Masquer Hunter: Adversarial Occlusion-aware Face Detection

Yujia Chen¹,², Lingxiao Song¹,²,³, Ran He¹,²,³*,
¹National Laboratory of Pattern Recognition, CASIA
²Center for Research on Intelligent Perception and Computing, CASIA
³University of Chinese Academy of Sciences, Beijing 100190, China

Abstract

Occluded face detection is a challenging detection task due to the large appearance variations incurred by various real-world occlusions. This paper introduces an Adversarial Occlusion-aware Face Detector (AOFD) by simultaneously detecting occluded faces and segmenting occluded areas. Specifically, we employ an adversarial training strategy to generate occlusion-like face features that are difficult for a face detector to recognize. Occlusion mask is predicted simultaneously while detecting occluded faces and the occluded area is utilized as an auxiliary instead of being regarded as a hindrance. Moreover, the supervisory signals from the segmentation branch will reversely affect the features, aiding in detecting heavily-occluded faces accordingly. Consequently, AOFD is able to find the faces with few exposed facial landmarks with very high confidences and keeps high detection accuracy even for masked faces. Extensive experiments demonstrate that AOFD not only significantly outperforms state-of-the-art methods on the MAFA occluded face detection dataset, but also achieves competitive detection accuracy on benchmark dataset for general face detection such as FDDB.

1. Introduction

Face detection has been well studied in recent years, since it is an essential step of many subsequent face-related applications, including face alignment, face verification and recognition. From the pioneering work of Viola-Jones face detector (Viola and Jones 2001) to recent state-of-the-art CNN-based methods, the performance of face detectors has been improved remarkably. For example, the average precision has been boosted to over 96% (Hu and Ramanan 2017; Najibi et al. 2017; Zhang et al. 2017; Wang et al. 2017) in the unconstrained FDDB dataset.

Although face detection algorithms have obtained quite good results, detecting faces in specific scenarios is still worth studying. For instance, one of the remaining challenges is partially occluded face detection. Facial occlusions occur frequently in the real world, e.g. facial accessories including sunglasses, masks and scarfs. The detection of occluded faces is indispensable in some applications, such as video surveillance and criminal investigation. Occluded faces are only partially visible, and occluded regions have arbitrary appearances that may diverse from normal face regions. Hence occluded faces have significant intra-class variation, leading to difficulties in learning discriminative features for detection. A standard paradigm to solve this problem is to enlarge the training dataset of occluded faces, but it can’t solve this problem in essence. Moreover, the lack of large-scale occluded face dataset makes it harder to handle this obstacle.

In this paper, we propose a framework for occluded face detection, aiming at formulating a new strategy to tackle the problem of limited occluded face training data, and exploiting the power of CNN representations for the faces with occlusions as far as possible.

Firstly, motivated by the remarkable success achieved by adversarial learning in recent years, a deep adversarial network is proposed in our approach to generate face samples with occlusions. A compact constraint is adopted to reinforce the realness of generated masks. Secondly, we introduce an occlusion-aware model by predicting the occlusion segments at the same time with detecting faces. Since occlusions not only impact the face detection process, but also severely deteriorate the performance of subsequent face alignment and recognition processes. It will be very meaningful to determine whether occlusion exists in a detected face for subsequent face-related applications. The predicted...
occlusion segment can be an important supervisory signal in occluded face alignment as well as occluded face recognition. For detection task, a segmentation branch can be of great help in locating heavily-occluded face area. Intuitively, jointly solving these two tasks can be reciprocal.

To sum up, we make contributions in the following aspects:

- A novel adversarial framework is proposed to alleviate the lack of occluded training face images by generating occluded or masked face features that are hard for detectors to handle. We employ a compact constraint to get more realistic occlusions.
- Mask prediction is conducted simultaneously while detecting occluded faces, which is pretty meaningful for not only face detection but also other subsequent face-related tasks. Comprehensive analysis shows that mask prediction is helpful for developing robust occluded face detectors.
- The proposed AMOD achieves competitive performance on the unconstrained face detection benchmark. Besides, experimental evaluations on the MAFA dataset also demonstrate the superiority of our method for partially occluded face detection.

2. Related Work

We first briefly survey face detection algorithms, followed by a review of the state-of-the-art unconstrained face detection researches.

Face detection can be considered as a special task of object detection. Successful general face detection algorithms often show great performance on face recognition. The Viola-Jones (Viola and Jones 2001) detector can be recognized as a milestone in the field of face detection. They innovatively adopted AdaBoost to train cascade classifier with Haar-like features, which first makes it possible to apply face detection in real-time applications. Following their work, lots of boosting-based models were proposed (Li and Zhang 2013; Dollár et al. 2014; Mathias et al. 2014; Zhu and Peng 2016), focusing on designing more sophisticated hand-crafted features or improving the boosting strategy. More Recently, CNN features (Yang et al. 2015a) were utilized in this boosting framework. Another famous category of face detectors is DPM-based. Deformable part models (Felzenszwalb et al. 2010) were proposed for object detection at first, which acquired impressive accuracy in complex environment. Inspired by this model, many extensions of DPM were developed to face detection (Ghiasi and Fowlkes 2014) by modeling potential deformations among facial parts. However, DPM models suffered from the high computational complexity, making it difficult to be applied in real-world applications such as digital cameras, phones or other mobile devices.

Generally speaking, boosting-based methods and DPM-based methods design features and optimize classifiers separately. The pipeline of these methods is divided into two stages, which is not an end-to-end architecture. Recently, benefitting from the prosperity of social network and big data, numerous deep learning based object detection algorithms have been proposed (Girshick 2015; Ren et al. 2015). CNN-based detectors therefore have become the mainstream in face detection gradually (Farfade, Saberian, and Li 2015; Li et al. 2015). CNN-based face detectors directly learn robust face representations from data and optimize classifiers in an end-to-end style. For example, (Zhang et al. 2016) developed a deep cascaded multi-task framework that predict face and landmark location in a coarse-to-fine manner, and (Yu et al. 2016a) further improved the performance of cascade models by optimizing feature selection algorithms.

Although many efforts have been made in face detection, the performance of occluded face detection is still far from satisfactory, and there are few works on occluded face detection as far as we know. (Yang et al. 2015b) explicitly inferred facesness score through local part responses via an attribute-aware model. But additional face-specific attribute annotations needed in this method were very difficult to collect. (Opitz et al. 2016) introduced a specific grid loss layer into CNNs that minimized the error rates on each sub-block of the feature map independently, thus every sub-part is discriminative on its own. (Mahbub et al. 2016) introduced a partial face detection approach based on detection of facial segments. They mainly focused on detecting incomplete faces that captured by the front camera of smart phones. Recently, (Ge et al. 2017) combined pre-trained CNN features with local linear embedding (LLE-CNN) to get similarity-based descriptors for partially visible faces. They built a dataset for masked face detection specifically, named the MAFA that contains 35K occluded faces.

As mentioned above, our work is also related to adversarial learning. Generative Adversarial Network (GAN) (Goodfellow et al. 2014) has shown great performance in numerous computer vision applications including image style transfer (Zhu et al. 2017; Isola et al. 2017), image generation (Shrivastava et al. 2017; Huang et al. 2017) and so on. Adversarial learning provides a simple yet efficient way to train powerful models via the min-max two-player game between the generator and the discriminator. Most of the previous work focused on promoting generators. Recently, researchers began to pay attention to increase the capacity of discriminator by adversarial learning. (Wang, Shrivastava, and Gupta 2017) used adversarial learning in generating hard examples for object detection. (Li et al. 2017) employed Perceptual GAN to enhance the representations for small objects. Inspired by these applications, we develop an adversarial occlusion-aware model, which can synthesize occlusion-like face features for boosting occluded face detectors.

3. Methods

The proposed model AOFD provides an effective and iconic method to tackle one of the most common and vicious problems in face detection-occlusion. This section first analyzes the problem (Sec 3.1) and summarizes the overall architecture (Sec. 3.2), and then introduces our mask generation method (Sec. 3.3) and segmentation method (Sec. 3.4) separately.
3.1 Problem Analysis

In real situations, we can classify face occlusion problem into three categories: facial landmark occlusion, occluded by faces and occluded by objects. As demonstrated in Figure 3, facial landmark occlusion includes conditions like wearing glasses and gauze masks. Occluded by faces is a complicated situation since detector easily mis-recognize several faces into one or only detect part of the face. The segmentation method is proposed in order to mitigate this problem. Occluded by object means occluded by things other than faces. Under this circumstance, around half of a face will be directly masked by things like clothes, scarves or books. An original masking strategy is used to mimic these in-the-wild situations.

We also delve into features of occluded faces, finding that occluded area rarely responds. For some heavily occluded faces, useful information in feature maps is too scarce for the detector to identify. To tackle this problem, we may need to enhance the expression ability of exposed area. On the other hand, recognition of occluded area can also bespeak that “there is a face” on condition that sufficient context information is provided.

3.2 Overall Architecture

In order to detect faces with heavy occlusion, a robust detector needs not only to find the distinctive part like eyes, nose and mouth, but also to transfer the interference of the occlusion into beneficial information. For the former problem, we find that undetected faces are typically those with their characteristic part of face occluded, such as eyes and mouth. In some cases, for example, sun glasses and gauze mask cover even all the face area, making it harder to detect. One feasible way is to mask the distinctive part of face in training set, forcing the detector to learn what possibly a face looks like even if there is less exposed area. To this end, a mask generator is designed in an adversarial way to generate a mask for each positive sample. It will generate different masks with faces of different poses. A masking strategy is applied for a better utilization of mask generator as well. More details are illustrated in Sec. 3.3.

For the latter problem, we introduce a segmentation branch to segment the occluded areas including hair, glasses, scarves, hands and other objects. This is not an easy task since no training set is available. Therefore, we labeled 374 training samples downloaded from the internet and came up with an original training strategy, more details are listed in Sec. 4.1. The name SFS (small dataset for segmentation) will be used in this paper to denote our training images for segmentation. One thing to be noted is that we are not aiming to get a segmentation result with extremely high accuracy, the meaning of this branch is to let the detector know what possibly an occlusion in front of a face looks like, and thus facilitates heavily occluded face detection.

As is demonstrated in Figure 2, a mask generator is added after a region of interest (RoI) pooling layer, followed by a classification branch and a bounding box regression branch. Finally, a segmentation branch is in responsible to segment the occluded area inside each bounding box. The final re-
The third type is randomly dropping half of the pixels. This half of the features, whether left, right, top or bottom, and to facial landmark occlusion. The second type is to mask type is using mask generator, which specifically corresponds posed and jointly training with the original features. The first extremely hard source for training, making the model diffic-

Loss function: The overall loss is a multitask loss:
\[ L = \alpha L_c + \beta L_b + \mu L_s \]  
where \( L_c \) denotes a binary softmax loss for classification, \( L_b \) denotes a smooth L1 loss for bounding box regression. We apply a binary softmax loss for segmentation branch, which is \( L_s \). During training, the coefficients \( \alpha, \beta \) and \( \mu \) are set 1, 1, and \( 1e^{-5} \) respectively.

3.3 Mask Generator

Mask generator: Since human face is very structural, facial features tend to appear in similar locations. However, with different poses, expressions and occlusions, distinguished facial area varies. Our aim is to find this distinguished area and generate a customized mask. We visualize some examples in Figure 4. As we have observed, occluded area in features rarely respond in real images. To simulate this characteristic, masks are directly operated on RoIs. Therefore, the generator, which contains four convolutional layers with a straight mapping, is designed very simple as it can be regarded as a binary prediction problem. Besides, the peculiar-

Masking strategy: The generated mask is a one-channel heat map where 0 represents masked area and 1 otherwise. During training, each pixel value will be squeezed to zero or one. We select a quarter of the minimum value as the mask when training the generator and one-third when training the overall model.

Heavily occluded samples after masking will become an extremely hard source for training, making the model difficult to converge. To this end, three types of masks are pro-

Loss function: When training the mask generator, an adversarial training method is applied. We aim to increase classification loss as much as possible. Since masked area is limited and distinguished facial area is comparatively salient in feature maps, the model can easily converge. However, we find it not enough because the occluded area is sometimes strip-like or sporadic, while it is supposed to be more compact in real situations. Recall that areas with longer or irregular edges will have larger value for each pixel using a kernel of edge detector. A kernel to make the occluded area sleeker and more circular is designed as a compact constraint for generated masks. The loss function is:
\[ L_g = \gamma L_{com} - \eta L_c \]  
where \( L_g \) denotes the loss for generator, \( L_{com} \) denotes a compact loss, and \( \gamma \) and \( \eta \) are coefficients. \( \gamma \) is set \( 1e^{-6} \) and \( \eta \) is set 1 in order to balance the derivatives. The compact loss is computed with a convolutional layer in a way as follows:
\[ L_{com} = \sum ((1 - mask) * kernel) \]  
where * denotes a convolutional operation, mask is the first type of mask generated by the mask generator and the last item is the designed kernel, which is
\[
\begin{bmatrix}
  -1 & -1 & -1 \\
  -1 & 8 & -1 \\
  -1 & -1 & -1 
\end{bmatrix}
\]

In this way, strip-like or sporadic areas will get very high penalty and more reasonable masks can be obtained.

3.4 Segmentation

Design: Lots of excellent works on segmentation have proved that CNNs is capable of comprehending the semantic information of a picture and elaborately conduct a pixel-wise classification. When combining detection with segmentation, it is usually designed in RoI level to achieve higher accuracy. Considering segmenting each RoI, one problem in occluded face situation is that the overlap of two bounding boxes will have different meanings. For example, if one face is occluded by another face, part of the front face should be regarded as an occlusion for the face in the back, while there shouldn’t be any occluded area for the front face. Since our
destination is to utilize the effective information contained in the occluded area to first confirm if there is a face behind and then make the exposed area more distinguished, ample context information is required. With reasons above, segmentation is conducted in image level to affect features. Therefore, the detector is able to find faces with more informative features embodying image-level signals like the appearance of a person. We call this method an occlusion-aware method.

The segmentation branch is designed in a fully convolutional way like FCN. In order to obviate noise, it follows a bounding box regression branch and only areas inside bounding boxes are maintained (Figure 5). As a matter of fact, mapping bounding boxes into features impairs useful information ascribe to computation accuracy, so bounding boxes are enlarged in scale before dropping the noise.

Although the final results have proved the feasibility of this method, the edges of segmentation seem to be a bit rough. This is caused by the limited size of the training set. Nevertheless, we have verified the possibility to deal with very limited training samples and achieved our original goal.

**Loss function:** We choose softmax loss instead of L1 or L2 loss used in some segmentation and image generation tasks because it helps stabilize the training process as we have observed.

**4. Experiments**

In this section, we first introduce detailed information during training (Sec. 4.1), and then conduct a series of ablative studies to verify the effectiveness of our methods (Sec. 4.2). We also test AOFD on several comparative benchmarks to verify its superior performance (Sec. 4.3). As a matter of fact, Faster RCNN based methods generally have poor performance on small object detection. Simply applying more anchor scales and training image scales will doubtlessly increase the accuracy but also more computation and it is less insightful, so we are not doing this. However, the ability to detect large faces under all the conditions can also be considered a success. So the competitive results on FDDB (Jain and Learned-Miller 2010) and MAFA (Ge et al. 2017) are capable enough to bespeak the superiority of AOFD. Figure 6 shows some qualitative results of our model.

**4.1 Training Details**

Based on Faster RCNN (Ren et al. 2015), we first train a mask generator with settings mentioned above. The detector and the segmentation branch are trained jointly with the mask generator fixed in the second stage. Due to the limitation of training data for segmentation, an unordinary training strategy is needed. We first train on SFS for 10000 iterations with original settings, then on the combination of WIDER FACE and SFS for 50000 iterations with loss weights for segmentation set $1e^{-7}$ and finally tune the model on SFS for 3 epochs with original settings. Derivatives from WIDER FACE training set will be set zero for segmentation branch. The basic learning rate is 0.001. AOFD runs 5 FPS on a TITAN X GPU, which is similar to original Faster RCNN.

**Experiment settings:** AOFD is based on a Faster RCNN with a VGG16 backbone (Simonyan and Zisserman 2014). For anchors of PRN, we use three aspect ratios (1.7, 1 and 1.3) and four scales ($64^2$, $128^2$, $256^2$ and $512^2$). Each convolutional layer contains a convolution operation followed by a rectified linear unit (ReLU). Batch size is set 1. An RoI is treated as foreground if its intersection over union (IoU) with any ground truth bounding box is higher than 0.5. To balance the number of foreground and background training samples, the ratio of foreground RoIs to background RoIs is set 1:3. During training, short side of input image is resized to either 512 or 1024 on condition that long side is no longer than 1024.
4.2 Model Analysis

To better understand the function of each part of our model, we ablate each component to observe AOFD’s performance. In this way, the mask generator and segmentation branch are removed one after the other. We delve into the optimal area of the mask as well and find that the mask area is crucial for the functioning of mask generator. Results on FDDB and MAFA test set are reported in Table 1 and Figure 8.

**Mask facilitates detecting:** State-of-the-art detectors are able to detect some of occluded faces, but with lower confidence. As shown in Figure 7, AOFD can increase the confidence of occluded faces by a large margin. Without mask generator, the detector pays less attention to exposed area or face structure, and the recall rate at 1000 false positives on FDDB drops by 0.57% (Table 1). The sharp decline (1.4%) of average precision on MAFA test set in Figure 8a reveals the value of mask generator as well.

**Segmentation increases recall:** With segmentation branch, the result witnesses an increase of 0.44% (Table 1), confirming not only the effectiveness of this multitask method, but also our training strategy with limited data. We can also observe that average precision achieves 76.3% without segmentation, which is comparable with state-of-the-art detectors. We should credit the mask generator for this commendable result.

**Mask area is crucial:** When training mask generator, we find that the mask would vitiate the detector if mask area is too large. Nevertheless, it would be of no use if it is too small. Table 1 and Figure 8 gives a brief overview of our experiments, from which we find occluding one-third of features is an ideal area for mask. Noting that experiments with different settings all obtain state-of-the-art results. It is because only one of the three types of masks is influenced in the training strategy, and the other two remain effective.

4.3 Evaluation on benchmarks

Our model is trained on the WIDER FACE (Yang et al. 2016) training set that contains 12880 images and 159424 face annotations categorized to 61 events, and evaluated on the FDDB and MAFA databases. WIDER FACE is not a desired source for testing since it contains half of tiny faces (the height range from 10 to 50 pixels) and our aim is not detecting tiny faces.

**FDDB (Face Detection Data Set and Benchmark)** is an unconstrained dataset for face detection. It has 2, 845 images with 5, 171 faces that are collected from the news articles on Yahoo websites. The detection results of different methods are shown in Figure 9. Noted that FDDB uses ellipse bounding boxes while ours are rectangle, so continuous score is negatively influenced but still fine. (Liu et al. 2017) and
Recall rate at 1000 FPs

| Settings            | Recall rate at 1000 FPs |
|---------------------|-------------------------|
| AOFD                | 97.12%                  |
| w/o segmentation    | 96.68%                  |
| w/o generator       | 96.55%                  |
| Mask Area ($\frac{1}{6}$) | 96.54% |
| Mask Area ($\frac{1}{4}$) | 97.12% |
| Mask Area ($\frac{1}{2}$) | 95.33% |

our adversarial training method and the occlusion-aware design. In some case studies, we find that the main type of false positives incurred by traditional methods are incomplete or unbalanced bounding boxes as shown in Figure 2. Our AOVD alleviates this problem to a great extend since it is aware of the occluded areas and can estimate preferable detections.

Furthermore, we have studied the main obstacle of our model to achieve a higher AP. In Figure 2, the minimum IoU threshold for a true positive proposal is modified, from which we observe that a slight decrease of IoU threshold can boost AP to a great extent. This explains that the precision of bounding boxes still needs to be improved even though it has already exceeded traditional methods.

**Conclusion**

This paper proposed a model named AOVD to solve the long-standing problem of face occlusion which has been rarely discussed in recent studies. Sufficiently utilizing CNN’s capacity, the original masking strategy increases training complexity, and can plastically mimic different situations of face occlusion including facial landmark occlusion, occluded by faces and by objects. The proposed multitask training method with a segmentation branch also proves to be a feasible solution and verifies the possibility to train an auxiliary task with very limited data. The superior performance on both general face detection and masked face detection benchmarks demonstrates the effectiveness of AOVD. Future works include transferring this method to one-stage detection method and combining with state-of-the-art tiny face detectors.

**Table 1: Results of the ablative studies on FDDB**

| Settings            | Recall rate at 1000 FPs |
|---------------------|-------------------------|
| AOFD                | 97.12%                  |
| w/o segmentation    | 96.68%                  |
| w/o generator       | 96.55%                  |
| Mask Area ($\frac{1}{6}$) | 96.54% |
| Mask Area ($\frac{1}{4}$) | 97.12% |
| Mask Area ($\frac{1}{2}$) | 95.33% |

**Table 2: Average precision (%) on the test set of MAFA**

| Methods                  | Average Precision |
|--------------------------|-------------------|
| AOFD                     | 77.3%             |
| LLE-CNNs (Ge et al. 2017) | 76.4%             |
| Faster RCNN (Ren et al. 2015) | 74.89%       |
| MT (Zhang et al. 2016)    | 60.8%             |
| HPM (Ghiasi and Fowlkes 2014) | 60.0%    |
| HH (Mathias et al. 2017)  | 50.9%             |
| ZR (Zhu and Ramanan 2012) | 41.6%             |
tion with discriminatively trained part-based models. *IEEE TPAMI* 32(9):1627–1645.

[Ge et al. 2017] Ge, S.; Li, J.; Ye, Q.; and Luo, Z. 2017. Detecting masked faces in the wild with lle-cnns. In *IEEE CVPR*.

[Ghiasi and Fowlkes 2014] Ghiasi, G., and Fowlkes, C. C. 2014. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE CVPR*, 2385–2392.

[Girshick 2015] Girshick, R. 2015. Fast r-cnn. In *IEEE ICCV*, 1440–1448.

[Goodfellow et al. 2014] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

[Huang et al. 2015] Huang, R.; Zhang, S.; Li, T.; and He, R. 2017. Beyond face rotation: Global and local perception gan for photo realistic and identity preserving frontal view synthesis. In *IEEE ICCV*.

[Ge et al. 2017] Ge, S.; Li, J.; Ye, Q.; and Luo, Z. 2017. Detecting masked faces in the wild with lle-cnns. In *IEEE CVPR*.

[Ozair et al. 2017] Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

[Ge et al. 2017] Ge, S.; Li, J.; Ye, Q.; and Luo, Z. 2017. Detecting masked faces in the wild with lle-cnns. In *IEEE CVPR*.
[Zhang et al. 2017] Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; and Li, S. Z. 2017. S3fd: Single shot scale-invariant face detector. In IEEE ICCV.

[Zhu and Peng 2016] Zhu, C., and Peng, Y. 2016. Group cost-sensitive boosting for multi-resolution pedestrian detection. In AAAI, 3676–3682.

[Zhu and Ramanan 2012] Zhu, X., and Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. In IEEE CVPR, 2879–2886.

[Zhu et al. 2017] Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In IEEE ICCV.