appearance structure and context prior of the object, denoted as global FCN. We first use the ImageNet pre-trained ResNet-101 released in [21] to initialize our network. For our detection task, we remove the last average pooling layer and the *fc* layer. Therefore, the whole network is fully convolutional and it could input images of any size for training and test. For inference, given an input image, the cascade attention FCN first generates several hierarchical attention maps, which force the network to locate the foreground regions gradually. We also extract candidate proposals by using the Region Proposal Network (RPN), which shares convolution features with ACoupleNet following [5]. Then each proposal flows to two different branches: the local FCN and the global FCN. Finally, the output of global and local FCN are coupled together as the final score of the object. We also perform class-agnostic bounding box regression in a similar way.

B. Cascade Attention FCN

Our cascade attention FCN consists of base feature extraction network and several attention blocks. As shown in Figure 3, the attention blocks are plugged into the base feature extraction network regularly so that they are able to automatically learn the spatial regions of objects from coarse to fine. The learned attention map in that block is then directly encoded back into the original convolutional features, where object-related features are strongly enhanced by suppressing the background interference in the convolutional maps.

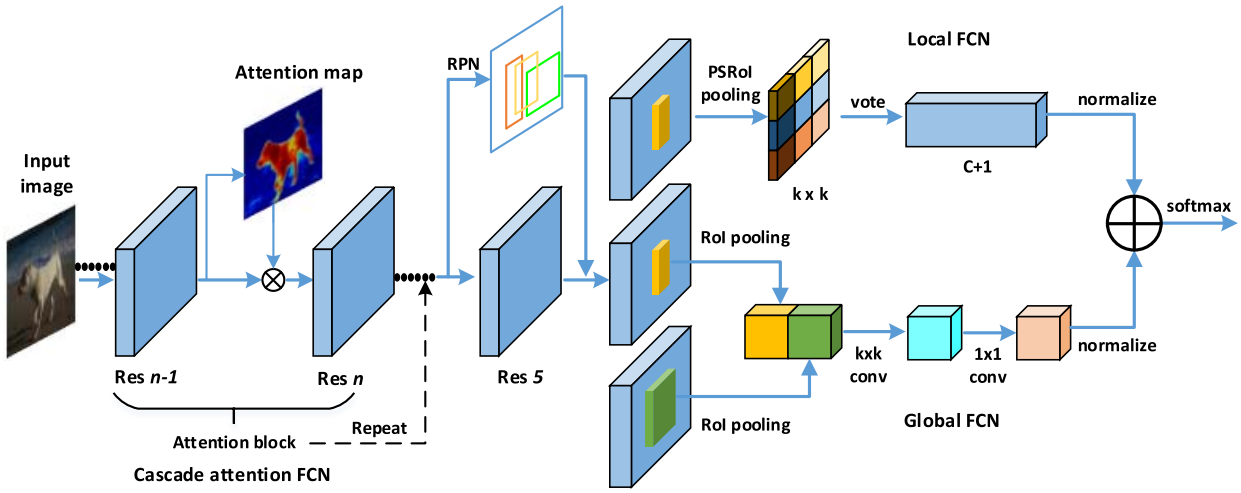

Fig. 2. The architecture of the proposed ACoupleNet. We use ResNet-101 as the basic feature extraction network. Given an input image, we first extract object-aware convolutional features using cascade attention FCN. Then the candidate proposals are generated by Region Proposal Network (RPN) [5] and each proposal flows to two different branches: local FCN and global FCN, in order to extract the global structure information and learn the object-specific parts of object respectively. Finally the output of the two branches are coupled together to predict the object categories.

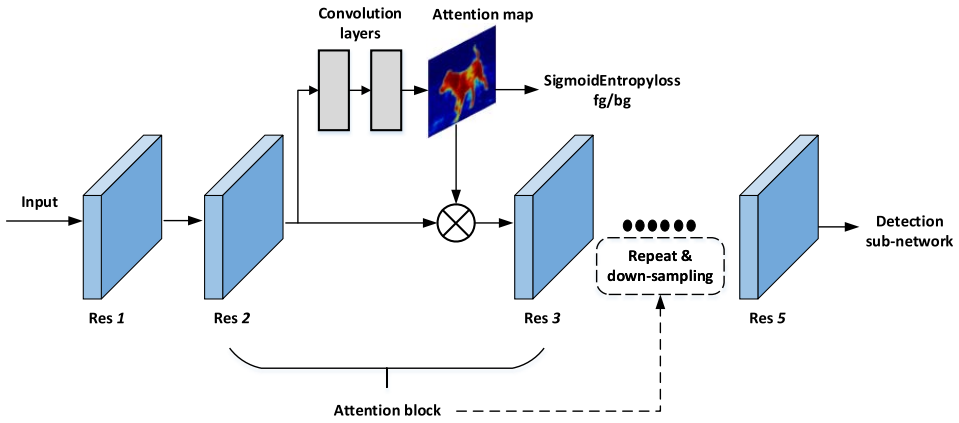

Fig. 3. An overview of our cascade attention structure. The attention block is inserted in the main body of the network repeatedly only when the feature map under down-sampling. Different from traditional unsupervised attention methods, we formulate the generation of attention map as the binary classification task and exploit the ground-truth instance segmentation to supervise the learning process. The learned attention map is then dot with the original feature maps in order to focus on the target areas.

Therefore, this cascade attention structure could be regarded as a kind of representation from global to local with respect to the image.

More specifically, for ResNet-101 the attention block is only inserted in each stage's last residual block, which has strides of {4, 8, 16} pixels (corresponding to conv2, conv3 and conv4) with respect to the input image. Here we do not include conv1 due to the lack of semantic information. For each attention block, we predict a probability heatmap A indicating how likely the pixels belong to the object of interest. This probability heatmap is referred as the attention map which has the same size of the input feature map. The attention block is achieved by applying two 3×3 convolutional layers initialized from a zero-mean Gaussian distribution with standard deviation 0.01. Then the attention map is generated by using a sigmoid activation function on the final convolution features. Let $f_{I,k}$ denote the input feature map of kth attention block, W_a and b_a is the overall weights and bias of transform function (here we omit the non-linearities for simplicity), the attention score $a_{i,j}$ at position (i, j) is calculated as:

$$g = W_a^T f_{I,k} + b_a \tag{1}$$

$$a_{i,j} = \frac{1}{1 + exp(g_{i,j})} \tag{2}$$

where $i = [1, 2, \ldots, H_k]$ and $j = [1, 2, \ldots, W_k]$. H_k, W_k are the height and width of attention map. Note that the attention map is class-agnostic and further encoded into the network by simply performing dot product with input convolutional map channel by channel:

$$f_{O,k}^d = A \odot f_{I,k}^d, \quad d = 1, 2 \ldots, D \tag{3}$$

where D means the number of channels of input feature map and the operation $\odot$ denotes dot product. The output features $f_{O,k}$ are used as the input for next stage of network.

Consequently, compared to the original convolutional features, the attention feature maps can be strongly enhanced by suppressing the background interference gradually.

Actually, the prediction of attention map is a pixel-wise classification task, therefore the pixel-level annotations are needed for training the attention block to generate accurate attention maps. Here we use the instance segmentation annotations to supervise the attention block during training, since most of images in PASCAL VOC [40] and every image in MS COCO have provided instance annotations. During inference, extra segmentation labels are not taken into account. In addition, we do not exploit the instance category information to produce any segmentation results and the predicted attention map is class-agnostic. Therefore, the attention map can be regarded as an indicator that explicitly guides the feature learning of base network and improve the final performance.

C. Local FCN

To effectively capture the specific fine-grained parts in local FCN, we construct a set of part-sensitive score maps by appending a 1×1 convolutional layer with $k^2(C+1)$ channels, where k means we divide the object into $k\times k$ local parts (here k is set to the default value 7) and $C+1$ is the number of object categories plus background. For each category, there are totally k^2 channels and each channel is responsible for encoding a specific part of the object. The final score of a category is determined by voting the k^2 responses. Here we use position-sensitive RoI pooling layer in [4] to extract object-specific parts and we simply perform average pooling for voting. Then, we obtain a $(C+1)$-d vector which indicates the probability that the object belongs to each class. This procedure is equivalent to dividing a strong object category decision into the sum of multiple weak classifiers, which serves as the ensemble of several part models. Here we refer this part ensemble as local structure representation.

As shown in Figure 4(a), for the truncated person, one can hardly get a strong response from the global description of the person due to truncation, on the contrary, our local FCN can effectively capture several specific parts, such as human nose, mouth, *etc.*, which correspond to the regions with large responses in the feature map. We argue that the local FCN is much concerned with the internal structure and components, which can effectively reflect the local properties of visual object, especially when the object is occluded or the whole boundary is incomplete. However, for those having simple spatial structure and encompassing considerable background in the bounding box, *e.g.* dining table, it is difficult to make robust predictions by only using local information. Thus it is necessary to add the global structure information to enhance the discrimination.

D. Global FCN

For the global FCN, we aim to describe the object by using the whole region-level features. Firstly, we attach a 1024-d 1×1 convolutional layer after the last convolutional block in ResNet-101 for reducing the dimension. Due to the diverse size of the object, we insert a RoI pooling layer in [2] to extract a fixed-length feature vector as the global structure description of the object. Secondly, we use two convolutional layers with kernal size $k\times k$ and 1×1 respectively (k is set to the default value 7) to further abstract the global representation of RoI. Finally, the output of 1×1 convolution is fed into the classifier whose output is also a $(C+1)$-d vector.

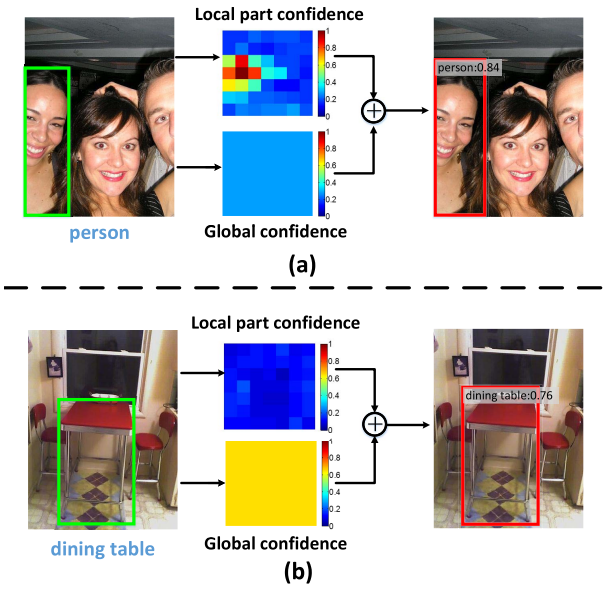

Fig. 4. An intuitive description of coupling structure for object detection. (a) It is difficult to determine the target by using the global structure information alone for objects with truncations. (b) Moreover, for those having simple spatial structure and encompassing considerable background in the bounding box, *e.g.* dining table, it is also not enough to use local parts alone to make robust predictions. Therefore, an intuitive idea is to simultaneously couple global structure with local parts to effectively boost the confidence. Best viewed in color.

In addition, context prior is the most basic and important factor for visual recognition tasks. For example, the boat usually travels in the water while is unlikely to fly in the sky. Despite the higher layers in deep neural network can involve the spatial context information around the objects due to the large receptive field, Zhou *et al.* [41] have shown that the practical receptive field is actually much smaller than the theoretical one. Therefore, it is necessary to explicitly collect the surrounding information to reduce the chance of misclassification. To enhance the feature representation ability of the global FCN, here we introduce the contextual information as an effective supplement. Specifically, we extend the context region by 2 times larger than the size of original proposal. The context region is adjusted to equal to the side of image when it is outside the bounds of image. Then the features RoI pooled from the original region and context region are concatenated together and fed into the latter RoI-wise subnetwork. As shown in Figure 2, the context region is embedded into the global branch to extract a more complete appearance structure and discriminative prior representation, which will help the classifier to better identity the object categories.

Due to the RoI pooling operation, the global FCN describes the proposal as a whole with CNN features, which can be

seen as a global structure description of the object. Therefore, it can easily deal with the objects with intact structure and finer scale. As shown in Figure 4(b), our global FCN shows a large confidence for the dining table. However, in most cases, natural scenes consist of considerable objects with occlusions or truncations, making the detection more difficult. Figure 4(a) shows that using the global structure information alone can hardly make a confident prediction for the truncated person. By adding local part structural supports, the detection performance can be significantly boosted. Therefore, it is essential to combine both local and global descriptions for a robust detection.

E. Coupling Structure

To match the same order of magnitude, we apply a normalization operation to the output of local and global FCN before they are combined together. We explored two different methods to perform normalization: an L2 normalization layer or a 1×1 convolutional layer to model the scale. Meanwhile, how to couple the local and global output is also a problem that needs to be explored. Here, we investigated three different coupling methods: element-wise sum, element-wise product and element-wise maximum. Our experiments show that using 1×1 convolution along with element-wise sum achieves the best performance and we will discuss it in Section IV-A.

With the coupling structure, ACoupleNet simultaneously exploits the local parts, global structure and context prior for object detection. The whole network benefits from approximate joint training and multi-task learning. We also note that the global branch can be regarded as a lightweight Faster R-CNN, in which all learnable parameters are from convolutional layers and the depth of RoI-wise subnetwork is only two. Therefore, the computational complexity is far less than the subnetwork in ResNet-based Faster R-CNN system whose depth is ten. As a consequence, our ACoupleNet can perform the inference efficiently, which runs slightly slower than R-FCN but much more faster than Faster R-CNN.

F. Loss Function

A multi-task loss function of binary classification, detection classification and bounding box regression is used to jointly optimize the model parameters:

$$L = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{j=1}^{N_k} L_{att}(a_j, a_j^*) + \lambda_1 \frac{1}{N} \sum_{i=1}^{N} L_{cls}(p_i, p_i^*) + \lambda_2 \frac{1}{N} \sum_{i=1}^{N} [p_i^* \geq 1] L_{reg}(t_i, t_i^*) \quad (4)$$

where K is the number of attention block (here K equals to 3), and N_k is the normalization factor which represents the area of feature map corresponding to the attention map in that block. The ground-truth label a_j^* is 1 if the pixel belongs to objects, 0 otherwise. a_j is the predicted classification confidence. $L_{att}(a_j, a_j^*)$ is the pixel-wise sigmoid cross-entropy loss.

The detection classification loss $L_{cls}(p_i, p_i^*)$ is softmax cross-entropy loss over $C+1$ object classes (add background). N is the number of training RoIs (128). p_i^* is the true category label, while $p_i = [p_{i,0}, \ldots, p_{i,C}]$ is the predicted probability distribution and $p_{i,c}$ is the estimated probability for the cth class. $t_i = [\Delta t_x^i, \Delta t_y^i, \Delta t_w^i, \Delta t_h^i]$ and $t_i^* = [\Delta t_x^{i*}, \Delta t_y^{i*}, \Delta t_w^{i*}, \Delta t_h^{i*}]$ are the predicted offsets and the true offsets for bounding box regression respectively. For L_{reg}, we use smoothed L_1 loss in [2]. λ_1 and λ_2 are hyper-parameters to balance the loss of classification and regression. As these two terms use the same normalization factor in Equation (4), we simply set $\lambda_1 = \lambda_2 = 1$ which makes the classification loss and regression loss equally weighted after normalization. With this joint loss function, bounding box regression, object category and all cascaded attention classifiers are learned jointly through backpropagation.

IV. Experiments

We train and evaluate our method on three challenging object detection datasets: PASCAL VOC2007, VOC2012 and MS COCO. Since all these three datasets contain a variety of circumstances, which can sufficiently verify the effectiveness of each module in our method. Note that due to lack of enough instance segmentation annotations in VOC dataset, all of the comparison experiments of attention block are performed on COCO dataset and we also report the final results on VOC test set by using the existing instance segmentation annotations [40] to train the model. This does not affect the overall integrity of our method, since COCO contains richer categories and more complicated scenarios. All of the following experimental results are obtained under a single test scale, that's the shorter side of the input image is resized to 600 pixels and the longer side is restricted to 1000 pixels. We also do not use any extra test tricks, such as multi-scale testing, horizontal flipping, box voting [34] and model ensemble. Finally we demonstrate state-of-the-art results on all three datasets.

A. Ablation Studies

We first perform experiments on PASCAL VOC 2007 with 20 object categories for detailed analysis of our proposed coupling structure. We train the models on the union set of VOC 2007 trainval and VOC 2012 trainval ("07+12") following [5], and evaluate on VOC 2007 test set. Object detection accuracy is measured by mean Average Precision (mAP). Then we conduct the attention experiments on MS COCO with 80 object categories and the evaluation metric is more rigorous than VOC. All the ablation experiments here use *single-scale* training.

1) Normalization: Since features extracted form different layers of CNN show various of scales, it is essential to normalize different features before coupling them together. Bell *et al.* [33] proposed to use L2 normalization to each RoI-pooled feature and re-scale back up by a empirical scale, which shows a great gain on VOC dataset. In this paper, we also explore two different normalization ways to normalize the output of local and global FCN: an L2 normalization layer or a 1×1 convolutional layer to learn the scale.

As shown in Table I, we find that the use of L2 normalization decreases the performance greatly, even worse than the

TABLE I

EFFECTS OF DIFFERENT NORMALIZATION OPERATION AND COUPLING METHODS. METRIC: DETECTION MAP(%) ON VOC07 TEST. ELTWISE: COMBINE THE OUTPUT FROM GLOBAL AND LOCAL FCN DIRECTLY. L2+ELTWISE: USE L2 NORMALIZATION TO NORMALIZE THE OUTPUT. 1×1 CONV+ELTWISE: USE 1×1 CONVOLUTION TO LEARN THE SCALE

Normalization methods	SUM	PROD	MAX
eltwise	**81.1**	-	80.7
L2+eltwise	**80.3**	63.5	78.2
1x1 conv+eltwise	**81.7**	-	81.3

direct addition (without any normalization ways). To explain such a phenomenon, we measured the outputs of two branches before and after L2 normalization. We further found that L2 normalization reduces the output gap between different categories, which results in a smaller score gap. As we know, a small score gap between different categories always means the classifier can not make a confident prediction. Therefore, we assume that this is the reason for the performance degradation. Moreover, we also exploit a 1×1 convolution to adaptively learn the scales between the global and local branches. Table I shows that using 1×1 convolution increases by 0.6 points compared to the direct addition and 2.2 points over R-FCN. Therefore, we use 1×1 convolution to replace the L2 normalization in the following experiments.

2) Coupling Strategy: We explore three different response coupling strategies: element-wise sum, element-wise product and element-wise maximum. Table I shows the comparison results for the above three different implementations. We can see that the element-wise sum always achieves the best performance even though in different normalization methods. Generally, current advanced residual networks [21] also use element-wise sum as the effective way to integrate information from previous layers, which greatly facilitates the circulation of information and achieves the complementary advantages. For element-wise product, we argue that the system is relatively unstable and is susceptible to the weak side, which results in a large gradient to update the weak branch that makes it difficult to converge. For element-wise maximum, it equals to an ensemble model within the network to some extent, which losts the advantages of mutual support compared to element-wise sum when both two branches are failed to detect the object. Moreover, a better coupling strategy can be taken into consideration as the future work to further improve the accuracy, such as designing a more subtle nonlinear structure to learn the coupling relationship.

3) Model Ensemble: Model ensemble is commonly used to improve the final detection performance, since diverse initialization of parameters and the randomness of training samples both lead to different performance for the same model. Although the differences and complementarities will be more pronounced for different models, the promotion is often very limited. As shown in Table II, we also compare

TABLE II

COUPLENET *vs.* MODEL ENSEMBLE. ALL MODELS ARE EVALUATED ON VOC 07. *ReIm*: OUR REIMPLEMENTATION USING OHEM. GLOBAL FCN: ONLY THE GLOBAL BRANCH OF OUR NETWORK. *w/o Context*: WITHOUT ADDING THE CONTEXT PRIOR TO ASSIST THE GLOBAL BRANCH

Method	mAP(%)
Faster-*ReIm*	79.0
R-FCN-*ReIm*	78.6
Global FCN	78.5
Faster&R-FCN ensemble	79.6
Global FCN&R-FCN ensemble	79.4
CoupleNet *[w/o context]*	**81.7**

TABLE III

CONTEXT PRIOR. ALL MODELS ARE EVALUATED ON VOC 07. LOCAL+CONTEXT: ADDING CONTEXT PRIOR TO PSROI POOLING LAYER. GLOBAL+CONTEXT: ADDING CONTEXT PRIOR TO ROI POOLING LAYER

Method	CoupleNet		
w/o context	√		
local+context		√	
global+context			√
mAP(%)	81.7	81.5	**82.1**

our coupling structure with the model ensemble. For a fair comparison, we first re-implemented Faster R-CNN [21] using ResNet-101 and online hard example mining (OHEM) [42], which achieves a mAP of 79.0% on VOC07 (76.4% in original paper without OHEM). We also re-implemented R-FCN with approximate joint training using the public available code py-R-FCN,[1] which achieves a slightly lower result compared to [4] (78.6% vs. 79.5%). The main difference between our re-implementation and the official implementation [4] is that we use joint training for convenience while [4] uses existing well-trained RPN to provide proposals for training. In addition, the results are also subject to some random variations, such as the initialization parameters for new layers and the training order of images. These may be the reasons for degradation in performance. Therefore, we use our re-implementation models to conduct the comparisons for consistency. We found that the promotion brought by model ensemble is less than 1 point. As shown in Table II, it is far less than our method (81.7%).

On the one hand, we argue that the naive model ensemble just combines the results together and does not essentially guide the learning process of the network, while our coupling structure can simultaneously utilize the global and local information to update the network and to infer the final results. On the other hand, our method enjoys approximate joint training and there is no need to train multiple models, thus greatly reducing the training time.

4) Context Prior: By adding the context prior to PSRoI pooling instead of RoI Pooling, our method achieves a mAP of 81.5% in VOC07, 0.6 points lower than the global/context as shown in Table III, which indicates that context information

[1] https://github.com/Orpine/py-R-FCN

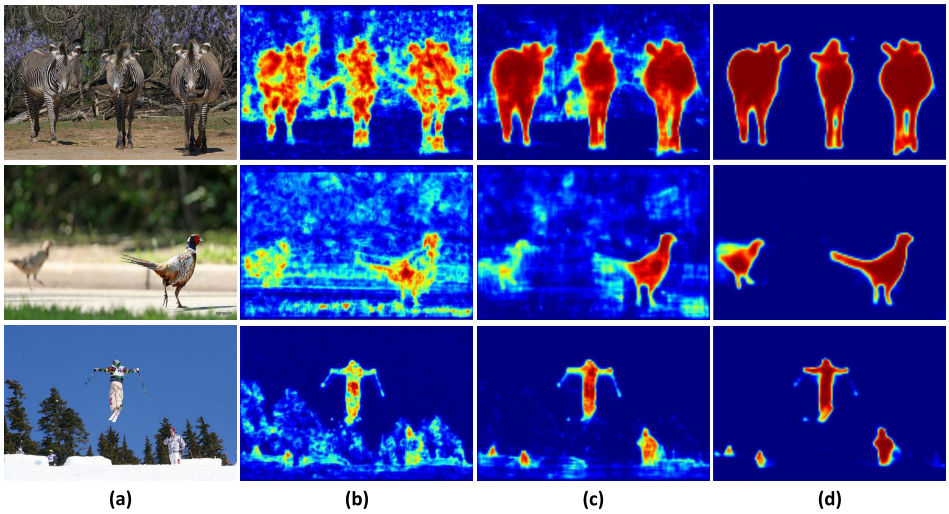

Fig. 5. Some qualitative results generated by our cascade attention FCN on COCO *minival* dataset. (a) the original image. (b) attention map estimated by the first attention block, where stride is 4. (c) attention map estimated by the second attention block, where stride is 8. (d) attention map estimated by the third attention block, where stride is 16. All the figures are zoomed to the same size for display. Best viewed in color.

TABLE IV

THE IMPACT OF THE AMOUNT OF PARAMETERS. ALL MODELS ARE EVALUATED ON VOC 07. SINCE ALL NEW BRANCHES OF OUR METHOD INTRODUCE MORE PARAMETERS, WE ISOLATE THEIR IMPACT BY ADDING A FEW MORE LAYERS TO THE PREDICTION HEAD ON BOTH TWO STANDARD DETECTORS. 3× PSROI POOLING: 3 POSITIVE-SENSITIVE ROI POOLING LAYERS

Method	mAP(%)
R-FCN-*ReIm*	78.6
+ one head	78.8
+ two head	78.7
Global FCN	78.5
+ one head	79.3
+ two head	79.0
CoupleNet	**82.1**
+ 3x PSRoI Pooling	79.7

TABLE V

RESULTS OF ATTENTION BLOCK COMPARISON EXPERIMENTS ON COCO *minival5k* SET. MMAP: THE MAP AVERAGED AT $IoU \in [0.5 : 0.05 : 0.95]$. ATTENTION-S4/8/16: THE ATTENTION BLOCK ADDED AT STRIDE 4, 8 AND 16

Component	CoupleNet			
attention-s4	√	√	√	
attention-s8	√	√		
attention-s16	√			
mmAP(%)	**29.9**	29.7	29.5	29.1

brings more assistance in global structure. The reason is that the purpose of context information in our approach is to assist the global branch (RoI) to extract a more complete appearance structure, while the local branch (PSRoI) pays more attention to the internal structure and object-specific parts.

5) Amount of Parameters: Since the CoupleNet [15] introduces a few more parameters compared with the single branch detectors, to further verify effectiveness of the coupling structure, here we increase the parameters of the prediction head for each single branch implementation to maintain the same amount of parameters with CoupleNet for comparison. In detail, we add a new residual variant block with three convolution layers, where the kernel size is $1\times1\times256$, $3\times3\times256$ and $1\times1\times1024$ respectively, to the prediction sub-network. As shown in Table IV, we found that the standard R-FCN with one or two extra heads got a mAP of 78.8% and 78.7% respectively in VOC07, which is slightly higher than our re-implemented version (78.6%) in [4]. Meanwhile, our global FCN, which performs the RoI pooling on top of conv5, got a relative higher gain (a mAP of 79.3% for one head, 79.0% for two heads).

The results indicate that simply adding more prediction layers obtains a very limited performance gain, while our coupling structure shows more discriminative power with the same amount of parameters. By using 3× PSRoI pooling layers, instead of our coupling structure, we got a mAP of 79.7% in VOC07, which is 2.4 points lower than our method (82.1%). Also, introducing more parameters to the network, the 3× PSRoI pooling structure does not effectively capture different structure information of the object, which also shows limited performance gain.

6) Attention Block: Some of our predicted attention maps can be found in Figure 5. We can see that the attention map becomes more and more clear with the gradual joining of attention block. We also quantitatively evaluate our cascade attention block on COCO dataset, since COCO contains instance segmentation annotations for each image. In order to accelerate the training speed, we train our model on *val35k* and evaluate it on *minival5k*. For the final model, we train it on *trainval* and evaluate on *test-dev*. The *val35k* includes 35504 images, we use horizontal image flipping as

TABLE VI

COMPARISONS WITH FASTER R-CNN AND R-FCN USING RESNET-101. 128 SAMPLES ARE USED FOR BACKPROPAGATION AND THE TOP 300 PROPOSALS ARE SELECTED FOR TESTING FOLLOWING [4]. THE INPUT RESOLUTION IS ABOUT 600 × 1000. WE ALSO NOTE THAT THE TITAN X USED HERE IS THE PASCAL ARCHITECTURE ALONG WITH CUDA 8.0 AND CUDNN-V5.1. "07+12": VOC07 TRAINVAL UNION WITH VOC12 TRAINVAL. "07+12+S": VOC07 TRAINVAL UNION WITH VOC12 TRAINVAL PLUS INSTANCE SEGMENTATION LABELS [40]. *w/o context*: WITHOUT ADDING THE CONTEXT PRIOR TO ASSIST THE GLOBAL BRANCH

	training data	training proposals	test proposals	mAP (%) on VOC07	GPU	test time (ms/img)
Faster R-CNN [21]	07+12	-	300	76.4	K40	420
R-FCN [4]	07+12	128	300	79.5	TITAN X	83
R-FCN *multi-sc train* [4]	07+12	128	300	80.5	TITAN X	83
CoupleNet *[w/o context]* [15]	07+12	128	300	81.7	TITAN X	102
CoupleNet [15]	07+12	128	300	82.1	TITAN X	122
CoupleNet *multi-sc train* [15]	07+12	128	300	**82.7**	TITAN X	122
ACoupleNet	07+12+S	128	300	82.6	TITAN X	145
ACoupleNet *multi-sc train*	07+12+S	128	300	**83.1**	TITAN X	145

TABLE VII

RESULTS ON PASCAL VOC 2007 TEST SET. THE FIRST FOUR METHODS USE VGG16 AND THE LATTER FOUR USE RESNET-101 AS THE BASE NETWORK. FOR FAIR COMPARISON, WE ONLY LIST THE RESULTS OF SINGLE MODEL WITHOUT MULTI-SCALE TESTING, ENSEMBLE OR ITERATIVE BOX REGRESSION TRICKS IN TESTING PHASE. "07+12": VOC07 TRAINVAL UNION WITH VOC12 TRAINVAL. "07+12+S": VOC07 TRAINVAL UNION WITH VOC12 TRAINVAL PLUS INSTANCE SEGMENTATION LABELS [40]. *: THE RESULTS ARE UPDATED USING THE LATEST MODELS. §: THIS ENTRY IS DIRECTLY OBTAINED FROM [21] WITHOUT USING OHEM

Method	Train	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
ION [33]	07+12+S	76.5	79.2	79.2	77.4	69.8	55.7	85.2	84.2	**89.8**	57.5	78.5	73.8	87.8	85.9	81.3	75.3	49.7	76.9	74.6	85.2	82.1
HyperNet [43]	07+12	76.3	77.4	83.3	75.0	69.1	62.4	83.1	87.4	87.4	57.1	79.8	71.4	85.1	85.1	80.0	79.1	51.2	79.1	75.7	80.9	76.5
SSD300* [25]	07+12	77.5	79.5	83.9	76.0	69.6	50.5	87.0	85.7	88.1	60.3	81.5	**77.0**	86.1	87.5	83.9	79.4	52.3	77.9	79.5	87.6	76.8
SSD512* [25]	07+12	79.5	84.8	85.1	81.5	73.0	57.8	87.8	88.3	87.4	63.5	85.4	73.2	86.2	86.7	83.9	82.5	55.6	81.7	79.0	86.6	80.0
Faster§ [21]	07+12	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	**89.8**	56.7	87.8	69.4	88.3	88.9	80.9	78.4	41.7	78.6	79.8	85.3	72.0
R-FCN [4]	07+12	80.5	79.9	87.2	81.5	72.0	69.8	86.8	88.5	89.8	67.0	**88.1**	74.5	**89.8**	**90.6**	79.9	81.2	53.7	81.8	81.5	85.9	79.9
CoupleNet [15]	07+12	82.7	85.7	87.0	84.8	75.5	73.3	88.8	**89.2**	89.6	**69.8**	87.5	76.1	88.9	89.0	**87.2**	86.2	59.1	83.6	**83.4**	**87.6**	80.7
ACoupleNet	07+12+S	**83.1**	**86.2**	**87.3**	**86.5**	**75.7**	**74.9**	**88.9**	89.1	89.3	**69.8**	**88.1**	75.9	88.4	89.3	87.0	**87.1**	**60.6**	**85.7**	82.9	**87.6**	**82.4**

the only form of data augmentation. In this part, we apply synchronized SGD over 4 GPUs with total of 8 images per minibatch (2 images per GPU). We train the models for 12 epochs with an initial learning rate of 0.001, which is then divided by 10 at 10 epochs. Weight decay of 0.0005 and momentum of 0.9 are used.

The attention block is plugged into the backbone network repeatedly, where the stride is 4, 8 and 16. We compare the baseline with and without attention structure. As shown in Table V, the baseline (CoupleNet) achieves 29.1 mmAP, which reaches the performance of R-FCN trained on *trainval* set (also 29.1% mmAP). While introducing the attention mechanism, ACoupleNet gets better results. Specifically, the mmAP was improved by 0.8 points when adding three attention blocks, which shows the cascade attention module helps to improve the performance by generating discriminative features and diminishing the regions without object.

B. Results on VOC2007

Using the public available ResNet-101 as the initialization model, we note that our method is easy to follow and the hyper-parameters for training are the same as in [4]. Similarly, we use the dilation strategy to reduce the effective stride of ResNet-101, just as [4] shows, thus both the global and local branches have a stride of 16. We also use a 1-GPU implementation, and the effective mini-batch size is 2 images by setting the $iter\_size$ to 2. The whole network is trained for 80k iterations with a learning rate of 0.001 and then for 30k iterations with a learning rate of 0.0001. In addition, the context prior is proposed to further boost the performance while keeping the iterations unchanged. Finally, we also perform multi-scale training with the shorter sides of images are randomly resized from 480 to 864.

Table VI shows the detailed comparisons with Faster R-CNN and R-FCN. As we can see that the single CoupleNet model has already achieved a mAP of 82.1%, which outperforms the R-FCN by 2.6 points. However, our ACoupleNet further improves mAP by 0.5% and rises up to 82.6%, which is the current best single model detector to our knowledge. Moreover, we also evaluate the inference time of our network using a NVIDIA TITAN X GPU (Pascal) along with CUDA 8.0 and cuDNN-v5.1. As shown in the last column of Table VI, our method is slightly slower than R-FCN, which also reaches a real-time speed (*i.e.* 8.2 fps for CoupleNet or 6.9 fps for ACoupleNet) and achieves the best trade-off between accuracy and speed. We argue that the sharing process of feature extraction between two branches and the design of lightweight RoI-wise subnetwork after RoI pooling both greatly reduce the model complexity.

As shown in Table VII, we also compared our method with other state-of-the-art single model. We found that our method outperforms the others with a large margin, including

TABLE VIII

RESULTS ON PASCAL VOC 2012 TEST SET. FOR FAIR COMPARISON, WE ONLY LIST THE RESULTS OF SINGLE MODEL WITHOUT MULTI-SCALE TESTING, ENSEMBLE OR ITERATIVE BOX REGRESSION TRICKS IN TESTING PHASE. THE INPUT RESOLUTION IS ABOUT 600 × 1000. "07++12": THE UNION SET OF VOC07 TRAINVAL+TEST AND VOC12 TRAINVAL. "07+12+S": VOC07 TRAINVAL UNION WITH VOC12 TRAINVAL PLUS INSTANCE SEGMENTATION LABELS [40]. *: RESULTS ARE UPDATED USING THE LATEST MODELS. §: THIS ENTRY IS DIRECTLY OBTAINED FROM [21] WITHOUT USING OHEM. †: *http://host.robots.ox.ac.uk:8080/anonymous/M5CQTL.html*. ‡: *http://host.robots.ox.ac.uk:8080/anonymous/AGEXBF.html*

Method	Train	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
ION [33]	07+12+S	76.4	87.5	84.7	76.8	63.8	58.3	82.6	79.0	90.9	57.8	82.0	64.7	88.9	86.5	84.7	82.3	51.4	78.2	69.2	85.2	73.5
HyperNet [43]	07++12	71.4	84.2	78.5	73.6	55.6	53.7	78.7	79.8	87.7	49.6	74.9	52.1	86.0	81.7	83.3	81.8	48.6	73.5	59.4	79.9	65.7
SSD300* [25]	07++12	75.8	88.1	82.9	74.4	61.9	47.6	82.7	78.8	91.5	58.1	80.0	64.1	89.4	85.7	85.5	82.6	50.2	79.8	**73.6**	86.6	72.1
SSD512* [25]	07++12	78.5	**90.0**	85.3	77.7	64.3	58.5	**85.1**	84.3	92.6	61.3	83.4	65.1	89.9	88.5	**88.2**	85.5	54.4	82.4	70.7	87.1	75.6
Faster§ [21]	07++12	73.8	86.5	81.6	77.2	58.0	51.0	78.6	76.6	93.2	48.6	80.4	59.0	92.1	85.3	84.8	80.7	48.1	77.3	66.5	84.7	65.6
R-FCN [4]	07++12	77.6	86.9	83.4	81.5	63.8	62.4	81.6	81.1	93.1	58.0	83.8	60.8	92.7	86.0	84.6	84.4	59.0	80.8	68.6	86.1	72.9
CoupleNet [15]	07++12	80.4†	89.1	86.7	81.6	71.0	64.4	83.7	83.7	94.0	**82.2**	84.6	**65.6**	92.7	**89.1**	87.3	87.7	64.3	84.1	72.5	88.4	75.3
ACoupleNet	07++12+S	**81.0**‡	**90.0**	**87.4**	**83.3**	**72.7**	**66.8**	84.2	**85.4**	**94.1**	62.0	**85.3**	65.2	**92.8**	**89.1**	87.7	**88.7**	**65.3**	**84.4**	71.0	**89.0**	**76.3**

the advanced SSD method [25], which requires complicated data augmentation and careful training skills. Just as discussed earlier, compared to Faster R-CNN and R-FCN, CoupleNet and ACoupleNet shows a large gain over the classes with occlusions, truncations and considerable background information, like sofa, person, table and chair, which verifies our analyses. We also observed a large improvement for airplane, bird, boat and potted plant, which usually have class-specific backgrounds, *i.e.* the sky for airplane and bird, water for boat and so on. Therefore, the context surrounding the objects also provides an extra auxiliary discrimination.

C. Results on VOC2012

We also evaluate our method on the more challenging VOC2012 dataset by submitting results to the public evaluation server. In this part, we only use VOC07 trainval, VOC07 test and VOC12 trainval as the training set, which consists of 21k images in total. We also follow the similar hyper-parameter settings in VOC07 but change the iterations, since there are more training images. We train our models with 4 GPUs, and the effective mini-batch size thus becomes 4 (1 per GPU). As a result, the network is trained for 60k iterations with a learning rate of 0.001 and 0.0001 for the following 20k iterations. Table VIII shows the results on the VOC2012 test set. Our CoupleNet has already obtained a top mAP of 80.4%, which is 2.8 points higher than R-FCN. By introducing the attention block to provide the object prior, our ACoupleNet improves the accuracy up to 81.0%. To our best knowledge, without using extra data, our single model ACoupleNet achieves the best accuracy among all methods. Similar promotions over the specific classes analysed in VOC07 are also observed, which once again validates the effectiveness of our method. Figure 6 shows some detection examples of ACoupleNet on VOC 2012 test set.

D. Results on MS COCO

Next we present more results on the MS COCO object detection dataset. The dataset consists of 80k training set, 40k validation set and 20k test-dev set, which involves 80 object categories. All our models are trained on the union set of 80k training set and 40k validation set, and evaluated on 20k test-dev set. The COCO standard metric denotes as AP, which is evaluated at $IoU \in [0.5 : 0.05 : 0.95]$. Here we also implement our method with and without attention structure. For the case of without attention, following the VOC2012, a 4-GPU implementation (batch size is 4) is used to train the model. We use an initial learning rate of 0.001 for the first 510k iterations and 0.0001 for the next 70k iterations. While adding the cascade attention block, we use 8 GPUs with a total of 16 images per mini-batch to accelerating the training process. The initial learning rate is set to 0.001 for the first 150k iterations and then dropped by 10 for the next 30k iterations. Weight decay of 0.0005 and momentum of 0.9 are used for both cases. In addition, we also conduct multi-scale training with the scales are randomly sampled from $\{480, 576, 600, 672, 768, 864\}$ while testing in a single scale.

Table IX shows our results. Our single-scale trained detector without attention has already achieved a result of 33.1%, which outperforms the R-FCN by 3.9 points. In addition, the multi-scale training improves the performance up to 34.4%. Interestingly, we observed that the more challenging the dataset, the more the promotion (*e.g.*, 2.2% for VOC07, 2.8% for VOC12 and 4.5% for COCO, all in multi-scale training), which directly proves that our approach can effectively cope with a variety of complex situations. Moreover, the accuracy can be further improved by 1 point with the addition of attention block and the best performance of single model is up to 35.4%.

In Figure 7, we also exhibit some detection examples on COCO test-dev set to qualitatively compare the differences between CoupleNet and ACoupleNet. As the introduction of attention structure, ACoupleNet improves the results in two cases. The first is the elimination of false positives as shown in the first two rows of Figure 7. Due to the strong similarity between the objects and some backgrounds, CoupleNet may be fooled under this situation, but ACoupleNet does work well. The second is the remediation of false negatives as shown in the last two rows of Figure 7. Attention not only serves to suppress the confusing backgrounds but also guides the feature

TABLE IX

RESULTS ON COCO 2015 TEST-DEV. THE COCO METRIC AP IS EVALUATED AT IOU THRESHOLDS RANGING FROM 0.5 TO 0.95. AP@0.5: PASCAL-TYPE METRIC, IOU=0.5. AP@0.75: EVALUATE AT IOU=0.75. "TRAIN+S": TRAIN SET PLUS INSTANCE SEGMENTATION LABELS. *mst Train*: MULTI-SCALE TRAINING. THE INPUT RESOLUTION IS ABOUT 600 × 1000

Method	train data	AP	AP @0.5	AP @0.75	AP small	AP medium	AP large	AR max=1	AR max=10	AR max=100	AR small	AR medium	AR large
SSD300* [25]	trainval35k	25.1	43.1	25.8	6.6	25.9	41.4	23.7	35.1	37.2	11.2	40.4	58.4
SSD512* [25]	trainval35k	28.8	48.5	30.3	10.9	31.8	43.5	26.1	39.5	42.0	16.5	46.6	60.8
ION [33]	train+S	24.9	44.7	25.3	7.0	26.1	40.1	23.9	33.5	34.1	10.7	38.8	54.1
Faster+++ [21]	trainval	**34.9**	**55.7**	-	15.6	38.7	50.9	-	-	-	-	-	-
R-FCN [4]	trainval	29.2	51.5	-	10.3	32.4	43.3	-	-	-	-	-	-
R-FCN *multi-sc train* [4]	trainval	29.9	51.9	-	10.8	32.8	45.0	-	-	-	-	-	-
CoupleNet [15]	trainval	33.1	53.5	35.4	11.6	36.3	50.1	29.3	43.8	45.2	18.7	51.4	67.9
CoupleNet *msc train* [15]	trainval	**34.4**	**54.8**	37.2	13.4	38.1	50.8	30.0	45.0	46.4	20.7	53.1	68.5
ACoupleNet	trainval+S	34.1	54.1	36.5	12.2	37.2	51.8	30.1	44.3	45.7	20.3	50.5	67.7
ACoupleNet *mst train*	trainval+S	**35.4**	**55.7**	37.6	13.2	38.6	52.5	30.3	46.1	48.2	21.9	53.6	70.2

Fig. 6. **Detection examples of ACoupleNet on PASCAL VOC 2012 test set.** The model was initialized from the MS COCO (corresponding to 35.4% mmAP) and fine-tuned on the union of VOC07 trainval+test and VOC12 trainval (corresponding to 84.3% mAP). A score threshold of 0.6 is used to draw the detection bounding boxes. Each color is related to an object category.

learning and enhances the representation of the object. Therefore, ACoupleNet can obtain more stable detection results.

E. From MS COCO to PASCAL VOC

Since the categories on MS COCO contains those on PASCAL VOC, we study how the MS COCO dataset helps the detection accuracy on PASCAL VOC by directly fine-tuning the detection model pre-trained on MS COCO. More specifically, both VOC 2007 and 2012 detectors are initialized from the COCO model with a mmAP of 35.4%. For model evaluated in VOC 2007 test, we use "07+12" (VOC07 trainval union with VOC12 trainval) as the training set, and for VOC 2012 test, "07++12" (the union set of VOC07 trainval+test and VOC12 trainval) is used for training. We randomly sample the scale from {480, 576, 600, 672, 768, 864} pixels, then resize the shorter side of the image into the sampled scale for training. During the inference, we still test a single

TABLE X

DETECTION MAP (%) ON PASCAL VOC DATASET. ALL MODELS ARE FINE-TUNED FROM COCO PRE-TRAINED MODEL.‡: *http://host.robots.ox.ac.uk:8080/anonymous/8M1JR6.html*

Method	model	2007 test	2012 test
Faster R-CNN [5]	VGG-16	78.8	75.9
R-FCN [4]	ResNet-101	83.6	82.0
SSD300 [25]	VGG-16	81.2	79.3
SSD512 [25]	VGG-16	83.2	82.2
ACoupleNet	ResNet-101	**85.7**	**84.3**‡

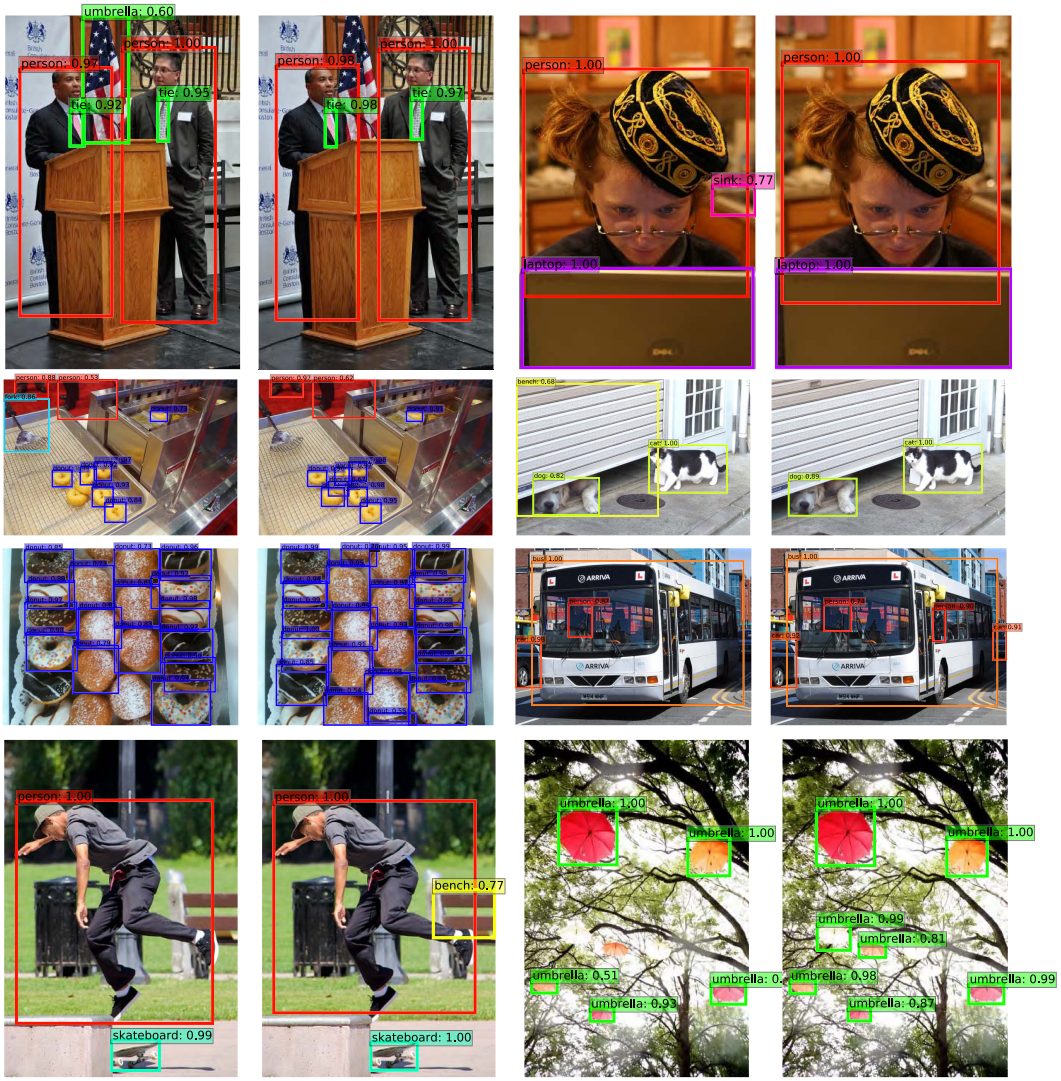

Fig. 7. **Detection examples on MS COCO test-dev set with CoupleNet and ACoupleNet.** For each pair of images, the left side is the result of CoupleNet model (corresponding to 34.4% mmAP) and the right side is the result of ACoupleNet model (corresponding to 35.4% mmAP). The first two rows indicate that ACoupleNet can eliminate some confusing false positives. The last two rows indicate that ACoupleNet can also help to detect more objects especially for dense cases. A score threshold of 0.5 is used to draw the detection bounding boxes.

scale of 600 pixels (the longer is constrained to 1000 pixels). As shown in Table X, our ACoupleNet obtains the top mAP scores of 85.7% and 84.3% respectively on both VOC 2007 and VOC 2012 test set. Last but not least, compared with the methods displayed on VOC 2012 Leaderboard, whose results are achieved by applying multi-scale testing, horizontal flipping and bounding box voting during inference, our single-model and single-scale result is still competitive.

V. Conclusion

In this paper, we present the Attention CoupleNet (ACoupleNet), a concise yet effective network that automatically learns the informative regions from the image and simultaneously couples global, local and context cues of objects for accurate object detection. Our system effectively exploits the cascade attention structure to gradually focus on the target regions by generating several class-agnostic attention feature maps. Based on the attention features, a novel coupling structure along with different coupling strategies is proposed to extract local part representation, global structural information and the context prior of target regions to cooperatively detect objects. A multi-task loss is utilized to jointly optimize the whole network. To validate our design decisions, we performed extensive experiments to evaluate choices like attention block, coupling strategies, amount of parameters and other variations. Experimental results show that the proposed ACoupleNet achieves state-of-the-art performance on the PASCAL VOC and COCO datasets. Currently, the ACoupleNet just divides the local parts and couple the local/global information in a naive way, the position relationship between local parts and the nonlinear coupling strategies will be investigated in a future work.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[2] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.
[3] 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.
[4] 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.*, 2016, pp. 379–387.
[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[6] 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.
[7] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.
[9] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.
[10] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.
[11] J. Gu, J. Zhou, and C. Yang, "Fingerprint recognition by combining global structure and local cues," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1952–1964, Jul. 2006.
[12] W.-Z. Nie, A.-A. Liu, Z. Gao, and Y.-T. Su, "Clique-graph matching by preserving global & local structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4503–4510.
[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[14] 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.
[15] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu, "Couplenet: Coupling global structure with local parts for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4146–4154.
[16] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.
[17] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[18] J. Wang, W. Fu, H. Lu, and S. Ma, "Bilayer sparse topic model for scene analysis in imbalanced surveillance videos," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5198–5208, Dec. 2014.
[19] J. Wang, L. Duan, Q. Liu, H. Lu, and J. S. Jin, "A multimodal scheme for program segmentation and representation in broadcast video streams," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 393–408, Apr. 2008.
[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[21] 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.
[22] H. Guo, J. Wang, Y. Gao, J. Li, and H. Lu, "Multi-view 3D object retrieval with deep embedding network," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5526–5537, Dec. 2016.
[23] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 346–361.
[25] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[26] 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.
[27] Y. Zhu, J. Wang, C. Zhao, H. Guo, and H. Lu, "Scale-adaptive deconvolutional regression network for pedestrian detection," in *Proc. ACCV*. Cham, Switzerland: Springer, 2016, pp. 416–430.
[28] W. Chu, Y. Liu, C. Shen, D. Cai, and X.-S. Hua, "Multi-task vehicle detection with region-of-interest voting," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 432–441, Jan. 2018.
[29] L. Huang, Y. Yang, Y. Deng, and Y. Yu. (2015). "DenseBox: Unifying landmark localization with end to end object detection." [Online]. Available: https://arxiv.org/abs/1509.04874
[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3007.
[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[32] J. Wang, W. Fu, J. Liu, and H. Lu, "Spatiotemporal group context for pedestrian counting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1620–1630, Sep. 2014.
[33] 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.
[34] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1134–1142.
[35] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 354–370.
[36] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.
[37] S. Sharma, R. Kiros, and R. Salakhutdinov. (2016). "Action recognition using visual attention." [Online]. Available: https://arxiv.org/abs/1511.04119
[38] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.
[39] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin. (2016). "Fully convolutional attention networks for fine-grained recognition." [Online]. Available: https://arxiv.org/abs/1603.06765
[40] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 991–998.
[41] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," *CoRR*, vol. abs/1412.6856, 2014.
[42] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.
[43] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 845–853.

Yousong Zhu received the B.E. degree from Central South University, Changsha, China, in 2014. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence systems with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include object detection, video object detection, pattern recognition and machine learning, and intelligent video surveillance.

Chaoyang Zhao received the B.E. and M.S. degrees from the University of Electronic Science and Technology of China in 2009 and 2012, respectively, and the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2016. He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include object detection, image and video processing, and intelligent video surveillance.

Haiyun Guo received the B.E. degree from Wuhan University in 2013 and the Ph.D. degree in pattern recognition and intelligence systems from the Institute of Automation, University of Chinese Academy of Sciences, in 2018. She is currently an Assistant Researcher with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her current research interests include pattern recognition and machine learning, image and video processing, and intelligent video surveillance.

Jinqiao Wang received the B.E. degree from the Hebei University of Technology, China, in 2001, the M.S. degree from Tianjin University, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2008. He is currently a Professor with the Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.

Xu Zhao received the B.E. degree from the Dalian University of Technology of China in 2014. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence systems with the National Laboratory of Pattern Recognition, Chinese Academy of Sciences. His research interests include object detection, scene text detection, image and video processing, and intelligent video surveillance.

Hanqing Lu received the B.E. and M.E. degrees from the Harbin Institute of Technology, in 1982 and 1985, respectively, and the Ph.D. degree from the Huazhong University of Sciences and Technology in 1992. He is currently the Deputy Director of the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include image and video analysis, medical image processing, object recognition, and so on.