Finding Tiny Faces in the Wild with Generative Adversarial Network

Yancheng Bai[1,3] Yongqiang Zhang[1,2] Mingli Ding[2] Bernard Ghanem[1]
[1] Visual Computing Center, King Abdullah University of Science and Technology (KAUST)
[2] School of Electrical Engineering and Automation, Harbin Institute of Technology (HIT)
[3] Institute of Software, Chinese Academy of Sciences (CAS)
baiyancheng20@gmail.com {zhangyongqiang, dingml}@hit.edu.cn bernard.ghanem@kaust.edu.sa

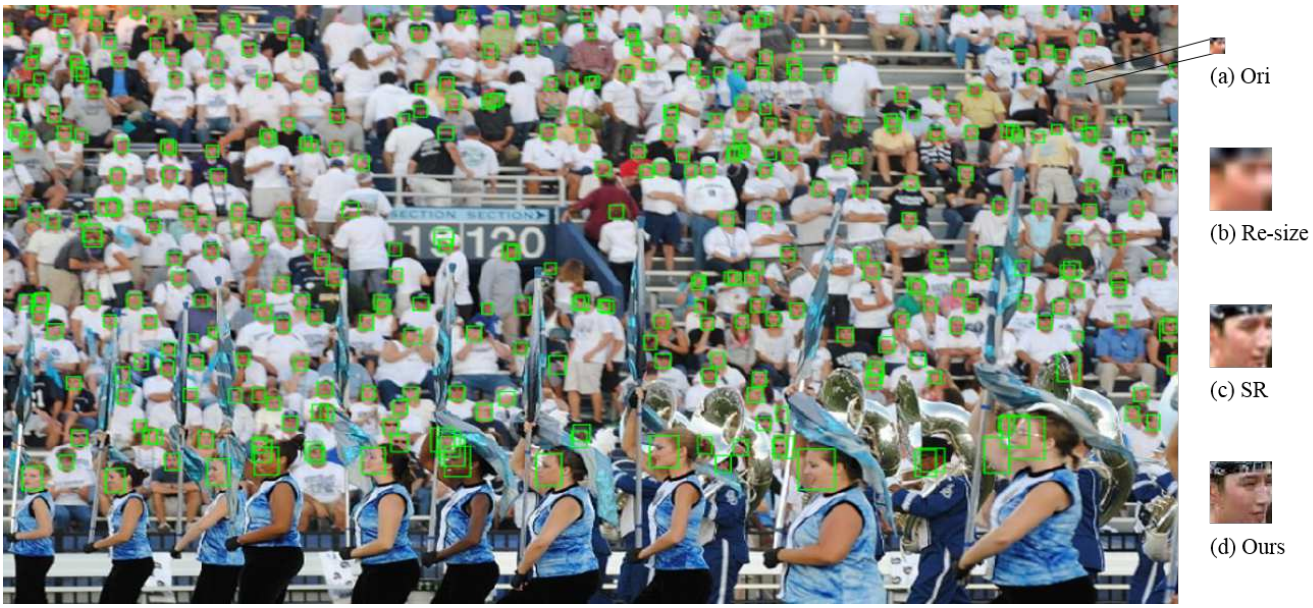

Figure1. The detection results of tiny faces in the wild. (a) is the original low-resolution blurry face, (b) is the result of re-sizing directly by a bi-linear kernel, (c) is the generated image by the super-resolution method, and our result (d) is learned by the super-resolution (×4 upscaling) and refinement network simultaneously. Best viewed in color and zoomed in.

Abstract

Face detection techniques have been developed for decades, and one of remaining open challenges is detecting small faces in unconstrained conditions. The reason is that tiny faces are often lacking detailed information and blurring. In this paper, we proposed an algorithm to directly generate a clear high-resolution face from a blurry small one by adopting a generative adversarial network (GAN). Toward this end, the basic GAN formulation achieves it by super-resolving and refining sequentially (e.g. *SR-GAN and cycle-GAN). However, we design a novel network to address the problem of super-resolving and refining jointly. We also introduce new training losses to guide the generator network to recover fine details and to promote the discriminator network to distinguish real* vs. *fake and face* vs. *non-face simultaneously. Extensive experiments on the challenging dataset WIDER FACE demonstrate the effectiveness of our proposed method in restoring a clear high-resolution face from a blurry small one, and show that the detection performance outperforms other state-of-the-art methods.*

1. Introduction

Face detection is a fundamental and important problem in computer vision, since it is usually a key step towards many subsequent face-related applications, including face parsing, face verification, face tagging and retrieval, etc. Face detection has been widely studied over the past few decades and numerous accurate and efficient methods have been proposed for most constrained scenarios. Recen-

t works focus on faces in uncontrolled settings, which is much more challenging due to the significant variations in scale, blur, pose, expressions and illumination. A thorough survey on face detection methods can be found in [32].

Modern face detectors have achieved impressive results on the large and medium faces, however, the performance on small faces is far from satisfactory. The main difficulty for small face (*e.g*. 10×10 pixels) detection is that small faces lack sufficient detailed information to distinguish them from the similar background, *e.g*. regions of partial faces or hands. Another problem is that modern CNN-based face detectors use the down-sampled convolutional (*conv*) feature maps with stride 8, 16 or 32 to represent faces, which lose most spatial information and are too coarse to describe small faces. To detect small faces, [28] directly up-samples images using bi-linear operation and exhaustively searches faces on the up-sampled images. However, this method will increase the computation cost and the inference time will increase significantly too. Moreover, images are often zoomed in with a small upscaling factors ($2\times$ at most) in [28], otherwise, artifacts will be generated. Besides, [1, 14, 25, 37] use the intermediate *conv* feature maps to represent faces at specific scales, which keeps the balance between the computation burden and the performance. However, the shallow but fine-grained intermediate *conv* feature maps lack discrimination, which causes many false positive results. More importantly, these methods take no care of other challenges, like blur and illumination.

To deal with the nuisances in face detection, we propose a unified end-to-end convolutional neural network for better face detection based on the classical generative adversarial network (GAN) framework. There are two sub-networks in our detector, a generator network and a discriminator network. In the generator sub-network, a super-resolution network (SRN) is used to up-sample small faces to a fine scale for finding those tiny faces. Compared to re-sizing by bi-linear operation, SRN can reduce the artifact and improve the quality of up-sampled images with a large upscaling factors ($4\times$ in our current implementation), as shown in Figure 1 (b) and (c). However, even with such sophisticated SRN, up-sampled images are unsatisfactory (usually blurring and lacking fine details) due to faces of very low resolutions (10×10 pixels). Therefore, a refinement network (RN) is proposed to recover some missing details in the up-sampled images and generate sharp high-resolution images for classification. In the discriminator sub-network, we introduce a new loss function that enforces the discriminator network to distinguish the real/fake face and face/non-face simultaneously. The generated images and real images pass through the discriminator network to JOINTLY distinguish whether they are real images or generated high-resolution images and whether they are faces or non-faces. More importantly, the classification loss is used to guide the generator to generate clearer faces for easier classification.

Contributions. To sum up, this paper makes following three main contributions. **(1)** A novel unified end-to-end convolutional neural network architecture for face detection is proposed, where super-resolution and refinement network are used to generate real and sharp high-resolution images and a discriminator network is introduced to classify faces *vs*. non-faces. **(2)** A new loss is introduced to promote the discriminator network to distinguish the real/fake image and face/non-face simultaneously. More importantly, the classification loss is used to guide the generative network to generate clearer faces for easier classification. **(3)** Finally, we demonstrate the effectiveness of our proposed method in restoring a clear high-resolution face from a blurry small face, and show that the detection performance outperforms other state-of-the-art approaches on the WIDER FACE dataset, especially on the most challenging Hard subset.

2. Related Work

2.1. Face Detection

As a classic topic, numerous face detection systems have been proposed during the past decades or so. Existing face detection methods can be broadly categorized as handcrafted feature based methods [24, 29, 30] and CNN-based methods [34, 2, 14, 25, 37, 1]. However, most of the detection systems based handcrafted features only train a single scale model, which is applied to each level of a feature pyramid, thus increasing the computation cost drastically, especially for complicated features. Moreover, the limited representation of handcrafted features restricts the performance of detectors, particularly in uncontrolled settings.

Inspired by the great success of Faster RCNN, several recent works [14, 25, 37] utilized this framework to detect faces and showed impressive performance on the FDDB benchmark [13]. However, performance drops dramatically on the more challenging WIDER FACE dataset [31], which contains a large number of faces with lower resolution. The main reason for this disparity is that deep conv feature maps with lower spatial resolution are used for representation, which is insufficient for handling small faces [34, 2]. To overcome this problem, detectors [14, 25, 37] have to up-sample by re-sizing input images to different scales during training and testing, which inevitably increases memory and computation costs. Furthermore, the re-size method often generates the images with large structural distortions, as shown in Figure1 (b). Compared to these methods, our method exploits the super-resolution and refinement network to generate clear and fine faces with high resolution ($4\times$ up-scaling), as shown in Figure1 (d), and then the discriminator is trained to distinguish faces from input images.

2.2. Super-resolution and Refinement Network

With the development of deep learning, great improvements have been achieved on super-resolution [5, 6, 15, 26]. However, when obtaining these promising results, there is a precondition that the down-sampling kernel is known, and most of these CNN-based super-resolution methods can not be applied to uncontrolled settings (*i.e*. in the wild).

There are different refinement networks for different tasks, and the most similar refinement method to our refinement network is the deblur method. Most existing deblur methods heavily rely on prior models to solve the ill-posed problem, and a prior assumes that gradients of natural images have a heavy-tailed distribution [21, 7]. Recently, conventional neural networks have also been used to deblur the blind image [36, 23, 3]. However, these deblurring methods still involve explicit kernel estimation, and the recovered images usually have significant ringing artifacts if the estimated kernels are inaccurate.

Although existing super-resolution methods and refinement methods are effective at up-sampling and refining images respectively, it is not easy to extend to jointly super-resolving and refining the low-resolution image. [28] proposed a method to simultaneously reconstruct a clear high-resolution image from a blurry low-resolution input. However, their blurry low-resolution images are obtained by using the bicubic interpolation down-sampling and a known blur kernel from the high-resolution images (*i.e*. synthetic). In this paper, we design a novel network to generate a clear super-resolution face from a small blurry face which is collected from the wild. We would like to note that our work is the first work trying to jointly super-resolve and refine the small blurry faces in the wild.

2.3. Generative Adversarial Networks

In the seminal work[8], generative adversarial network (GAN) is introduced to generate realistic-looking images from random noises. GANs have achieved impressive results in image generation [4], image editing [38], representation learning [18], image annotation [27], image super-resolving [17] and character transferring [12]. Recently, GAN has been applied to super-resolution (SRGAN) [17] and has obtained promising results. Compared to super-resolution on natural images, face images in the wild are of arbitrary poses, illumination and blur, so super-resolution on face images is much more challenging. More importantly, the high resolution images generated by SRGAN are blurry and lack fine details especially for low-resolution faces, which are unfriendly for the face classifier. Towards this end, we design a refinement sub-network to recover some detailed information. In the discriminator network, the basic GAN [17, 12, 8] is trained to distinguish the real and fake high resolution images. To classify faces or non-faces, we extend the discriminator network to classify the fake *vs*. real and face *vs*. non-face simultaneously. Furthermore, the classification loss is propagated back to the generator network, and guides generator network to reconstruct clearer super-resolution images for easier classification.

3. Proposed Method

In this section, we introduce our proposed method in details. First, we give a brief description on the classical GAN network. Then, the whole architecture of our method is presented, as shown in Figure 2. Finally, we introduce each part of our network in details and define the loss functions for training the generator network and discriminator network respectively.

3.1. GAN

GAN [8] learns a generative model G via an adversarial process. It trains a generator network G and a discriminator network D simultaneously. The training process alternately optimizes the generator and discriminator, which compete with each other. The generator G is trained for generating the samples to fool the discriminator D, and the discriminator D is trained to distinguish the real image and the fake image from the generator. The objective function can be defined as follows:

$$\begin{aligned}\mathcal{L}_{GAN}(G,D) = &\mathbb{E}_{x\sim p_{data}(x)}[\log D_\theta(x)]+ \\ &\mathbb{E}_{z\sim p_z(z)}[\log(1-D_\theta(G_\omega(z)))],\end{aligned} \tag{1}$$

where z is the random noise and x denotes the real data, θ and ω denote the parameters of G and D respectively. Here, G tries to minimize the objective function and adversarial D tries to maximize it as Eq(2):

$$\arg\min_G \max_D \mathcal{L}_{GAN}(G,D). \tag{2}$$

Similar to [8, 17], we further design a generator network G_{w_G} which is optimized in an alternative method along with a discriminator network D_{θ_D} to solve the small face super-resolution and classification problem, which is defined as follows:

$$\begin{aligned}\arg\min_{w_G}\max_{\theta_D} &\ \mathbb{E}_{(I^{HR},y)\sim p_{train}(I^{HR},y)}[\log D_{\theta_D}(I^{HR},y)]+ \\ &\mathbb{E}_{(I^{LR},y)\sim p_G(I^{LR},y)}[\log(1-D_{\theta_D}(G_{w_G}(I^{LR},y)))],\end{aligned} \tag{3}$$

where I^{LR} denotes face candidates with low-resolution, I^{HR} represents the face candidates with high-resolution, and y is the label (*i.e*. face or non-face). Unlike [8], the input of our generator is low-resolution images rather than the random noise. Different from [17], we up-sample and refine the input image simultaneously in the generator network. In the discriminator network, we distinguish the generated *vs*. true high-resolution images and faces *vs*. non-faces jointly.

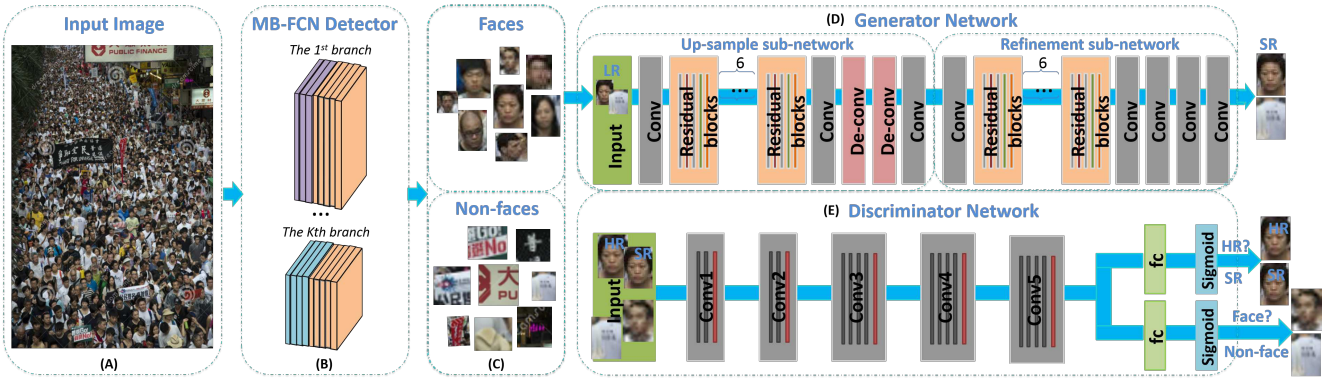

Figure 2. The pipeline of the proposed tiny face detector system. (A) The images are fed into the network; (B) MB-FCN detector is our baseline, it crops the positive data (*i.e*. faces) and negative data (*i.e*. non-faces) from the input images for training the generator network and the discriminator network, or generates the regions of interest (ROIs) for testing. (C) The positive data and negative data (or ROIs) are generated by the MB-FCN detector. (D) The generator network is trained to reconstruct a clear super-resolution image ($4\times$ upscaling) from the low-resolution input image, which includes the upsample sub-network and the refinement sub-network. (D) The discriminator network is the vgg19 architecture with two parallel fc layers, and the first fc layer is to distinguish the natural real images or the generated super-resolution images and the second one is to classify faces or non-faces.

3.2. Network Architecture

Our generator network includes two components (*i.e*. upsample sub-network and refinement sub-network), and the first sub-network takes the low-resolution images as the inputs and the outputs are the super-resolution images. Since the blurry small faces lack fine details and due to the influence of MSE loss Eq(4), the generated super-resolution faces are usually blurring. So we design the second sub-network to refine the super-resolution images from the first sub-network. Furthermore, we add the classification branch to the discriminator network for the purpose of detection, which means our discriminator can classify faces and non-faces as well as distinguish the fake and real images.

Generator network. As shown in Table 1 and Figure 2, we adopt a deep CNN architecture which has shown effectiveness for image super-resolution in [17]. There are two fractionally-strided convolutional layers [20] (*i.e*. de-convolutional layer) in the network, and each de-convolutional layer consists of learned kernels which perform up-sampling a low-resolution image to a $2\times$ super-resolution image. In contrast to their network, our generator network includes refinement sub-network which is also a deep CNN architecture. Similar to [20], we use the batch-normalization [11] and rectified linear unit (ReLU) activation after each convolutional layer except the last layer.

The up-sampling sub-network first up-samples a low-resolution image and outputs a $4\times$ super-resolution image, and this super-resolution image is blurring when the small faces are far from the cameras or under fast motion. Then, the refinement sub-network processes the blurring image, and outputs a clear super-resolution image, which is easier for the discriminator to classify the faces *vs*. non-faces.

Discriminator network. We employ VGG19 [22] as our backbone network in the discriminator, as shown in Table 1. To avoid too many down-sampling operations for the small blurry faces, we remove the max-pooling from the "conv5" layer. Moreover, we replace all the fully connected layer (*i.e*. $fc6$, $fc7$, $fc8$) with two parallel fully connected layers fc_{GAN} and fc_{clc}. The input is the super-resolution image, the output of fc_{GAN} branch is the probability of the input being a real image, and the output of the fc_{clc} is the probability of the input being a face.

3.3. Loss Function

We adopt the pixel-wise loss and adversarial loss from some state-of-the-art approaches [17, 12] to optimize our generator network. In contrast to [17], we remove the VGG feature matching loss due to the calculation cost and we introduce the classification loss to drive the generator network to recover fine details from the blurry small faces.

Pixel-wise loss. The input of our generator network is the small blurry images instead of random noise [8]. A natural way to enforce the output of the generator to be close to the super-resolution ground truth is through the pixel-wise MSE loss, and it is calculated as Eq(4):

$$L_{MSE}(w) = \frac{1}{N}\sum_{i=1}^{N}(||G_{1_{w1}}(I_i^{LR}) - I_i^{HR}||^2 + ||G_{2_{w_2}}(G_{1_{w1}}(I_i^{LR})) - I_i^{HR}||^2), \tag{4}$$

where I^{LR} and I^{HR} denote the small blurry images and super-resolution images respectively, G_1 means up-sampling sub-network, G_2 denotes the refinement sub-network and w is the parameters of generator network.

	Generator												Discriminator					
	Up-sample Sub-network						Refinement Sub-network											
Layer	conv	conv x8	conv	de-conv	de-conv	conv	conv	conv x8	conv	conv	conv	conv	conv	conv	conv	conv	conv	fc
Kernel Num.	64	64	64	256	256	3	64	64	64	256	256	3	64	128	256	512	512	2
Kernel Size	3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	-
Stride	1	1	1	2	3	1	1	1	1	1	1	1	2	2	2	2	1	-

Table 1. Architecture of the generator and discriminator network. "conv" represents a convolutional layer, "x8" denotes a residual block which has 8 convolutional layers, "de-conv" means a fractionally-stride convolutional layer, "2x" denotes up-sampling by a factor of 2, and "fc" indicates a fully connected layer.

However, while achieving less loss between the generated and the neutral high-resolution image in pixel level, the solution of the MSE optimization problem usually lacks the high-frequency content which results in perceptual unsatisfactory images with over-smooth texture. Also, this is one reason why the generated image is blurry.

Adversarial loss. To achieve more realistic results, we introduce the adversarial loss [17] to the objective loss, defined as Eq(5):

$$L_{adv} = \frac{1}{N}\sum_{i=1}^{N}\log(1 - D_\theta(G_w(I_i^{LR}))). \tag{5}$$

Here, the adversarial loss encourages the network to generate sharper high-frequency details for trying to fool the discriminator network. In Eq(5), the $D_\theta(G_w(I_i^{LR}))$ is the probability that the reconstruction image $G_w(I_i^{LR})$ is a natural super-resolution image.

Classification loss. In order to make the reconstructed images by the generator network easier to classify, we also introduce the classification loss to the objective loss. Let $\{I_i^{LR}, i = 1, 2, \ldots, N\}$ and $\{I_i^{HR}, i = 1, 2, \ldots, N\}$ denote the small blurry images and the high-resolution real natural images respectively, and $\{y_i, i = 1, 2, \ldots, N\}$ represents the corresponding labels, where $y_n = 1$ or $y_n = 0$ indicates the image is the face or non-face respectively. The formulation of classification loss is like Eq(6):

$$\begin{aligned} L_{clc} = \frac{1}{N}\sum_{i=1}^{N}(&\log(y_i - D_\theta(G_w(I_i^{LR})))+ \\ &\log(y_i - D_\theta(I_i^{HR}))). \end{aligned} \tag{6}$$

Our classification loss plays two roles, where the first is to distinguish whether the high-resolution images, including both the generated and the natural real high-resolution images, are faces or non-faces in the discriminator network. The other role is to promote the generator network to reconstruct sharper images.

Objective function. Based on above analysis, we incorporate the adversarial loss Eq(5) and classification loss Eq(6) into the pixel-wise MSE loss Eq(4). The GAN network can be trained by the objective function Eq(7):

$$\begin{aligned} &\max_\theta \min_w \frac{1}{N}\sum_{i=1}^{N}\alpha(\log(1 - D_\theta(G_w(I_i^{LR}))) + \log D_\theta(I_i^{HR})) \\ &+ (||G_{1_{w1}}(I_i^{LR}) - I_i^{HR}||^2 + ||G_{2_{w_2}}(G_{1_{w1}}(I_i^{LR})) - I_i^{HR}||^2) \\ &+ \beta(\log(y_i - D_\theta(G_w(I_i^{LR}))) + \log(y_i - D_\theta(I_i^{HR}))), \end{aligned} \tag{7}$$

where α and β are trade-off weights.

For better gradient behavior, we optimize the objective function in an alternative way as in [17, 12, 10] and modify the loss function of generator G and the discriminator D as:

$$\begin{aligned} &\min_w \frac{1}{N}\sum_{i=1}^{N}\alpha\log(1 - D_\theta(G_w(I_i^{LR})))+ \\ &(||G_{1_{w1}}(I_i^{LR}) - I_i^{HR}||^2 + ||G_{2_{w_2}}(G_{1_{w1}}(I_i^{LR})) - I_i^{HR}||^2)+ \\ &\beta\log(y_i - D_\theta(G_w(I_i^{LR}))), \end{aligned} \tag{8}$$

and

$$\begin{aligned} &\min_\theta \frac{1}{N}\sum_{i=1}^{N} -((\log(1 - D_\theta(G_w(I_i^{LR}))) + \log D_\theta(I_i^{HR}))+ \\ &(\log(y_i - D_\theta(G_w(I_i^{LR}))) + \log(y_i - D_\theta(I_i^{HR})))). \end{aligned} \tag{9}$$

The loss function of generator G in Eq(8) consists of adversarial loss Eq(5), MSE loss Eq(4) and classification loss Eq(6), which enforce the reconstructed images to be similar to the real natural high-resolution image on the high-frequency details, pixel, and semantic level respectively. The loss function of discriminator D in Eq(9) introduces the classification loss to classify whether the high-resolution images are faces or non-faces, which is parallel to the basic formulation of GAN [8] to distinguish whether the high-resolution images are fake or real. By introducing the classification loss, the recovered images from generator are more realistic than the results optimized by the adversarial loss and MSE loss. Further ablation analysis on the influence of each loss function is presented in Section 4.3.

4. Experiments

In this section, we experimentally validate our proposed method on two public face detection benchmarks (*i.e.*

WIDER FACE [31] and FDDB [13]). First, we conduct an ablation experiment to prove the effectiveness of GAN. Then, we give a detailed analysis on the importance of each loss in the generator and discriminator network. Finally, our proposed face detector is evaluated on both of these public benchmarks, while comparing the performance against other state-of-the-art approaches.

4.1. Training and Validation Datasets

We use a recently released large-scale face detection benchmark, the WIDER FACE dataset [31]. It contains 32,203 images, which are selected from the publicly available WIDER dataset. 40%/10%/50% of the data is randomly selected for training, validation, and testing, respectively. Images in WIDER FACE are categorized into 61 social event classes, which have much more diversities and are closer to the real-world scenario. Therefore, we use this dataset for training and validating the proposed generator and discriminator networks.

The WIDER FACE dataset is divided into three subsets, Easy, Medium, and Hard, based on the heights of the ground truth faces. The Easy/Medium/Hard subsets contain faces with heights larger than 50/30/10 pixels respectively. Compared to the Medium subset, the Hard one contains many faces with a height between $10-30$ pixels. As expected, it is quite challenging to achieve good detection performance on the Hard subset.

4.2. Implementation Details

In the generator network, we set the trade-off weights $\alpha = 0.001$ and $\beta = 0.01$. During training, we use the Adam optimizer [16] with momentum term $\beta_1 = 0.9$. The generator network is trained from scratch and the weights in each layer are initialized with a zero-mean Gaussian distribution with standard deviation 0.02, and biases are initialized with 0. To avoid undesirable local optima, we first train an MSE-based SR network to initialize the generator network. For the discriminator network, we employ the VGG19 [22] model pre-trained on ImageNet as our backbone network and we replace all the fc layers with two parallel fc layers. The fc layers are initialized by a zero-mean Gaussian distribution with standard deviation 0.1, and all biases are initialized with 0.

Our baseline MB-FCN detector is based on ResNet50 network [9], which is pre-trained on ImageNet. All hyper-parameters of the MB-FCN detector are the same as [1]. For training our generator and discriminator network, we crop face samples and non-face samples from WIDER FACE [31] training set with our baseline detector. The corresponding low-resolution images are generated by down-sampling the high-resolution images using the bicubic interpolation with a factor 4. During testing, 600 regions of interest (ROIs) are cropped and these ROIs are fed to our GAN network to give the final detection performance.

Method	Easy	Medium	Hard
Baseline[1]	0.932	0.922	0.858
w/o Refinement Network	0.940	0.929	0.863
w/o adv loss	0.935	0.925	0.867
w/o clc loss	0.936	0.927	0.865
Ours(Baseline+MES+adv+clc)	**0.944**	**0.933**	**0.873**

Table 2. Performance of the baseline model trained with and without GAN, refinement network, adversarial loss and classification loss on the WIDER FACE invalidation set. "adv" denotes adversarial loss Eq(5), "clc" represents classification loss Eq(6) and "MES" means pixel-wise loss Eq(4).

All the GAN variants are trained with first 3 epochs at a learning rate of 10^{-4} and another 3 epochs at a lower learning rate of 10^{-5}. We alternately update the generator and discriminator network, which is equivalent to $k = 1$ as in [8]. Our implementation is based on tensorflow, and all the experiments are done on an NVIDIA TITAN X GPU.

4.3. Ablation Studies

We first compare our proposed method with the baseline detector to prove the effectiveness of GAN. Moreover, we perform the ablation study by removing the refinement network to validate the effectiveness of refinement network. Finally, to validate the contribution of each loss, including adversarial loss and classification loss in the loss function of generator network, we also conduct ablation studies by cumulatively adding each of them to the pixel-wise loss.

Influence of the GAN. Table 2 (the 1^{st} and the 5^{th} row) shows the detection performance (AP) of the baseline detector and our method on WIDER FACE validation set. Our baseline detector is a multi-branch RPN face detector with skip connection of feature maps, and please refer to [1]for more detailed information . From Table 2 we observe that the performance of our detector outperforms the baseline detector by a large margin (1.5% in AP) on the Hard subset. The reason is that the baseline detector performs the down-sampling operations (*i.e.* convolution with stride 2) on the small faces. The small faces themselves contain limited information, and the majority of the detailed information will be lost after several convolutional operations. For example, the input is a 16×16 face, and the result is 1×1 on the C4 feature map and nothing is reserved on the C5 feature map. Based on those limited features, it is normal to get the poor detection performance. In contrast, our method first learns a super-resolution image and then refines it, which solves the problem that the original small blurry faces lack detailed information and blurring simultaneously. Based on the super-resolution images with fine details, the boosting of the detection performance is inevitable.

Influence of the refinement network. From Table 2

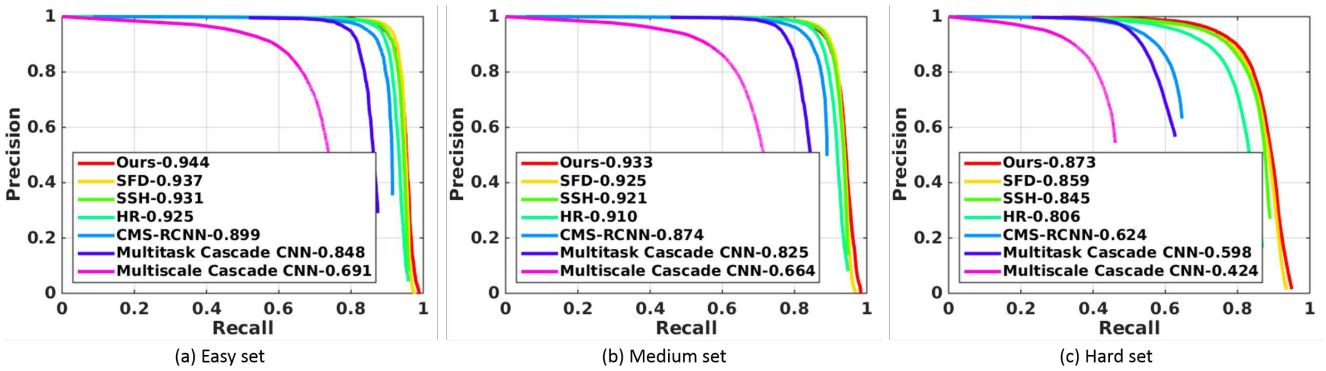

Figure 3. On the WIDER FACE validation set, we compare our method with several state-of-the-art methods: MSCNN[31], MTTCNN[33], CMS-RCNN[37], HR[10], SSH[19], SFD[35]. The average precision (AP) is reported in the legend. Best viewed in color.

(the 2^{nd} and 5^{th} row), we see that the AP performance increases by 1% on the Hard subset by adding the refinement sub-network to the generator network. Interestingly, the performances of Easy and Medium subset also have an improvement (0.4%). We visualize the reconstructed faces from the generator network and find that our refinement network can reduce the influence of illumination and blur as shown in Figure 4. In some cases, the baseline detector fails to detect the faces if those faces are heavily blurred or illuminated. However, our method reduces influence of such attributions and can find these faces successfully. Here, we would like to note that our framework is not specific and any off-the-shelf face detectors can be used as our baseline.

Influence of the adversarial loss. From Table 2 (the 3^{rd} and 5^{th} row), we see that the AP performance drops by about 1% without the adversarial loss. The reason is that the generated images derived by pixel-wise loss and classification loss are over smooth. Upon close inspecting the generated images, we find that the fine details around eyes are of low quality. Since these details are not important features for the discriminator, the generator can still fool the discriminator when making mistakes in this region. To encourage the generator to restore the high-quality images, we include the adversarial loss in our generator loss function.

Influence of the classification loss. From Table 2 (the 4^{th} and 5^{th} row), we see that the AP performance increases by about 1% with the classification loss. This is because the classification loss promotes the generator to recover the fine details for easier classification. We find that the generated faces have clearer contour when adding the classification loss. We think the contour information may be the most important evidence for the discriminator to classify face/non-face when faces are too small and heavily blurred.

4.4. Comparison with the State-of-the-Art

We compare our proposed method with state-of-the-art methods on two public face detection benchmarks (*i.e.*

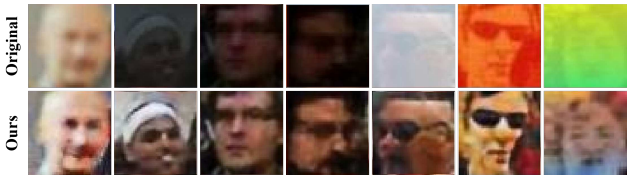

Figure 4. Some examples of the clear faces generated by our generator network from the blurry ones. The top row shows the small faces influenced by blur and illumination, and the bottom row shows the clearer faces generated by our method. The low-resolution images in the top row are re-sized for visualization.

WIDER FACE [31] and FDDB [13]).

Evaluation on WIDER FACE. We compare the our method with the state-of-the-art face detectors [31, 33, 37, 10, 19, 35]. Figure 3 shows the performance on WIDER FACE validation set. From Figure 3, we see that our method achieves the highest performance (*i.e.* 87.3%) on the Hard subset, outperforming the state-of-the-art face detector by more than 2%. Compared to these CNN-based methods, the boosting of our performance mainly comes from three contributions: (1) our up-sampling sub-network in the generator learns a super-resolution image, which reduces too much information loss caused by down-sampling while implementing convolution operations on small faces; (2) the refinement sub-network in the generator learns finer details and reconstructs clearer images. Based on the clear super-resolution images, it is easier for the discriminator to classify faces or non-faces than depending on the low-resolution blurry images; (3) the classification loss Eq(6) promotes the generator to learn a clearer face contour for easier classification. Furthermore, we also get the highest performance (94.4%/93.3%) on the Easy/Medium subset, outperforming the state-of-the-art face detector by 0.7% and 0.9% respectively. This is because some big faces are heavily influenced by illumination and blur, as shown in Figure 4. As a result,

Figure 5. Qualitative detection results of our proposed method. Green bounding boxes are ground truth annotations and red bounding boxes are the results from our method. Best seen on the computer, in color and zoomed in.

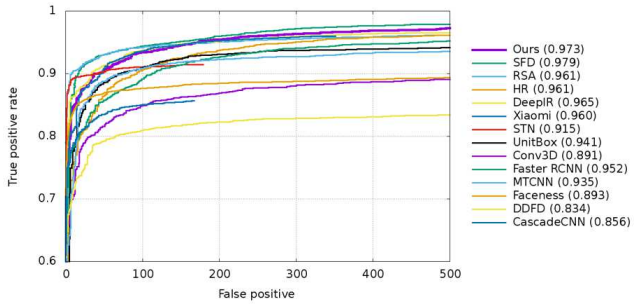

Figure 6. On the FDDB dataset, we compare our method against many state-of-the-art methods. The precision rate with 500 false positives is reported. Best viewed in color and zoomed in.

CNN-based methods fail to detect these faces. However, our method alleviates the influence of these attributions and finds these faces successfully.

Evaluation on FDDB. We follow the standard metrics (*i.e.* precision at specific false positive rates) of the FDDB [13] and use this metric to compare with other methods. There are many unlabeled faces in FDDB, making precision not accurate at small false positive rates. Hence, we report the precision rate at 500 false positives. Our face detector achieves a superior performance (0.973) over all other state-of-the-art face detectors except SFD [35] detector, as shown in Figure 6. We would like to note that the performance of SFD [35] is achieved after manually adding 238 unlabeled faces on the test set. However, we test our model on the original labeled test set. Under such an unfair condition, our method still gets the comparable performance, which further proves the effectiveness of our method.

4.5. Qualitative Results

In Figure 5, we show some detection results generated by our proposed method. It can be found that our face detector successfully finds almost all the faces, even though some faces are very small and blurred. However, Figure 5 also shows some failure cases including some false positive results. These results indicate that more progress is needed to further improve the small face detection performance. Future work will address this problem by adding the context to detecting these more challenging small faces.

5. Conclusion

In this paper, we propose a new method by using GAN to find small faces in the wild. In the generator network, we design a novel network to directly generate a clear super-resolution image from a blurry small one, and our up-sampling sub-network and refinement sub-network are trained in an end-to-end way. Moreover, we introduce an extra classification branch to the discriminator network, which can distinguish the fake/real and face/non-face simultaneously. Furthermore, the classification loss is brought to generator network to restore a clearer super-resolution image. Extensive experiments on WIDER FACE and FDDB demonstrate the substantial improvements of our method in the Hard subset, as well as in the Easy/Medium subset, when compared to previous state-of-the-art face detectors.

Acknowledgments

This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research and by Natural Science Foundation of China, Grant No. 61603372.

References

[1] Y. Bai and B. Ghanem. Multi-branch fully convolutional network for face detection. *CoRR*, abs/1707.06330, 2017. 2, 6

[2] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. *A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection*, pages 354–370. Springer International Publishing, Cham, 2016. 2

[3] A. Chakrabarti. *A Neural Approach to Blind Motion Deblurring*, pages 221–235. Springer International Publishing, Cham, 2016. 3

[4] E. L. Denton, S. Chintala, a. szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems 28*, pages 1486–1494. Curran Associates, Inc., 2015. 3

[5] C. Dong, C. C. Loy, K. He, and X. Tang. *Learning a Deep Convolutional Network for Image Super-Resolution*, pages 184–199. Springer International Publishing, Cham, 2014. 3

[6] C. Dong, C. C. Loy, and X. Tang. *Accelerating the Super-Resolution Convolutional Neural Network*, pages 391–407. Springer International Publishing, Cham, 2016. 3

[7] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. *ACM Trans. Graph.*, 25(3):787–794, July 2006. 3

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 3, 4, 5, 6

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[10] P. Hu and D. Ramanan. Finding tiny faces. *CoRR*, abs/1612.04402, 2016. 5, 7

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 4

[12] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. 3, 4, 5

[13] V. Jain and E. Learned-miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, a, 2010. 2, 6, 7, 8

[14] H. Jiang and E. Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 650–657, May 2017. 2

[15] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[17] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. 3, 4, 5

[18] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems 29*, pages 5040–5048. Curran Associates, Inc., 2016. 3

[19] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. SSH: single stage headless face detector. *CoRR*, abs/1708.03979, 2017. 7

[20] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 4

[21] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. *ACM Trans. Graph.*, 27(3):73:1–73:10, Aug. 2008. 3

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4, 6

[23] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[24] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004. 2

[25] S. Wan, Z. Chen, T. Zhang, B. Zhang, and K. Wong. Bootstrapping face detection with hard negative examples. *CoRR*, abs/1608.02236, 2016. 2

[26] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[27] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu. Tagging like humans: Diverse and distinct image annotation. In *CVPR*. IEEE, 2018. 3

[28] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang. Learning to super-resolve blurry face and text images. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3

[29] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[30] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Real-time high performance deformable model for face detection in the wild. In *2013 International Conference on Biometrics (ICB)*, pages 1–6, June 2013. 2

[31] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 6, 7

[32] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138(Supplement C):1 – 24, 2015. 2

[33] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 7

[34] L. Zhang, L. Lin, X. Liang, and K. He. *Is Faster R-CNN Doing Well for Pedestrian Detection?*, pages 443–457. Springer International Publishing, Cham, 2016. 2

[35] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S^3fd: Single shot scale-invariant face detector. *CoRR*, abs/1708.05237, 2017. 7, 8

[36] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *CoRR*, abs/1612.08534, 2016. 3

[37] C. Zhu, Y. Zheng, K. Luu, and M. Savvides. *CMS-RCNN: Contextual Multi-Scale Region-Based CNN for Unconstrained Face Detection*, pages 57–79. Springer International Publishing, Cham, 2017. 2, 7

[38] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. *Generative Visual Manipulation on the Natural Image Manifold*, pages 597–613. Springer International Publishing, Cham, 2016. 3