Face Mask Detection Based on SSDv2 Network

Zheng Shanshan, Wu Kaisheng, Lan Yuxiang, Lin Jin
School of Informatics Xiamen University
No.422 Siming South Road, Siming District, Xiamen City, Fujian Province
{shanshanzheng, wukaisheng, lanyuxiang, linjin}@stu.xmu.edu.cn

Abstract

2020 is an emergency year for epidemic prevention and control. In order to reduce the risk of cross-infection of people in public places, by detecting whether face masks are worn or not, people can be reminded in time to wear masks correctly and further protect themselves and others' lives. We proposed a multi-scale detection algorithm to aim at the difficulty of detecting obscured targets and small targets in the mask wearing detection tasks, which is called SSDv2. The algorithm is based on the SSD network, not only can obtain feature maps of different scales, but also uses the feature pyramid strategy for feature fusion to enhance detailed feature information. At the same time, prediction is made on different feature maps, which realizes the full use of different levels of feature information, and further improves the accuracy of algorithm detection. Experiments show that the accuracy of the SSDv2 algorithm for detecting mask wearing targets and unmask targets in different scenarios reaches 80.97% and 64.61%,respectively. And the detection accuracy of mask targets is improved by 3.37% compared with the YOLOv3 algorithm, which shows our method has its advantage.

Introduction

With the continuous development of society, face recognition, as a kind of modern biological information recognition, is obtained by its technical convenience and information security in the development process of ensuring personal information security and realizing rapid identification and authentication of human identity. It is widely recognized and used by people. Among them, face detection is the basis of other applications of face. Only after obtaining the accurate location information of human face and face, other applications can continue to develop. Therefore, face detection is still the focus and difficulty in the research of face recognition technology. Because of its importance, it has become a subject of widespread concern and active research in the field of pattern recognition and computer vision in the past 20 years.

The current research on face recognition has made remarkable progress. Especially in practical applications, excellent results have been achieved, such as in school, hospital, subway and etc. Most mainstream face recognition focuses on the two major tasks of face detection and recognition, and recent studies have shown surprisingly high precision and accuracy.

Face detection is the problem of positioning a box to bound each face in a photo. Facial landmark detection seeks to localize specific facial features: e.g. eye centers, tip of the nose. Together, these two steps are the cornerstones of many face-based reasoning tasks, most notably recognition (Deng et al. 2019) (Masi et al. 2017) (Masi et al. 2016) (Masi et al. 2019) (Wang et al. 2018) (Wolf, Hassner, and Maoz 2011) and 3D reconstruction (Feng et al. 2018) (Hernandez et al. 2017) (Tuan Tran et al. 2017) (Tran et al. 2017)Processing typically begins with face detection followed by landmark detection in each detected face box. Detected landmarks are matched with corresponding ideal locations on a reference 2D image or a 3D model, and then an alignment transformation is resolved using standard means (Dementhon and Davis 1995) (Lepetit, Moreno-Noguer, and Fua 2009) The terms face alignment and landmark detection are thus sometimes used interchangeably (Browatzki and Wallraven 2020) (Dapogny, Bailly, and Cord 2019) (Kumar et al. 2020).

Although this approach was historically successful, it have to turn to the task of face mask detection. At the beginning of the new year of 2020, the sudden outbreak of COVID-19 severely disrupted our pace of life. Masks play an important role in the prevention and control of the epidemic, so they have already been a necessity in people's lives. Wearing a mask is the most essential and effective measure to prevent novel Coronavirus from spreading. However, in real life, wearing a mask will inevitably lead to respiratory discomfort, so some people with insufficient awareness of prevention will unconsciously take off masks or directly do not wear masks. This kind of behavior is a great threat to public safety. Therefore, in certain places, such as hospitals, buses, subways and subway stations and other public places with high crowd density, it is right to test whether to wear masks. Therefore, we turn the face recognition task into a recognition task of whether the face wears a mask. In particular, we carry out the detection task and the recognition task at the same time, which greatly improves the timeliness of the network. Face mask detection is essentially a specific target detection task. Current state-of-the-art

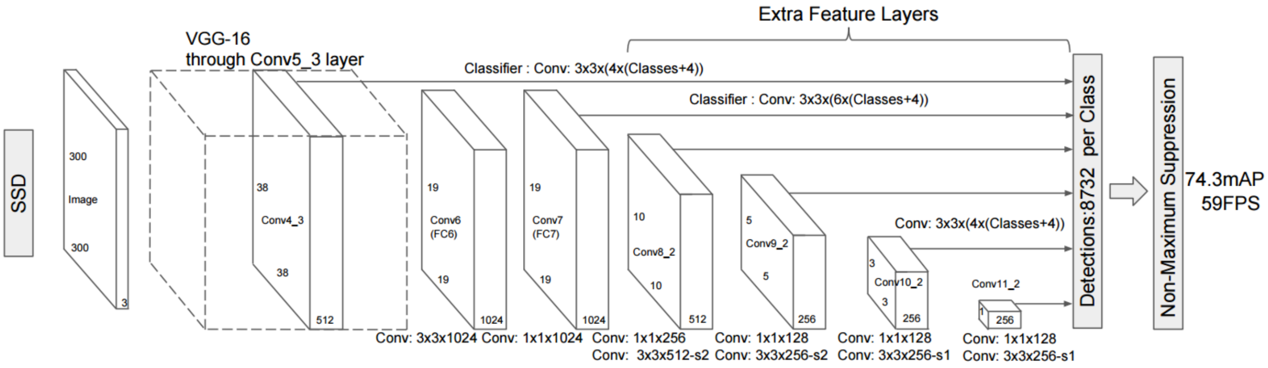

Figure 1: Overall structure of SSD.

object detection systems are variants of the following approach: hypothesize bounding boxes, resample pixels or features for each box, and apply a high-quality classifier.

SSD (Liu et al. 2016), a single-shot detector for multiple categories, is faster than the previous state-of-the-art for single shot detectors (YOLO), and significantly more accurate, in fact as accurate as slower techniques that perform explicit region proposals and pooling (including Faster R-CNN). In this paper, we improve the SSD (Single Shot MultiBox Detector) network and recognize different face targets with higher recognition efficiency.On these tasks, the experimental results show that the SSD-improved-NET works well on the dataset.The main contributions of this article are as follows:

- We use the improved SSD network as the backbone network of the algorithm in this paper. The improved SSD proposed in this paper can effectively and quickly recognize different face-mask targets and the algorithm has certain robustness for different posesangles and light occlusion.
- The test accuracy of the model has reached a high level.

Related Work

Feature-based face detection method

The feature-based method mainly uses the obvious features in the face image to determine the position and size of the face, and treats the face image as a high-dimensional vector, thereby turning the face detection problem into a detection problem of distributed signals in a high-dimensional space.

The representative result is the method proposed by Rowley et al. (Rowley, Baluja, and Kanade 1998a)(Rowley, Baluja, and Kanade 1998b). They used neural networks for face detection and trained a multi-layer perceptron model with 20x20 face and non-face images. The method of (Rowley, Baluja, and Kanade 1998a) is used to solve the problem of face detection that is approximately frontal. The method of (Rowley, Baluja, and Kanade 1998b) solves the problem of multi-angle face detection. The entire system is composed of two neural networks. The first network is used to estimate the angle of the face, and the second is used to judge whether it is a face. The angle estimator outputs a rotation angle, and then uses the entire angle to rotate the detection window, and then uses the second network to judge the rotated image to determine whether it is a face.

Image-based face detection method

The idea of the image-based face detection method is to treat the face detection problem as a generalized pattern recognition problem, and let the computer summarize the facial features through learning. Through the research and time in recent years, it has been found that the image-based method is superior to the basic feature method. The current mainstream is to use the image-based method.

Viola and Jones (Li et al. 2002) proposed a face detection method based on Adaboost. The core idea of this method is to use exhaustive methods to find the Haar-Like features of face images that are different from non-face images, and then use cascade to improve the detection rate.

With the development of deep learning, people use more convolutional neural networks to solve the problem of face detection. Cascade CNN (Li et al. 2015) can be considered as a representative of the combination of traditional technology and deep network. Like the VJ face detector, it contains multiple classifiers. These classifiers are organized in a cascade structure. However, the difference lies in , Cascade CNN uses convolutional network as the classifier for each level. MTCNN (Zhang et al. 2016) , is a multi-tasking method. It combines face area detection and face key point detection. Like Cascade CNN, it is also based on the cascade framework, but the overall idea is more clever and reasonable. Face detection The alignment with the face is integrated into a framework, and the overall complexity is well controlled. Face R-CNN (Wang et al. 2017) This method is based on the Faster R-CNN framework for face detection and is optimized for the particularity of face detection. For the final two categories, center loss is added on the basis of softmax. By adding center loss, the feature difference within the class is smaller (playing a clustering role), and the difference between positive and negative samples in the feature space is improved to improve the performance of the classifier.

The two most famous real-time detectors are YOLOv3 (Redmon and Farhadi 2018) and SSD (Liu et al. 2016).

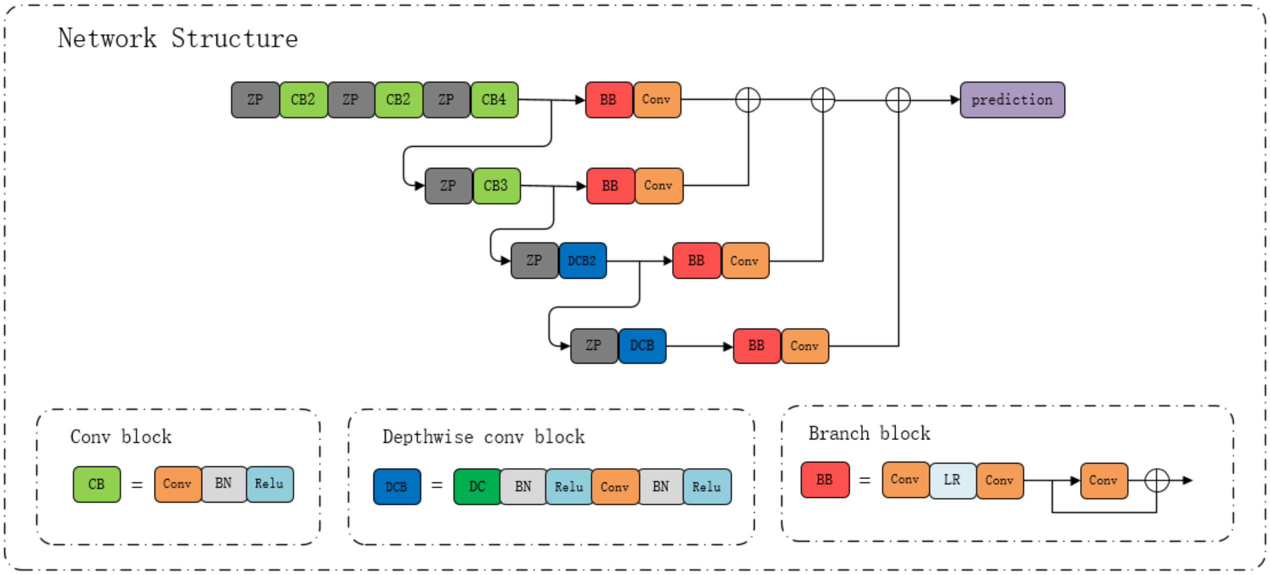

Figure 2: The overall frame structure of SSDv2 and it structural blocks.

Yolov4 (Bochkovskiy, Wang, and Liao 2020) established an efficient and powerful target detection model. The researchers added the SPP block to CSPDarknet53 because it can greatly increase the experience. Wild, isolate the most significant contextual features, and hardly reduce the network speed. They chose PANet from different backbone layers as the parameter aggregation method for different levels of detectors, and abandoned the FPN network used in YOLOv3.

Our Method

SSD Introduction

The SSD algorithm, its full English name is Single Shot MultiBox Detector. SSD uses VGG16 as the basic model, and then adds a new convolutional layer based on VGG16 to obtain more feature maps for detection.

The SSD detector differs from others single shot detectors due to the usage of multiple layers that provide a finer accuracy on objects with different scales. And the SSD normally start with a VGG on Resnet pre-trained model that is converted to a fully convolution neural network. Then we attach some extra conv layers, which will actually help to handle bigger objects. The SSD architecture can in principle be used with any deep network base model. The specific structure is shown in Figure 1.

SSDv2 Introduction

Overall Framework. Here we introduce a simple but effective network framework. It predicted the final result after concatenating a plurality of inflow and outflow branches. The overall framework is shown in Figure 2. In the figure, ZP module denotes Zero Padding layer, Conv module denotes convolution layer which kernel size is (3,3), CB denotes the Conv block, BB denotes the Branch block, DCB denotes the Depthwise conv block, CBn module means containing n CB modules and DCBn module means containing n DCB modules which are also showed in the Fig.2.

Detailed Introduction. SSDv2 can be regarded as composed of four sub-branches. The first sub-branch is first connected in series with multiple ZP modules and CB blocks to extract the mask face information in the picture, and the obtained output is passed to this branch and the next sub-branch, as shown in Figure 2. The output of the first sub-branch is passed to the current branch starting with BB block and the next sub-branch starting with ZP module. The output of the first sub-branch is passed to the current branch starting with BB block and the next sub-branch starting with ZP module. Similarly, subsequent branches continue in the same way. Finally, it constitutes the total four sub-branch of the SSDv2 algorithm. Obviously, these four branches gradually increase in scale from top to bottom. At last, concatenating the results of these four sub-branches to get the final results, and finally make predictions.

Therefore, each sub-branch in the framework transfers the current learning results to the current branch and the next branch to continue learning, which constitutes a multi-scale detection algorithm. Large-scale branches extract information from the low-resolution feature map, which has a larger receptive field, which is conducive to detecting large targets in the picture; small-scale branches extract information from the high-resolution feature map and have a smaller receptive field, which is conducive to detecting smaller targets in the picture. This is also in line with the feature fusion strategy of the feature pyramid, which can effectively improve the final prediction accuracy to a certain extent.

Loss Function. We use the multi loss as our training objective likely to SSD. The overall objective loss function is a weighted sum of the localization loss (loc) and the confi-

Figure 3: Example of dataset images.

dence loss (conf):

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

where N is the number of matched default boxes. If N = 0, wet set the loss to 0. The localization loss is a Smooth $L1$ loss between the predicted box and the ground truth box parameters. Similar to Faster R-CNN, we regress to offsets for the center (cx, cy) of the default bounding box and for its width and height. And the confidence loss is the softmax loss over multiple classes confidences and the weight term is set to 1 by cross validation.

Implementation Details. In our model, an image with the size of 240×320 is fed into our model. We use SGD with initial learning rate as 1×10^{-2} to optimize our model, the weight decay coefficient and momentum are set to 5×10^{-4} and 0.9, respectively. The maximum number of alternate training are set to 100. We train our model with 100 epochs on google colab with a single Tesla T4 GPU in a Tensorflow2.0 framework. In the inference phase, our model can also perform real-time inference. Our model only uses an AMD Ryzen 4800U CPU, and can use the notebook camera to achieve a detection speed of 14FPS.

Experiments

Experiment Settings

Datasets. We carry out experiments on the Face Mask Data, source data from AIZOOTech, which is a great job. The dataset contains 7,954 pictures, of which 3,893 pictures are people without masks, and the remaining 4,061 pictures are people with masks. The ratio of the two classes is close to 1:1. The pictures without masks are from WIDER Face and MAFA, and the pictures with masks are from the web. We divide the dataset into a training set and a test set according to the ratio of masks to those without masks. The training set contains 6,153 pictures and the test set contains 1,839 pictures.An example of the data set picture is shown in Figure 3.

Evaluation Metrics. The detection accuracy of each category in the detection of mask wearing is very important. Especially in the situation of epidemic prevention and control, false detections and missed detections may cause the risk of epidemic spread. Therefore, this paper chooses Average Precision (AP) and Mean Average Precision (mAP) as the evaluation indicators of the target detection algorithm. These two evaluation indicators take into account precision (Precision, P) and recall (Recall, R).

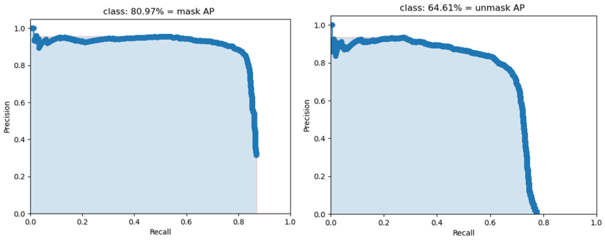

Figure 4: The P-R curve of the mask and unmask categories.

$$P(classes) = \frac{T_P}{T_P + F_P} \times 100\% \quad (2)$$

$$R(classes) = \frac{T_P}{T_P + F_N} \times 100\% \quad (3)$$

Take the mask category of the target detected in this paper as an example. In the above formula, T_P stands for the number of the target correctly worn by the model detected as a mask; F_P stands for the number of the target wrongly worn or without a mask detected as a mask; F_N stands for the number of the target incorrectly worn as an unmask detected. Take the values of recall rate and accuracy rate as the abscissa and ordinate respectively, draw a P-R curve, and the area under the curve is AP. If AP is obtained for all categories and the mean value is taken, then the mean value of precision is obtained. MAP is an important index to evaluate the performance of the model, which can reflect the overall performance of the network model and avoid the problem of extreme performance of some categories while weakening the performance of others in the evaluation process.

Experiment Results

Test results. We use the trained model to test the face mask on the test set. The P-R curve of the detection result of our algorithm is shown in Figure 4. The abscissa is the recall rate, and the ordinate is the precision rate. The AP value can be obtained by calculating the area of the shaded part of the P-R curve. The AP values of the mask and unmask categories reached 80.97% and 64.61%, respectively, and the mAP value can be calculated to be 72.79%.

We compare the experimental results with the experimental results of the original YOLOv3, Faster R-CNN, SSD and other mainstream target detection algorithms. The comparison results are shown in Table 1.

Table 1: Performance comparison results of four algorithm

algorithm	mask/%	unmask/%	mAP/%
RetinaFace	87.3	76.5	81.9
Attention-RetinaFace	90.6	84.7	87.7
YOLOv3	77.6	80.5	79.1
Ours	80.97	64.61	72.79

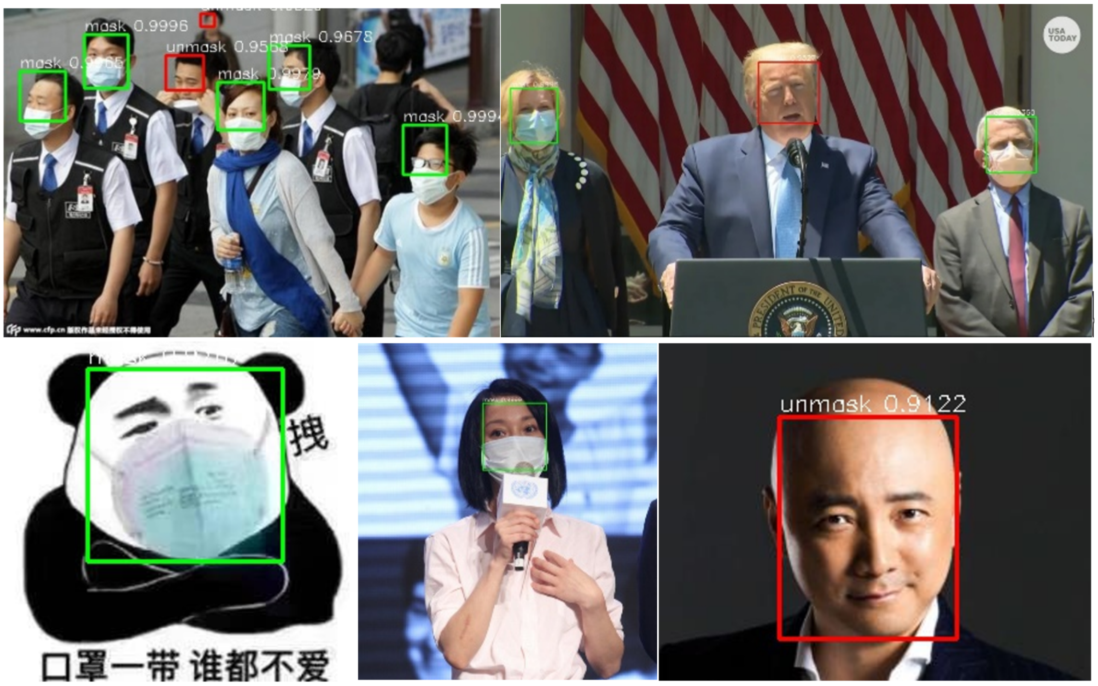

Figure 5: The test results of some images.

Table 2: Results of ablation experiments that gradually remove larger-scale branches.

ablation	mAP/%
1-branch	62.35
2-branch	67.12
3-branch	70.93
4-branch	72.79

Figure 5 shows some of our test results. We have selected five samples, a single person without mask, multiple people without mask, a single person with mask, multiple people with mask, and a mixed situation. It can be seen from the detection results that our algorithm has achieved a good effect in the detection of faces and the classification of whether to wear a mask.

Ablation study. As seen in Table 2, when only the branch with the smallest scale is used,which we call it 1-branch, the test accuracy is 62.35%; when the second scale branch is added(2-branch), the test accuracy can be increased by 67.12%; when another scale branch is added(3-branch), the test accuracy can be increased by 8.58% than when only using one branch; when all four branches are used, the test accuracy can reach 72.79%, which is the highest value of all combinations.

Conclusion

In this paper, we proposed a multi-scale detection algorithm to aim at the difficulty of detecting obscured targets and small targets in the mask wearing detection tasks, which is called SSDv2. The application of multi-scale features is conducive to the network detection of targets of different sizes, so that it can be applied to face mask detection in both distant and close shots. Experiments show that the network has good accuracy in the detection of masks worn by different people in different complex scenes. The network can also be used for real-time face mask wearing detection, which has a good application prospect.

Acknowledgments

Thanks to Mr. Lu Yang for giving us a semester deep learning course. From this course we learned a lot of deep learning knowledge, and we also have a logical understanding of deep learning. And by completing the coursework to consolidate the knowledge learned. Thank you teacher for your contribution, we will continue to learn and improve in the future.

References

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934* .

Browatzki, B.; and Wallraven, C. 2020. 3FabRec: Fast Few-shot Face alignment by Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6110–6120.

Dapogny, A.; Bailly, K.; and Cord, M. 2019. DeCaFA: Deep Convolutional Cascade for Face Alignment in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Dementhon, D. F.; and Davis, L. S. 1995. Model-based object pose in 25 lines of code. *International journal of computer vision* 15(1-2): 123–141.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 534–551.

Hernandez, M.; Hassner, T.; Choi, J.; and Medioni, G. 2017. Accurate 3D face reconstruction via prior constrained structure from motion. *Computers & Graphics* 66: 14–22.

Kumar, A.; Marks, T. K.; Mou, W.; Wang, Y.; Jones, M.; Cherian, A.; Koike-Akino, T.; Liu, X.; and Feng, C. 2020. LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8236–8246.

Lepetit, V.; Moreno-Noguer, F.; and Fua, P. 2009. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision* 81(2): 155.

Li, H.; Lin, Z.; Shen, X.; Brandt, J.; and Hua, G. 2015. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5325–5334.

Li, S. Z.; Zhu, L.; Zhang, Z.; Blake, A.; Zhang, H.; and Shum, H. 2002. Statistical learning of multi-view face detection. In *European Conference on Computer Vision*, 67–81. Springer.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Masi, I.; Hassner, T.; Tran, A. T.; and Medioni, G. 2017. Rapid synthesis of massive face sets for improved face recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 604–611. IEEE.

Masi, I.; Trn, A. T.; Hassner, T.; Leksut, J. T.; and Medioni, G. 2016. Do we really need to collect millions of faces for effective face recognition? In *European conference on computer vision*, 579–596. Springer.

Masi, I.; Trn, A. T.; Hassner, T.; Sahin, G.; and Medioni, G. 2019. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision* 127(6-7): 642–667.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* .

Rowley, H. A.; Baluja, S.; and Kanade, T. 1998a. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence* 20(1): 23–38.

Rowley, H. A.; Baluja, S.; and Kanade, T. 1998b. Rotation invariant neural network-based face detection. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*, 38–44. IEEE.

Tran, A. T.; Hassner, T.; Masi, I.; Paz, E.; Nirkin, Y.; and Medioni, G. 2017. Extreme 3D Face Reconstruction: Seeing Through Occlusions. *arXiv preprint arXiv:1712.05083* .

Tuan Tran, A.; Hassner, T.; Masi, I.; and Medioni, G. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5163–5172.

Wang, H.; Li, Z.; Ji, X.; and Wang, Y. 2017. Face r-cnn. *arXiv preprint arXiv:1706.01061* .

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5265–5274.

Wolf, L.; Hassner, T.; and Maoz, I. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, 529–534. IEEE.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23(10): 1499–1503.