CoupleNet: Coupling Global Structure with Local Parts for Object Detection

Yousong Zhu[1,2]   Chaoyang Zhao[1,2]   Jinqiao Wang[1,2]   Xu Zhao[1,2]   Yi Wu[3,4]   Hanqing Lu[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences
[3]School of Technology, Nanjing Audit University, Nanjing, China
[4]Department of Medicine, Indiana University, Indianapolis, USA

{yousong.zhu, chaoyang.zhao, jqwang, xu.zhao, luhq}@nlpr.ia.ac.cn, ywu.china@gmail.com

Abstract

The region-based Convolutional Neural Network (CNN) detectors such as Faster R-CNN or R-FCN have already shown promising results for object detection by combining the region proposal subnetwork and the classification subnetwork together. Although R-FCN has achieved higher detection speed while keeping the detection performance, the global structure information is ignored by the position-sensitive score maps. To fully explore the local and global properties, in this paper, we propose a novel fully convolutional network, named as CoupleNet, to couple the global structure with local parts for object detection. Specifically, the object proposals obtained by the Region Proposal Network (RPN) are fed into the the coupling module which consists of two branches. One branch adopts the position-sensitive RoI (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. Next, we design different coupling strategies and normalization ways to make full use of the complementary advantages between the global and local branches. Extensive experiments demonstrate the effectiveness of our approach. We achieve state-of-the-art results on all three challenging datasets, i.e. *a mAP of 82.7% on VOC07, 80.4% on VOC12, and 34.4% on COCO. Codes will be made publicly available*[1].

1. Introduction

General object detection requires to accurately 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 large inter-class appearance differences. The variations arise not only from changes in a va-
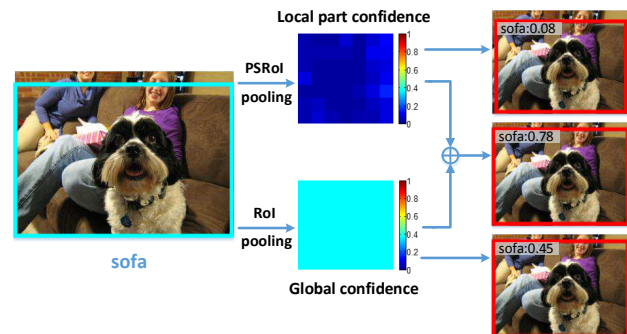
---

[1]https://github.com/tshizys/CoupleNet

Figure 1. A toy example of object detection by combing local and global information. Only considering the local part information or global structure leads to low confidence score. By coupling the two kinds of information together, we can detect the sofa accurately with a confidence score of 0.78. Best viewed in color.

riety of non-rigid deformations, but also due to the truncations, occlusions and inter-class interference. However, no matter how complicated the objects are, when humans identify a target, 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 that fused both global and local information.

With the revival of Convolutional Neural Networks [15] (CNN), CNN-based object detection pipelines [8, 9, 16, 21] have been proposed consecutively and made impressive improvements in generic benchmarks, *e.g.* PASCAL VOC [5] and MS COCO [17]. As two representative region-based CNN approaches, Fast/Faster R-CNN [8, 21] uses a certain subnetwork to predict the category of each region proposal while R-FCN [16] 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 0.08 for the sofa detection since the local responses of sofa are disturbed by a women and a dog (they are also the categories that need to be detected). Conversely, the global structure of sofa could be extracted by the RoI pooling, but the confidence score is 0.45, which is also very low for the incomplete structure of sofa. By coupling the global confidence with the local part confidence together, we can obtain a more reliable prediction with the confidence score of 0.78.

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*. [10] combined the global orientation field and local minutiae cue to largely improve the performance. In clique-graph matching, Nie *et al*. [19] 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*. [27] 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*. [6] 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.

Therefore, to fully explore the global and local clues, in this paper, we propose a novel full convolutional network named as CoupleNet, to couple the global structure and local parts to boost the detection accuracy. Specifically, 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 have a more powerful representation capacity and generalization ability. Extensive experiments demonstrate that CoupleNet can significantly improve the detection performance. Our detector shows 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. We propose a unified fully convolutional network to jointly learn the local, global and context information for object detection.

2. We design different normalization methods and cou-

pling strategies to mine the compatibility and complementarity between the global and local branches.

3. We achieve the state-of-the-art results on all three challenging datasets, *i.e.* a mAP of 82.7% on VOC07, 80.4% on VOC12, and 34.4% on MS COCO.

2. Related work

Before the arrival of CNN, visual tasks have been dominated by traditional paradigms [4, 6, 23, 26, 24]. As one of an outstanding framework, DPM [6] 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 [3], 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.

In order to leverage the great success of deep neural networks for image classification [12, 15], considerable object detection methods based on deep learning have been proposed [9, 11, 18, 20, 29]. Although there are end-to-end detection frameworks, like SSD [18], YOLO [20] and DenseBox [13], region-based systems (*i.e.* Fast/Faster R-CNN [8, 21] and R-FCN [16]) still dominate the detection accuracy on generic benchmarks [5, 17].

Compared to the end-to-end framework, the region-based systems have several advantages. Firstly, by exploiting a divide-and-conquer strategy, the two-step framework is more stable and easier to converge. Secondly, 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 [8, 16, 21] 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 [8, 21] or a position-sensitive RoI pooling layer [16] is used to achieve the translation variance in region-based systems. However, all the existing region-based systems 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.

In addition, context [25] is known to play an important role in visual recognition. Considerable works have been proposed for exploiting context in object detection. Bell *et al*. [1] explored the use of recurrent neural networks to mod-
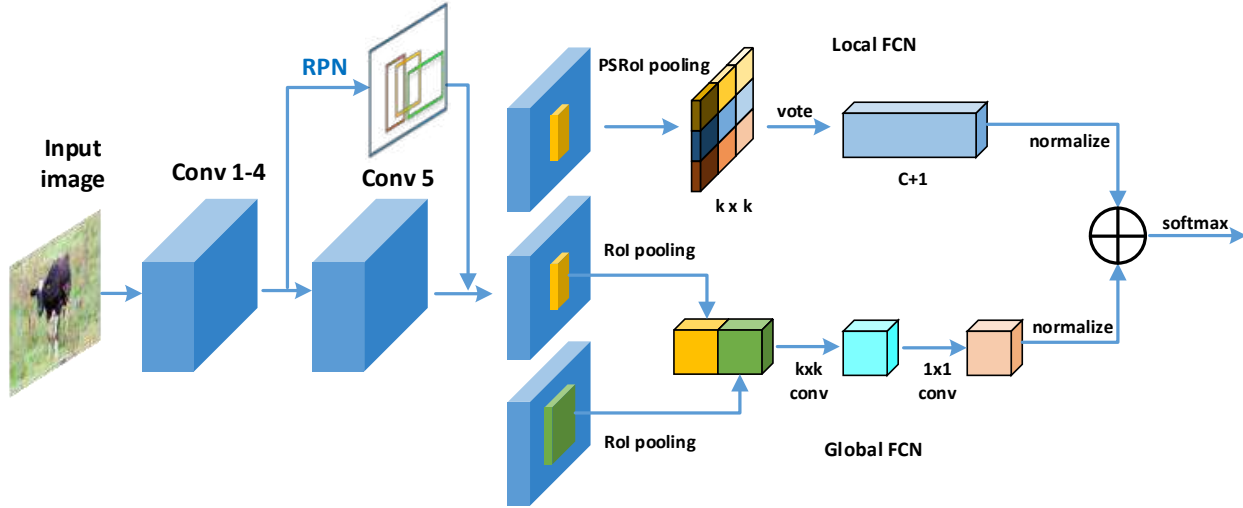
Figure 2. The architecture of the proposed CoupleNet. We use ResNet-101 as the basic feature extraction network. Given an input image, we first exploit Region Proposal Network (RPN) [21] to generate candidate proposals. Then each proposal flows to two different branches: local FCN and global FCN, in order to extract the global structure information and learn the object-specific parts respectively. Finally the output of the two branches are coupled together to predict the object categories.

el the contextual information. Gidaris *et al*. [7] proposed to utilize multiple contextual regions around the object. Cai *et al*. [2] 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.

3. CoupleNet

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

3.1. Network architecture

The architecture of our proposed CoupleNet is illustrated in Figure 2. Our CoupleNet includes two different branches: a) a local part-sensitive fully convolutional network to learn the object-specific parts, denoted as local FCN; b) a global region-sensitive fully convolutional network to encode the whole appearance structure and context prior of the object, denoted as global FCN. We first use the ImageNet pre-trained ResNet-101 released in [12] to initialize our network. For our detection task, we remove the last average pooling layer and the *fc* layer. Given an input image, we extract candidate proposals by using the Region Proposal Network (RPN), which also shares convolution features with CoupleNet following [21]. 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.

3.2. 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 1x1 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 [16] 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 3(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

4128

Local part confidence

Global confidence

person

(a)

Local part confidence
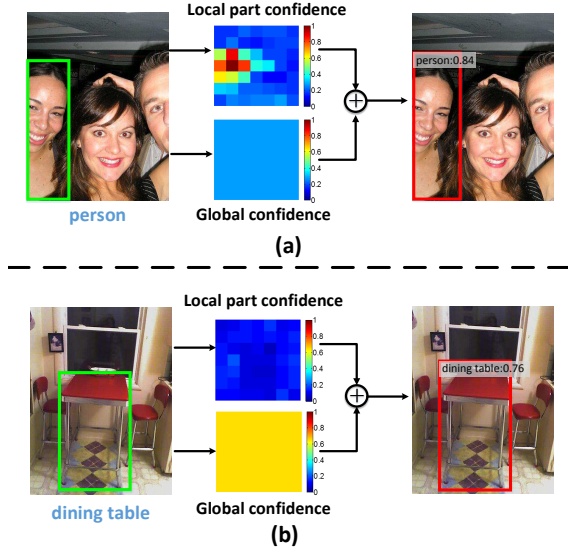
Global confidence

dining table

(b)

Figure 3. An intuitive description of CoupleNet 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.

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, the local FCN alone is difficult to make robust predictions. Thus it is necessary to add the global structure information to enhance the discrimination.

3.3. 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 1x1 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 [8] 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 1x1 convolution is fed into the classifier whose output is also a $(C + 1)$-d vector.

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 ob-

jects due to the large receptive field, Zhou *et al*. [28] 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. 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 3(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 3(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.

3.4. 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 1x1 convolutional layer to model the scale. Meanwhile, how to couple the local and global output is also a problem that needs to be researched. Here, we investigated three different coupling methods: element-wise sum, element-wise product and element-wise maximum. Our experiments show that using 1x1 convolution along with element-wise sum achieves the best performance and we will discuss it in Section 4.1.

With the coupling structure, CoupleNet simultaneously exploits the local parts, global structure and context prior for object detection. The whole network is fully convolutional and 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 CoupleNet can perform the inference efficiently, which runs slightly slower than R-FCN but much more faster than Faster R-CNN.

4. 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 our method. We demonstrate state-of-the-art results on all three datasets without bells and whistles.

4.1. Ablation studies on VOC2007

We first perform experiments on PASCAL VOC 2007 with 20 object categories for detailed analysis of our proposed CoupleNet detector. We train the models on the union set of VOC 2007 trainval and VOC 2012 trainval ("07+12") following [21], and evaluate on VOC 2007 test set. Object detection accuracy is measured by mean Average Precision (mAP), all the ablation experiments use *single-scale* training and testing, and we did not add the context prior.

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*. [1] 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 1x1 convolutional layer to learn the scale.

As shown in Table 1, we find that the use of L2 normalization decreases the performance greatly, even worse than the 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 1x1 convolution to adaptively learn the scales between the global and local branches. Table 1 shows that using 1x1 convolution increases by 0.6 points compared to the direct addition and 2.2 points over R-FCN. Therefore, we use 1x1 convolution to replace the L2 normalization in the following experiments.

Coupling strategy. We explore three different response coupling strategies: element-wise sum, element-wise product and element-wise maximum. Table 1 shows the comparison results for the above three different implementations. We can see that the element-wise sum always achieves

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

Table 1. **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. 1x1 conv+eltwise: use 1x1 convolution to learn the scale.

the best performance even though in different normalization methods. Generally, current advanced residual networks [12] 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.

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 4, we also compare our CoupleNet with the model ensemble. For a fair comparison, we first re-implemented Faster R-CNN [12] using ResNet-101 and online hard example mining (OHEM) [22], which achieves a mAP of 79.0% on VOC07 (76.4% in original paper without OHEM). We also re-implemented R-FCN with appropriate joint training using the public available code py-R-FCN[2], which achieves a slightly lower result compared to [16] (78.6% vs. 79.5%). We use our reimplementation 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 4, 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 CoupleNet can simultaneously utilize the global and local infor-

---

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

	training data	mAP (%) on VOC07	GPU	test time (ms/img)
Faster R-CNN [12]	07+12	76.4	K40	420
R-FCN [16]	07+12	79.5	TITAN X	83
R-FCN *multi-sc train* [16]	07+12	80.5	TITAN X	83
CoupleNet	07+12	**81.7**	TITAN X	102
CoupleNet *context*	07+12	**82.1**	TITAN X	122
CoupleNet *context multi-sc train*	07+12	**82.7**	TITAN X	122

Table 2. **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 [16]. The input resolution is 600x1000. We also note that the TITAN X used here is the new Pascal architecture along with CUDA 8.0 and cuDNN-v5.1. "07+12": VOC07 trainval union with VOC12 trainval. *context*: add the context prior to assist the global branch.

Method	Train	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
ION [1]	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 [14]	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* [18]	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* [18]	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§ [12]	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 [16]	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 [ours]	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

Table 3. **Results on PASCAL VOC 2007 test set.** The first four methods use VGG16 and the latter three 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 segmentation labels. *: the results are updated using the latest models. §: this entry is directly obtained from [12] without using OHEM.

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	**81.7**

Table 4. **CoupleNet *vs.* model ensemble.** *ReIm*: our reimplementation using OHEM. Global FCN: only the global branch of our network.

mation to update the network and to infer the final results. On the other hand, our method enjoys end-to-end training and there is no need to train multiple models, thus greatly reducing the training time.

Amount of parameters. Since our CoupleNet 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 1x1x256, 3x3x256 and 1x1x1024 respectively, to the prediction sub-network. 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 [16] as shown in Table 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.

4.2. 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 [16]. Similarly, we use the dilation strategy to reduce the effective stride of ResNet-101, just as [16] 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 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 2 shows the detailed comparisons with Faster R-

Method	Train	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
ION [1]	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 [14]	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* [18]	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* [18]	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§ [12]	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 [16]	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 [ours]	07++12	**80.4**†	89.1	**86.7**	81.6	71.0	64.4	83.7	83.7	**94.0**	62.2	84.6	65.6	92.7	89.1	87.3	**87.7**	64.3	84.1	72.5	**88.4**	75.3

Table 5. **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. "07++12": the union set of VOC07 trainval+test and VOC12 trainval. "07+12+S": VOC07 trainval union with VOC12 trainval plus segmentation labels. *: results are updated using the latest models. §: this entry is directly obtained from [12] without using OHEM. †: http://host.robots.ox.ac.uk:8080/anonymous/M5CQTL.html.

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* [18]	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* [18]	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 [1]	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+++ [12]	trainval	**34.9**	**55.7**	-	15.6	38.7	50.9	-	-	-	-	-	-
R-FCN [16]	trainval	29.2	51.5	-	10.3	32.4	43.3	-	-	-	-	-	-
R-FCN *multi-sc train* [16]	trainval	29.9	51.9	-	10.8	32.8	45.0	-	-	-	-	-	-
CoupleNet	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 *multi-sc train*	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

Table 6. **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 segmentation labels.

CNN and R-FCN. As we can see that our single model achieves a mAP of 81.7%, which outperforms the R-FCN by 2.2 points. However, while embedding the context prior to the global branch, our mAP rises up to 82.1%, 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 2, our method is slightly slower than R-FCN, which also reaches a real-time speed (*i.e.* 8.2 fps or 9.8 fps without context) 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 3, 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 the advanced end-to-end SSD method [18], which requires complicated data augmentation and careful training skills. Just as discussed earlier, CoupleNet 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 pottedplant,

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 provides an extra auxiliary discrimination.

4.3. Results on VOC2012

We also evaluate our method on the more challenging VOC2012 dataset by submitting results to the public evaluation server. We 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 5 shows the results on the VOC2012 test set. Our method obtains a top mAP of 80.4%, which is 2.8 points higher than R-FCN. We note that without using the extra tricks in the testing phase, our detector is the first one with a mAP higher than 80%. Similar promotions over the specific classes analysed in VOC07 are also observed, which once again validates the effectiveness of our method. Figure 4 shows some detection examples on VOC 2012 test set.

Figure 4. **Detection examples of CoupleNet on PASCAL VOC 2012 test set.** The model was trained on the union of VOC07 trainval+test and VOC12 trainval (80.4% mAp). Our method works well with the occlusions, truncations, inter-class interference and clustered background. CoupleNet also shows good performance for the categories with class-specific backgrounds, *e.g.* airplane, bird, boat, *etc*. A score threshold of 0.6 is used to draw the detection bounding boxes. Each color is related to an object category.

4.4. Results on MS COCO

Next we present more results on the Microsoft 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 $IoU \in [0.5 : 0.05 : 0.95]$. Following the VOC2012, a 4-GPU implementation is used to accelerate the training process. We use an initial learning rate of 0.001 for the first 510k iterations and 0.0001 for the next 70k iterations. In addition, we conduct multi-scale training with the scales are randomly sampled from $\{480, 576, 672, 768, 864\}$ while testing in a single scale.

Table 6 shows our results. Our single-scale trained detector has already achieved a result of 33.1%, which outperforms the R-FCN by 3.9 points. In addition, the multi-scale training further 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.

5. Conclusion

In this paper, we present the CoupleNet, a concise yet effective network that simultaneously couples global, local and context cues for accurate object detection. Our system naturally combines the advantages of different region-based approaches with the coupling structure. With the combination of local part representation, global structural information and the contextual assistance, our CoupleNet achieves state-of-the-art results on the challenging PASCAL VOC and COCO datasets without using any extra tricks in the testing phase, which validates the effectiveness of our method.

References

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*, pages 354–370. Springer, 2016.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[4] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014.

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[7] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1134–1142, 2015.

[8] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[10] J. Gu, J. Zhou, and C. Yang. Fingerprint recognition by combining global structure and local cues. *IEEE Transactions on Image Processing*, 15(7):1952–1964, 2006.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.

[14] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: towards accurate region proposal generation and joint object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 845–853, 2016.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] Y. Li, K. He, J. Sun, et al. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[19] W.-Z. Nie, A.-A. Liu, Z. Gao, and Y.-T. Su. Clique-graph matching by preserving global & local structure. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4503–4510, 2015.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[22] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[23] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[24] 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 Transactions on Multimedia*, 10(3):393–408, 2008.

[25] J. Wang, W. Fu, J. Liu, and H. Lu. Spatiotemporal group context for pedestrian counting. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(9):1620–1630, 2014.

[26] J. Wang, W. Fu, H. Lu, and S. Ma. Bilayer sparse topic model for scene analysis in imbalanced surveillance videos. *IEEE Transactions on Image Processing*, 23(12):5198–5208, 2014.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[29] Y. Zhu, J. Wang, C. Zhao, H. Guo, and H. Lu. Scale-adaptive deconvolutional regression network for pedestrian detection. In *Asian Conference on Computer Vision*, pages 416–430. Springer, 2016.