Improvement of YOLOv3 network based on ROI

Shijin Li[1,a] , Fushou Tao[1,b],Ting Shi[2,c,*],Jinyun Kuang[1,d]
1. Yunnan University of Finance and Economics
2. Yunnan School of Business, Information and Engineering
*Corresponding author
[a]sothink1984@126.com,[b]shuilifang1985@qq.com, [c]shiting1988@126.com, [d]853051783@qq.com

***Abstract*—The paper adopts the YOLOv3 network, by removing the calculation of non-interest areas to improve the operation speed of YOLOv3. It increases the operating speed in proportion to the size of the non-interest area. The author studies the characteristics of the convolutional layer and the YOLO layer, and proposes a new processing method from the experimental results. In this paper, the ROI image processing technology in the video detection sequence is used to filter the ROI region, eliminate the calculation outside the ROI, and improve the detection speed and accuracy of the YOLOv3 network.**

Keywords—YOLOv3; ROI; Target detection; RONI; convlayer

I. BACKGROUND

At present, the deep tracking-based target tracking technology has a great improvement in robustness and generalization ability compared with the traditional tracking method [1, 2, 3]. There are two main ideas for target detection on tracking. First, the target is directly detected from the video image without using the prior theory, and the tracking is achieved after the target is determined; The other is to use the prior knowledge of the target to model the target, and then find the target matching the feature from the video image for tracking.

But in complex scenarios, there will be some changes in the environment and goals, and these changes may lead to detection target drift or target loss [4, 5]. Currently, the main factors affecting the target are as follows:

(1) Variable environment: mainly reflected in the background of targets during tracking and detection of complexity and uncertainty, such as changes in light intensity, target occlusion problems, etc[6].

(2) Variable target background: When the target moves, its background image will change, which will affect the tracking effect of the target.

(3) Photography problems. As the picture may be shaken, and the target is mostly in size changes and blurred video sequences. In addition, because the target is small and blurred in the picture and far away. These problems greatly limit the robustness of the tracking algorithm to some extents.

As the deep learning has powerful feature extraction and ability to express against the target. It can use a variety of deep learning methods such as convolutional neural networks, combined with traditional classifiers, to perform automatic feature extraction on the processed images. In addition, due to CNN's weight sharing mechanism and local receptive domain mechanism, the interconnection redundancy between parameters is reduced, and the algorithm time and space complexity are relatively low.

Since 2012, several CNN algorithms have been proposed, such as YOLO and R-CNN. R-CNN is a region-based CNN proposed by Girshick et al. It Combines regional proposal algorithms with CNN. It extracts 2000 regions by selective search and then classifies them on the selected regions instead of processing the entire image. In 2016, Joseph Redmon proposed YOLO. It is different from the region-based approach. YOLO uses the Full Convolutional Neural Network (FCN) to deliver only n×n images once, making it very fast and real-time. It splits the image into a grid of m times m, and generates a bounding box and its class probability. Compared to region-based technology, YOLOv2 [8] achieves relatively high positioning error and low recall rate by producing batch standardized and higher resolution classifiers.

In 2018, YOLOv3 [9] replaced the softmax function with logistic regression and thresholds to make it more accurate. As the YOLOv3 has the high detection accuracy and detection speed , the higher map and recall rate, Therefore, this article selects the YOLOv3 network and improves it.

In this paper, by removing the calculation of non-interest regions to improve the operation speed of YOLOv3. It not only improves the accuracy of the corresponding detection, but also improves the loss of the training process and the final map value.

II. RELATED PRINCIPLES AND IMPROVEMENTS

In order to improve the detection speed of YOLOv3 network, this paper adopts ROI (region of interest) image processing technology in video detection sequence[14-16]. Figure 1 shows an example of the change in ROI size for a continuous convolutional layer without sampling. The left side of the figure is the input of the YOLO network, the ROI is the shaded area.The ROI is convoluted with a 3x3 kernel, and the shading in the figure represents the new ROI generated by the feature map. After the convolution operation, the ROI in the feature map is bigger than the ROI in the input image. This is because the convolution operation causes a single pixel in the input image to affect the value of 9 pixels (3 x 3 pixels) in the output image. In Figure 1, the ROI of feature map 2 is bigger than the ROI of the previous layer. As the convolution operation continues to the end of the network, the ROI becomes larger and eventually it can become the same as the input image. This means that all the data in the input image will be considered the region of interest, which will result in no

reduction in the calculation of the final ROI. Therefore, it is necessary to derive the appropriate ROI selection after the convolution operation.

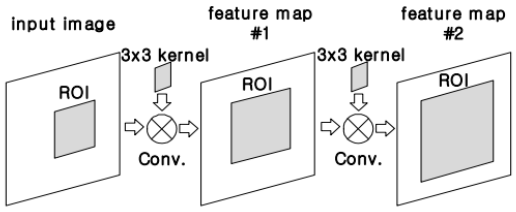

Fig. 1. the ROI changes without sampling

Figure 2 shows an example of determining the ROI in a convolutional layer. Each box corresponds to one pixel, "N" and "I" represents RONI data and ROI data, The kernel size is 3×3. When the number of ROI data in the kernel window is greater than a predetermined threshold, the output is ROI. In Fig. 2(a), the thick square at the bottom right of the input image indicates the sliding window for the current convolution operation. The output of this convolution operation is treated as ROI data, and the output after convolution depends on the amount of ROI data in the input window. In the figure, 3 data are in the ROI and the remaining 6 are in RONI. The predefined threshold is used to determine the ROI area in the output layer. When the data value is 0, 1 or 2, it is output as the ROI area, and the area exceeding 2 is output as the RONI area. In FIG. 2(b), when the threshold is "0", it means that when the kernel window includes at least one ROI data, the output is the ROI. Figure 2(b) shows when this condition is TH0. the RONI data adjacent to the boundary of the ROI becomes the ROI data, so the ROI size of the output of the convolutional layer is larger than the ROI of the input. Figure 2(c) shows when his condition is TH3 and the threshold is 3 ,the ROI will not increase.

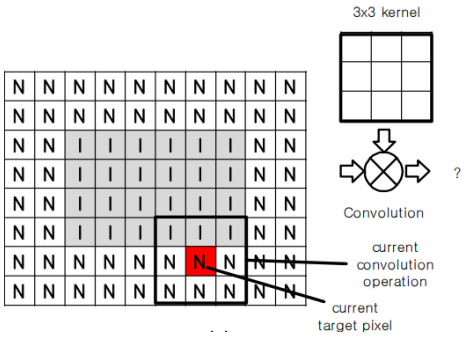

Fig. 2. (a)Example of ROI and convolution operations

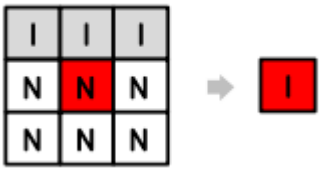

Fig. 2(b) the output ROI of the TH0

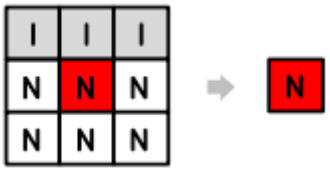

Fig. 2(c) the output ROI of the TH3

Once the ROI area is output, the calculation can be reduced by avoiding the operation of generating data in the RONI area. Only the operation of generating ROI data is processed in each layer, and the result is transferred to the next layer. To investigate the effects of RONI, the data in RONI was replaced with a predefined value. In the YOLOv3 network, even data stored in RONI may affect the accuracy of object detection in the ROI. Therefore, an experiment is required to obtain an appropriate value for the RONI data. Therefore, the value of each layer is studied by using an input image that is fed back by a black image. In the experimental results, all values are zero. Obviously, there are no objects to detect in this image. Table 1 shows the distribution of data values for the output of some layers.

TABLE 1. THE DISTRIBUTION OF DATA VALUES FOR THE OUTPUT OF SOME LAYERS.

Input Data	1st conv layer(%)	81st conv layer(%)	82st YOLO layer(%)
Below -1.2	0.00	95.32	0.02
-1.2 ~ -0.7	6.25	0.32	0.06
-0.7 ~ -0.2	3.13	1.27	0.69
-0.2 ~ -0.3	15.63	2.29	96.90
0.3 ~ 0.8	18.75	0.64	2.30
0.8 ~ 1.2	12.50	0.14	0.03
Above 1.2	43.75	0.02	0.00

The first column represents the range of data values in the input image. The second column shows the data portion corresponding to the data value after convolution with the first convolutional layer. From the table, most of the output data exceeds "1" in the first convolutional layer. The third column shows the results after the 81st layer of convolution. This layer is followed by the YOLO layer, that is, the output of the 81st layer is directly used for detection. Table I shows that the data in this layer is higher than '0'. Therefore, the RONI area data is set to "0" for experimental verification. First set the RONI data to '0' and record the experimental results, secondly set the RONI data to 0 and record the experimental results.

As shown in table 2. The results show that the accuracy of TH3 is better than TH0, so the threshold is set to 3 in the experiment. As shown in Figure 2(c), this means that the data in RONI does not affect the calculation of the ROI. This is because the ROI calculation is not affected by RONI and the RONI data of the input feature map of the YOLO layer is set to zero, so the RONI need not be included in the input image. Therefore, only the ROI of the cropped image from the input image is used for the input image. The resolution of the proposed ROI input image is lower than the resolution of the input image, which obviously increases the operation speed.

III. METHOD AND IMPROVEMENT

A. Improvement measures

In order to process ROI input, we modify the following layers:

Residual layer： In the remaining layers of the original network, the resolution of the two input feature maps is the same. When performing calculations only for the ROI, this layer adds two ROIs with the same resolution.

Upsampling layer： The input ROI of this layer is upsampled in the same way as the original network.

Fully connected layer： In the original network, the resolution of the two input feature maps is the same in that layer. Therefore, the ROI is handled in the same way as the original network.

YOLO layer: The resolution of the input feature map is set to the resolution of the ROI, and then the ROI is processed in the same way as the original network. The output resolution of the YOLO layer is restored to the resolution of the original network. It generates an empty feature map with the output resolution of the YOLO layer. The ROI data processed by the YOLO layer is patched at the estimated position on the output feature map. The prediction is performed by using the recovered data.

B. Implementation Method

The experiment randomly takes 90% of the above data to be set as the training set, and the remaining 10% as the verification set. At the beginning of the training, the YOLOv3 preloading model was downloaded from the yolo official website for training. The training process is divided into two parts. The first part of the training trains 50 generations on the training set. Batch size is 32. The training optimizer uses the adam optimization algorithm, Learning rate=0.0001， beta_1 = 0.9， beta_2 = 0.999 ; In the second part, the training set retrained 50 generations， in this part batch size=16, Learning rate=0.0001， beta_1 = 0.9， beta_2 = 0.999

MAP is the average AP value for multiple verification sets, which is an indicator of the accuracy of detection in target detection. The key value is the area and coordinate axis enclosed by the precision curve. The exact curve is accurate for both dimension curves, recalled for vertical and horizontal axis coordinates. It gets different thresholds by choosing the appropriate precision and recall rate. The accuracy is calculated as shown in Equation 1:

$$precision = \frac{TP}{TP + FP} \qquad (1)$$

Where TP is a numerical example of correct division, and FP is a wrong numerical example. The calculation formula of the recall rate is as follows

$$recall = \frac{TP}{TP + FN} = \frac{TP}{precision} \qquad (2)$$

Where FN is a negative example of the number of errors. Through the above formula, P-R can draw a curve to calculate the AP value of a single class. And the average of the AP values for all categories can be calculated for the mAP value of the entire model.

C. Experimental results

Compare the accuracy of TH0 and TH3 by experiment, The first 50 images of the COCO dataset 2017 test image are used for evaluation. In these images, the ROI is manually defined because the ROI detection algorithm in the system can only be applied to video sequences. Only objects in the ROI are considered to be objects to be detected. Table 2 shows that the third line shows the number of real result detections.the 4th line and the 5th line respectively represent the number of false positives and false negatives caused by the proposed method with TH0 or TH3 conditions, the 6th line and the 7th line respectively represent improvements compared to YOLOv3. As shown in these lines, Under the condition of TH3, one false positive and three false negatives were improved compared to YOLOv3. This table shows that TH3 is better than TH0 in accuracy.

TABLE 2.COMPARISON OF DETECTION PERFORMANCE IN TH0 AND TH3

Threshold type		TH0	TH3
Test results		Number of objects	
Real result		62	66
Degradation	false positives	0	0
	false negatives	7	2
Improvement	false positives	0	1
	false negatives	2	3

The operation time and accuracy of the proposed method were evaluated by experiments. Use a single GPU and reuse the pre-training weights in the COCO dataset[18]. The first 300 images of the COCO dataset test image were used to test the image. The input resolution of the network is set to 416 x 416. In YOLOv3, the 416 x 416 image is divided into 13 x 13 blocks, and the ROI is defined as a rectangular shape. The ROI is set manually. All objects in the test image cannot be included in the ROI in order to select some objects for the ROI.

Figure 3 shows the enhancement of the operating time in the proposed method compared to the original YOLOv3. The vertical axis represents the ratio of the proposed method to the operating time of YOLOv3, The horizontal axis represents the ratio of the number of pixels in the ROI to 416 x 416 pixels. The figure shows that the operating time is reduced in proportion to the ROI size. In the experiment, we can see that the operating speed has increased by 3.29 times.

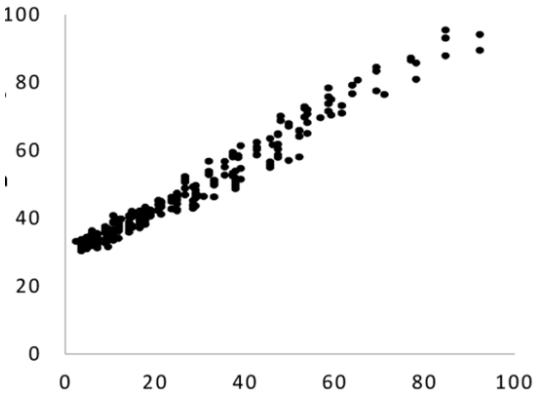

Fig. 3. the enhancement of the operating time in the proposed method

Table 3 shows the accuracy of the proposed method compared to the original YOLOv3. Line 2 shows the method and the original YOLOv3 correctly detected 356 objects. Lines 3 and 4 show incorrect detection by the proposed method, and the correct detection is performed by the original YOLOv3. As shown in the table, 10 false positives and 25 false negative tests were performed by the proposed method. They are detected correctly in the original YOLOv3. Lines 5 and 6 show the opposite, making the original YOLOv3 produce 8 false positives and 11 false negatives. And correctly detect them by the proposed method. As a result, the number of false positives and false negatives increased by 2 and 14 respectively compared to YOLOv3.

TABLE 3 COMPARISON OF TEST METHODS PERFORMANCE

Threshold type		Number of objects
Real result		356
Degradation	false positives	10
	false negatives	25
Improvement	false positives	8
	false negatives	11

The training loss curve and the test loss curve are shown in Figure 4. The horizontal axis is the training time and the vertical axis is the loss Figure 5 shows the mAP curve of the model value. The loss value of YOLOv3 after removing non-interest area is lower than YOLOv3. The horizontal axis is the training time , the vertical axis is the mAP value. The highest mAP value of YOLOv3 after removing non-interest areas is 66.5.but the highest mAP value of the YOLOv3 model is 62.0. Obviously, the improved YOLOv3 detection after removing non-interest areas is better than the original YOLOv3.

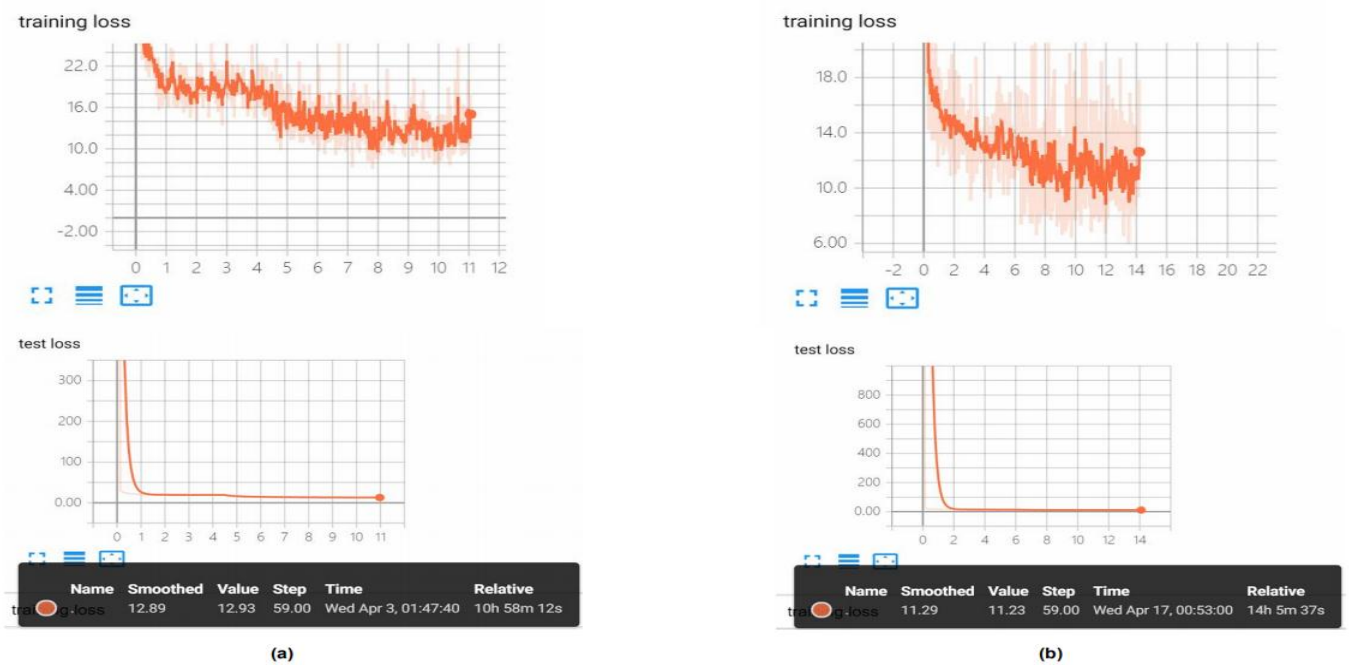

Fig. 4. training loss curve and test loss curve

(a)Training loss curve and test loss curve in tiny- yolov3
(b)Improved training loss curve and test loss curve in tiny-yolov3

Figure 6 shows the test results of the original YOLOv3 and the proposed method. Column 1 shows the results of YOLOv3 and column 2 shows the results of the proposed method. In these figures, the dashed green rectangle represents the manually determined ROI. Since the object outside the ROI is not detected, the detection rate is measured only for the object in the ROI. In the first line, the ROI includes the small person detected by YOLOv3 and the proposed method.

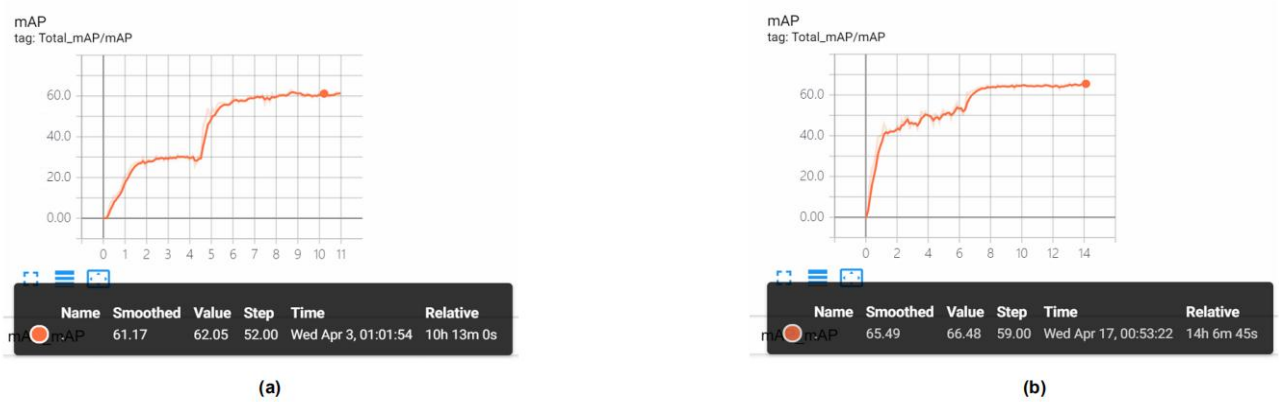

Fig. 5. map curve

(a) map curve in tiny- yolov3
(b)map curve in improvedTiny- yolov3

In the first row, two smaller pedestrians were not successfully detected in YOLOv3. But it was correctly detected in the proposed method. In line 2, the extra "television monitor" is detected, which means that a false positive has occurred in the proposed method. Line 3 "Banana" is included in the ROI, but YOLOv3 is not detected, but is detectable in the proposed method.In Figure 6, The first column shows the results of YOLOv3 and the second column shows the results of the proposed method.

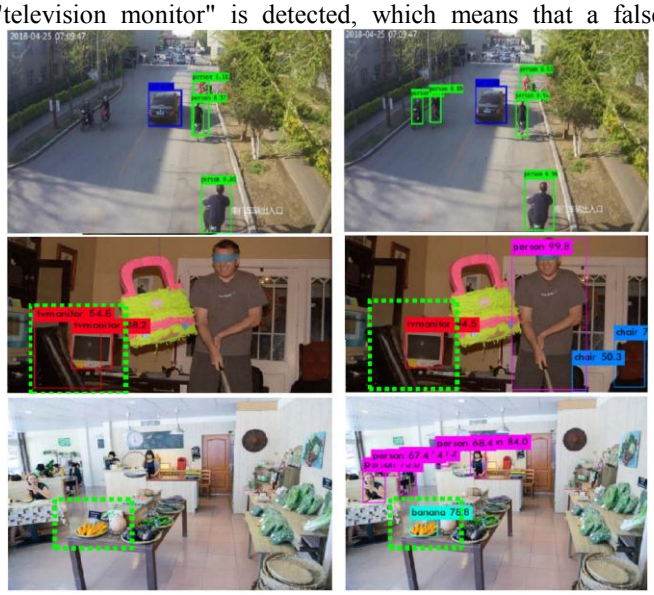

Figure 6.the results of YOLOv3 and the proposed method

IV. CONCLUSION

The paper adopts the YOLOv3 network, by removing the calculation of non-interest areas to improve the operation speed of YOLOv3. It increases the operating speed in proportion to the size of the non-interest area. The author studies the characteristics of the convolutional layer and the YOLO layer, and proposes a new processing method from the experimental results. In this paper, the ROI image processing technology in the video detection sequence is used to filter the ROI region, eliminate the calculation outside the ROI, and improve the detection speed and accuracy of the YOLOv3 network.Construction

V. ACKNOWLEDGMENTS

This work was supported by information construction of higher education management under Internet plus model under Grant No.41610200001/005. and by the design and Application of Intelligent Teaching Model Based on Rain Classroom Grant No. 41620200001/002. Data Mining Research on MOOC Learning Behavior Based on Internet Environment under Grant 80059900227.

REFERENCES

[1] Poblete V, Espic F, King S, et al. Regional deep learning model for visual tracking[J]. Neurocomputing, 2015, 175(PA):310-323.
[2] Grabner H, Leistner C, Bischof H. Semi-supervised On-Line Boosting for Robust Tracking[C]// European Conference on Computer Vision. Springer-Verlag, 2008:234-247.
[3] Avidan S. Ensemble Tracking[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2005:494-501.
[4] Zhang K, Zhang L, Yang M H, et al. Fast Tracking via Spatio-Temporal Context Learning[J]. Computer Science, 2013.
[5] Kalal, Zdenek. Tracking learning detection[D]. University of Surrey, 2011.
[6] Ojala T. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 24(7):971-987.
[7] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 779–788, 2016.
[8] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 6517–6525, 2017.
[9] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” CoRR, vol. abs/1804.02767, 2018

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 580–587, 2014.

[11] R. Girshick, “Fast R-CNN,” in Proceedings of the IEEE International Conference on Computer Vision, 2015.

[12] R. Girshick, “Fast R-CNN,” in Proceedings of the IEEE International Conference on Computer Vision, 2015.

[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards RealTime Object Detection with,” IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2017.

[14] C. L. Zitnick, and P. Dollar, "Edge Boxes: Locating object proposals from edges," in Proceeding of European Conference on Computer Vision, 2014

[15] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," International Journal of Computer Vision, 2013.

[16] Yong Li, Bin Sheng, Lizhuang Ma, Wen Wu, and Z. Xie, "Temporally coherent video saliency using regional dynamic contrast," IEEE Transactions on Circuits and Systems for Video Technology, vol. 23 , issue. 12, pp. 2067-2076 December 2013.

[17] 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 Proceeding of European Conference on Computer Vision, 2014.

[18] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv, 2018.