Rich Features and Precise Localization with Region Proposal Network for Object Detection

Mengdie Chu[1] , Shuai Wu[1], Yifan Gu[1] and Yong Xu[1],

[1] Computer Science and Technology, Harbin Insitute of Technology, China
chumengdd@126.com

Abstract. Deep Network greatly accelerates the development of object detection. Recent advances in object detection are mainly attributed to the combination of deep network and region proposal methods [1][2][3]. However, the accuracy of object detection on the complicated datasets is still not satisfied, especially on small object detection. This is mainly because of the coarseness of the convolution feature maps. In this paper, we design a new strategy for generating region proposals and propose a new localization method for object detection. Compared with previous baseline detectors such as Fast R-CNN[4] and Faster R-CNN [5]， Our method makes use of the adjacent-level feature maps at all scales to generate region proposals and also adopts the cascaded region proposal network (RPN) to fine-tune the location of the bounding box. Compared with other state-of-the-art methods, our method achieves the best recall and object detection accuracy.

Keywords: Object detection, Proposal, Features, Localization, Cascaded

1 Introduction

Object detection is one of the most fundamental researches in computer vision [6][7]. The purpose of object detection is to detect and localize all instances of pre-defined classes in one image [8][9]. Basically, most object detection methods localize the instances via bounding boxes. Object detection problem can also be treat as the classification problems with the sliding window strategy [10][11]. However, sliding window is very time-consuming because the windows are generated from all possible locations with different scales and aspect ratios. Recently, a two-stage approach has been proposed to integrate recognition and localization stages [12] into a region-based convolution neural network. It firstly generates a series of object proposals by using a proposal generator, and then determines whether an instance with exact class label exists in the ROI. Both stages are on basis of a deep neural network.

Recently, Faster R-CNN [5] is proposed to integrate proposal generation stage, classification stage and bounding boxes regression stage into a unified process through a deep convolutional network. It applies the RPN substructure to generate proposals and uses fast-RCNN to perform classification and bounding boxes regression. Faster R-CNN [5] achieve impressive performance on public benchmarks and

has become the baseline framework. However, Faster R-CNN[5] suffers from small object detection, this is mainly because it applies the deep convolution feature map which contains more semantic information while has limited effect on localization to perform the final classification and bounding boxes regression. The target region of small object after mapping on the feature map is too coarse to get expectable performance [13]. A satisfactory object detection system needs a proposal generator that can obtain a small number of region boxes with high recall [14][15][16]. Deeper convolutional layers usually contain powerful semantic information and are benefit for finding the region of interests with high recall. However, deeper layers are not appropriate for localization because of the limited feature map size [17]. In comparison, the lower layers could effectively localize the region of interests [18].

Considering the two situations above, we apply a multi-scale feature extraction strategy which combines the advantage of both deeper layers and lower layers. In this paper, we apply the strategy proposed in [19], which designs a top-down pathway that increasingly expand the top feature map and merge them with the corresponding layers in the backbone Network. Then all the merged feature map with multi-scale sizes will be used to generate proposals. The advantage of the top-down pathway is that it can make full use of the low-resolution feature map with powerful semantic information and high resolution feature map with efficient localization capability [20]. According to the empirical experience in classification problem, multiple representations of objects are very beneficial for recognition. Our method predicts the specific-scale proposals independently on every new feature map with multi-scale sizes.

In addition, the object detection task needs precise localization which is free effect from translation variant. For example, translation of an object inside a candidate box should produce meaningful responses for describing how well the candidate box overlaps the object [21][22]. In order to obtain high-quality proposals, we apply a more precise localization strategy [23]. After firstly generating proposals by region proposal network(RPN), we treat these proposals as new anchors to be sent to the RPN again, which can finely tune the location of these proposals. We called this operation cascaded RPN. We evaluate our method based on the VOC detection benchmark [24] with three different baseline models. More convincingly, we verify the effect of these two improvements respectively. First, we use the top-down pathway to generate proposals without fine-tuning the location, our model significantly increases the accuracy by 0.8 points on the ResNet101 baseline network. And then, we add the cascaded region proposal network without using new features, the final result also improves 0.9 points. So, we can find that our improvements are both effective. For the final object detect, we add these two improvements together into the original Faster R-CNN, and improves the mAP by 2.0 point on ResNet101 baseline network. These results suggest that our method is an effective way for improving object detection accuracy.

2 Our Approach

In this paper, we apply the strategy in [19] that merges the adjacent-level semantic feature maps to generate proposals. However, our model is different with [19] in

terms of region proposals network (RPN). We apply the cascade-RPN which could be considered as two-stage RPN after each merged feature map in [19]. The cascade-RPN treats the proposals in first stage as new anchors and sends them into the second-stage RPN. Finally, these new proposals are classified and adjusted based on the detection module. We explain our two improvements in detail respectively.

2.1 Feature Production

In order to exploit the advantage of different layers and generate richer features, we apply the strategy in [19] which is illustrated in Figure 1. Because feature maps of different-levels have different implications, it designs a top-down pathway to generate powerful semantic feature maps of multi-scales and merge them with the corresponding layers in the backbone network. In real application, we take a single-scale image of arbitrary size as input, and regard the corresponding at multiple level feature map as outputs. The top-down pathway is independent of backbone convolutional architectures which in this paper we apply ResNet model. In ResNet, there are many output layers with the same size and we say these layers are in the same network stage, ResNet has five stages in all. As for the strategy in [19], we use the output of the last layer of each stage as our reference set of feature maps and this strategy enriches feature representation. Because every layer of the same stage has the same feature map size and the deepest layer of each stage should represent the strongest features. For conv2_x, conv3_x, conv4_x, conv5_x, we remark the output of these last residual blocks as {F2, F3, F4, F5} and each of them has strides of {4, 8, 16, 32} respectively. In our paper, we do not include the output of conv1 into the reference set, owing to its large memory footprint.

Higher-level feature maps contain powerful semantic information while coarse boundary information. Lower-level feature maps have weak semantic information but are more beneficial in localization because of the limited subsampling times. The top-down pathway deconvolutes the spatial resolution of F5 by a factor 2 to powerful semantic feature maps with multi-scales. As shown in Fig.1. the deconvoluted [25] feature maps are merged with the corresponding feature map in the backbone network (In this paper, ResNet). Before merging, the lower-level feature map undergoes a 1*1 convolutional layer to reduce channel dimensions. This process is iterated until the finest resolution map is generated. Because F5 is the highest-level feature map, there is no higher-level feature map to merge with it. To start the iteration, we simply attach a 1*1 convolutional layer after F5 to produce the coarsest resolution map. Expect for those operations, in order to reduce the aliasing effect of up-sampling, we also connect a 3*3 convolution on each merged map to generate the final feature map. So, we regard these final merged feature maps as {M2, M3, M4, M5}, corresponding to {F2, F3, F4, F5} and they all have the same spatial sizes.

Region Proposal Network is a huge improvement for generating proposals in Faster R-CNN [5]. It is a sliding-window class-agnostic object detector. We adapt RPN to multi-scale feature maps {M2, M3, M4, M5} instead of the single-scale feature map. Like Faster R-CNN [3], we also attach a 3*3 convolutional layer and two siblings 1*1 convolution to every new feature map. For our new generated feature maps, every

new merged feature represents special scales respectively. It is not necessary to have multi-scale anchors on these merged feature maps. So [19] chooses a single scale for each new feature map. Formally, they set the anchors to have scales of {32, 64, 128, 256, 512} pixels on {M2, M3, M4, M5, M6} respectively. M6 is simply a stride of two subsampling of M5. It is used only for covering a larger anchor scale of 512, and still different aspect ratios {1:2, 1:1, 2:1} are used at each level. So, there are 15 anchors over these features in total.

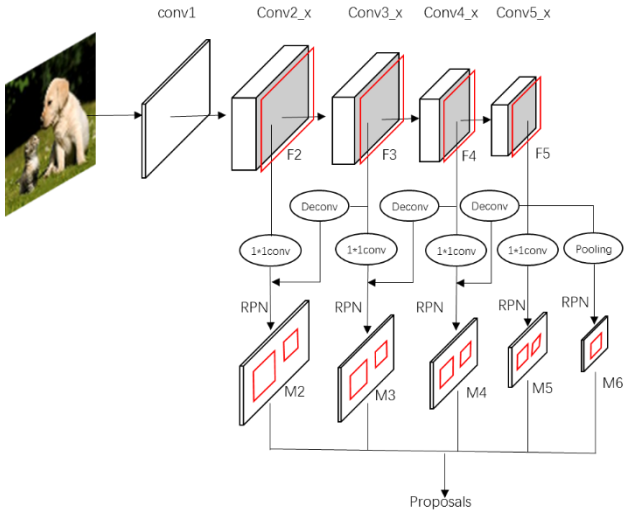

Fig. 1. Feature production architecture

2.2 Fine-tuning of Proposals

An ideal proposals generator should generate as few proposals as possible while covering almost all object instances. With the help of strong abstraction ability of CNN, RPN could generate limited number of proposals with high recall. However, the output of general object proposal algorithms still contains a large proportion of background regions [26]. The existence of many negative samples makes the representation feature less sensitive to identify between the object category and background, causing many false positives on ambiguous object categories. In addition, due to the resolution loss caused by CNN pooling operation and the fixed aspect ratio of sliding windows, RPN is weak at covering objects with extreme scales and shapes. In this paper, we finely tune the location of proposals using cascade-RPN, which makes the localization more precise and is illustrated in Figure 2.

In our paper, after each new merged feature maps, we apply a two-stage cascade RPN, we treat the proposals generated in first stage as new anchors, and then input these new anchors into the second stage RPN to generate the final proposals. This work makes the object proposals more compact and better localized.

As illustrated in Fig 2, the first stage RPN is trained regularly in a sliding window manner to generate a series of anchors at special scales and various aspect ratios in an image, with the same parameters as in [5]. For every generated anchor, it is followed by a region of interest (ROI) pooling layer to extract a fixed-length feature vector from the feature map, which the ROI pooling layer is simply the special-case of the spatial pyramid pooling layer used in SPPnets [27] in which there is only one pyramid level. And each feature map is fed into two sibling fully-connected layers, one for estimating whether this anchor enwrap an object or not, and another one for estimating four real-valued numbers for localization of the object. In details applications, we use three different aspect ratios for a special scale on each new merged feature, so the Bbox regression outputs has 4*3 channels encodes the 4 coordinates of 3 anchors, and the Softmax-layer outputs has 2*3 channels that estimate probability of object or non-object for each proposal. When we get the proposals from the first-stage RPN, we feed these proposals into the second stage RPN again without any operation, treat these proposals as new anchor location for fine-tuning the location of these anchors and generated final proposals.

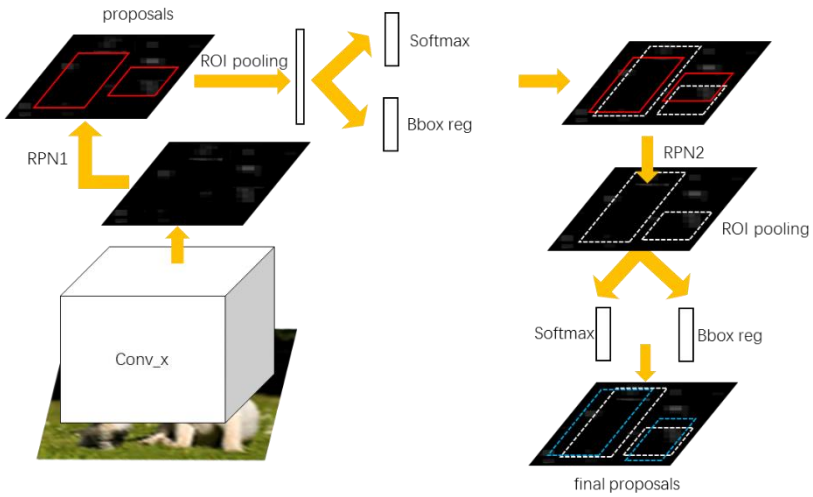

Fig. 2. Cascade RPN architecture

For each image, there are about 30K candidate boxes with different sizes and aspect ratios. After each candidate box is scored and adjusted, some region proposals extremely overlap with each other. To reduce redundancy, we adopt greedy non-maximum suppression(NMS) [28] on the regions in terms of their scores. Formally, an anchor is assigned a positive label if it has the highest IoU for a given ground-truth box or an IoU over 0.7 with any ground-truth box, and a negative label if it has IoU lower than 0.3 for all ground-truth boxes. We select the top-k ranked region proposals for detection after NMS. We train the detection network using top-300 region proposals.

2.3 Implementation

We train the RPN net and detection network with shared convolution network. We use the NMS method to restrict the number of positive and negative samples for the generated proposals. And then input these proposals into the detection network.

When we train the detection network, we use the same convolutional network as in RPN, so it can achieve shareable parameter, which have 13 shareable convolutional layers in VGG16 [29], 49 and 100 convolutional layers in ResNet50[30] and ResNet101[30] respectively, all of three basic models drop the full connected network. For fair comparisons with original RPN[5], we set the same parameter with it.

3 Experimental Evaluation

We evaluate our approach on PASCAL VOC 2007 and 2012 challenges [24] detection benchmarks. These datasets consist of about 20k trainval images and 5k test image over 20 categories. We train our models bases on VGG16[29], ResNet-50 [30]and ResNet-101[30] which is pre-trained on the ILSVRC CLS-LOC dataset [31], and compare results with other state-of-the-art methods [32][20][18][21]. Object detection accuracy is measured by mean Average Precision (mAP). We also provide deep analysis of our approach affection to object proposal and detection performances.

We resize the shortest side to 600 pix, and the longest side to 1000 pix, which are same with Faster R-CNN [5]. We fine-tune the resulting model using SGD with a weight decay of 0.0005 and a momentum of 0.9. We train the model with 1 GPUs. We fine-tune our model using a learning rate of 0.001 for 20k mini-batches and 0.0001 for 10k mini-batches on VOC.

3.1 On the Impact of Feature Representation

In order to verify our merged feature strategy is effective, we compute the recall of proposals at different IoU ratios with ground-truth boxes. The results are shown in Figure 3.

We set the IoU ratio is 0.5 as baseline. The plots show that our merged feature method behaves gracefully when the number of proposals drops from 2000 to 300. This explains why our method has a good ultimate detection Map when using as few as 300 proposals. As we can see, this property is mainly attributed to the classify of the RPN. The recall of selective search drops quickly than our new method when the proposals are fewer. So, our new feature represent is feasible.

We compare our first improvement based on the new feature generation with basic Faster R-CNN on PACAL VOC 2007 and 2012. We test this method on three different models (VGG-16, ResNet-50, ResNet-101). The comparative results are shown in Table 1.

Faster R-CNN based in VGG-16 achieves a mAP of 73.8%, and 76.4% on ResNet-50 and 76.7% on ResNet-101. And our improvement method achieves a mAP of 74.8% on VGG-16, 1 points higher than Faster R-CNN, for the Res50 and Res101, are 0.9 points higher and 0.8 points higher respectively than original feature. As we

shown in Table 3, this is because proposals generated by our new feature are more accurate than traditional single-scale feature map. Reasonable resolution of our new feature makes for better object localization, especially for small object. These results demonstrate that our feature merge strategy can give excellent performance for high-quality proposals and can get better accuracy.

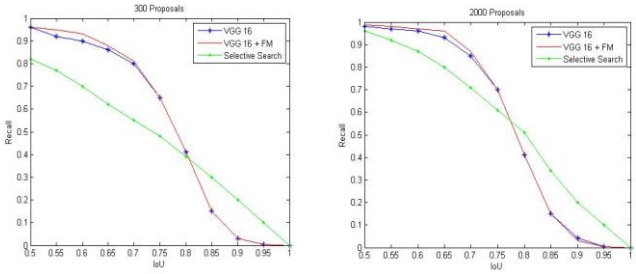

Fig. 3. Recall vs IoU overlap ratio on the PACAL VOC 2007 test set

Table 1. Results on PASCAL VOC 2007 test set. Baseline denote the original Faster R-CNN method, FM denotes our new feature merged method. We set IoU = 0.5

Approach	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	…
Baseline+VGG16	73.8	78.2	79.2	73.3	62.5	57.7	84.0	84.1	86.9	56.0	81.4	64.9	83.7	…
Baseline+Res50	76.4	79.1	79.9	76.5	69.2	63.1	85.1	85.4	89.2	58.4	83.5	67.5	87.1	…
Baseline+Res101	76.7	78.7	79.6	78.8	66.4	64.4	86.0	86.4	87.4	62.2	83.8	68.4	88.0	…
FM+VGG16	74.8	86.5	83.1	77.4	59.1	57.3	79.1	78.7	91.2	54.7	79.6	58.8	89.7	…
FM+Res-50	77.3	86.8	83.8	76.6	65.8	59.6	81.8	82.7	90.8	60.0	81.1	64.2	88.1	…
FM+ Res-101	77.5	88.7	85.2	76.7	64.9	61.3	85.1	84.0	90.1	59.8	82.7	61.8	88.6	…

3.2 On the Impact of Region Proposals

We take the same train strategy as above to verity that whether cascade RPN is efficient or not. We add the cascade strategy on original Faster R-CNN without using our feature improvement. we also test this method on three different models (VGG-16, ResNet-50, ResNet-101). The comparative results are shown in Table 2

Table 2. . Results on PASCAL VOC 2007 test set. Baseline denote the original Faster R-CNN method, C-RPN denote our cascade RPN method. We set IoU = 0.5.

Approach	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	…
Baseline+VGG16	73.8	78.2	79.2	73.3	62.5	57.7	84.0	84.1	86.9	56.0	81.4	64.9	83.7	…
Baseline+Res50	76.4	79.1	79.9	76.5	69.2	63.1	85.1	85.4	89.2	58.4	83.5	67.5	87.1	…
Baseline+Res101	76.7	78.7	79.6	78.8	66.4	64.4	86.0	86.4	87.4	62.2	83.8	68.4	88.0	…
C-RPN+VGG16	75.0	86.8	83.2	77.7	59.2	57.4	79.4	79.0	91.3	54.8	79.9	59.1	89.8	…
C-RPN +Res50	77.2	86.8	83.6	76.7	65.8	59.4	81.8	82.5	90.8	60.0	81.1	64.0	88.1	…
C-RPN + Res101	77.4	88.9	85.2	76.9	64.9	61.5	85.1	84.2	90.1	60.0	82.7	62.0	88.6	…

As we shown in Table 2, three different basic models are all obtain improvement compare with original Faster R-CNN. Our cascade RPN method gets 1.2 points, 0.8 points and 0.7 point higher respectively on basic method than original Faster R-CNN. In the fact, in the matching step, most of the default boxes are negatives, especially when the number possible and negative training is large. This introduces a significant imbalance between the positive and negative training examples. Instead of using all the negative examples, we sort them using the highest confidence loss for each default box and pick the top ones so that the ratio between the negative and positives is at most 1.5:1. We found that this leads to faster optimization and a more stable training. We show the results in Table 3. We can observe that when we restrict the ratio between positive and negative examples, the results have improvements more or less.

Table 3. Results on PASCAL VOC 2007 test set. Baseline denote the original Faster R-CNN method, CRPN denote our cascade RPN method. CRPN-ex denote the method add neg-pos example restrict. We set IoU = 0.5

Approach	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	…
Baseline+VGG16	73.8	78.2	79.2	73.3	62.5	57.7	84.0	84.1	86.9	56.0	81.4	64.9	…
Baseline+Res50	76.4	79.1	79.1	79.9	76.5	69.2	63.1	85.1	89.2	58.4	83.5	67.5	…
Baseline+Res101	76.7	78.7	79.6	78.8	66.4	64.4	86.0	86.4	87.4	62.2	83.8	68.4	…
CRPN+VGG16	75.0	86.8	83.2	77.7	59.2	57.4	79.4	79.0	91.3	54.8	79.9	59.1	…
CRPN +Res50	77.2	86.8	83.6	76.7	65.8	59.4	81.8	82.5	90.8	60.0	81.1	64.0	…
CRPN + Res101	77.4	88.9	85.2	76.9	64.9	61.5	85.1	84.2	90.1	60.0	82.7	62.0	…
CRPN-ex+VGG16	75.2	87.0	83.4	77.9	59.4	57.6	79.6	79.2	91.5	55.0	80.1	59.3	…
CRPN-ex+Res50	77.5	88.8	85.1	76.8	64.8	61.4	85.0	84.1	90.0	59.9	82.6	61.9	…
CRPN-ex+Res101	77.6	79.7	78.6	80.8	67.4	65.4	87.0	87.4	86.4	64.2	84.8	69.4	…

3.3 Final result

We merged all elevated methods and the results are shown in Table 4, we compare our method with other object detection method.

Table 4. Results on PASCAL VOC 2007 test set. C-FM denotes our final method

Approach	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	…
Faster+VGG16 [5]	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	…
Faster+Res101	76.4	76.4	79.1	76.5	69.2	63.1	85.1	85.4	89.2	58.4	83.5	67.5	…
ION+VGG16[19]	74.6	78.2	79.1	76.8	61.5	54.7	81.9	84.3	88.3	53.1	78.3	71.6	…
ION+R+VGG16[19]	75.6	79.2	83.1	77.6	65.6	54.9	85.4	85.1	87.0	54.4	80.6	73.8	…
Hyper+VGG16 [35]	76.3	77.4	83.3	75.0	69.1	62.4	83.1	87.4	87.4	57.1	79.8	71.4	…
SDD512+VGG [20]	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1	70.2	…
R-FCN+Res101[25]	79.5	82.5	83.7	80.3	69.0	69.2	87.5	88.4	88.4	65.4	87.3	72.1	…
C-FM+VGG16	**75.9**	87.4	83.6	76.8	62.9	59.6	81.9	82.0	91.3	54.9	82.6	59.0	…
C-FM+Res101	**78.4**	79.1	80.3	79.7	69.8	68.2	86.9	87.4	88.4	65.5	84.7	67.7	…

Some detection results are shown in Fig 4

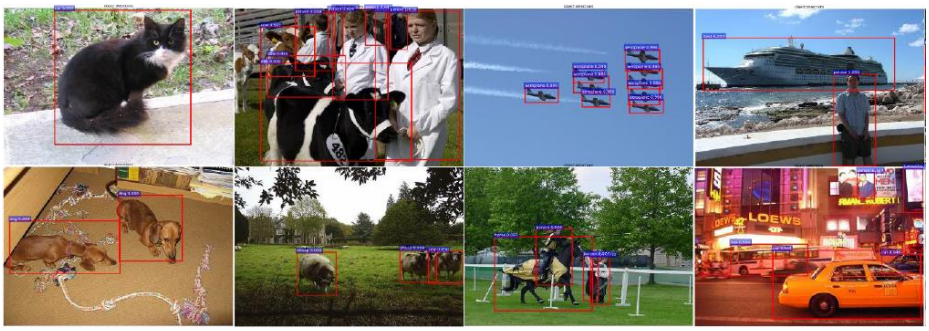

Fig. 4. Some detection results

4 Conclusion

We presented two improvements for feature representation and location of object proposals. We use the new feature representation to generate high-quality proposal, which we can get better classification and proposals. For the location task, we apply the cascade RPN strategy which is a simple but accurate and efficient way to fine-tuning of the location of proposals. Our method achieves consistent and considerable improvements over state-of-the-art methods on PASCAL VOC benchmark, while being complementary to many other advanced in object detection.

References

1. J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders.: Selective search for object recognition. International Journal of Computer Vision (2013)
2. K. He, X. Zhang, S. Ren, and J. Sun.: Spatial pyramid pooling in deep convolutional networks for visual recognition. ECCV (2014)
3. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi.: You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640(2015)
4. R. Girshick.: Fast R-CNN. ICCV (2015)
5. S. Ren, K. He, R. Girshick, and J. Sun.: Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS (2015)
6. J. Hosang, R. Benenson, P. Doll ́ar, and B. Schiele.: What makes for effective detection proposals? IEEE Transactions on Pattern Analysis and Machine Intelligence (2015)
7. C. Szegedy, S. Reed, D. Erhan, and D. Anguelov.: Scalable, high-quality object detection. arXiv:1412.1441 (v1) 2015.
8. S. Gidaris and N. Komodakis.: Object detection via a multi-region & semantic segmentation-aware CNN model. ICCV (2015)
9. B. Hariharan, P. Arbel ́aez, R. Girshick, and J. Malik.: Hyper columns for object segmentation and fine-grained localization. CVPR (2015)
10. P. Viola and M. J. Jones.: Robust real-time face detection. IJCV (2004)

11. M.-Y. Liu, A. Mallya, O. Tuzel, and X. Chen.: Unsupervised network pretraining via encoding human design. In: 2016 IEEE Winter Conference on Applications of Computer Vision. pp. 1–9. IEEE (2016)
12. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun.: Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR (2014)
13. A. Ghodrati, M. Pedersoli, T. Tuytelaars, A. Diba, and L. V. Gool.: Deep proposal: Hunting objects by cascading deep convolutional layers. ICCV (2015)
14. Y. Hua, K. Alahari, and C. Schmid.: Online object tracking with proposal selection. ICCV (2015)
15. Y. Jia and M. Han.: Category-independent object-level saliency detection. ICCV (2013)
16. Guo, Kai, S. Wu, and Y. Xu.: Face recognition using both visible light image and near-infrared image and a deep network. Caai Transactions on Intelligence Technology vol. 2.1, pp. 39--47 (2017)
17. Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos.: A unified multi-scale deep convolutional neural network for fast object detection. ECCV (2016)
18. Bell S, Zitnick C L, Bala K, et al.: Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. J.2874--2883(2016)
19. Lin, Tsung Yi, et al.: Feature Pyramid Networks for Object Detection. (2016).
20. Y. Xu, B. Zhang, Z. Zhong.: Multiple representations and sparse representation for image classification. Pattern Recognition Letters, vol. 68, pp. 9--14. (2015)
21. Dai J, Li Y, He K, et al.: R-FCN: Object Detection via Region-based Fully Convolutional Networks (2016)
22. P. Doll ́ar, R. Appel, S. Belongie, and P. Perona.: Fast feature pyramids for object detection. PAMI, vol. 36(8), pp. 1532--1545 (2014)
23. Yang B, Yan J, Lei Z, et al.: CRAFT Objects from Images.pp. 6043--6051 (2016)
24. M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I.Williams, J. Winn, and A. Zisserman.: The pascal visual object classes challenge: A retrospective. IJCV, pp. 98—136 (2015)
25. Lin G, Milan A, Shen C, et al. Refine Net: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation (2016)
26. A. Ghodrati, M. Pedersoli, T. Tuytelaars, A. Diba, and L. Van Gool.: Deep boxes: Hunting objects by cascading deep convolutional layers. In: Proceedings ICCV (2015)
27. K. He, X. Zhang, S. Ren, and J. Sun.: Spatial pyramid pooling in deep convolutional networks for visual recognition. ECCV (2014)
28. A. Shrivastava, A. Gupta, and R. Girshick.: Training region based object detectors with online hard example mining. CVPR (2016)
29. Simonyan K, Zisserman A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science (2015)
30. K. He, X. Zhang, S. Ren, and J. Sun.: Deep residual learning for image recognition. CVPR (2016)
31. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV (2015)
32. Kong T, Yao A, Chen Y, et al.: HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. pp. 845--853(2016)