Residual Joint Attention Network with Graph Structure Inference for Object Detection

Chuansheng Xu[1,2], Gaoyun An[1,2(✉)], and Qiuqi Ruan[1,2]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
{chshxu,gyan,qqruan}@bjtu.edu.cn
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China

Abstract. Most object detectors include three main parts, CNN feature extraction, proposal classification, and duplicate detection removal. In this work, focusing on the improvement of the feature extraction, we propose Residual Joint Attention Network, a convolutional neural network using a residual joint attention module which is composed of a spatial attention branch, a channel attention branch, and a residual learning branch within an advanced object detector with graph structure inference. An attention map generated by the joint attention mechanism is used to weight the original features extracted from a specific layer of VGG16 aiming at performing feature recalibration. Besides, the residual learning mechanism is complementary to the joint attention mechanism and keeps good attributes of the original features. Experimental results show that different branches of our residual joint attention module do not contradict each other. By combining them together, the proposed network obtains higher mAP than many advanced detectors including the baseline on VOC dataset.

Keywords: Joint attention · Residual learning · Graph structure inference · Object detection

1 Introduction

In recent years, thanks to the advances of deep convolutional neural networks, a large number of computer vision tasks have enjoyed significant progress, including segmentation [7,8], image classification [1–3], object detection [4–6]. Among them, object detection is one of the fundamental problems that has been widely studied. Currently, there are two mainstream frameworks to solve the problem of object detection: the one-stage frameworks such as SSD [9] and YOLO [10], which directly transform the problem of object border positioning into a regression problem without extracting proposals; and the two-stage frameworks such as Fast R-CNN [5] and Faster R-CNN [6] which generate proposals by RPN layers [6] and then apply classification and regression to each proposal.

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11901, pp. 210–221, 2019.
https://doi.org/10.1007/978-3-030-34120-6_17

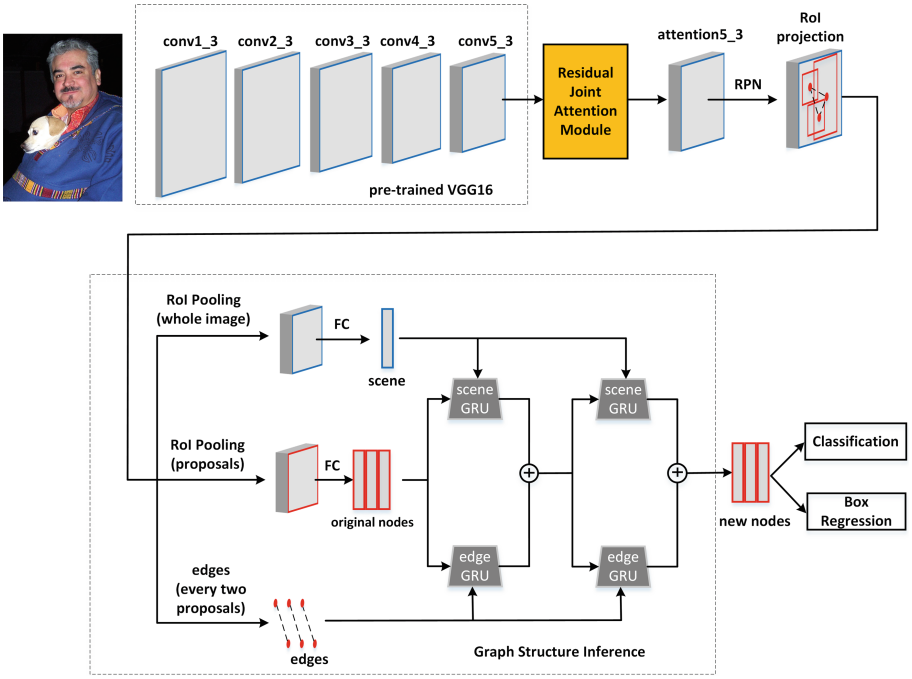

Fig. 1. The framework of our method. We feed an image into VGG16 pre-trained on ImageNet dataset to get a feature map named conv5_3. Then the residual joint attention module recalibrates conv5_3 feature map. Next, the feature map named attention5_3 is passed to an RPN layer followed by the graph structure inference part which involves two contextual information into the inference of node state. Eventually, the final state of each node is used to predict the category and refine the location of the corresponding RoI.

Most object detectors include three main parts, CNN feature extraction, proposal classification, and duplicate detection removal. For these three parts, improving the quality of the features of ConvNet backbones is a straightforward idea through which a lot of algorithms have made major breakthroughs [12–15]. Most of them use effective methods to increase the receptive field or semantic information of the feature maps extracted from ConvNet backbones. However, all of them do not consider utilizing the spatial and channel information of the feature maps when improving the detection accuracy.

Motivated by the success of the attention modules in image classification field [16]. We consider the combination of the spatial attention and the channel attention. As indicated in SENet [16], the channel-wise features can be adaptively recalibrated by effectively modeling the interdependencies between the channels of the feature map extracted from a ConvNet backbone. Similar to SENet [16], we also model the interdependencies between the spatial features. Spontaneously, the joint and multiplicative result of the spatial attention map

and the channel attention map is applied to recalibrate the original features. Intuitively, We conjecture this adds the complementary and compatible information between the spatial attention and the channel attention to the proposed network which enhances the useful features and suppresses the less informative ones. In addition, we combine the residual learning [3] with the joint attention to form a residual joint attention module. All of these lead to boost model's discriminative power. The proposed object detection network is shown in Fig. 1.

In this paper, the proposed module incorporates into an advanced object detector [11] with graph structure inference only increasing a small number of parameters. In principle, The residual joint attention module is universal and not restricted to object detection.

2 Related Work

With the rise of deep convolutional neural networks, the two-stage detectors have rapidly dominated object detection over the past few years [4–6]. These advanced object detectors prevailingly follow the pioneering work R-CNN [4]. R-CNN first generates object proposals by Selective Search [21] and then operates classification and bounding box regression on every proposal. But the biggest problem of R-CNN is repetitive convolutional operation consuming too much time. To speed up, Fast R-CNN [5] introduces a novel RoI pooling layer to extract features for each proposal from the shared ConvNet feature map of the whole image. Whereas proposal generators are still not trained together with Fast R-CNN. To solve this problem, Faster R-CNN [6] develops RPN which can generate precise proposals and be trained together with detection subnetwork. Different from the two-stage detectors, the one-stage detectors remove proposal generators and directly operate classification and regression on a series of pre-computed anchors for real-time detection. Anyway, these state-of-the-art methods only consider the appearance features of the objects without considering the connections between the context and the objects in an image. Consequently, it is natural to utilize contextual information to improve object detection.

Many papers have proposed that scene information or relations between objects help object detection [17–19]. However, After the rise of deep learning, There haven't been significant breakthroughs in using contextual information to explore object detection until the emergence of [11,20]. In SIN [11], two kinds of contextual information are introduced: one is scene-level context, the other is instance-level relationships. These two complementary contextual information are combined through GRU [22] to help detection. Hu *et al.* [20] proposes an object relation module for object detection. By modeling the interdependencies between object appearance features and object geometry features, the object relation module can be used for instance recognition.

Most object detectors include CNN feature extraction, proposal classification, and duplicate detection removal. Actually, Using contextual information is working in the proposal classification part. Another way to improve object detection is promoting the quality of the features of ConvNet backbones. At present,

many works are focusing on increasing the receptive field and semantic information of the features extracted from ConvNet backbones [12–15]. To involve multi-scale features, FPN [12] utilizes the hierarchical feature maps from different depths of CNN. DES [13] augments the low-level feature maps of VGG16 with strong semantic information which is trained by week bounding-box level segmentation ground-truth. In order to make the feature maps own higher resolution and larger receptive field at the same time, DetNet [14] designs a new backbone. RFB [15] adds dilated convolution layers on the basis of SSD [9] to effectively increase the receptive field of the feature maps.

Attention can be seen as a way of allocating limited computational resources to the most useful parts of an image. Therefore, attention can be used to improve the quality of the features of ConvNet backbones by selectively emphasizing the informative features and suppressing noises. However, as far as we know, there is only one work [13] that applies attention mechanism to ConvNet backbones in object detection.

3 Method

In this section, we present the details of the proposed network. Firstly, we describe the graph structure inference part, next elaborate the residual joint attention module.

3.1 Graph Structure Inference

Contextual information plays an important role in accurate object detection. Therefore advanced detectors not only consider object visual appearance, but also take advantage of two kinds of structured contextual information: scene-level information and object relationship information. SIN [11] is one of them which considers object detection as the problem of graph structure inference. Given an image, the objects will be treated as graph nodes while the relationships between the objects will be regarded as graph edges jointly under the supervision of the scene context formed by the whole image. More specifically, an object will receive information passed from other objects and scene which is closely related to it. By this way, the object state is finally confirmed by both its appearance features and the contextual information. For encoding different information into objects, SIN chooses Gated Recurrent Units (GRU) [22] as the tool of graph structure inference. The graph structure inference part is shown in Fig. 1. The specific operation steps are described as follows.

Initially, an image is passed through pre-trained VGG16 and the residual joint attention module. The features map named attention5_3 is extracted and then sent to the graph structure inference part. After RPN, a fixed number of RoIs (Region of Interest) are obtained. To get the descriptors about the graph nodes of 4096 dimension, operation of RoI pooling followed by an FC layer is performed on per-RoIs. The conv5_3 feature map is extracted as the scene by the same layer as the graph nodes. As for the descriptors of the graph edges

of 4096 dimension, object-object relationships are modeled by both the spatial features and the visual features of the objects. Eventually, GRU whose input and initial state are respectively the 4096-dimension scene or the edge vectors and the 4096-dimension object vectors iteratively updates two steps to determine the node final state.

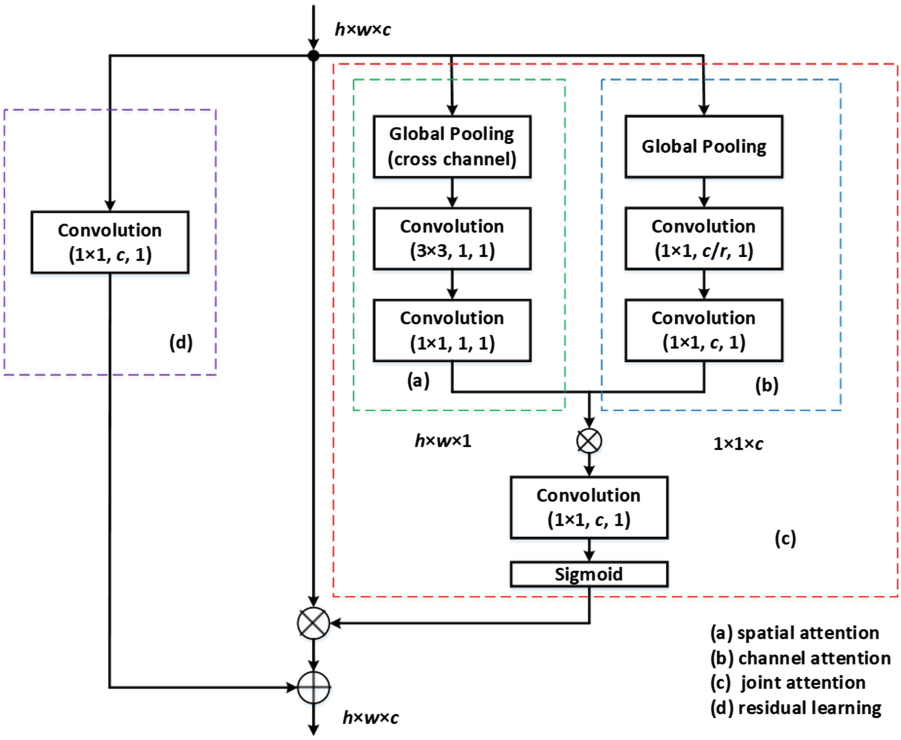

Fig. 2. The structure of residual joint attention module. The three items in the block of a convolution layer are filter shape, filter number, and stride.

3.2 Residual Joint Attention Module

It can be seen from Fig. 2 that our residual joint attention module is the union of the spatial attention, the channel attention, and the residual learning. The spatial attention aims at choosing spatially important features (not related to channels), while the channel attention is dedicated to seeking vital channels for our task. The ultimate goal of them is promoting the quality of the features of ConvNet backbones by performing feature recalibration. Intuitively, if they can be compatible and complementary to each other in functionality, the combination of the spatial attention and the channel attention should apply attention

mechanism to every pixel of a specific feature map leading to better performance than any single attention. At the same time, the residual learning is proposed to keep good attributes of the original features. To the end, we package the spatial attention, the channel attention, and the residual learning into a module which can conveniently be embedded everywhere in a CNN with only a small number of additional parameters.

Mathematically, let $X \in \boldsymbol{R}^{h \times w \times c}$ be the input to a residual joint attention module where h, w, c respectively denotes dimension in height, width, channel of the input feature map. Then X goes through three branches: the channel attention branch which produces a weight map $C \in \boldsymbol{R}^{1 \times 1 \times c}$, the spatial attention branch which produces a weight map $S \in \boldsymbol{R}^{h \times w \times 1}$, and the residual learning branch. Eventually, we select a natural way to combine two weight maps together to get the final weight map $A \in \boldsymbol{R}^{h \times w \times c}$:

$$A = C \times S \tag{1}$$

Next, we describe the designs of three branches in details.

Channel Attention Branch. Our channel attention branch is derived from SENet [16] aiming at promoting the quality of features of its convolutional neural networks by effectively modeling the interactions between the channels from a specific layer of a network. There are two main steps in it. First is squeeze operation (0 parameters) which produces a channel descriptor by squeezing global spatial information named GAP. Squeeze stage will produce $Z \in \boldsymbol{R}^{1 \times 1 \times c}$:

$$Z_i = \frac{1}{h \times w} \sum_{h,w} X_{hwi} \tag{2}$$

Next is excitation operation ($\frac{2c^2}{r}$ parameters) which aims to capture the interactions between the channels. We use two convolutions to generate $C \in \boldsymbol{R}^{1 \times 1 \times c}$:

$$C = \mathrm{RELU}(W_2 \times \mathrm{RELU}(W_1 Z)) \tag{3}$$

where $W_1 \in \boldsymbol{R}^{1 \times 1 \times \frac{c}{r}}$, $W_2 \in \boldsymbol{R}^{1 \times 1 \times c}$ (to simplify the notation, bias terms are omitted). In our work, we keep $r = 4$ to reduce parameters.

Spatial Attention Branch. To explicitly modeling the interactions between the spatial features of a convolution layer, we imitate SENet [16] to build a spatial attention branch. First step is to eliminate channel information by a global cross-channel averaging pooling operation (0 parameters) producing $M \in \boldsymbol{R}^{h \times w \times 1}$. The operation is defined as follows:

$$M_{i,j} = \frac{1}{c} \sum_{c} X_{ijc} \tag{4}$$

A convolution operation with the kernel size of 3×3 (9 parameters) is applied to M next. This filter purposes to model the interactions between the spatial

features. Lastly, we use a harmonic convolution operation of a 1×1 filter (1 parameter) to obtain $S \in \boldsymbol{R}^{h \times w \times 1}$.

After getting S and C two attention maps, we adopt tensor multiplication to combine them. But the union of the spatial attention and the channel attention is not inherent. So it is necessary to further use a $1 \times 1 \times c$ convolution (c^2 parameters) to make the combination more harmonious. A sigmoid activation function maps the combination into the range between 0.5 and 1.

Residual Learning Branch. In the experiment, we notice that the dot production of the original features and the joint attention map who ranges from 0.5 to 1 will degrade the values of the original features. In fact, the values of the useless features decrease more significantly than the useful features. But to ease the situation, we apply residual learning to the joint attention mechanism. According to ResNet [3], if the attention module can be built as identical mapping, the performance should be no worse without attention. The new output is expressed as:

$$X^{'} = X + A \times X \tag{5}$$

To harmoniously fusing the original features with the weighted features, we deploy a convolutional operation with a $1 \times 1 \times c$ filter (c^2 parameters) for the original features before fusing.

4 Experiments

In our experiments, we evaluate our model on VOC dataset [23]. At the same time, several ablation studies are conducted on our various branches to verify the effectiveness of our method. All experiments are evaluated by using VOC metric with IOU = 0.5.

4.1 Experimental Settings

During training and testing, the proposal number is set to 128 because too many proposals lead to out of memory when inferencing graph structure. Specifically, we follow the popular split which takes the combination of VOC2007 trainval and VOC2012 trainval as the train data, and takes VOC2007 test as the test data. The training steps are set to 130000. In the previous 80,000 iterations, we use a learning rate of 0.0005 while the learning rate is reduced by 10 times for the next 50000 iterations. We use momentum gradient descent with momentum 0.9 and batch size of 1 to train the parameters of our network.

4.2 Overall Performance

The results are shown in Table 1. To illustrate the superiority of our method, the ConvNet backbone for all methods in the Table 1 is VGG16. Comparing the results of the baseline, our mAP is higher than SIN [11], which proves that the residual joint attention module really helps our detector to achieve better

Table 1. Overall performance on VOC2007 test.

Method	Faster R-CNN [6]	ION [24]	SIN [11]	Shrivastava *et al.* [25]	Ours
Backbone	VGG16	VGG16	VGG16	VGG16	VGG16
mAP	73.2	75.6	76.0	76.4	76.7
aero	76.5	79.2	77.5	79.3	79.5
bike	79.0	83.1	80.1	80.5	80.4
bird	70.9	77.6	75.0	76.8	76.4
boat	65.5	65.6	67.1	72.0	68.4
bottle	52.1	54.9	62.2	58.2	63.4
bus	83.1	85.4	83.2	85.1	86.0
car	84.7	85.1	86.9	86.5	86.9
cat	86.4	87.0	88.6	89.3	88.3
chair	52.0	54.4	57.7	60.6	59.8
cow	81.9	80.6	84.5	82.2	85.5
table	65.7	73.8	70.5	69.2	71.4
dog	84.8	85.3	86.6	87.0	86.1
horse	84.6	82.2	85.6	87.2	86.5
mbike	77.5	82.2	77.7	81.6	77.1
perpon	76.7	74.4	78.3	78.2	78.6
plant	38.3	47.1	46.6	44.6	50.2
sheep	73.6	75.8	77.6	77.9	77.3
sofa	73.9	72.7	74.7	76.7	74.1
train	83.0	84.2	82.3	82.4	82.8
tv	72.6	80.4	77.1	71.9	75.1

detection accuracy. Interestingly, on some specific classes, it is found that our model performs very well including *aero, bird, bus, chair, plant* and so on. Our method is also better than ION [24] which is a network with explicitly modeling of contextual information using RNN, and Shrivastava *et al.* [25] which exploits segmentation information in the framework of Faster R-CNN. We show some detection examples in Fig. 3. The top column is the results of the original SIN, and the bottom column is the results of our network. From these examples, it can see that our method is good at detecting objects in complex situations like a dog only with a head, a blurry ship, obscured cows and obscured sheep. These also directly indicate that our residual joint attention module makes the original inapparent features more powerful and differentiated through feature recalibration.

Table 2. Ablation studies on VOC2007 test.

Method	Baseline	Baseline+SA	Baseline+CA	Baseline+SA+CA	Ours
mAP	76.0	76.2	76.3	76.5	76.7
aero	77.5	78.8	78.0	78.7	79.5
bike	80.1	79.9	80.2	79.5	80.4
bird	75.0	76.3	76.2	75.6	76.4
boat	67.1	67.5	65.7	69.5	68.4
bottle	62.2	61.9	61.2	61.5	63.4
bus	83.2	85.9	86.4	85.6	86.0
car	86.9	87.0	86.9	86.7	86.9
cat	88.6	89.3	87.8	89.2	88.3
chair	57.7	60.2	61.1	59.7	59.8
cow	84.5	83.1	84.2	84.5	85.5
table	70.5	70.9	71.0	70.0	71.4
dog	86.6	84.2	86.3	86.8	86.1
horse	85.6	87.6	87.1	86.0	86.5
mbike	77.7	77.4	77.6	78.8	77.1
perpon	78.3	78.1	78.3	78.4	78.6
plant	46.6	51.0	47.9	47.6	50.2
sheep	77.6	78.6	77.7	76.3	77.3
sofa	74.7	72.2	73.4	74.2	74.1
train	82.3	78.8	83.1	83.4	82.8
tv	77.1	75.8	76.4	77.0	75.1

4.3 Ablation Studies

In order to verify the effectiveness of each branch in our proposed method, we conduct several ablation studies which still use the same dataset settings as above. Table 2 shows the results of different branches, where Baseline stands for SIN, SA stands for the spatial attention branch, CA stands for the channel attention branch. Comparing with the baseline, there is a slight increase by adding any kind of attention mechanism to the baseline. This shows that the feature recalibration through attention mechanisms is effective. What's more, The combination of the spatial attention and the channel attention improves more than any single attention which proves our preliminary conjecture that the spatial attention and the channel attention are complementary and compatible. Similarly, the residual learning continues to optimize our model that demonstrates the validity of the residual learning.

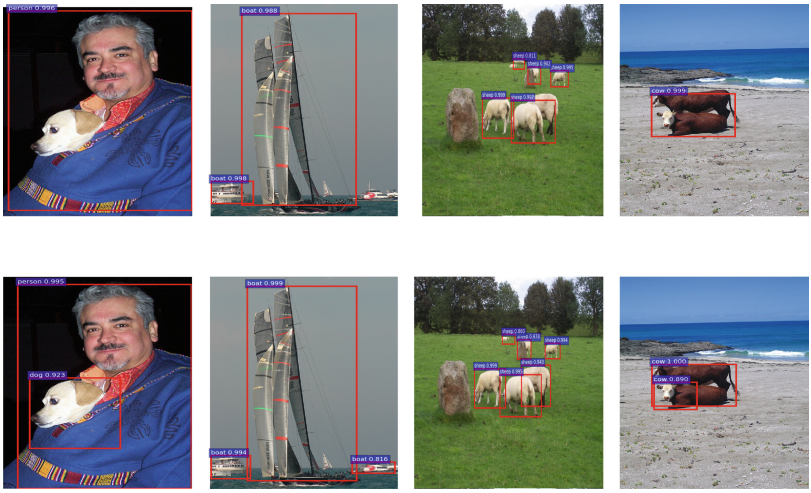

Fig. 3. Examples of detection results. Top: SIN. Bottom: ours.

5 Conclusion

In this paper, we proposed a residual joint attention module embedded in an advanced network with graph structure inference. The graph structure inference part is used for the detection subnetwork of the detector and the residual joint attention module composed of the spatial attention, the channel attention and the residual learning follows VGG16. Due to the complementarity and compatibility of the spatial attention and the channel attention, the joint attention mechanism more significantly improves the representational power of a network by performing feature recalibration than any single attention. Moreover, the residual learning keeps good attributes of the original features. Quantitative evaluations show that our residual joint attention module boosts model's discriminative power. We hope that this paper can provide reference for researchers to use attention.

Acknowledgment. This work was supported partly by the National Natural Science Foundation of China (61772067, 61472030, 61471032).

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, NIPS 2012, pp. 1097–1105. Curran Associates Inc. (2012)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016 (2016)

4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 142–158 (2016)
5. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference On Computer Vision, ICCV 2015, pp. 1440–1448 (2015)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, NIPS 2015, pp. 91–99 (2015)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017 (2017)
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 3431–3440 (2015)
9. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
10. Redmon, J., Divvala, S., Girshick, R.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 779–788 (2015)
11. Liu, Y., Wang, R., Shan, S., Chen, X.: Structure inference net: object detection using scene-level context and instance-level relationships. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018 (2018)
12. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (2017)
13. Zhang, Z., Qiao, S., Xie, C., Wei, S.: Single-shot object detection with enriched semantics. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018 (2018)
14. Li, Z., Chao, P., Gang, Y., Zhang, X., Jian, S.: DetNet: a backbone network for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018 (2018)
15. Liu, S., Huang, D., Wang, Y.: Receptive field block net for accurate and fast object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018 (2018)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arxiv:1709.01507 (2017)
17. Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A.A., Hebert, M.: An empirical study of context in object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (2009)
18. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008 (2008)
19. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2003 (2003)
20. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018 (2018)
21. Uijlings, J.R., Van, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)

22. Cho, K., Merrienboer, B.V., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: SSST-8 (2014)
23. Everingham, M., Gool, L.V., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
24. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016 (2016)
25. Shrivastava, A., Gupta, A.: Contextual priming and feedback for faster R-CNN. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 330–348. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_20