ATTENTIVE LAYER SEPARATION FOR OBJECT CLASSIFICATION AND OBJECT LOCALIZATION IN OBJECT DETECTION

Jung Uk Kim and Yong Man Ro

Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

ABSTRACT

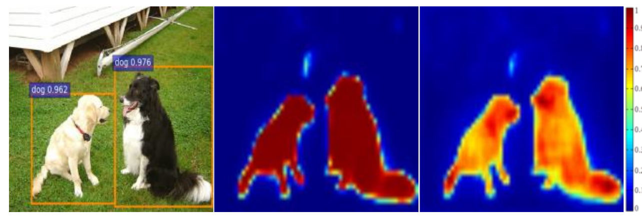
Object detection became one of the major fields in computer vision. In object detection, object classification and object localization tasks are conducted. Previous deep learning-based object detection networks perform with feature maps generated by completely shared networks. However, object classification focuses on the most discriminative object part of the feature map. Whereas, object localization requires a feature map that is focused on the entire area of the object. In this paper, we propose a novel object detection network by considering the difference between the two tasks. The proposed deep learning-based network mainly consists of two parts; 1) Attention network part where task-specific attention maps are generated, 2) Layer separation part where layers for estimating two tasks are separated. Comprehensive experimental results based on PASCAL VOC dataset and MS COCO dataset showed that proposed object detection network outperformed the state-of-the-art methods.

Index Terms**—** Object detection, Attention network, Object classification, Object localization, Layer separation

1. INTRODUCTION

Object detection is one of the most actively studied areas in computer vision [1, 2, 3, 4, 24]. In object detection, object classification and object localization are performed. The purpose of object classification is to separate objects from the background and categorize the object class (e.g. human and car). Whereas the purpose of object localization is to locate objects by drawing bounding-box around the object.

With the development of a deep convolutional neural network (CNN), object detection performance has remarkably improved. CNN based object detection networks can be categorized into two types [17]: two-stage detection network and one-stage detection network. In the two-stage detection network, the object detection process consists of two steps. The first step is for generating RoI regions which are likely to be the objects in the image. Fast R-CNN [5], Faster R-CNN [6], and R-FCN [7] are representatives of the two-stage detection networks. In the one-stage detection network, object classification and object localization are performed using pre-defined anchors. YOLO [8] and SSD [9] are representatives of the one-stage detection networks. A common characteristic of all CNN based object detection is that feature maps are encoded within a sharing network. In other words, object classification and object localization are conducted using the same feature maps of the sharing network.



(a) (b) (c)

Fig. 1. (a) Example of the detection result on PASCAL VOC 2007 *test* set [13]. Example of the proposed attention map of the (b) object localization and (c) object classification. Attention map of the object localization is activated on the overall object area, whereas attention map of the object classification is activated on the part of the object.

However, it is obvious that there are different characteristics of object classification and object localization. As introduced in [10, 11, 23], object classification concentrates on the most discriminative parts of objects to predict a correct object category. On the other hand, object localization concentrates on the whole object region so that it can draw a bounding-box containing the entire objects. For this reason, it is reasonable to devise an object detection network which takes the different characteristics of two tasks into account.

In this paper, we propose a new object detection network considering the differences between the two tasks mentioned above. First, we devise a network in which sharing layers are less semantic in both object classification and object localization. Based on the feature maps of the last sharing layer, we deploy two attention networks. Two types of attention maps are generated to fit object classification and object localization. Fig. 1 shows an example of the proposed two types of generated attention maps for two tasks. Fig. 1 (b) shows the generated attention map for object localization, which concentrates on the entire area of an object, which is

This work was partially supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01779, A machine learning and statistical inference framework for explainable artificial intelligence). The corresponding author of this project is Yong Man Ro.

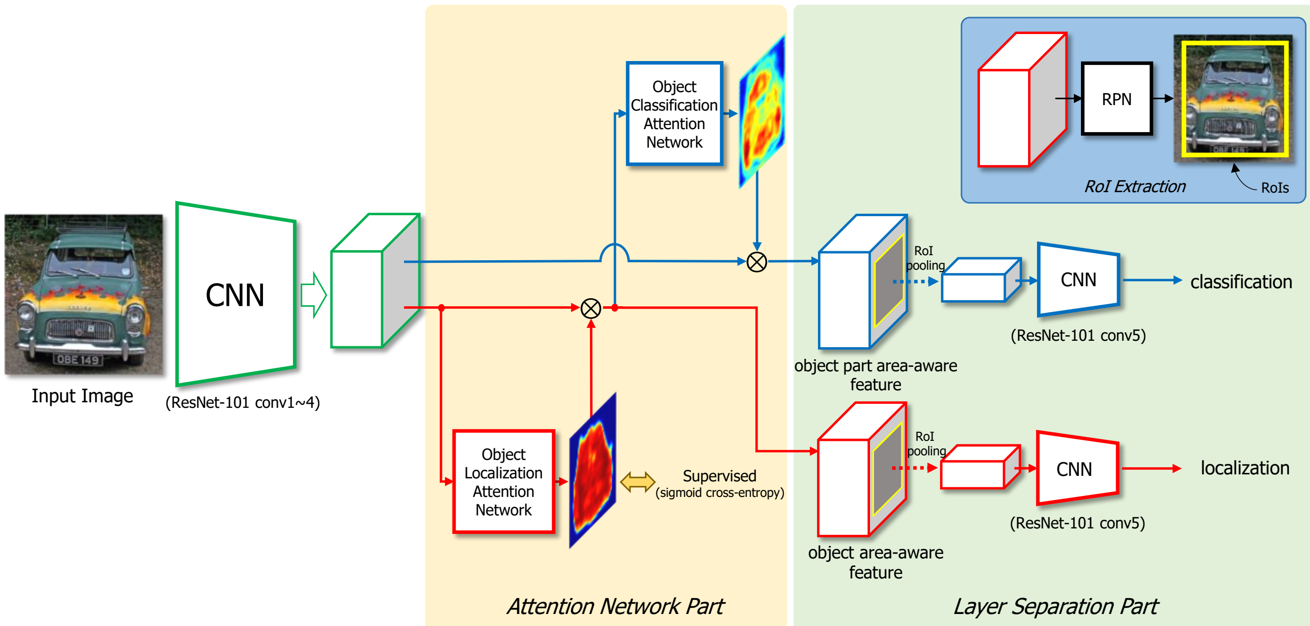


Fig. 2. Overview of the proposed object detection network.

applied (multiplication) to the feature maps of the last sharing layer to perform object localization. Note that the feature under consideration of the object localization attention map is fed into the Region Proposal Network (RPN). Fig. 1 (c) shows the generated attention map for object classification, which concentrates on partial areas of the object. It is also applied to the feature maps of the last sharing layer to perform object classification.

Then, we separate the layers which have respective semantics in object classification and object localization. Object localization is performed within the object localization semantic layer using the feature multiplied to the object localization attention map. In a similar manner, object classification is performed within the object classification semantic layer using the feature multiplied to the object classification attention map. As a result, the proposed method improves the object detection performance.

The rest of this paper is organized as follows: We describe the proposed network in section 2, experimental results are performed to verify the performance of the proposed object detection network in section 3, and conclusions are drawn in section 4.

2. PROPOSED METHOD

Fig. 2 shows an overview of the proposed object detection network. As shown in the figure, when the input image is given, the feature maps are encoded in the less semantic layers of the backbone network. With the feature maps, two attention maps for two tasks (object classification and object localization) are generated in the attention network part of the Fig. 2. Next, two attention maps are multiplying the feature maps. With two feature maps, in the layer separation part, object classification and object localization are separately conducted. There are two separated semantic layers of the backbone network with the same structure. In this paper, we used a backbone network as ResNet-101 [12]. To perform the object classification and object localization, the conv1~conv4 and conv5 of ResNet-101 are used as less semantics layers and semantics layers, respectively. Detailed descriptions of each part are given in the following sections.

2.1. Attention network part

In this section, we describe the attention networks for two tasks in detail. In the first step, based on the feature maps of the last block of ResNet-101 conv4, attention map for object localization is preferentially generated in the object localization attention network. More specifically, object localization attention network has consisted of three 1×1 convolutional operations, two are 2,048 channel layers and the last one is 1 channel layer. Finally, the attention map for object localization is generated by applying the sigmoid activation function.

In order to allow the attention map for object localization to focus on the entire area of the object, we employ 1 channel binary segmentation map; 1 for pixels of objects and 0 otherwise. Using a binary segmentation map, the object localization attention network is trained in a supervised manner. Attention map for object localization is simply multiplied channel by channel on the last block of conv4 of ResNet-101. We denote the feature map which is multiplied by object localization attention map as object area-aware feature.

In the next step, object area-aware feature feed into object classification attention network. Note the structure of the object classification attention network is same as object

Table 1. Results on VOC 2007 *test* set [13]. All methods were trained on union of VOC 2007 *trainval* and VOC 2012 *trainval*. The backbone network of first five methods is VGG16 [19] and latter four methods is ResNet-101 [12]. *: we re-implemented network.

method	backbone	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster R-CNN [6]		73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
SSD 300 [9]		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
MR-CNN [16]	VGG16	78.2	80.3	84.1	78.5	70.8	68.5	**88.0**	85.9	87.8	60.3	85.2	73.7	87.2	86.5	85.0	76.4	48.5	76.3	75.5	85.0	81.0
LED 300 [22]		78.7	82.7	86.5	76.9	71.7	51.7	87.1	88.0	89.9	60.8	84.0	74.9	88.2	87.9	**85.1**	81.3	52.5	79.5	80.8	87.6	76.8
SSD 512 [9]		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 R-CNN [12]		76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	89.8	45.7	**87.8**	69.4	**88.3**	**88.9**	80.9	78.4	41.7	78.6	79.8	85.3	72.0
Faster R-CNN * [12]	ResNet-101	77.2	78.3	83.4	77.4	65.6	68.4	81.4	87.7	84.7	61.7	80.5	73.6	83.6	87.8	82.6	78.7	51.6	79.3	79.5	79.1	76.4
R-FCN [7]		79.5	82.5	**93.7**	80.3	69.0	69.2	87.5	88.4	88.4	65.4	87.3	72.1	87.9	88.3	81.3	79.8	54.1	79.6	78.8	87.1	79.5
Proposed Method		**80.1**	82.1	88.7	**81.6**	**74.2**	**69.4**	86.8	**88.5**	**89.9**	**67.1**	85.1	**77.2**	84.4	85.1	84.0	76.1	53.2	79.6	**80.9**	**87.7**	**82.2**

localization attention network. As a result, the attention map for object classification is generated by focusing the most discriminative object part in term of object classification. The attention map for object classification is also multiplied channel by channel by the last block of conv4 of ResNet-101. We denote the feature map which is multiplied by object classification attention map as object part area-aware feature.

2.2. Layer separation part for two tasks

As shown in Fig. 2, in the layer separation part, object classification and object localization tasks are conducted with the aforementioned two types of features (i.e. object area-aware feature and object part area-aware feature). Note that object area-aware feature highlights objects while suppress non-objects (i.e. background). The object area-aware feature feed into the RPN, which extracts candidate the region of interests (RoIs) of the object. This is because, in order to extract the RoIs, the entire object region activated feature is required [12].

Based on the predicted RoIs in the RPN, RoI pooling is conducted on two types of features. Finally, two RoI pooled features are conducted two tasks through two different layers. RoI pooled features in object area-aware feature and object part area-aware feature are used for object localization and object classification, respectively. Structure of two different layers is the same as the conv5 of the ResNet-101. However, these layers do not share parameters.

2.3. Training objectives

When training our network, objective function is used as

$$L = L_{RPN} + \lambda_1 L_{cls} + \lambda_2 L_{loc} + \lambda_3 L_{att-loc}\,, \tag{1}$$

where λ_1, λ_2, and λ_3 are hyper-parameters to control the objective function. L_{RPN}, L_{cls}, and L_{loc} are the objective function of the RPN [12], object classification, and object localization, respectively. L_{cls} is a softmax cross-entropy loss over (C+1), which is the number of object classes and background. We used L_{loc} as smoothed L_1 loss to learn the bounding-box location (x, y, w, h) of the object. In $L_{att-loc}$, sigmoid cross-entropy loss is used to make attention map for object localization to activate the entire areas of the object. We simply set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in Eq. (1) and jointly learn the proposed network through backpropagation.

3. EXPERIMENTS

3.1. Experimental setup

In the experiment, we evaluated the proposed object detection network on three publicly available datasets; PASCAL VOC 2007 [13], VOC 2012 [14], and MS COCO [15]. These datasets contain images which can be seen in a general environment, which are well-known datasets for measuring object detection performance [21]. Specifically, in the VOC 2007 [13] and VOC 2012 [14], they have 20 object categories. However, as in [4], because VOC 2007 and VOC 2012 had lack of the number of instance segmentation annotation, we additionally used [18] which has PASCAL VOC instance segmentation annotations. In the MS COCO [15], it has 80 object categories. It has fully bounding-box annotations and instance segmentation annotations.

In the VOC 2007 and VOC 2012, we adopted mean Average Precision (mAP) to compare the performance, following [12]. In the MS COCO, we reported AP on objects, following the definitions in [15]. We also evaluated AP when the overlap threshold was 0.5 (i.e. AP_{50}), AP on objects of the small, medium, and large sizes (i.e. AP_S, AP_M, and AP_L). We used the backbone network as ResNet-101 [12].

3.2. Results on PASCAL VOC

We first trained our proposed network without the attention network part using the training and validation sets (*trainval*) of VOC 2007 and VOC 2012, and fine-tuned the proposed network with the attention network part using [18]. All tests with PASCAL VOC were conducted with VOC 2007 *test* set.

Table 1 showed the results of the proposed method. We compared the state-of-the-art object detection networks [6, 7, 9, 12, 16, 22]. As shown in the Table 1, performance of the proposed method was 80.1 mAP. It was 0.6 mAP higher than the R-FCN [7] with ResNet-101 as the backbone. By generating attention maps to focus the different area according to two tasks and separating semantic layers, object detection performance was improved.

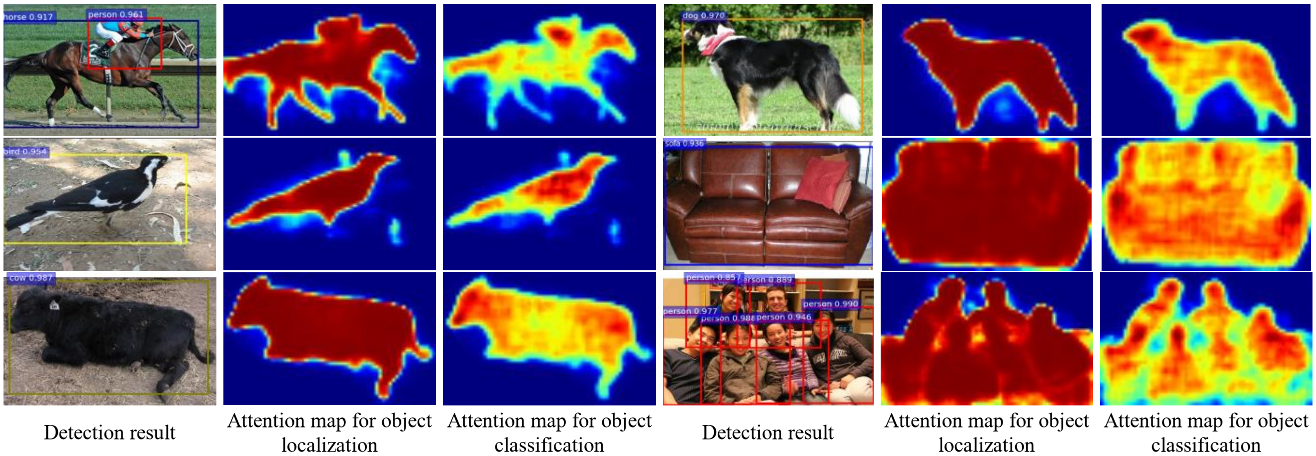


Fig. 3. Visualization of the detection results on VOC 2007 *test* set [13], attention maps for object localization and object classification. To display the results, all the figures are resized to the same size.

Table 2. Ablation studies on VOC 2007 *test* set [13]. All methods were trained on union of VOC 2007 *trainval* and VOC 2012 *trainval*. 'att' and 'split' denote attention network part and layer separation part for two tasks (object classification and object localization), respectively.

method	att	split	mAP
Faster R-CNN [12]	-	-	76.4
Proposed Method	✗	✗	77.2
	✗	✓	79.1
	✓	✓	**80.1**

To the next, we performed ablation studies to demonstrate the effectiveness of the attention network part and layer separation part of Fig. 2. It is shown in Table 2. We compared the Faster R-CNN [12]. We re-implemented [12], which has no attention network part and layer separation part. As shown in Table 2, when there was only a layer separation part, performance was 79.1 mAP. It improved 2.7 mAP compared to [12] and 1.9 mAP compared to the result of re-implementation. When the attention network part was additionally introduced, performance was 80.1 mAP which was 1.0 mAP additional improvement. The performance improved by separating the layers to perform the two tasks. In addition, attention network enhanced performance by emphasizing areas according to the two tasks.

We additionally visualized the detection results and the attention maps for two tasks. It is shown in Fig. 3. As shown in the figure, as the attention network for object localization was learned supervised manner, attention map for object localization was activated on the entire region of the object. Therefore, object area-aware feature could highlight the object region and suppress the background region. Also, since the attention map for object localization filtered the feature to focus the object area, the attention map for object classification could focus on the part of the object. Interestingly, the attention map for object classification focused on the most sensible object part (e.g. near the head of the animals). It verified that object part area-aware feature could be highlighted most discriminative part, similar with [10, 11, 23].

Table 3. Results on MS COCO *minival* [15]. All methods were trained on the *trainval35k*.

method	AP	AP_{50}	AP_S	AP_M	AP_L
Faster R-CNN [12]	31.6	53.1	13.2	35.6	47.1
Faster R-CNN*	32.8	54.0	12.8	35.9	46.8
R-FCN [7]	34.4	56.9	17.8	38.7	47.3
Proposed Method	**36.2**	**58.3**	**19.7**	**40.1**	**48.1**

3.3. Results on MS COCO

To show the effectiveness of the proposed object detection network, we also conducted on MS COCO dataset. All models were trained on the union of a training set and a subset of validation images (*trainval35k*), following [20]. We evaluated the performance on COCO *minival*. Table 3 showed the results of the detection results. Since our proposed method is two-stage object detection network, we compared with the state-of-the-art two-stage object detection networks; Faster R-CNN and R-FCN. As shown in Table 3, the proposed method achieved 36.2 AP. It was 4.6 AP and 1.8 AP higher than Faster R-CNN and R-FCN, respectively. All metrics (i.e. AP_{50}, AP_S, AP_M, and AP_L) were improved at least 0.8 points. It also verified that the proposed object detection network improved the overall object detection accuracy.

4. CONCLUSIONS

In this paper, we proposed a new object detection network considering the distinct difference in object classification and object localization. We pointed out that object classification focuses on the discriminative part and object localization focuses on entire object areas. Therefore, we introduced attention network to generate the two task-specific attention maps. Features that were multiplied by each two attention maps were fed into the layer separation part and two tasks are performed. As a result, proposed methods were effectively encoded each layer by considering two tasks properties. The experimental results verified that proposed method outperformed state-of-the-art methods.

5. REFERENCES

[1] T. –Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117-2125.

[2] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. “Scalable, high-quality object detection,” *arXiv preprint arXiv: 1412.1441*, 2014.

[3] J. U. Kim, J. Kwon, H. G. Kim, H. Lee, and Y. M. Ro, “Object bounding box-critic networks for occlusion-robust object detection in road scene,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 1313-1317.

[4] Y. Zhu, C. Zhao, H. Guo, and J. Wang, “Attention couplenet: fully convolutional attention coupling network for object detection,” *IEEE Trans. Image Processing*, vol. 28, no. 1, pp. 113-126, 2019.

[5] R. Girshick, “Fast r-cnn,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1440-1448.

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[7] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” *arXiv preprint arXiv: 1605.06409*, 2016.

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

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed, “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21-37.

[10] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2921-2929.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618-626.

[12] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv: 1512.03385*, 2015.

[13] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Sep. 2009.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. “The pascal visual object classes challenge 2011 (VOC2011) results,” http://host.robots.ox.ac.uk/pascal/VOC/voc2011/results/index.html.

[15] T. –Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740-755.

[16] S. Gidaris, N. Komodakis, “Object detection via a multi-region & semantic segmentation-aware cnn model,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1134-1142.

[17] T. –Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” *arXiv preprint arXiv: 1708.02002*, 2017.

[18] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 991–998.

[19] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprints arXiv: 1409.1556*, 2015.

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

[21] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, “Object detection networks on convolutional feature maps,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1476-1481, 2017.

[22] S. Zhang, X. Zhao, L. Fang, H. Fei, and H. Song, “LED: Localization-quality estimation embedded detector,” *in Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 584-588.

[23] F. Wang, M. Jiang, C. Qian, S. Yang, and C. Li, “Residual attention network for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3156-3164.

[24]. J. U. Kim, J. Kwon, H. G. Kim, and Y. M. Ro, “BBC Net: Bounding-box critic network for occlusion-robust object,” *IEEE Trans. Circuits Syst. Video Technol.*, 2019.