Robust person head detection based on multi-scale representation fusion of deep convolution neural network

Yingying Wang[1,2], Yingjie Yin[1], Wenqi Wu[1,2], Siyang Sun[1,2], Xingang Wang[1]
1 Research Center of Precision Sensing and Control,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2 University of Chinese Academy of Sciences, Beijing 101408, China.

***Abstract*—Person head detection is still a challenge due to the large variability in heads' sizes and orientations, lighting conditions and strong occlusions. Small heads require local information contained in low level layers instead of semantic features of upper layers. But most of these fine details are lost in the early convolutional layers of the deep convolution neural networks (DCNN). In order to improve the overall detection accuracy, it is important to utilize local information from lower layers into the detection framework. In this letter, we use multi-scale representation fusion of DCNN as a way to incorporate lower layers with upper layers for detection. Our proposed model is based on the recent object detection network Single Shot MultiBox Detector (SSD). VGG16 is used as the base network. Batch normalization (BN) layers are used in our proposed multi-task learning method to accelerate training process and improve the robustness. Compared to state-of-the-art methods, our proposed detector achieves superior person head detection performance on the HollywoodHeads dataset (81.0 AP) and Casablance dataset (78.5 AP).**

Keywords—Person head detection; deep convolution neural network; multi-scale representations; multi-task learning.

I. INTRODUCTION

Person head detection in still images is the task of localizing head in the wild robustly, regardless of human poses and head orientations. It's important for many tasks, such as attributes recognition, person identification, human tracking, action recognition, autonomous driving, person count and many others. In recent years, face detection and pedestrian based on deep Convolution Neural Networks (CNN) have made significant achievements. However, person head detection is still a difficult challenge due to the large variability in appearance, body postures, lighting conditions and strong partial occlusions in the wild.

Significant progress has been made in object detection because of recent advances in deep CNNs. In particular, Single Shot MultiBox Detector (SSD) [1] achieves high object detection accuracy on PASCAL VOC, COCO, and ILSVRC and this network runs at real-time. SSD can localize objects better because it learns to regress the object shape and classify object categories instead of repurposing classifiers to perform detection. However, this detector has worse performance on smaller object categories than bigger object categories mainly due to the coarseness of its feature maps. As demonstrated in [2], features in the CNN are hierarchically distributed. Lower layers respond to corners and edges and hence contain more local information. Features of upper layers are more semantic and class-specific. Most of existing methods only use the very top layer for objects detection and ignore the importance of lower layers.

In this paper, we introduce a novel framework that based on SSD for end-to-end head detection. Multi-scale represents are used to enhance the detection accuracy especially for small-size human heads. The practice of fusing the multiple convolutional layers of deep CNNs has been applied successfully in many object classification and detection methods [3, 4, 5], the differences and details of our proposed method will be stated in part III. In our detection model, VGG16 [6] is used as the base network to extract features. Then some convolutional layers are added to the end of the base network for detection. In the next, multiple scale layers from the base network are combined. This combined layer and the added convolutional feature layers are selected to detect person heads. We add a batch normalization layer [18] after each of those convolution feature layers. At training time, these feature layers produce a set of default boxes of different scales and aspect ratios and are matched to the ground truth boxes. At prediction time, both the shape offsets and the confidences of the head for each of this default box are produced by the detector. Finally, a non-maximum suppression method is adopted to produce the final detections. Fig 1 shows some results of the proposed detector. It can ban seen that Our model has good performance on small-size human heads and achieves better results in complex conditions such as strong partial occlusions.

The remainder of this paper is organized as follows. In Section II, the related work for object detection is reviewed. Human head detection method based on deep neural network is described in Section III. Section IV introduces datasets and experimental results. Finally, in Section V, we draw conclusions and discuss future research directions of this work.

This work is partly supported by The National Natural Science Foundation of China (Grant No. 61573349), The National High Technology Research and Development Program of China (863 Program) (Grant No. 2015AA042308).

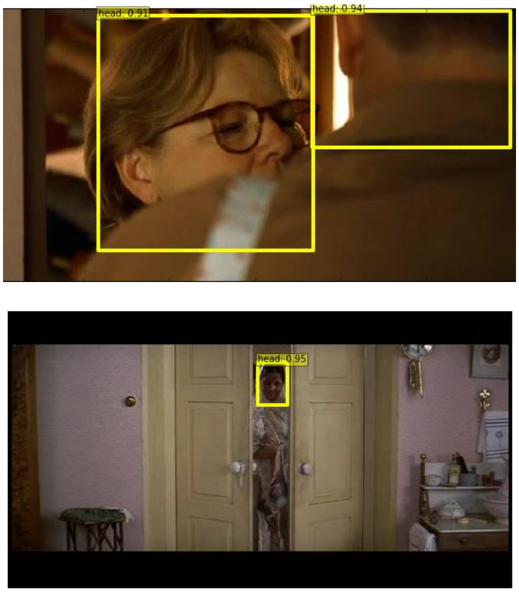


Fig. 1. The results of person head detection using our proposed model.

II. Related Work

In recent years, image classification and object detection have achieved significant improvement due to the advances in deep CNNs [6, 7, 8, 9, 27, 28]. Object detection is more challenging compared to image classification. Firstly, it needs to judge if the input image contains the target or not. And then, localizes the objects if they are contained in the image.

Traditional object detection methods are based on feature extraction methods and classifiers, including obtaining sub regions of the whole images and extracting their features, then classifiers are used on the features to determine whether the sub windows contain the target objects or not. The classical face detection method, proposed by Viola and Jones [10], uses Haar-like features and cascade boosting classifiers. HOG feature-based deformable part model (DPM) [11] is widely adopted among conventional object detection.

Most traditional object detection methods are time-consuming and have poor performance and robustness. And with the development and successful of deep neutral networks, object detection methods based on deep CNNs become the first choice. Region-based convolutional neural network (R-CNN) [12] is the first efficient deep learning based method. Firstly, R-CNN extracts around 2000 region proposals from an input image using Selective Search (SS) method [13]. In the second, features of region proposals are extracted by CNNs. Finally, liner SVM is produced on the features to classify these proposals. This original region-based CNNs method is computationally expensive. To reduce the cost, SPPnet [14] and Fast R-CNN [15] shares convolutions among region proposals. With the advances in deep learning and object detection, many efficient and real-time detectors are proposed. Faster R-CNN [16] uses a Region Proposal Network (RPN) to generate high-quality region proposal instead of using SS method and this network shares full-image convolutional features. The pipeline of region-based method is to produce potential bounding boxes firstly and then apply a classifier on them. This practice repurposes classifiers to perform detection. Differently, YOLO [17] uses regression method to detection objects. This network is much faster and can be trained end-to-end. Since YOLO predicts detections only on the convolutional layer, its detection accuracy is worse than Faster R-CNN. SSD is a fully convolutional and end-to-end deep neural network. It produces a set of defaults boxes and makes detection predictions at multi-scale feature layers. Defaults boxes in SSD are similar to the anchor boxes that used in Faster R-CNN. SSD is faster than Faster R-CNN and is more accurate than YOLO.

Although significant process has been made by CNNs-based object detection methods, they still have trouble with localizing smaller objects than larger objects. It mainly because most of object detection methods only use the output of the last layer of a feature extraction network as a feature representation. However, this representation is too coarse and detail representations which are importance to small object categories are lost in the early convolutional layers. To improve the detection accuracy especially for small-size human heads, outputs of low-level and high-level layers are combined as multi-scale representations. [23] combines features of different layers to improve accuracy for pedestrian detection, [3, 4] use fusion outputs of multi-scale layers for object detection and [5] combines features from two pooling layers and one convolutional layer for multi-task learning. Multi-scale representations contain both of global and local information and can improve overall accuracy for object detection and classification.

III. Proposed Person Head Detector

The framework of our model is illustrated in Fig. 2. This section will describe our proposed head detection model based on multi-scale representation fusion of deep convolution neural networks (DCNN) and its effective training method.

A. Head detection Model based on multi-scale representation fusion of DCNN

Our head detection method is based on the Single Shot MultiBox Detector [1]. The early layers in the network are based on VGG16. This image classification architecture is pre-trained on the ILSVRC dataset and then converted to a fully convolutional network. Dropout layers and the fc8 layer are discarded in our model. Then, a series of smaller convolutional layers are added after the base network. Each of the added feature layers is used to predict confidences and shape offsets for a fixed set of default boxes.

Some human heads are very small in images so that they have little information at the very top layers. In this case, these small-size heads are more likely to be ignored at predicting time. In order to improve the overall detection accuracy, the multi-scale representation fusion of DCNN is proposed. We fuse the Conv3_3, Conv4_3 and Conv5_3 layers of VGG16 using three separate networks as shown in Fig. 3. The information gained from multi-scale layers is important for accurate visual recognition and especially for small objects which require the higher spatial resolution that provided by lower convolutional layers. The dimensions of these fused layers are 75×75×256, 38×38×512, 19×19×512. Because of the different dimensions,

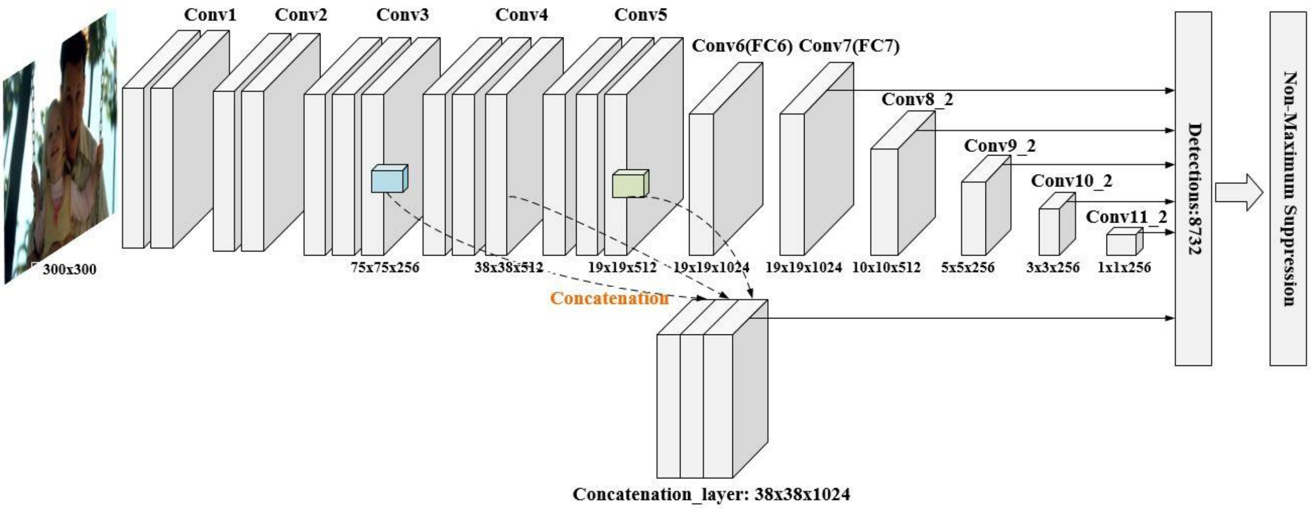


Fig. 2. The results of person head detection using our proposed model.

they can't be concatenated directly. Therefore, various sampling strategies are applied to different layers. Unlike some existing methods [3,4,5], we add a convolutional layer to Conv3_3 to carry out subsampling. For Conv5_3, we add a deconvolutional layer [25] to conduct upsampling. And then L2 normalization method [26] is adopted to normalize activations from multiple layers before combining them into a uniform space. Finally, we concatenate these layers along the channel axis. The architecture details of our proposed model are showed in Fig. 2. We use the Concatenation_layer as well as Conv7 (FC7), Conv8_2, Conv9_2, Conv10_2 and Conv11_2 to predict location and confidence of human head and concatenate the outputs of these layers finally. Compare to the other layers to predict detections, Concatenation_layer has a larger feature scale. In order to enhance robustness we add an L2 normalization layer after Concatenation_layer to scale the feature norm at each location in this feature map。

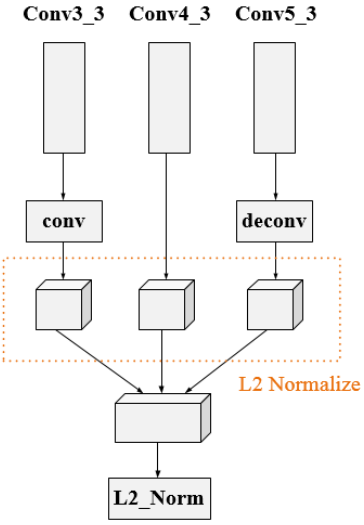


Fig. 3. Concatenation of multi-scale layers.

For a feature layer of size m×n with p channels, predictions are produced by two 3×3×p small convolutional filters. One filter is applied to predict confidences for person head and background and the other is used for offsets relative to some predefined default bounding boxes. Take conv8 as example, the details of predictions are showed in the Fig. 4.

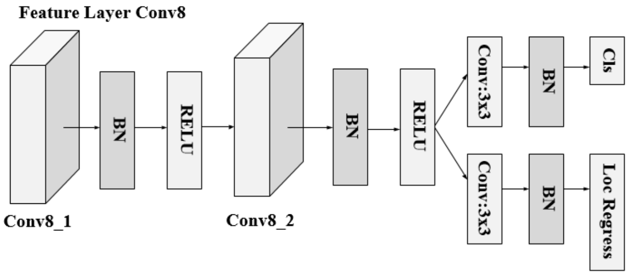


Fig. 4. Prediction Module.

Before each added feature layers and Concatenation_layer predict confidences and shape offsets for the default bounding boxes, a batch normalization layer is added after each of them. As illustrated in [18], it's important to normalize layers' inputs to solve internal covariate shift phenomenon. For a layer with d-dimensional input $x=\{x^1, x^2, \ldots, x^d\}$, the normalization is applied to each activation independently. So, one dimension is taken as example to show the process of BN Transform. Let the normalized value of $x = \{x_{1\ldots m}\}$ over a mini-batch (B) is $\hat{x}=\{\hat{x}_{1\cdots m}\}$ and the output value is $y = \{y_{1\ldots m}\}$. The overall BN Transform can be referred to as

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B{}^2 + \varepsilon}} \tag{1}$$

$$y_i = \gamma \hat{x}_i + \beta \tag{2}$$

where $\sigma_B{}^2$ is mini-batch variance, μ_B is the mini-batch mean and ε. is a constant for numerical stability. γ, β are the parameters to be learned during training.

This methodology allows using of much higher learning rates to speed up training process and eliminates the need for dropout layers.

B. *Training*

Our proposed model can be trained end-to-end and only needs input images and ground truth boxes of human heads. Conv7, Conv8_2, Conv9_2, Conv10_2, Conv11_2 and the Concatenated layer are used for prediction. Each location in these feature maps is defined to produce a small set of default boxes of different aspect ratios.

At training time, default boxes are matched to the ground truth firstly to determine which default boxes responding to ground truth bounding boxes. To be specific, for each ground truth bounding box, its matching boxes are picked from default boxes that vary over location, aspect ratio and scale. Default boxes with overlap higher than 0.5 correspond to this ground truth bounding box during training. Then, two small convolutional filters are applied to Conv7, Conv8_2, Conv9_2, Conv10_2, Conv11_2 and the Concatenated layer to predict category confidences and shape offsets for these default bounding boxes. The multi-task learning method is used for training the whole networks, and the total loss [1] is a weighted sum of the localization loss (L_{conf}) and the confidence loss (L_{loc})

$$L(x,c,l,g)=\frac{1}{N}\left(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)\right) \qquad (3)$$

where x means the input sample and its class confidence is c. l means the predicted box which is matched to any ground truth box (g) with jaccard overlap higher than 0.5. N is the number of the matched default boxes (defined as *Pos*).

The confidence loss is computed using softmax loss function and the weight parameter α is set to 1 by cross validation. The localization loss is a Smooth L1 loss [16] between l and g. A default bounding box can be characterized by $\{cx, cy, w, h\}$, where (cx, cy) are the coordinates of the center and w, h are width and height respectively. Offsets are regressed to adjust the default box to match the head shape better. Let $x_{ij} = \{0, 1\}$ to represent whether i-th default box matches to j-th ground truth or not.

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}\,smooth_{L1}(l_i^m-\hat{g}_j^m) \qquad (4)$$

where,

$$\hat{g}_j^{cx}=(g_j^{cx}-d_i^{cx})/d_i^w \qquad \hat{g}_j^{cy}=(g_j^{cy}-d_i^{cy})/d_i^h$$

$$\hat{g}_j^{w}=\log(\frac{g_j^w}{d_i^w}) \qquad \hat{g}_j^{h}=\log(\frac{g_j^h}{d_i^h})$$

Hard negative mining and additional data augmentation strategies are also adopted at training time. The proposed head detection model is fine-tuned using SGD with initial learning rate 0.0001, 0.9 momentum, 0.0005 weight decay, and batch size 8. The weights of all the new layers are initialized with "Xavier" method [24].

IV. Datasets And Experiments

Our head detection model is evaluated on Casablance dataset [19] and HollywoodHeads dataset [20] respectively. This section will take a brief introduce of two datasets and present our experimental results.

A. *Datasets for evaluation*

Casablance dataset contains 1,466 frames. These frames were collected from the film "Casablance" and each of them is annotated with head bounding boxes. The original dataset has some disadvantages and was perfected by [20].

HollywoodHeads dataset contains 224,740 frames from 21 Hollywood movies and 369,846 person heads annotated in total. The training set of this dataset is composed of 216,719 frames from 15 movies, the validation set contains 6,719 frames from 3 movies and the test set has 1,302 frames that from the rest movies. Frames with poor quality such as low lighting conditions and strong occlusions are labeled by "difficult".

B. *Results and Comparison*

The standard Average Precision (AP) [21] is used to evaluate the performance of our detection model. Detections having higher intersection-over-union score than 0.5 with the ground-truth bounding box are considered to be correct. We compare our model with other detectors: the DPM-base model (DPM) [22], R-CNN-based model (R-CNN) [12] and Contex-aware CNNs model [20]. Fig. 6 and TABLE I present the results of detection performance and comparison with other methods. It can be seen that detection accuracy has an obvious improvement using our method. It achieves 81.0 AP compared to 72.7 AP in state-of-the-art method proposed in [10] on HollywoodHeads dataset and achieves 78.5 AP compared to 72.6 AP in [10] on Casablance dataset.

Some results of head detection using our model are showed in Fig. 1 and Fig. 5. Our model has good performance on small-size human heads and achieves better results in complex conditions such as of body postures variation and strong partial occlusions.

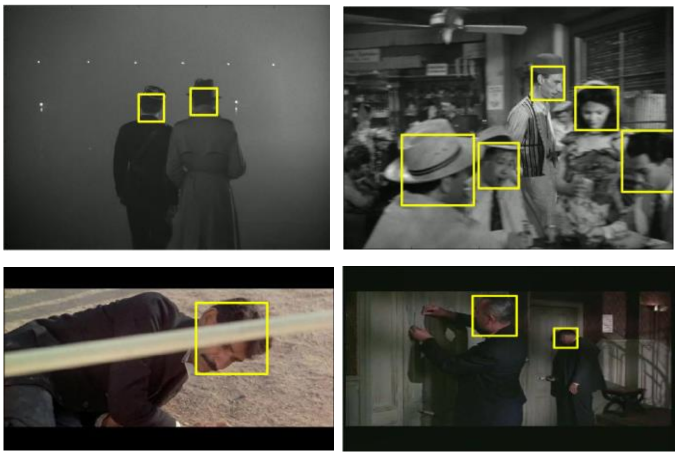
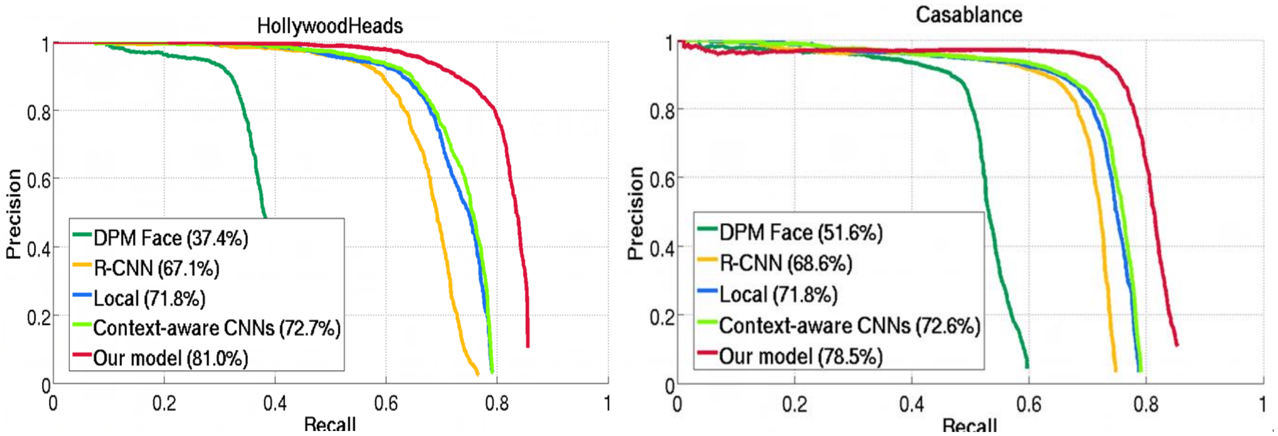


Fig. 5. The results of detection using our detector. Our model has good performance on small-size human heads, poor lighting conditions, body postures and strong partial occlusions conditions.



(a) Comparison on HollywoodHeads dataset

(b) Comparison on Casablance dataset

Fig. 6. The results of our method compared to the state-of-the-art on the two datasets.

TABLE I
Preformance (% AP) of our detection model and results of our method compared with the other state-of-the-art methods on two datasets.

	models	DPM [22]	R-CNN [12]	Local [20]	Context-aware CNNs［20］	**Our model**
Datasets	HollywoodHeads	37.4	67.1	71.8	72.7	**81.0**
	Casablance	51.6	68.6	71.8	72.6	**78.5**

V. Conclusion

In this paper, we presented an efficient and robust deep learning method to detect person heads on the real-time. Multi-scale representations are used to improve the detection accuracy. We also add additional respect ratio according to the real shape of human heads. The experimental results show that our detector can achieve higher recall and accuracy in real scenes. Our method can extend to other networks such as AlexNet [27] and ResNet [28]. In the future, we will do some experiments using the presented person head detection method in real environment and make more improvement.

References

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," European Conference on Computer Vision. Springer International Publishing, 2016.

[2] Zeiler, D. Matthew and R. Fergus, "Visualizing and understanding convolutional networks," European conference on computer vision. Springer International Publishing, 2014.

[3] S. Bell, C. L. Zitnick, K. Bala and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[4] T. Kong, A. Yao, Y. Chen and F. Sun, "HyperNet: towards accurate region proposal generation and joint object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[5] R. Ranjan, V. M. Patel, and R Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," arXiv preprint arXiv:1603.01249 (2016).

[6] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).

[7] W. Chen, T. Qu, Y. Zhou, K. Wang, et al, "Door recognition and deep learning algorithm for visual based robot navigation," Robotics and Biomimetics (ROBIO), 2014 IEEE International Conference on. IEEE, 2014.

[8] J. Yu, J. Chen, Z. Q. Xiang and Y. X. Zou, "A hybrid convolutional neural networks with extreme learning machine for WCE image classification," Robotics and Biomimetics (ROBIO), 2015 IEEE International Conference on. IEEE, 2015.

[9] Z Zhu, W. Zou, Q. Wang, and F. Zhang, "STD: A Stereo Tracking Dataset for evaluating binocular tracking algorithms," Robotics and Biomimetics (ROBIO), 2016 IEEE International Conference on. IEEE, 2016.

[10] P. Viola, and M. J. Jones, "Robust real-time face detection," International journal of computer vision 57.2 (2004): 137-154.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE transactions on pattern analysis and machine intelligence 32.9 (2010): 1627-1645.

[12] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[13] J. R. Uijlings, T. Gevers, et al, "Selective search for object recognition," International journal of computer vision 104.2 (2013): 154-171.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," European Conference on Computer Vision. Springer International Publishing, 2014.

[15] R. Girshick, "Fast r-cnn," Proceedings of the IEEE International Conference on Computer Vision. 2015.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems. 2015.

[17] J. Redmon, S. Diwala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[18] S. Ioffe, and C Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167 (2015).

[19] X. Ren, "Finding person in archive films through tracking," Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

[20] T. H. Vu, A Osokin, and I Laptev, "Context-aware CNNs for person head detection," Proceedings of the IEEE International Conference on Computer Vision. 2015.

[21] M Everingham, L. Van Gool, et al, "The pascal visual object classes (voc) challenge," International journal of computer vision 88.2 (2010): 303-338.

[22] M Mathias, R. Benenson, M. Pedersoli and L. Van Gool, "Face detection without bells and whistles," European Conference on Computer Vision. Springer International Publishing, 2014.

[23] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[24] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Aistats. Vol. 9. 2010.

[25] J. Long, E Shelhamer, and T Darrell, "Fully convolutional networks for semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[26] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," arXiv preprint arXiv:1506.04579 (2015).

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems. 2012.

[28] K. He, X Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.