mathematics

MDPI

Article

SE-IYOLOV3: An Accurate Small Scale Face Detector for Outdoor Security

Zhenrong Deng [1], Rui Yang [2], Rushi Lan [2,3,*], Zhenbing Liu [2] and Xiaonan Luo [2]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; zhrdeng@guet.edu.cn

[2] Guangxi Key Laboratory of Images and Graphics Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China; 1803304025@mails.guet.edu.cn (R.Y.); zbliu@guet.edu.cn (Z.L.); luoxn@guet.edu.cn (X.L.)

[3] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510000, China

* Correspondence: rslan@guet.edu.cn

Received: 10 December 2019; Accepted: 27 December 2019; Published: 7 January 2020

Abstract: Small scale face detection is a very difficult problem. In order to achieve a higher detection accuracy, we propose a novel method, termed SE-IYOLOV3, for small scale face in this work. In SE-IYOLOV3, we improve the YOLOV3 first, in which the anchorage box with a higher average intersection ratio is obtained by combining niche technology on the basis of the k-means algorithm. An upsampling scale is added to form a face network structure that is suitable for detecting dense small scale faces. The number of prediction boxes is five times more than the YOLOV3 network. To further improve the detection performance, we adopt the SENet structure to enhance the global receptive field of the network. The experimental results on the WIDERFACEdataset show that the IYOLOV3 network embedded in the SENet structure can significantly improve the detection accuracy of dense small scale faces.

Keywords: small scale face; SENet; face detection; SE-IYOLOV3

1. Introduction

Face detection refers to the detection of the relative position and size information of all face targets in the image through the computer intelligence system. Small scale face detection means that on the basis of face detection, the small face information of the target can be accurately detected. This subject has a wide range of application prospects, including security [1], traffic statistics [2], digital cameras [3], pattern recognition [4], and other aspects.

Traditional face detection methods are mostly used for single face matching in a simple background [5]. For example, the PCA method [6] is used to extract facial features; serial and parallel methods are used to combine the extracted facial features [7]; and the LBP pattern is widely used for face recognition [8–10]. Due to the limitations of traditional face detection algorithms, it is usually effective to detect a single face in a specific environment, but the accuracy of face recognition for a dense small scale is low.

Since the emergence of the AlexNet network structure model in 2012 [11], the application of convolutional neural networks in face detection has been greatly developed [12]. The powerful learning ability of convolutional neural networks can greatly improve the accuracy of image detection; among them from R-CNN [13] generated by region proposal using selective search technology, spatial pyramid pooling network [14], single stage training Fast R-CNN [15], to improved Faster R-CNN [16] based on a fully convolutional neural network [17]. Researchers found that the corresponding improvement

of the general target detection method applied to face detection tasks can achieve better results than traditional methods [18]. Jiang Het al. used the face dataset to retrain Faster R-CNN [19] for face detection. Wan S et al. improved the Faster R-CNN model [20] and iteratively trained for face detection on the FDDB dataset [21]. Li used the cascaded Faster R-CNN structure [22] to improve detection accuracy. However, the above network adopted a two stage detection method, and the speed was slow. To solve this problem, Redmon et al. proposed the YOLO (You Only Look Once) model [23]. Using the whole graph as the input of the network, the position of the bounding box and the category of the bounding box were directly regressed in the output layer, which greatly improved the detection speed, but the detection accuracy was low. Later, he proposed the YOLOV2 [24] and YOLOV3 [25] detection algorithms successively in 2017 and 2018. Among them, YOLOV3 had a better detection effect, achieving an MAP effect of 57.9 percent within 51 ms on the COCO dataset [26]. Therefore, YOLOV3 could guarantee the accuracy and detection rate at the same time in the target detection field.

Face detection is a major issue in target detection. Many scholars have made significant progress in related fields [27–29]. For faces of different sizes, Guo et al. [30] proposed MSFD, which is a multi-scale face detector in the reception domain and can detect faces of different scales. For face clustering, Wang [31] proposed using graph convolutional networks [32] for face clustering to improve the recall rate of multiple faces. Luo et al. [33] added two residual units to the original YOLOV3 to detect smaller targets. Wu proposed that SENet [34] be embedded into the DenseNet [35] network prediction model, which can realize feature re-calibration in the process of feature extraction and improve the accuracy of network prediction.

To improve the speed and accuracy of dense small scale face detection, a detection method for embedding the squeeze-and-excitation networks (SENet) structure into an improved YOLOV3 network is proposed. Based on the k -means algorithm [36], we used the niche technology [37] to calculate the anchor box with higher average intersection over union (IOU) [38], which reduced the impact of the random initialization anchor box on detection accuracy. In order to make the algorithm more suitable for detecting smaller dense faces, the width of the prediction layer was changed, the number of prediction frames was increased by more than five times, and the small scale face information was captured. Finally, the SENet structure was fused to enlarge the perception field of the network and improve the score of a face that was not easy to recognize, so as to obtain higher precision and recall. The experimental results showed that the proposed network structure could significantly improve the detection of dense small scale faces on WIDERFACE [39] datasets, and the speed and accuracy of face detection achieved good results. The contributions of this paper are as follows: (1) A prediction frame calculation method that combines the small niche technology with K-means is proposed. (2) For small face detection, the YOLOV3 prediction layer scale is improved. (3) The SENet structure is embedded in the YOLOV3 network model.

The remainder of this article is organized as follows. Section 2 describes the improvement of YOLOV3 and introduces the specific composition structure of SE-IYOLOV3. Section 3 presents the experimental results in detail. Finally, the article is summarized in Section 4.

2. SE-IYOLOV3 Model

2.1. Improved YOLOV3 Model

YOLOV3 is a new end-to-end target detection model after R-CNN, Fast R-CNN, and Faster R-CNN. It combines the target classification and detection training, directly regresses the position and category of the target detection frame in the output layer, and converts the detection problem into a regression problem. At the same time, the detection task is concentrated in a convolutional neural network, which completes the output from the input of the original image to the target category and location.

2.1.1. Improved Anchor Box Algorithms

In the process of detecting dense faces, the accuracy of the detection depends on the coordinates of the last prediction frame of each grid, and the coordinate values of the anchor box are randomly initialized when the network starts training. Therefore, the result of random initialization of the anchor box has an important impact on the accuracy of network prediction. The YOLOV3 algorithm uses the K-means algorithm to cluster data. The K-means algorithm has low accuracy in selecting initial points and needs many attempts to get a better solution. Based on the K-means algorithm, this paper uses the niche technology to adjust the fitness of individuals in a population by sharing functions reflecting the similarity between individuals. The fitness between individuals is embodied in the similarity of the individual genotype or individual phenotype. When individuals are comparatively similar, the value of their shared function is relatively large; thus, the anchor box with a higher intersection ratio can be obtained. The distance function between each prediction box and the reference standard box is defined as Formula (1), where IOU represents the ratio of the intersection and union sets of "predicted borders" and "real borders".

$$d(x) = \sum_i \sum_j 1 - IOU(box_i, truth_j) \tag{1}$$

The specific steps are as follows: Step 0: Set the maximum number of iterations; set the initial particle flying speed v = 0; and use the K-means algorithm to cluster the data to obtain m initial cluster centers. Step 1: Calculate the sharing degree of individuals in the group. The shared function of this paper is calculated by the distance Formula (1). The smaller the distance, the larger the shared value. Step 2: After calculating the sharing degree of each individual in the group, adjust the fitness of each individual according to the following formula:

$$F_i = \sum_{i=1}^{m} d_i \qquad i = (1, 2, 3 \ldots m) \tag{2}$$

Step 3: Arrange them in ascending order according to the fitness of each individual; remember the first n individuals ($n < m$); carry out proportional selection operation on population $P(m)$ to obtain $P(t)$; and then, do cross selection and uniform variation calculation on $P(t)$ to get $P_i(t)$. Step 4: Combine n and t individuals in memory into a new clustering $n + t$. Compare the fitness of the individuals in the clustering, and impose penalty function $Fmin(x_i, x_j) = Penalty$ on the individuals with higher fitness. Step 5: Repeat Step 3 to update the evolutionary algebraic memory $e = e + 1$ until the highest number of iterations, and the population with the least fitness is the output.

By combining the K-means algorithm and the niche technology, the influence of the random initial point on the prediction result can be reduced. By finding the cluster group with the highest fitness, that is the higher similarity, the anchor box with the higher IOU can be obtained.

2.1.2. Change the Loss Function

The loss function used by YOLOV3 is a binary cross entropy loss (*BCELoss*), which is represented as:

$$BCEloss = -\frac{1}{n} \times \sum_i (t_i \times log(o_i) + (1 - t_i) \times log(1 - o_i)) \tag{3}$$

where o_i is the output value and t_i is the target value. Since the structure of the network layer needs to be changed after that, in order to prevent the predicted value from being too large, the negative predicted value causes the loss function to take too long to converge or have difficulty converging, so a sigmoid layer is added before the *BCEloss* loss function is used; the variable is mapped between zero and one; and then, the value is transferred to the loss function for calculation. Therefore, replace the

loss function with the *BCEWithLogitsLoss* loss function with better numerical stability, as shown in the following formula:

$$BCEWithLogitsLoss = -\frac{1}{n} \times \sum_i (t_i \times log(sigmoid(o_i)) + (1 - t_i) \times log(1 - sigmoid(o_i))) \quad (4)$$

The *BCEWithLogitsLoss* loss function integrates the sigmoid layer into the BCELoss class and uses the log-sum-exp technique to achieve numerical stability.

2.1.3. Improved Prediction Layer Scale

The YOLOV3 algorithm uses the DarkNet-53 network, which contains 53 convolutional layers. It combines three different scale feature maps, using a high resolution of low level features and high semantic information of high level features. By upsampling the features of different layers, objects are detected on three different scale feature layers. As shown in Figure 1, the bottom level downsampling feature map is $13 * 13$, and the two upsampling feature maps are $26 * 26$, $52 * 52$, respectively. The YOLOV3 network has 32 times downsampling of the input detection image. The downsampling factor is high; the receptive field of the feature map is relatively large; and the shallow information is not fully used, which will cause some information to be lost after multi-layer convolution. Therefore, it is suitable for detecting relatively large sized objects in an image.

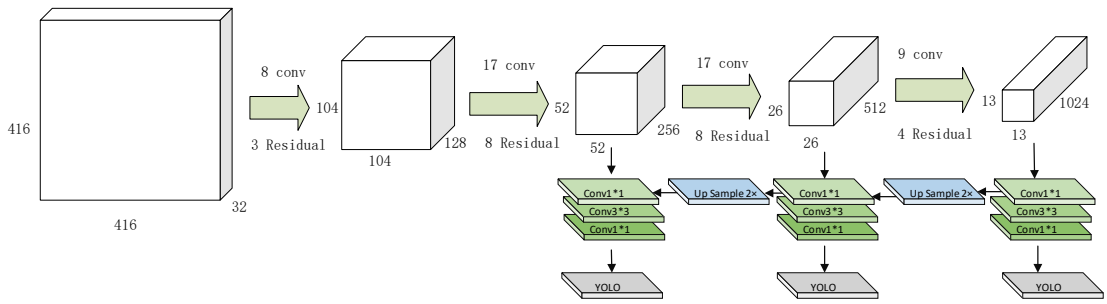

Figure 1. Original YOLOV3 network architecture.

Consider that when there are dense small scale faces in the input image, the detection effect on small scale faces is not ideal. We improved the scale detection module in YOLOV3 and expanded the scale of the original detection from three to four. As shown in Figure 2, when performing multi-scale fusion, an upsampling fusion operation is used, and a feature map with an upsampling size of 104 * 104 is added. For larger feature maps, we assigned a more accurate anchor box to the target. By taking 12 different sizes of anchor boxes to predict faces of different scales, the sizes were (12, 16), (16, 24), (21, 32), (24, 41), (24, 51), (33, 51), (28, 62), (39, 64), (35, 74), (44, 87), (53, 105), (64, 135). When the original YOLOV3 had three scales, it could predict a total of 3549 bounding boxes. When performing multi-scale fusion, an upsampling fusion operation was used. After adding a scale for the fusion operation, the training model could predict 14,365 bounding boxes, which was closer to five times the original three scale model. It could greatly improve the recognition rate of small targets.

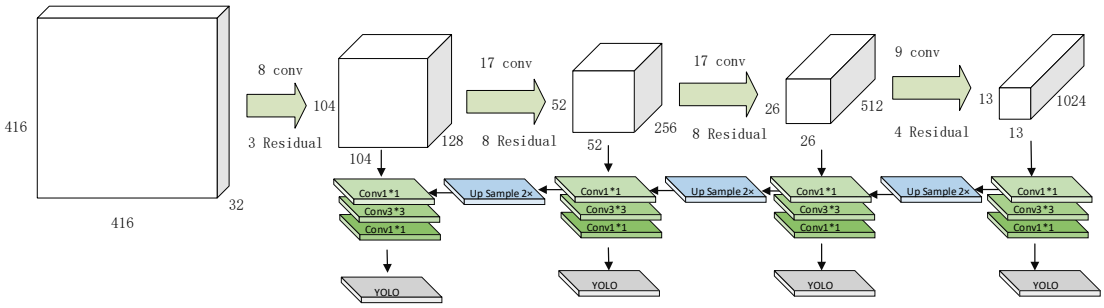

Figure 2. Improved YOLOV3 network structure of the prediction layer.

2.2. SE-IYOLOV3

SENet is a convolutional neural network structure proposed in 2017. It was the champion of the Image Classification task in the last ImageNet Competition. It proposes a method to emphasize information features selectively and suppress less useful features by learning to use global information. The core is squeeze and excitation operations. The structure is shown in Figure 3, which is a repetitive unit composed of the conventional shortcut layer and SE structure. The squeeze operation uses a global average pooling. The results showed the numerical distribution of C feature maps in this layer, also known as global information. The excitation operation uses a gating mechanism and sigmoid activation function to describe the weight of C feature maps in the tensor. The function of two Fully Connected layers (FC) is to fuse the feature map information of each channel.

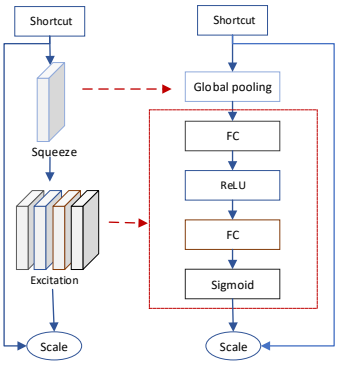

Figure 3. SENet structure. FC, Fully Connected layer.

In densely distributed images, conventional YOLOV3 often erroneously detects or misses face detection, which is due to misrecognition caused by an unbalanced confidence distribution. In order to make the network learn global features and improve the detection accuracy of dense faces, the weight of each feature channel is automatically calibrated.

SENet structure is embedded in the improved YOLOV3 network, and a feature map is transformed into a number with global receptive fields. The robustness of the whole neural network can be enhanced by retaining the global information under the condition of greatly reducing the computational parameters. In YOLOV3, there is a shortcut layer whenever a $1 * 1$ conv and $3 * 3$ conv combination is ended, so the shortcut layer aggregates multiple layers of features. Embedding the SENet structure into the shortcut layer will expand the range of perception of the global information by the feature map. In the YOLOV3 network, there are 23 shortcut layers. Therefore, the improved YOLOV3 network will be changed from the original 107 layer to the 130 layer, as shown in Figure 4.

The feature map of $W * H * C$ is transmitted from the shortcut layer, where W is the width, H is the height, and C is the number of channels. After the global average pooling, the feature map of $1 * 1 * C$ is obtained. After that, the dimension reduction of the first fully connected layer becomes $1 * 1 * C/r$,

where r is the dimension reduction parameter, and $r = 16$ was taken in this paper. The dimension reduction becomes $1 * 1 * C$ after the second fully connected layer, and after the sigmoid function, the dimension reduction becomes the weight value of $1 * 1 * C$. Finally, the input feature map is multiplied by the weight value as the input to the next layer. Therefore, the feature map size of the network layer output that added the SENet block is shown in Table 1, where the CSR module is a submodule composed of a convolutional + shortcut + SENet layer.

The number before the multiplier represents the number of modules with the same size of the feature map, for example $4 * CSR$, $13 * 13 * 1204$, indicating that there are four CSR modules with an output feature size of $13 * 13 * 1204$. The YOLOV3 network embedded in the SENet structure can fuse the shallow information with the deep information and efficiently utilize the multi-dimensional feature information, thereby expanding the global receptive field of the information, and it can slow down the attenuation of the error items of each hidden layer and ensure the stability of the gradient weight information.

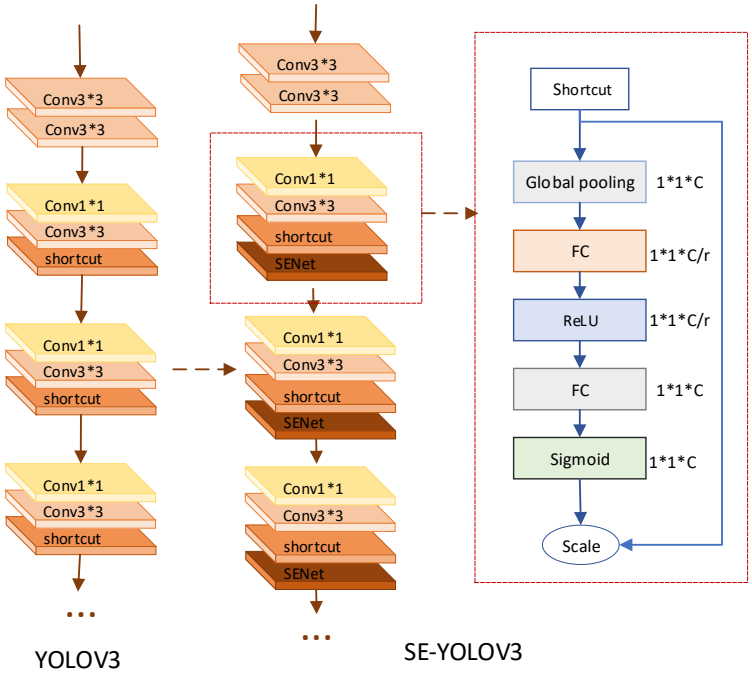

Figure 4. SE-IYOLOV3 structure.

Table 1. Feature map information output from the SENet layer.

CSRModule	Output Feature Map Size
1 * CSR	208 * 208 * 64
2 * CSR	104 * 104 * 128
8 * CSR	52 * 52 * 256
8 * CSR	26 * 26 * 512
4 * CSR	13 * 13 * 1024

3. Experimental Results

In order to speed up the convergence of the network and to avoid over-fitting, the impulse constant was set to 0.9, the weight attenuation coefficient to 0.0005, and the initial learning rate to 0.0005. The experimental environment was the Ubuntu 14.04 operating system, Intel (R) Xeon (R) CPU

E5-2698 v4 @ 2.20 GHz processor, 16 GB running memory (RAM), GPU for NVIDIA Tesla K80, and 16 G memory.

3.1. Datasets

In YOLOV3, the features of the image were extracted mainly through the Darknet53 network, and the facial features needed to be learned from a large number of samples. Therefore, in order to learn better feature representation, it was necessary to adopt a dataset with obvious facial features. In this paper, the WIDERFACE dataset with obvious facial features was used for training and testing.

The WIDERFACE detection dataset contained 32,203 images and 393,703 face images, which showed great changes in scale, posture, occlusion, expression, dressing, and care. WIDER FACEwas based on 61 event categories. For each event category, 50 percent of them were selected as the training set, 10 percent for cross-validation, and 40 percent for the test set.

3.2. Convergence Verification of Improved YOLOV3 Embedded SENet Structure Model

Based on the improved YOLOV3 structure and embedded SENet structure, a training intensive face detection model was built. The results showed that the model could converge to a stable state quickly in the training process. The performance of the trained model on the test dataset was better than that of the original YOLOV3 model.

In the process of training with the WIDERFACE dataset, the log information of each iteration of training of the improved SE-YOLOV3 model was collected, including the accuracy of face detection, the average IOU value, the accuracy of correct classification, the total number of detected faces, and the recall rate. By visualizing the information, as shown in Figure 5, the loss function converged steadily in the first 2000 iterations as the number of iterations increased.

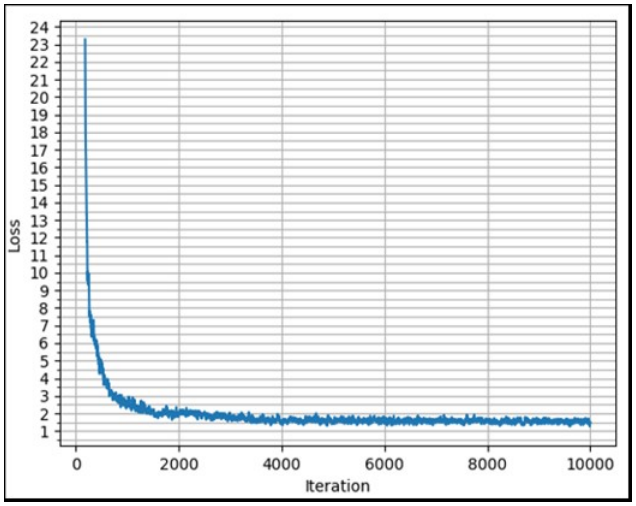

Figure 5. Loss curve of model training.

3.3. The Impact of Different Improvement Strategies on the Average IOU

The three improved strategies proposed above are respectively calculated for the accuracy of the model, and the original YOLOV3 is used as a reference, as shown in Table 2.

Table 2. Average intersection:parallel ratio of the prediction box and real box. b, anchor box; P, prediction layer.

Program Name	Percentage
YOLOV3	78.56
IYOLOV3-B	84.12
IYOLOV3-P	82.88
IYOLOV3-E	81.37
SE-IYOLOV3	85.98

Table 2 shows: (1) YOLOV3, the original YOLOV3 model; (2) IYOLOV3-B, the improved anchor box algorithm is added to the original YOLOV3 model; (3) IYOLOV3-P, the structure of the Prediction layer of the original YOLOV3 model is improved; (4) IYOLOV3-E , only the SENet module is introduced to the original YOLOV3 model; (5) SE-IYOLOV3, the face detection model proposed in this paper. As can be seen from Table 2, each of the improved strategies used in this paper improved the performance of the original YOLOV3 detection network to varying degrees. Among them, the improvement of the anchor box algorithm had the most significant improvement on the accuracy of the model, with the mean value of IOU increased by nearly six percentage points; the improvement of the prediction layer structure of the network raised the mean value of IOU by nearly four percentage points; and the addition of the SENet structure raised the mean value of IOU by nearly three percentage points. Each improvement strategy was integrated, and the final average IOU was nearly eight percentage points higher than the original YOLOV3 network.

3.4. Comparison of Different Detection Models

Taking the Precision rate (P) and Recall rate (R) as evaluation indexes, the method was compared with R-CNN, FAST-RCNN, FASTER-RCNN, and YOLOV3 with different improvement strategies. In order to accelerate the convergence speed of the network and avoid over-fitting, the impulse constant was set to 0.9, the weight attenuation coefficient to 0.0005, and the initial learning rate to 0.0005. Moreover, the multi-step strategy was adopted, and the dataset was WIDERFACE. The detection results are shown in Table 3. The precision and recall rate of the YOLOV3 network embedded in SENet was the highest, because the SENet structure enhanced the global receptive field of the feature map, so that the information learned by the network was more comprehensive. Therefore, the face features that were not easily recognized had higher scores, which made the network's precision and recall rate higher. IYOLOV3-B performed better than the original YOLOV3 because it used the improved anchor box algorithm to get an anchor box with a higher average IOU. IYOLOV3-P had higher performance than the original YOLOV3 because it changed the prediction layer structure and increased the number of prediction frames by more than six times, which was more accurate for capturing dense face images. Therefore, by embedding SENet into the improved YOLOV3 structure, the precision and recall rate were increased by 17 percent and 26 percent respectively compared with the original YOLOV3.

The detection results are shown in Figure 6. (a) is the effect of the YOLOV3 model detecting dense small scale faces, and (b) is the effect of the method in this paper.

From the comparison of the first picture, it can be seen that YOLOV3 incorrectly recognized the fingers of a man in green clothes as a human face in the case of a complicated background. The middle comparison chart shows that YOLOV3 did not detect the man on the far right. In the last picture, the face detection effect of this method was significantly better than the original YOLOV3.

Table 3. Feature map information output from the SENet layer.

Models	Improved Anchor Box	Improved Prediction Layer	SENet Added	Precision %	Recall %	Detection Speed (ms)
R-CNN [13]	✗	✗	✗	68.5	54.2	1300
FAST-RCNN [15]	✗	✗	✗	82.6	69.4	700
Faster RCNN [16]	✗	✗	✗	86.4	76.3	350
Single Stage Joint [40]	✗	✗	✗	92.1	87.8	510
YOLOV3 [25]	✗	✗	✗	75.6	63.4	230
IYOLOV3-B	✓	✗	✗	90.5	86.1	360
IYOLOV3-P	✗	✓	✗	90.1	88.2	340
SE-IYOLOV3	✓	✓	✓	92.3	89.6	460

(a) YOLOV3

(b) SE-IYOLOV3

Figure 6. Face detection results via YOLOV3 and the proposed SE-IYOLOV 3.

4. Conclusions

In order to solve the problem of dense face detection, this paper firstly used the niche technology to calculate the anchor box with higher average IOU based on the K-means algorithm, which reduced the impact of the randomly initialized anchor box on the detection accuracy. To make the algorithm more suitable to detect smaller dense faces, the width of the prediction layer was changed, changing three dimensions of the original network to four. Finally, the SENet structure was fused to enlarge the perception field of the network and improve the score of the face that was not easy to recognize. The experimental results showed that the proposed network structure could significantly improve the detection accuracy of dense small scale faces. In future research, we will consider reducing the parameters and network layers to improve the detection speed of the network and using a densely connected upper sampling layer to improve detection accuracy.

Author Contributions: Conceptualization, Z.D.; Methodology, R.Y.; Formal analysis, R.L.; Project administration, Z.L.; Investigation, X.L.; Supervision, Z.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Natural Science Foundation of China (Nos. 61772149, 61702129, 61762028, and U1701267), the China Postdoctoral Science Foundation (No. 2018M633047), and the Guangxi Science and Technology Project (Nos. 2018GXNSFAA138132, AD18216004, AD18281079, and 2018GXNSFAA294132).

Acknowledgments: The authors would like to sincerely thank the anonymous reviewers for their valued comments and constructive suggestions, which significantly improved the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liao, S.; Jain, A.K.; Li, S.Z. A Fast and Accurate Unconstrained Face Detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *38*, 211–223.
2. Zhao, x.; Liang, X.; Zhao, C.; Tang, M.; Wang, J. Real-Time Multi-Scale Face Detector on Embedded Devices. *Sensors* **2019**, *19*, 2158. doi:10.3390/s19092158.
3. Dong, W.; Jing, Y.; Deng, J.; Liu, Q. FaceHunter: A multi-task convolutional neural network based face detector. *Signal Process. Image Commun.* **2016**, *47*, 476–481.
4. Zhang, C.; Zhang, Z. Improving multiview face detection with multi-task deep convolutional neural networks. In Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), Steamboat Springs, CO, USA, 24–26 March 2014.
5. Kramer, R.S.; Mohamed, S.; Hardy, S.C. Unfamiliar Face Matching With Driving Licence and Passport Photographs. *Perception* **2019**, *48*, 175–184.
6. Wu, Y.; Lim, J.; Yang, M.H. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848.
7. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
8. Lu, Z.; Jiang, X.; Kot, A. A novel LBP-based Color descriptor for face recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA, 5–9 March 2017; pp. 1857–1861.
9. Lan, R.; Zhou, Y.; Tang, Y. Quaternionic Local Ranking Binary Pattern: A Local Descriptor of Color Images. *IEEE Trans. Image Process.* **2016**, *25*, 566–579.
10. Lan, R.; Zhou, Y. Quaternion-Michelson Descriptor for Color Image Classification. *IEEE Trans. Image Process.* **2016**, *25*, 5281–5292.
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 3–8 December 2012; pp. 1097–1105.
12. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503.
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 7–12 December 2015 ; pp. 91–99.
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

18. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the BMVC 2015, Swansea, UK, 7–10 September 2015. ; p. 6.
19. Jiang, H.; Learned-Miller, E. Face detection with the faster R-CNN. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 650–657.
20. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. How far are we from solving pedestrian detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1259–1267.
21. Najibi, M.; Samangouei, P.; Chellappa, R.; Davis, L.S. Ssh: Single stage headless face detector. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4875–4884.
22. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 7263–7271.
25. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767 .
26. Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; Zhu, J. Efficient Decision-based Black-box Adversarial Attacks on Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 7714–7722.
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
28. Lan, R.; Zhou, Y.; ; Liu, Z.; Luo, X. Prior Knowledge-Based Probabilistic Collaborative Representation for Visual Recognition. *IEEE Trans. Cybern.* **2018**, doi:10.1109/TCYB.2018.2880290.
29. Lan, R.; Sun, L.; Liu, Z.; Lu, H.; Su, Z.; Pang, C.; Luo, X. Cascading and Enhanced Residual Networks for Accurate Single Image Super-resolution. *IEEE Trans. Cybern.* **2019**, doi:10.1109/TCYB.2019.2952710.
30. Guo, Q.; Dong, Y.; Guo, Y.; Bai, H. MSFD: Multi-Scale Receptive Field Face Detector. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1869–1874.
31. Wang, Z.; Zheng, L.; Li, Y.; Wang, S. Linkage Based Face Clustering via Graph Convolution Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 1117–1125.
32. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
33. Moran, J.; Haibo, L.; Zhongbo, W.; Miao, H.; Zheng, C.; Bin, H. Improved YOLO V3 Algorithm and Its Application in Small Target Detection. *Acta Opt. Sin.* **2019**, *39*, 0715004.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, OR, USA, 18–22 June 2018; pp. 7132–7141.
35. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 4700–4708.
36. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137.
37. Tjin, G.; Flores-Figueroa, E.; Duarte, D.; Straszkowski, L.; Scott, M.; Khorshed, R.A.; Purton, L.E.; Celso, C.L. Imaging methods used to study mouse and human HSC niches: Current and emerging technologies. *Bone* **2018**, *119*, S8756328218301765.
38. Nowozin, S. Optimal Decisions from Probabilistic Models: The Intersection-over-Union Case. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014.

39. Yang, S.; Ping, L.; Chen, C.L.; Tang, X. WIDER FACE: A Face Detection Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
40. Deng, J.; Guo, J.; Zafeiriou, S. Single-Stage Joint Face Detection and Alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Long Beach, CA, USA, 15–21 June 2019.