A Dual-branch CNN Structure for Deformable Object Detection

Jianjun Li[1], Kai Zheng[1], Zhenxing Luo[2], Zhuo Tang[2], Ching-Chun Chang[3]

[1] School of Computer Science and Engineering, Hangzhou Dianzi University, China
[2] Science and Technology on Communication and Information Security Control Laboratory of the 36th Institute of China, Electronics Technology Group Corporation, China
[3] Department of Computer Science, University of Warwick, United Kingdom
jianjun.li@hdu.edu.cn, Kzheng@gmail.com, ZhenxingL@gmail.com, Tangz@gmail.com, c.chang.2@warwick.ac.uk

Abstract

Object detectors based on CNN are now able to achieve satisfactory accuracy, but their ability to deal with some targets with geometric deformation or occlusion is often poor. This is largely due to the fixed geometric structure of the convolution kernel and the single inflexible network structure. In our work, we use dual branch parallel processing to extract the different features of the target area to coordinate the prediction. To further enhance the performance of the network, this study rebuilds the feature extraction module. Finally, our detector learns to adapt to a variety of different shapes and sizes. The proposed method achieves up to 81.76% mAP on the Pascal VOC2007 dataset and 79.6% mAP on the Pascal VOC2012 dataset.

Keywords: Dual-branch structure, Convolution neural network, Deformable object detection

1 Introduction

Object recognition, such as fingerprint recognition and face recognition, plays an important role in the security field. As we know, the first step in object recognition is object detection. Therefore, object detection is also essential in the field of security. Object detection is widely used to precisely locate and classify kinds of targets in an image or video. To perform properly, these detectors need to "learn" as many different categories of feature representations as possible. Also, they need to learn the different scales, poses, viewing angles, and even non-rigid body deformations which the same individual may present differently in different images.

In this work, we propose to create an object detector that can view images in the same way that humans do. When a human identifies an object, the first thing to be observed is the overall shape of the target object, which represents the global information about the target. Based on the shape and the scale of the object, a human can preliminarily judge the general characteristics of the object. Next, a human will gradually identify the specific features of the target based on the upper, lower, left, and right sides of the object. This is known as local information. By considering the local and global information combinedly, humans can make precise judgments about the target object. Our objective in this study is to design a detection model that can adapt to different shapes of objects, in other words, obtain both local and global information simultaneously. Natural language processing; Dialogue act recognition; Inference model.

We have assessed existing object detectors based on Convolutional Neural Networks (CNN). Fast/Faster R-CNN [5, 19] and Region-based Fully Convolutional Networks (RFCN) [1] are two representative region-based CNN approaches. Fast/Faster R-CNN uses a subnetwork to predict the category and bounding box of each region proposal. Unlike Fast/Faster R-CNN, R-FCN proposes the concept of a position-sensitive score map and conducts the inference with a position-sensitive region of interest pooling (PSRoIPooling). The reason why R-FCN achieves more accuracy is that it focuses on details of the target object and sensitivity of the location. Inspired by their work [1], this study uses position-sensitive RoI pooling to extract the local information of the object and employing Faster R-CNN to extract global information. Besides, a dual branch structure based on Faster R-CNN has been used to rebuild the network in the proposed method, which improves the performance significantly.

Besides, we employ deformable convolution and deformable position-sensitive RoI pooling to enable the detection model to handle the geometric transformations, which is inspired by [2]. These two new modules can greatly enhance the capability of the model to handle different shapes or poses of the same object. Therefore, our model can achieve performance

*Corresponding Author: Jianjun Li; E-mail: jianjun.li@hdu.edu.cn

DOI: 10.3966/160792642020052103018

beyond the current level of some detectors for objects with deformation, occlusion, and overlap. We have also achieved outstanding results on objects with high similarities, such as chairs and tables.

The remainder of the paper is organized as follows. In Section II, the related works are presented. Section III describes the methodology of the proposed approach. Section IV presents the experimental results. Section V concludes our work with a summary.

2 Related Work

With the rapid development of deep convolutional networks [9, 12, 18, 23, 24-28], many object detection methods have been proposed and made substantial improvements in this field. Some of the recent object detection works are introduced as follows.

R-CNN: Region-based CNN (R-CNN) [6] achieves high accuracy by extracting features via deep neural networks. Following R-CNN, a large number of variants of R-CNN have evolved. In Faster R-CNN [19], Region Proposal Network is introduced to generate proposals and RoIPooling is adopted in the subnetwork on each proposal. Following Faster R-CNN [19], R-FCN [1] proposes the Position-sensitive RoIPooling, which also speeds up the detection when dealing with a large number of proposals. Feature Pyramid Networks (FPN) [15] construct feature pyramids with the inherent multi-scale pyramidal hierarchy of deep convolutional networks. Mask R-CNN [8] further implements a mask predictor by adding an extra branch in parallel based on Faster R-CNN. It also incorporates a RoIAlign layer that removes the harsh quantization of RoIPooling. RetinaNet [16] is an FPN-based single stage detector that employs FocalLoss to address the class imbalance issue caused by extreme foreground-background ratio.

STN: Spatial Transform Networks (STN) [11] is the first work that learns spatial transformation from image datasets by deep neural networks. STN warps the feature map employing a global parametric transformation, for example, affine transformation. Such warping is expensive, and it is quite difficult to infer the transformation parameters. STN is effective to solve the classification problem on small-scale images. The inverse STN method [14] is then proposed to address the aforementioned problems by replacing the expensive feature warping with efficient transformation parameter propagation.

DPM: Deformable Part Models [4] is a shallow model that can maximize the classification score by learning the spatial deformation of object parts. Its inference process equals to CNNs [7] when treating the distance transform as a special pooling operation. However, the training process is not end-to-end and involves manual adjustments, including the selection of components and hyper-parameter.

Deformable Convolution: Convolutional neural networks (CNNs) are inherently restricted to model geometric transformations, due to the fixed geometric structure in the module construction. The Deformable Convolution Network [2] indicates that the pixels in a receptive field have a different impact on the output response. The pixels on the object contribute greater than others. Therefore, Deformable Convolution Networks enable the receptive field to distinguish objects from the background. With this improvement, a convolution filter can autonomously attain the ability to sense the object. It rebuilds the convolution filter, introducing two new modules (Deformable Convolution and Deformable PSRoIPooling) which greatly enhance CNN's capability of modeling geometric transformation. During network training, offsets can be learned by additional convolution layers so that different locations can be sampled more flexibly.

Multi-Branch: Single-branch networks often have limitations, such as fixed receptive field and simple hierarchy, which can result in the network not being able to make full use of feature diversity. However, in recent network structure, like CoupleNet [28] and Mask R-CNN [8], which both adopted multi-branch structure. CoupleNet uses multi-branch to obtain feature maps of different inclination which then merged. Mask R-CNN uses two-branch to collaboratively perform segmentation and detection of two related tasks and achieves mutually beneficial results.

Our model draws on the idea of feature fusing that is easily overlooked in many of recent work, continuing the multi-branch structure of CoupleNet [28] and MASK RCNN [8]. Based on the multi-branch of CoupleNet, we carefully verified its theory and experimental details. Unlike the CoupleNet, we choose the structure of dual-branch while retaining some of the fully-connected layers. With experimental support, we design Branch B which keeps the fully-connected layer (the experimental data in Table 4 proves that our determination's correction). In Branch A, the convolution layer is used instead of a fully-connected layer. Combining two different branches, our network can not only take into account the number of model parameters but also ensure accuracy. At the same time, we add a Deformable module in Branch B, which greatly improved the ability of the Branch to adapt to the deformation of an object (experiment results are shown in Table 1 and Table 2).

3 The Proposed Approach

In this section, we first present the architecture of the proposed detection model and then describe design details.

Table 1. Results on PASCAL VOC 2007 test set (trained on VOC 2007 trainval and VOC 2012 trainval)

Method	SSD 512	Faster R-CNN	R-FCN	D-R-FCN	Mask R-CNN	Ours*	our§	ours†	ours
mAP	78.5	77.1	79.2	81.8	80.5	80.0	80.5	81.1	**81.6**
aero	90.0	79.2	78.5	81.2	84.6	81.2	83.3	85.6	83.4
bike	85.3	84.1	85.8	87.5	87.6	87.6	86.1	86.1	88.2
bird	77.7	77.4	80.1	84.3	79.4	78.4	79.0	80.3	80.5
boat	64.2	68.7	70.8	79.4	78.2	74.4	74.6	74.8	75.5
bottle	58.4	59.1	68.4	69.4	68.4	65.5	67.8	68.5	67.3
bus	85.3	86.1	85.1	89.0	88.5	86.9	87.2	88.7	87.4
car	84.4	84.9	86.9	88.7	85.4	87.9	87.9	88.1	88.3
cat	92.5	86.0	88.4	92.4	90.3	88.2	88.3	88.9	88.6
chair	61.3	60.9	65.6	66.5	63.2	64.8	66.7	67.4	**69.1**
cow	83.4	86.4	86.8	89.4	88.4	87.2	87.1	87.2	88.1
table	65.0	72.8	73.1	72.3	70.1	73.9	73.8	74.0	**74.6**
dog	89.8	88.0	88.4	91.3	90.2	89.7	89.5	88.4	88.9
horse	88.5	85.1	88.6	90.1	87.3	87.9	87.8	88.0	88.9
mbike	88.2	83.8	80.4	83.2	84.4	83.1	84.5	85.1	84.0
person	85.5	79.0	80.6	80.4	80.5	83.3	84.2	85.6	**85.8**
plant	54.4	49.4	52.3	57.3	54.3	56.8	57.0	57.3	**57.4**
sheep	82.4	80.9	80.3	85.0	84.9	82.6	82.9	83.3	85.6
sofa	70.7	76.0	80.1	80.6	75.3	78.2	78.5	79.2	**81.2**
train	87.1	78.3	84.7	89.2	88.3	84.3	84.3	86.2	88.4
tv	75.4	76.4	78.5	78.6	79.7	76.9	78.7	80.3	81.2

Note. Ours*: only includes Branch A and Branch B. Ours§: includes Branch A and Branch B with RoIAlign. Ours†: includes Branch A and Branch B with RoIAlign and deformable module.

Table 2. Results on PASCAL VOC 2012 test set (trained on VOC 2007 trainval+test and VOC 2012 trainval)

Method	SSD512	Faster R-CNN	R-FCN	D-R-FCN	Mask R-CNN	ours
mAP	77.7	73.8	77.6	79.4	78.3	79.6
aero	88.9	86.5	86.9	86.9	88.4	88.1
bike	84.3	81.6	83.4	86.4	83.2	85.7
bird	76.9	77.2	81.5	78.6	78.3	81.6
boat	63.2	58.0	63.8	72.2	67.8	71.2
bottle	57.8	51.0	62.4	64.6	62.8	64.0
bus	85.0	78.6	81.6	83.4	83.2	83.2
car	83.4	76.6	81.1	84.3	84.3	83.9
cat	91.8	93.2	93.1	94.2	90.3	93.1
chair	60.7	48.6	58.0	60.0	59.3	61.2
cow	83.1	80.4	83.8	81.2	80.3	83.0
table	64.0	59.0	60.8	64.3	63.5	65.6
dog	88.2	92.1	92.7	93.0	93.9	92.3
horse	88.1	85.3	86.0	90.1	89.3	88.1
mbike	87.9	84.8	84.6	83.2	86.4	85.6
person	84.5	80.7	84.4	84.3	86.3	87.5
plant	53.8	48.1	59.0	59.9	58.8	62.0
sheep	81.6	77.3	80.8	85.0	80.0	83.0
sofa	69.7	66.5	68.6	71.5	68.4	71.6
train	86.5	84.7	86.1	88.4	86.0	87.2
tv	74.4	65.6	72.9	76.4	74.5	74.3

3.1 Network Architecture

Figure 1 shows the architecture of our model. The proposed detector is mainly composed of two components. The first one is RPN (Region Proposal Network), which is in charge of generating candidate proposals. The second is R-CNN subnetwork which makes the final location and classification. The R-CNN subnetwork involves two separate branches: (A) a RoIAlign to gather global feature information for an RoI, and (B) Deformable PSRoIPooling to encode details and local feature information of an RoI. Our network is initialized in the same way as the pre-trained ImageNet model, ResNet101 [9]. To realize the detection task, we remove the last average pooling layer and FC (Fully-Connected) layer. All proposals produced by RPN are fed to Branch A and Branch B. Finally, the outputs of the two branches are fused to do classification and box regression.

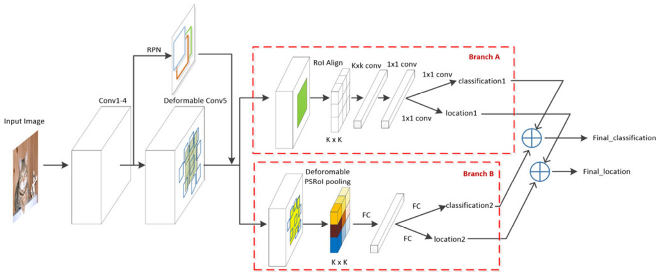

Figure 1. Our detector: In Feature Extractor, this study uses Deformable Conv5. In the R-CNN subnetwork, we employ a dual branch (branch A and branch B). Branch A: for the 2000 proposals produced by RPN, we process all the proposals by RoIAlgin. We replace the last two FC layers by a full convolution layer. Branch B: for the 2000 proposals produced by RPN, we use Deformable PSRoIPooling and then add one fully connected layer. Finally, we fuse the two feature maps produced by branch A and branch B to predict the final classification and location

3.2 Regional Proposal Networks (RPN)

The RPN is based on a sliding-window class-agnostic object detector and it uses features extracted from the 4th stage, following [19]. Specifically, the RPN pre-defines a set of anchors that are related to scales and aspect ratios. In the proposed approach, three aspect ratios are set to {1: 2, 1: 1, 2: 1} and three scales are set to {1282, 2562, 5122} separately to cover objects of different shapes, to the extent possible. Besides, many proposals are redundant with each other, which is addressed by non-maximum suppression (NMS), and the intersection-over-union (IoU) threshold of NMS is set to 0.7 in our approach. The anchor is assigned with a label if its IoU is the highest or exceeds 0.7 with any ground-truth box. On the other hand, if an anchor's IoU is smaller than 0.3 with all the ground-truth boxes, it will be assigned with a negative label. After processing by RPN, each input image outputs 2000 proposals, which is then used in the R-CNN subnet.

3.3 R-CNN Subnetwork

In this section, we discuss the R-CNN subnetwork in our approach in four aspects. Also, we provide the topology of our network for readers in Figure 2.

Maintain Spatial Information. For Faster R-CNN which is presented in Figure 3, we find that two 1024-d FC layers bring plenty of computations. Meanwhile, the FC layer forces the features to be compressed to a one-dimensional vector. The reduction of dimension is not conducive to a detection task that is sensitive to spatial location. Inspired by [21], we replace the two FC layers by two Conv (convolution) layers and use a 1×1 convolution to do classification and location.

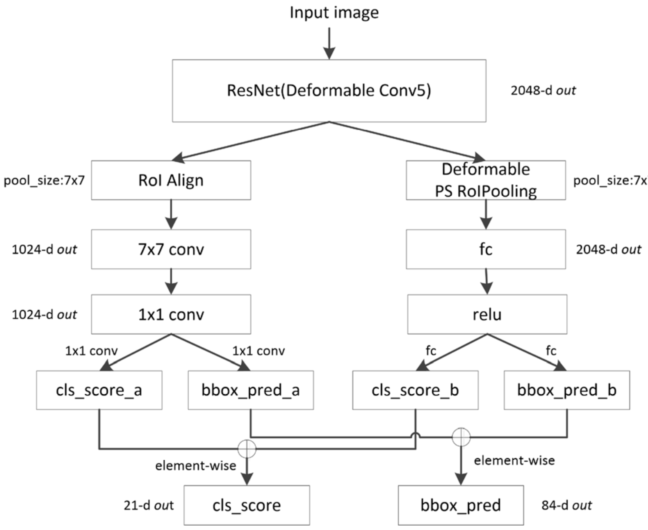

Figure 2. The Topology of our network

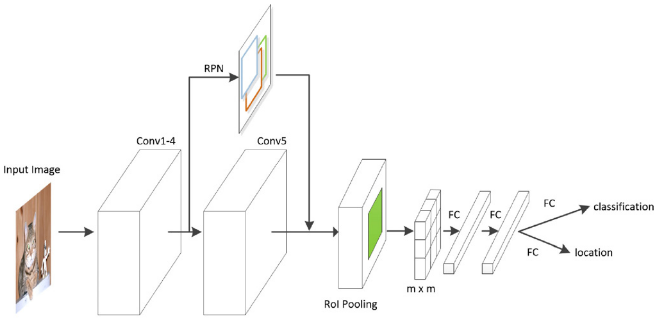

Figure 3. Faster R-CNN [19]: the whole network can be divided into two parts: RPN and R-CNN subnet. After feature extracting network (like VGG or ResNets), the feature map flows into RPN. Then RPN produces 2000 proposals, which is then used in the R-CNN subnet. In the R-CNN subnet, each proposal is used to make classification and location regression

Accurate Quantification. Faster R-CNN [19] uses RoIPooling to extract features from each candidate box, and the extracted features are then used in classification and box regression. As discussed in [8], RoIPooling inevitably involves quantization two times. The quantization introduces misalignments between the RoI and the extracted features. To a large extent, it affects the accuracy of localization. Inspired by the work [8], we remove the RoIPooling to reduce the misalignments caused by quantization. Instead, we use bilinear interpolation [11] in RoIAlign to compute the exact values of the input features sampled at four regularly locations in each RoI bin and then aggregate the output with max or average pooling strategies. The experimental results show that the RoIAlign operator brings substantial improvement to our detector.

Feature Combination. Based on our analysis of Faster R-CNN [19], RoIPooling transforms an arbitrary-sized input rectangular region into specified features (e.g., 7×7). The output features from RoIPooling tend to describe overall information of an RoI while ignoring the details of objects such as internal orientation and hierarchy.

Therefore, we extend an extra branch in parallel like [8] to extract features in detail. Inspired by the work [1, 13], we attach a 490-dim 1×1 convolutional layer to construct a position-sensitive score map. Considering computational efficiency, we choose 490dim (10 × 7 × 7) rather than 1029-dim (21 × 7 × 7) for the output channel of our score map, and yet achieve equal accuracy.

For the position-sensitive score map, our detector conducts PSRoIPooling on the 72 scores and then vote on the RoI, producing a 10-dim vector for each RoI. The score map can be divided into 7×7 small grids, which can encode the cases of {top-left, top-center, top-right, ..., bottom-right} of an object category. Through the analysis above, PSRoIPooling represents the local and detailed feature information. We then attach two simple FC layers to do the classification and regression, in which the dimension is raised to 21 and 84 (4 × 21). Note that we use a layer rather than a convolutional layer here. The experimental results show that this branch gets lower mAP when using the convolutional layer.

Finally, the detector uses element-wise to fuse class probability and box prediction value, which are produced by Branch A and Branch B and then outputs the final class and box prediction.

Effective Receptive Field. For the object detection task, we find that the pixels in a receptive field have a different impact on the output response, and the pixels on the object contribute greater than others. Based on this information, we hope the receptive field to distinguish objects against the background. In other words, we utilize the receptive field to sense the target autonomously. Inspired by the work [2], we find that the deformable convolution can make an adaptive adjustment based on the scale and shape of the objects, and therefore greatly enhances the capability of modeling geometric transformations for the convolution kernel.

Motivated by [2], we add deformable convolution to the last stage of ResNet. Then, the 2D offsets extracted from the preceding feature map via convolutional layers are added to the regular grid sampling locations in the standard convolution. As for PSRoIPooling, an offset is also added to each bin position in the regular bin partition of the previous PSRoIPooling [1-2]. Therefore, our detector can accommodate geometric variations or model geometric transformations in object scale, pose, viewpoint, and part deformation. Meanwhile, the localization capability is improved, especially for non-rigid objects. By using the receptive field, the detector can learn semantic information adaptively and effectively and handle the overlapping and semi-occlusion issues.

4 Experiments

We first evaluate Faster R-CNN [19], and it achieves a 77.1% mAP on the VOC2007 test set and 73.8% mAP on the VOC2012 test set. Our detector, by contrast, achieves 81.6% mAP on the VOC2007 test set and 79.6% mAP On the VOC2012 test set, which largely surpasses our baseline model 81.6% vs. 77.1% and 79.6% vs. 73.8%, and other detection models, such as R-FCN [1] (81.6% vs. 79.2% and 79.6% vs. 77.6%). In the meantime, we also compare our detector with SSD [17], D-R-FCN [2] (R-FCN with Deformable Convolution) and Mask R-CNN [8]. The results in Table 1, Table 2 and Table 3 validate the effectiveness of our method over Faster R-CNN. In the meantime, we also compare our model with the SSD method [17]. In the following discussion, we present the training details as well as ablation experiments.

Table 3 Results on the test set of PASCAL VOC 2007

Method	mAP	
	mAP@0.5 (%)	mAP@0.7 (%)
Faster R-CNN [19]	77.1	61.0
R-FCN [1]	79.2	62.8
D-R-FCN [2]	81.8	68.0
Mask R-CNN [8]	80.5	65.7
ours	81.6	68.3

Table 4. Results on comparing using a fully-connected (fc) layer and convolutional (conv) layer in Branch B

Branch B	mAP
fc layer	81.6
conv layer	78.8

4.1 Implementation Details

Our approach is trained on 2 NVIDIA TITAN X GPUs, where the weight decay is set to 0.0005 and the momentum is set to 0.9. The batch size for each GPU is 2 and each image has 2000/1000 ROIs for the training and testing phase. Besides, the learning rate is 0.001 for the first 80k iterations and decrease to 0.0001 for later 30k iterations. During the training phase, the image scale is randomly sampled from {480, 570, 670, 760, 860}, and the shorter edge of the image is resized to the sampled scale. We adopt online hard example mining (OHEM) [22] techniques. The backbone network in our approach is initialized according to the pre-trained ImageNet [20] unless explicitly noted. Besides, the parameters of stages 2, 3 and 4 in the base model are also adjusted and batch normalization is also fixed to improve the training speed.

Next, a series of ablation experiments are conducted to validate the effectiveness of the proposed approach. All the ablation experiments use single-scale training and testing.

4.2 Ablation Experiments

We conduct experiments on PASCAL VOC 2007 [3], which includes 20 object categories for a detailed evaluation of the proposed approach. Table 1 lists the detailed comparison results of Faster R-CNN [19], R-FCN [1], SSD [17], D-R-FCN [2], Mask RCNN [8] and the proposed method. Specifically, the proposed detector is trained on the union set of VOC 2007 trainval and VOC 2012 trainval, and is evaluated on VOC 2007 test set. To be fair, we only provide the experimental results of a single model without multi-scale testing. All the methods use ResNet-101 as a base network (except SSD512, which uses VGG16). Note that Faster means Faster R-CNN method. D-R-FCN means R-FCN with the Deformable Convolution. Mask R-CNN reimplemented in our experiment is the one without FPN. We find that our method achieves 81.6% mAP, which outperforms R-FCN by 1.6 percent point. The standard mAP (mean Average Precision) score [3] is adopted to evaluate the performance of all methods.

Besides, we also perform experiments on PASCAL VOC 2012. The models are trained on the union set of VOC 2007 trainval+test and VOC 2012 trainval, and are evaluated on VOC 2012 test set as shown in Table 2. To be fair, we only provided the experimental results of a single model without multi-scale testing. The present comparison on PASCAL VOC 2012 dataset, our method achieves 79.6% mAP, which outperforms R-FCN by 2.0 percent point.

As discussed above, our model can sense the objects of different geometric transformations, which can also adapt to part deformation, occlusion, and partial overlap. As shown in Table 1 and Table 2, the proposed model achieves better performance on the sofa, person, chair, and table, which validates the effectiveness of our model.

In Table 3 we compare three models' mAP scores with IoU thresholds being set to 0.5 or 0.7. Our detector gets a higher accuracy even at a high IOU threshold, which strongly verifies the efficiency.

For visual comparison, we present detected images in Figure 4 and Figure 5. In Figure 4, our model (top) shows great robustness for geometric deformation and semi-occluded objects. For example, the proposed model successfully detects the regions for the sofa and the animal lies on the sofa in the first image, the red bus in the third image and the chair in the fourth image. However, Faster R-CNN fails to capture these objects with geometric deformation or occlusion. The images in Figure 4 include overlapping objects which should be detected separately. Compared to R-CNN, our model can distinguish overlapping objects.

Demonstrated by the quantitative metric assessment and visual comparison, the proposed model is effective for the object detection task. We extract features both globally and locally and combinedly take advantage of deformable convolution and pooling to make our model adaptive to objects with geometric deformation, occlusion and overlapping.

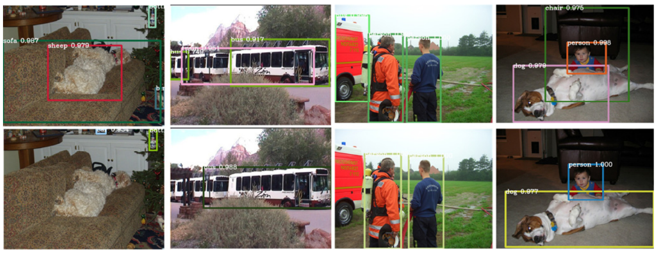

Figure 4. Visualization detection results of our detector (top) and Faster R-CNN (bottom). It is obvious that our detector has great robustness for geometric deformation and semi-occluded objects. It verifies our model's capability to learn geometric deformation and the ability to fuse global and local features to make predictions

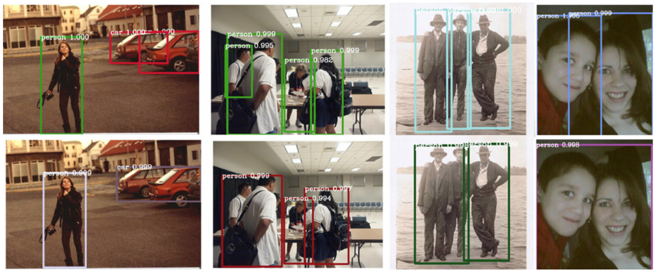

Figure 5. Comparison of overlapping object detection. Images in the first row (top) show the detection results by our proposed model. Images in the second row (bottom) are the detection results produced by R-CNN. The comparison demonstrates the superior ability of our model to distinguish overlapping objects

5 Conclusions

In this paper, we present a novel and adaptive object detection model. The proposed approach can adaptively adjust geometric variations and model geometric transformations in object scale, pose and part deformation. Especially, we adopt the state-of-the-art image classification methods as backbones and also add deformable modules into our framework to enhance the capability of transformation modeling. Besides, our method fuses the feature produced by R-FCN and Faster R-CNN, then generates the final feature for accurate prediction. In general, our detector provides a novel idea that utilizes a dual branch that combines global and local feature information to make further classification and location. Benefiting from the architecture, our detector can well adapt to objects with occlusion or deformation, even at a high IoU threshold, which strongly verifies that the approach we propose is robust and efficient. In the future, we will try to optimize the structure and efficiency of the proposed approach to obtain real-time detection results.

Acknowledgments

This work was supported by the National Natural Science Fund of China (No. 61871170), the National Equipment Development Pre-Research Fund: 6140137050202 and the National Defense Basic Research Program of JCKY2017210A001.

References

[1] J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object Detection via Region-based Fully Convolutional Networks, *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, 2016, pp. 379-387.

[2] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable Convolutional Networks, *IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, 2017, pp. 764-773.

[3] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes (VOC) Challenge, *International Journal of Computer Vision*, Vol. 88, No. 2, pp. 303-338, June, 2010.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part-based Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, pp. 1627-1645, September, 2010.

[5] R. Girshick, Fast R-CNN, *IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1440-1448.

[6] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus, OH, USA, 2014, pp. 580-587.

[7] R. Girshick, F. Iandola, T. Darrell, J. Malik, Deformable Part Models are Convolutional Neural Networks, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, MA, USA, 2015, pp. 437-446.

[8] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, *IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, 2017, pp. 2980-2988.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[10] K. He, X. Zhang, S. Ren, J. Sun, Identity Mappings in Deep Residual Networks, *Computer Vision - ECCV 2016 - 14th European Conference*, Part IV, Amsterdam, The Netherlands, 2016, pp. 630-645.

[11] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial Transformer Networks, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, 2015, Montreal, Quebec, Canada, pp. 2017-2025.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet Classification with Deep Convolutional Neural Networks, *International Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, 2012, pp. 1097-1105.

[13] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, J. Sun, Light-head R-CNN: In Defense of Two-stage Object Detector, CoRR abs/1711.07264, November, 2017.

[14] C. Lin, S. Lucey, Inverse Compositional Spatial Transformer Networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, USA, 2017, pp. 2252-2260.

[15] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature Pyramid Networks for Object Detection, Honolulu, HI, USA, 2017, pp. 936-944.

[16] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, *IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, 2017. pp. 2999-3007.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. C. Berg, SSD: Single Shot Multibox Detector, *Computer Vision - ECCV 2016 - 14th European Conference*, Part I, Amsterdam, The Netherlands, 2016, pp. 21-37.

[18] Q. Lu, C. Liu, Z. Jiang, A. Men, B. Yang, G-CNN: Object Detection via Grid Convolutional Neural Network, *IEEE Access*, Vol. 5, pp. 24023-24031, November, 2017.

[19] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, June, 2017.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, F. Li, Imagenet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211-252, December, 2015.

[21] E. Shelhamer, J. Long, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 4, pp. 640-651, April, 2017.

[22] A. Shrivastava, A. Gupta, R. Girshick, Training Region-based Object Detectors with Online Hard Example Mining, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, 2016. pp. 761-769.

[23] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, CoRR abs/1409.1556, September, 2014.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper with Convolutions, *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR 2015), Boston, MA, USA, 2015, pp. 1-9.

[25] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, USA, 2017, pp. 5987-5995.

[26] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, Q. Dai, Supervised Hash Coding with Deep Neural Network for

Environment Perception of Intelligent Vehicles, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 19, No. 1, pp. 284-295, January, 2018.

[27] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices, CoRR abs/1707.01083, December, 2017.

[28] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, H. Lu, Couplenet: Coupling Global Structure with Local Parts for Object Detection, *IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, 2017, pp. 4146-4154.

Biographies

Jianjun Li received the B.Sc. degree in information engineering from Xi'an University of Electronic Science and Technology, Xi'an, China, and the M.Sc. and Ph.D. degrees in electrical and computer from The University of Western Ontario and University of Windsor, Canada separately. He is currently working at HangZhou Dianzi University as a chair professor. His research interests include micro-electronics, audio, video and image processing algorithms and implementation.

Kai Zheng (In Zhejiang Province, China, born in 1997) is an undergraduate in School of Computer Science and Engineering, Hangzhou Dianzi University. He entered the computer science and technology major of Hangzhou Dianzi University in 2015. His research interests include object detection, face recognition, image processing, and machine learning.

Zhenxing Luo received the B.S. degree in telecommunication engineering from Hangzhou Dianzi University, Hangzhou, Zhejiang, in 2006 and the M.S. degree in Signal and Information Processing from Hangzhou Dianzi University, Hangzhou, Zhejiang, in 2009. He is currently pursuing the Ph.D. degree in Communication and Information System at Xian Dianzi University, Xi'an, Shaan, China. From 2009 to 2018, he was a Research Assistant with the Science and Technology on Communication Security Control Laboratory, Jiaxing, Zhejiang. His research interest includes the development and fundamental study of signal processing and machine learning techniques used in spectrum management.

Zhuo Tang received the B.S. degree in electronics engineering from Xidian Dianzi University, Xi'An, ShanXi, in 1990. She is now a senior researcher with the Science and Technology on Communication Security Control Laboratory, Jiaxing, Zhejiang. His research interest includes the development and fundamental study of signal processing and machine learning techniques used in spectrum management.

Ching-Chun Chang received his Ph.D. candidate in the Department of Computer Science, University of Warwick, UK. He received the BBA degree in information management from National Central University, Taiwan, in 2015. He was granted the Marie-Curie fellowship in 2017. He engaged in a short-term scientific mission supported by European cooperation in science and technology in the Faculty of Computer Science, Otto von Guericke University Magdeburg, Germany, in 2016. He participated in a research and innovation staff exchange scheme supported by Marie Skłodowska-Curie in the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, USA, during 2017. He has started research activities at the School of Computer and Mathematics, Charles Sturt University, Australia, since 2018. His research interests include information hiding, digital watermarking, steganography, secret sharing, applied cryptography, digital forensics, multimedia security, image processing, and machine learning.