Recurrent Exposure Generation for Low-Light Face Detection

Jinxiu Liang, Jingwen Wang, Yuhui Quan, Tianyi Chen, Jiaying Liu, *Senior Member, IEEE*, Haibin Ling, and Yong Xu, *Senior Member, IEEE*

***Abstract*—Face detection from low-light images is challenging due to limited photons and inevitable noise, which, to make the task even harder, are often spatially unevenly distributed. A natural solution is to borrow the idea from *multi-exposure*, which captures multiple shots to obtain well-exposed images under challenging conditions. High-quality implementation/approximation of multi-exposure from a single image is however nontrivial. Fortunately, as shown in this paper, neither is such high-quality necessary since our task is *face detection* rather than *image enhancement*. Specifically, we propose a novel *Recurrent Exposure Generation (REG)* module and couple it seamlessly with a *Multi-Exposure Detection (MED)* module, and thus significantly improve face detection performance by effectively inhibiting non-uniform illumination and noise issues. REG produces progressively and efficiently intermediate images corresponding to various exposure settings, and such pseudo-exposures are then fused by MED to detect faces across different lighting conditions. The proposed method, named *REGDet*, is the first 'detection-with-enhancement' framework for low-light face detection. It not only encourages rich interaction and feature fusion across different illumination levels, but also enables effective end-to-end learning of the REG component to be better tailored for face detection. Moreover, as clearly shown in our experiments, REG can be flexibly coupled with different face detectors without extra low/normal-light image pairs for training. We tested REGDet on the DARK FACE low-light face benchmark with thorough ablation study, where REGDet outperforms previous state-of-the-arts by a significant margin, with only negligible extra parameters.**

***Index Terms*—Gated recurrent networks, low-light face detection, multi-exposure.**

I. Introduction

As THE cornerstone for many face-related systems, face detection has been attracting long-lasting research attention [22], [25], [44], [53], [55]. It has extensive applications in human-centric analysis such as face recognition [10], [60]–[64], person re-identification [8], [21], and human parsing [14]. Despite great progress in recent decade, face detection remains challenging particularly for images under bad illumination conditions. Images captured in low-light conditions typically have their brightness reduced and intensity contrast compressed, and thus confuse feature extraction and hurt the performance of face detection. Poor illumination also causes annoying noise that further damages the structural information for face detection. To make things even worse, the illumination status may spatially vary a lot within a single image. For systematic evaluation of face detection algorithms under adverse lighting conditions, a challenging benchmark named DARK FACE [56] is recently constructed, which shows clear performance degradation of state-of-the-art face detectors. For example, DSFD [28] produces an mAP of 15.3%, in a sharp contrast to above 90% on the *hard* subset of the popular WIDER FACE [55] benchmark. The dramatic performance degeneration of modern face detectors on the DARK FACE dataset clearly shows that it remains extremely challenging to detect faces under low-light conditions, which is the main focus of this paper.

Naturally, one may seek help from low-light image enhancement as preprocessing, as evidenced clearly by the experiments shown in [56]. However, as illustrated in Fig. 1 (b-c), there is still a large room for improvement. For one reason, image enhancement aims to improve visual/perceptual quality for the entire image, which is not fully aligned with the goal of face detection. For example, the smoothing operations for enhancing noisy images could compromise the feature discriminability that is critical for detection. This suggests a close integration between the enhancement and detection components, and points to an end-to-end 'detection-with-enhancement' solution.

Another reason lies in that the illumination in the original image may vary greatly in different regions. Consequently, it is hard to expect a single light-enhanced image to handle well

Manuscript received July 21, 2020; revised January 12, 2021, February 24, 2021, and March 11, 2021; accepted March 11, 2021. Date of publication March 25, 2021; date of current version March 29, 2022. This work was supported in part by the National Natural Science Foundation of China under Grants 62072188 and 61872151, in part by the Natural Science Foundation of Guangdong Province under Grants 2017A030313376 and 2020A1515011128, in part by the Science and Technology Program of Guangdong Province under Grant 2019A050510010, and in part by the Science and Technology Program of Guangzhou under Grant 201802010055. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. M. Murshed. *(Corresponding author: Yong Xu.)*

Jinxiu Liang, Yuhui Quan, and Tianyi Chen are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: cssherryliang@mail.scut.edu.cn; csyhquan@scut.edu.cn; csttychen@mail.scut.edu.cn).

Jingwen Wang is with Tencent AI Laboratory, Tencent, Shenzhen, Guangdong 518057, China (e-mail: jaywongjaywong@gmail.com).

Jiaying Liu is with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China (e-mail: liujiaying@pku.edu.cn).

Haibin Ling is with the Department of Computer Science, Stony Brook University, Strony Brook, NY 11794 USA (e-mail: hling@cs.stonybrook.edu).

Yong Xu is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China, with Peng Cheng Laboratory, Shenzhen, 518066, China, and also with the Communication and Computer Network Laboratory of Guangdong, Guangzhou 510006, China (e-mail: yxu@scut.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3068840.

Digital Object Identifier 10.1109/TMM.2021.3068840

(a) Low-light image (b) KinD [59] (c) LIME [16] (d) Ours

Fig. 1. Detection results of DSFD [28] on a low-light image (a) and its enhanced versions using KinD [59] (b) and LIME [16] (c). Green and red boxes indicate true positives and missed targets, respectively. It can be seen that the improvement brought by lighting enhancement is very limited. By contrast, our result in (d) (plotted on the same image of (c) for better visibility) show clear advantages.

facial regions under different lighting conditions in terms of detection. This suggests the use of a multiple enhancement strategy and brings our attention to the *multi-exposure* technique. In particular, when it is difficult to obtain a well-exposed image with a single shot, the technique takes multiple shots with varying camera settings. Such multi-exposure images are then fused for light enhancement. Similarly and intuitively, we may generate multi-exposure images and then detect faces from them to cover different exposure conditions. However, automatically deriving high quality multi-exposure images from a single image is non-trivial [48], let alone a low-light image – but such high quality is not required for *face detection*. It is the mechanism for capturing information at different exposures that matters.

Driven by the above motivations, we propose a novel end-to-end low-light face detection algorithm named *REGDet*. REGDet contains two sequentially connected modules, a *Recurrent Exposure Generation* (REG) module and a *Multi-Exposure Detection* (MED) module. From an input image, REG generates a sequence of pseudo-exposures to loosely mimic the effect of the highly non-linear process of in-camera multi-exposure. This is done by assembling a set of ConvGRUs marching in two directions: one direction points progressively and recurrently to the degree of exposure, while the other guides encoder-decoder structures to produce exposure compensated images. Then, these pseudo-exposures are fed into MED, which adapts generic face detectors so as to fuse 'multi-exposure' information of different pseudo-exposures smoothly. With the two modules collaborated together, REGDet not only encourages rich interaction and feature fusion across different illumination levels, but also enables end-to-end learning of effective low-light processing tailored for face detection. Moreover, as shown in our experiments, REG can be flexibly coupled with different face detectors without extra low/normal-light image pairs. We tested REGDet on the DARK FACE low-light face benchmark with thorough ablation study. In the experiments, REGDet outperforms previous state-of-the-arts by a significant margin, with only negligible extra parameters.

To summarize, we make the following contributions:

- The first end-to-end 'detection-with-enhancement' solution, REGDet, for face detection under poor lighting conditions,
- A novel and lightweight recurrent exposure generation module to tackle the non-uniform darkness issue,
- A flexible framework compatible to existing face detectors,
- New state-of-the-art performance on the publicly available benchmark.

II. Related Work

The focus in this paper is on developing a learning solution for low-light face detection. In the following we describe previous studies from three aspects: low-light image enhancement, low-light face detection, and gated recurrent networks.

A. Low-Light Image Enhancement

Low-light image enhancement has been a popular topic recently for improving the perceptual quality of images. Early solutions often rely on local statistics or intensity mapping, *e.g.*, histogram equalization [2] and gamma correction [9]. Later solutions are often based on the Retinex theory [26] which assumes an image as a combination of a reflectance map that reflects the physical characteristic of scene objects and a spatially smooth illumination map. Thus developed solutions focus on resolving the ambiguity between illumination and reflectance by imposing certain priors on a variational model based on empirical observations (*e.g.*, [11], [12], [16], [29], [47]). More recently, deep learning-based solutions boost further the image enhancement quality. These recent methods often produce impressive results for enhancing low-light images (*e.g.*, [46], [49], [51],

[59]). However, the performance gain, when applied to low-light face detection, is still far from saturated [56]. As discussed in previous section, this is partly due to their different goal with face detection, dealing with uneven illumination inside a single image, and weak collaboration with a face detection module.

The most related work to ours in low-light image enhancement is the multi-exposure fusion-based method BIMEF [57]. BIMEF first synthesizes a brighter image by a Brightness Transform Function (BTF) with fixed camera parameters, and then blends it with the original low-light image into a better one. Our method shares the idea of generating multi-exposure images, but is driven by a very different goal, *i.e.*, face detection. Consequently our model is learned end-to-end for the goal. Moreover, BIMEF does not consider the inevitable noise in low-light images and does not leverage the powerful data-driven modeling capacity of deep learning.

B. Low-Light Face Detection

With the advent of large-scale face detection datasets [22], [25], [55] and the proliferation of deep learning technologies [13], [31], [32], [38], face detection in unconstrained environments (a.k.a. 'in the wild') has made remarkable progress [18], [20], [28], [36], [37], [41], [43], [58]. Most recent technological developments have focused on robustness to geometric variance. Typical geometric distortion includes scale variation, deformation, occlusion and so on. To handle the pose variation, many effective techniques have been proposed. For example, synthesizing realistic profile faces for data augmentation [10], [63], [64], jointly normalizing profile face images to frontal pose and extracting pose invariant features [60]–[62]. For scale variation, researchers have proposed many effective strategies based on the idea of multi-scale analysis: designing image pyramids with different image scales [20], designing a pre-defined set of anchor boxes with different sizes and aspect ratios [23], [37], detecting at different layers of the network [36], [58] and so on. Deformable part-based model improves deformation invariance by decomposing the task of face detection into detecting different facial parts [54]. The idea of face calibration is explored to obtain deformation invariance in [41]. Spatial context aggregation is a modern strategy for obtaining invariant features. Existing context aggregation techniques include enlarging receptive field by dilated convolution [6], multi-layer fusion [42] and top-down feature fusion [28], [43].

Low-light face detection has been attracting research attention for a long time. In the era of hand-crafted features, enduring efforts have been made to understand and handle the non-uniform illumination issue [17], [27], [52]. Recently, there are increasing interests in data-driven approaches for face detection on low-quality images such as low-resolution images and low-light images [35], [56], [65]. Illumination variation is known to be a major challenge for modern face detection algorithms [1], [65]. Pioneering approaches preprocess images by intensity mapping such as logarithmic transform [1] and gamma transform [40]. Photometric normalization is another commonly adopted method that counteracts the varying lighting conditions in hand-crafted feature [5], [52] and deep learning-based methods [32], [65]. Hand-crafted feature based methods derive the illumination invariance from various priors such as image differences or gradients [1], [17], while deep learning-based methods use random photometric distortions as augmentation to implicitly enhance the illumination invariance [28], [43], [58]. Despite previous studies, face detection in extremely adverse light conditions has been under explored, due partly to the lack of high quality labeled data. Addressing this issue, Yang *et al.* present a large manually labeled low-light face detection dataset, DARK FACE, and show that existing face detectors perform poorly on the task [56]. Our work is thus motivated and evaluated on the benchmark, and outperforms clearly previous arts. Baseline experiments have shown that, despite of the outstanding success achieved nowadays, even the best well-trained face detectors are less than ideal if the images are simply pre-processed using existing low-light enhancement methods [56].

C. Gated Recurrent Networks

Gated Recurrent Networks are the most related work to ours from the learning aspect. Gated recurrent unit (GRU) in recurrent networks is a gating mechanism to adaptively control how much each unit remembers or forgets for sequence modeling [7]. It was first proposed and applied to task of machine translation. ConvGRU [3] extends the fully-connected layers in GRU with convolution operations to model correlations among image sequence. The design of the REG module is greatly inspired by [30]. However, the learning of the REG module is performed with a proposed pseudo-supervised pre-training strategy and the implicit guidance of a follow-up detection module instead of ground-truth data. Moreover, instead of predicting rain streak layer by residual learning, REB directly learns to generate various pseudo-exposures.

III. The Proposed Method

As shown in Fig. 2, the proposed REGDet involves two main modules, the Recurrent Exposure Generation module (REG) and the Multi-Exposure Detection module (MED). To loosely mimic the complex and highly non-linear in-camera multi-exposure process, REG generates progressively brighter images while encoding historical regional information. These pseudo-exposures are then fed into MED to produce face bounding boxes. The two modules are coupled together to form an end-to-end framework.

A. The Recurrent Exposure Generation Module

To progressively generate T pseudo-exposures from a low-light input image I_0, a natural solution is to generate the next image I_{t+1} by an NN conditioned on the previous image I_t. However, as there exists non-uniform darkness in low-light images, such strategy could lead to locally over-smoothed or over-exposed regions, and consequently hurt the face detection task that relies seriously on discriminative details.

To address the above issue, the proposed Recurrent Exposure Generation (REG) module leverages historical generated images to maintain critical region details in a Recurrent Neural

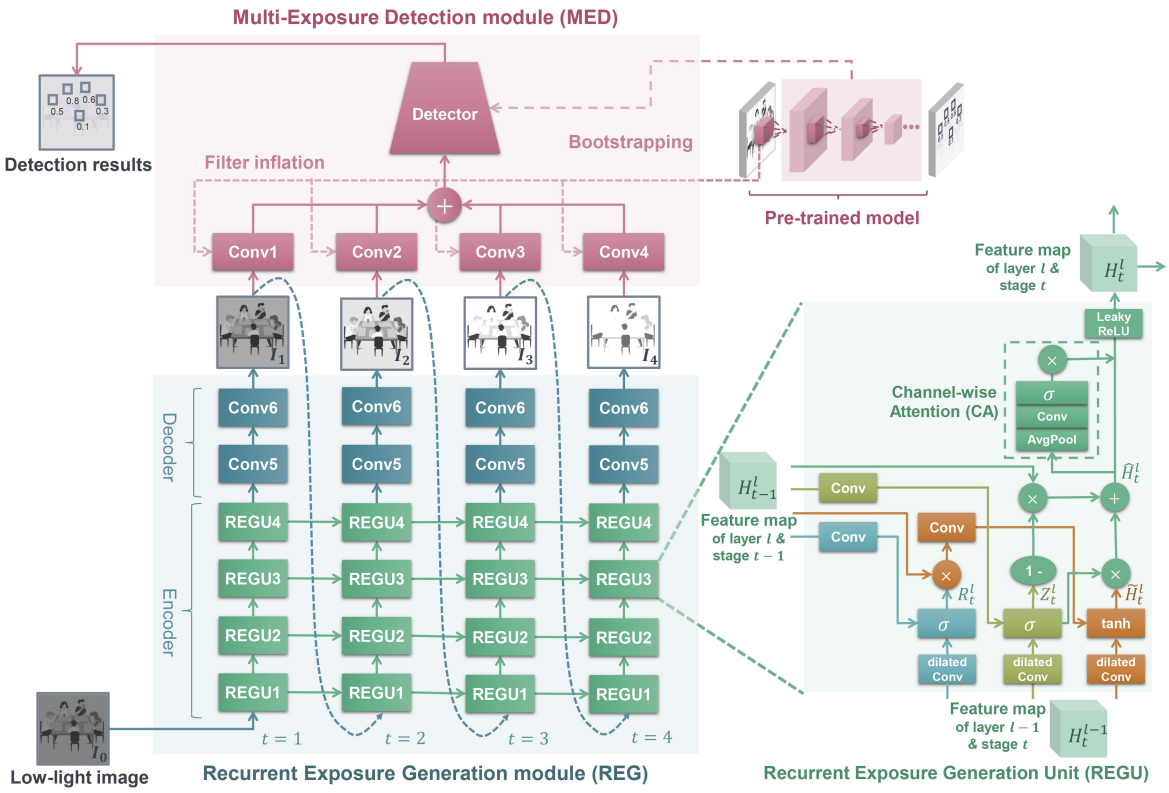


Fig. 2. The main framework of the proposed REGDet for low-light face detection.

Network (RNN) framework. Starting from I_0 and initial hidden state $H_0 = \mathbf{0}$, REG generates recurrently T intermediate pseudo-exposures $\mathbb{I} = \{I_t\}_{t=1}^{T}$ formulated as

$$(I_t, H_t) = \mathcal{G}_\omega(\mathcal{F}_\theta(I_{t-1}, H_{t-1})), \quad t = 1, 2, \ldots, T, \tag{1}$$

where $\mathcal{F}_\theta$ and $\mathcal{G}_\omega$ denote the encoder and the decoder of the proposed module, respectively, with corresponding parameters θ and ω. The encoder consisting of four cascaded convolutional recurrent layers is responsible for transforming the input image into features maps of multiple scales (layers), while the decoder consisting of two convolutional layers learns to decode the feature maps back to images, as shown in Fig. 2.

At stage $t > 0$, $H_t = \{H_t^l\}_{l=1}^{L}$ where H_t^l denotes feature map from the l-th layer. Initialized by $H_t^0 = I_{t-1}$, the feature maps are produced by our recurrent exposure generation unit (REGU) $\mathcal{F}^l$ as

$$H_t^l = \mathcal{F}^l(H_t^{l-1}, H_{t-1}^l), \quad l = 1, 2, \ldots, L. \tag{2}$$

In particular, REGU is designed based on the Convolutional Gated Recurrent Unit (ConvGRU) [3] for performance and memory consideration, as shown in the right part of Fig. 2. An REGU $\mathcal{F}^l$ in the l-th layer can be described by the following equations:

$$Z_t^l = \sigma\left(W_z^l * H_t^{l-1} + U_z^l * H_{t-1}^l\right), \tag{3}$$

$$R_t^l = \sigma\left(W_r^l * H_t^{l-1} + U_r^l * H_{t-1}^l\right), \tag{4}$$

$$\widetilde{H}_t^l = \tanh\left(W_h^l * H_t^{l-1} + U_h^l * (R_t^l \odot H_{t-1}^l)\right), \tag{5}$$

$$\hat{H}_t^l = (1 - Z_t^l) \odot H_{t-1}^l + Z_t^l \odot \widetilde{H}_t^l, \tag{6}$$

$$H_t^l = \xi(\mathcal{A}^l(\hat{H}_t^l)), \tag{7}$$

where Z and R are update and reset gates, respectively, which decide the degree to which the unit updates or resets its historical encoding information, $\sigma(x) = \frac{1}{1+e^{-x}}$ is sigmoid function, $\odot$ denotes the Hadamard product, $*$ denotes a convolution operator, filters W and U are for dilated and regular convolution respectively. ξ denotes leaky ReLU [33] activation function

$$\xi(x) = \begin{cases} \alpha x, & x < 0, \\ x, & x \geq 0, \end{cases} \tag{8}$$

where $\alpha = 0.2$ denotes the negative slope. Given a feature map $H \in \mathbb{R}^{X \times Y \times C}$, the channel-wise attention (CA) [45] $\mathcal{A}^l$ can be computed as

$$\mathcal{A}^l(H) = \mathcal{A}_s\left(\sigma(W_a^l * \mathcal{A}_g(H)), H\right), \tag{9}$$

where $\mathcal{A}_g(H) = \frac{1}{XY}\sum_{i=1,j=1}^{X,Y} H_{ij}$ is channel-wise global average pooling, W_a^l denotes a 1D convolution kernel with kernel size 3 and $\mathcal{A}_s$ denotes channel-wise multiplication between the feature map and the obtained channel weighting vector.

REGU has several extensions compared with the standard ConvGRU. First, an important component in our REGU is the channel-wise attention, which is integrated in each unit before activation except for the last one. Like in other vision tasks [45], such an efficient mechanism enables appropriate cross-channel interaction inside a feature map and therefore helps aggregate spatial global information and recalibrate the feature map at

each step. Second, REGU uses leaky ReLU [33] as the activation function to alleviate the 'dying ReLU' issue, *i.e.*, some neurons going through the flat side of zero slope stop being updated. Third, to tackle the issue of unevenly distributed darkness, different dilation rates (2^l in the l-th layer) are used in different convolutional layers of the encoder to obtain progressively larger receptive fields while maintaining small parameter cost.

B. Pseudo-Supervised Pre-Training of the REG Module

To enable good diversity and complementarity of the generated sequence, we adopt a pseudo-supervised pre-training strategy which leverages pseudo ground-truth images corresponding to different exposures. The pseudo ground-truth images $\{\hat{I}_t\}_{t=1}^{T}$ are generated from I_0 by a camera response model [57] that characterizes the relationship between pixel values and exposure ratios. A camera response model contains a camera response function (CRF), *i.e.*, the nonlinear function relating camera sensor irradiance with image pixel value, and a brightness transform function (BTF), *i.e.*, the mapping function between two images captured in the same scene with different exposures [39]. Once the parameters of CRF corresponding to a specific camera is known, the parameters of BTF can be estimated by solving the comparametric equation [34]. However, the information about the cameras to estimate accurate camera response models is often far from enough in the publicly available low-light face detection dataset. Therefore, we adopt the camera response model proposed in [57] that can characterize a general relationship between the pixel values and exposure ratios when no camera information is available. Its BTF is in the form of Beta-Gamma Correction

$$\mathcal{B}(P,k) = e^{b(1-k^a)}P^{(k^a)}, \tag{10}$$

where P and k denote the pixel value and the exposure ratio respectively, and the camera parameters $a = -0.3293, b = 1.1258$ are estimated by fitting the 201 real-world camera response curves in the DoRF database [15]. Specifically, the exposure ratios are $k, \ldots, k^T$, where the base ratio is empirically set as $k = 2.4$.

The REG module is then guided to generate images corresponding to diversified exposures. To measure the distance between the generated image I_t and the pseudo ground-truth $\hat{I}_t$ produced from I_0 with parameter k^t, we use a combination of ℓ_1 norm and the Structure Similarity (SSIM) index [50] that reflects the difference on luminance and contrast, which is formulated as

$$\mathcal{L}_{\text{reg}}(I,\hat{I}) = \frac{1}{TN}\sum_t (\|I_t - \hat{I}_t\|_1 + 1 - \text{SSIM}_t), \tag{11}$$

and the SSIM measure is defined as

$$\text{SSIM} = \frac{(2\mu_{p_t}\mu_{\hat{p}_t} + C_1)(2\sigma_{\hat{p}_t p_t} + C_2)}{(\mu_{p_t}^2 + \mu_{\hat{p}_t}^2 + C_1)(\sigma_{p_t}^2 + \sigma_{\hat{p}_t}^2 + C_2)}, \tag{12}$$

where means μ and deviations σ are computed by applying a Gaussian filter at pixel p_t of image I_t and N denotes the number of pixels in the image. Following common practice in image enhancement, we randomly crop 64×64 patches followed by random mirror, resize and rotation for data augmentation.

TABLE I
RESULTS OF ABLATION STUDY ON THE PROPOSED REG MODULE. THE MAP IS REPORTED AS PERCENTAGE (%)

Method	DSFD [28]		PyramidBox [43]		S3FD [58]	
	#Params	mAP	#Params	mAP	#Params	mAP
Baseline	47.49M	71.42	54.53M	72.48	21.42M	54.99
Ours-BEG	+ 0.09M	75.60	+ 0.09M	76.11	+ 0.09M	56.78
Ours-CEG	+ 0.09M	74.07	+ 0.09M	73.16	+ 0.09M	54.30
Ours-SEG	+ 0.03M	73.52	+ 0.03M	74.19	+ 0.03M	52.82
Ours-REG	+ 0.12M	**76.94**	+ 0.12M	**77.69**	+ 0.12M	**57.95**

TABLE II
RESULTS OF ABLATION STUDIES ON DIFFERENT COMPONENTS OF THE PROPOSED METHOD

pseudo-supervised pre-training	joint training with MED	channel-wise attention	filter inflation	mAP (%)
✔	✔	✔	✔	**77.69**
✗	✔	✔	✔	76.36
✔	✗	✔	✔	70.63
✔	✔	✗	✔	76.70
✔	✔	✔	✗	77.15

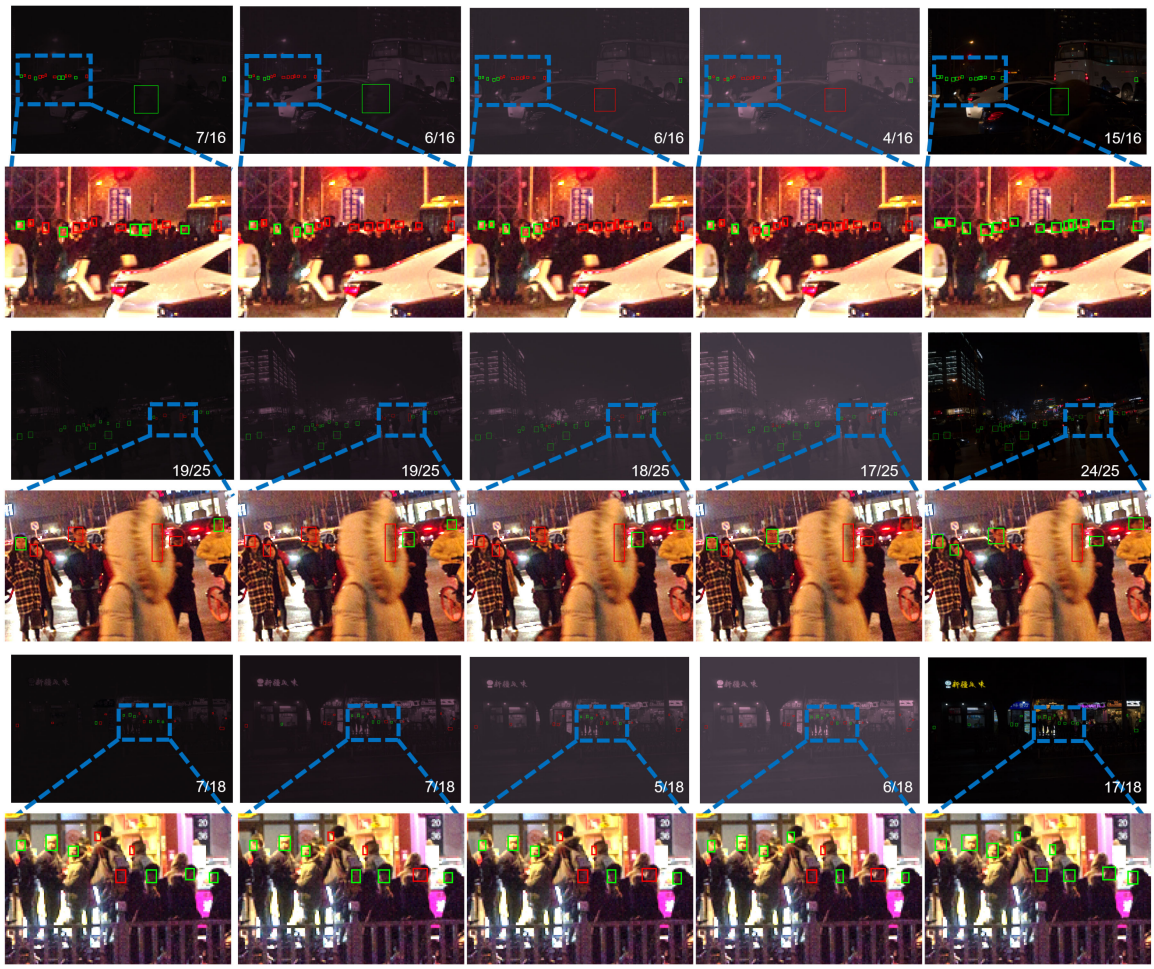
As the pseudo ground-truth images have inevitable noise and artifacts, we adopt the early stopping strategy to prevent overfitting to those noise and artifacts. Specifically, the pre-training stops when the average PSNR of I_t compared to $\hat{I}_t$ reaches around 25. We use the training split of the DARK FACE dataset to perform the pseudo-supervised pre-training. As our method does not rely on any external low/normal-light image pairs, it enjoys good scalability and can be fairly compared to other approaches. This pre-training practice can be expected to speedup the joint training process and boost the final detection performance. The performance comparison can be found in Table II.

To understand and verify the complementarity of the generated sequence from the REG module, we visualize them in Fig. 3. The detection results on the generated images using the pre-trained DSFD detector in the left four images show good *complementarity* between different generated images, indicating that the REGDet learns to generate a complementary detection-oriented image sequence to benefit subsequent face detection.

C. The Multi-Exposure Detection Module

Once the multiple pseudo-exposures $\mathbb{I}$ are created by the REG module, a straightforward strategy is to separately feed them into a face detector and fuse their corresponding detected bounding boxes, *i.e.*, *late fusion*. This is however computationally expensive as it requires multiple runs of the detection process. Instead, we introduce a resource efficient strategy to fuse the low-level features extracted from $\mathbb{I}$ in early stage of detection. Such strategy not only takes advantage of available pre-trained face detectors, but also allows the collaboration among different pseudo-exposures.

Specifically, the proposed Multi-Exposure Detector (MED) module integrates a generic pre-trained CNN-based face detection algorithm, named *base detector* with *early fusion*. We tailor



(a) Detect on I_1 (b) Detect on I_2 (c) Detect on I_3 (d) Detect on I_4 (e) REGDet

Fig. 3. The left four are detection results on intermediate I_1, I_2, I_3 and I_4 generated from the REG module, which show complementarity among the generated images, supporting the effectiveness of our proposed REG module. Note these 'images' are linearly normalized for visualization so that the minimum (maximum) value corresponds to 0 (255). The rightmost column shows our final detection result, where more faces (14 out of 15) are successfully localized, showing superiority of the proposed MED module. Green and red boxes indicate true positives and missed targets, respectively. The zoom-in versions on the second row are enhanced by LIME [16] for better visibility.

its first convolutional layer using *filter inflation* technique [4] in the channel dimension so that the detector can simultaneously process multiple images and perform adaptive integration, as shown in Fig. 2. The weights of the T convolutional layers are bootstrapped from the first layer in the pre-trained base detector, by duplicating and normalizing the pre-trained filter weights T times, which helps maintain better discriminative and complementary regional clues across different pseudo-exposures. Formally, MED $\mathcal{M}$ simultaneously predicts the confidences $p = \{p_i\}_{i=1}^{N_a}$ and the bounding box coordinates $g = \{g_i\}_{i=1}^{N_a}$ of anchor boxes indexed by $1, 2, \ldots, N_a$ as

$$(p, g) = \mathcal{M}(\mathbb{I}), \tag{13}$$

where N_a denotes the number of anchors, p_i measures how confident the i-th anchor is a face and g_i is a vector representing the 4 parameterized coordinates of the predicted face boxes. Following [32], we use weighted sum of the confidence loss and the localization loss:

$$\mathcal{L}(p, \hat{p}, g, \hat{g}) = \frac{1}{N_a} \sum_i \mathcal{L}_{\text{conf}}(p_i, \hat{p}_i) + \frac{\lambda}{N_p} \sum_i \hat{p}_i \mathcal{L}_{\text{loc}}(g_i, \hat{g}_i), \tag{14}$$

where N_p denotes the number of positive anchors, λ is used to balance the two loss terms, the ground-truth label $\hat{p}_i$ represents whether the i-th anchor is positive (a.k.a., is a face), and $\hat{g}_i$ is the ground-truth bounding box assigned to the anchor. The confidence (classification) loss $\mathcal{L}_{\text{conf}}(p_i, \hat{p}_i)$ is a two-class (face or background) softmax loss,

$$\mathcal{L}_{\text{conf}}(p_i, \hat{p}_i) = \hat{p}_i \log(p_i) + (1 - \hat{p}_i) \log(1 - p_i), \tag{15}$$

where the $\hat{p}_i$ in the second term means that the localization loss is only calculated for those positive anchors. Following [13], the localization loss $\mathcal{L}_{\text{loc}}(g_i, \hat{g}_i)$ is defined as the smooth ℓ_1 loss, *i.e.*, the distance between the predicted box g_i and the ground-truth

$\hat{g}_i$ measured by Huber norm

$$\mathcal{L}_{\text{loc}}\left(g_i, \hat{g}_i\right) = \sum_{j \in \{x,y,h,w\}} \mathcal{H}\left(g_i^{(j)} - \hat{g}_i^{(j)}\right), \tag{16}$$

where the Huber norm $\mathcal{H}(\cdot)$ is defined as

$$\mathcal{H}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise}. \end{cases} \tag{17}$$

The Huber norm is less sensitive to outliers than the ℓ_2 norm.

Being an end-to-end system, REGDet allows joint optimization of the REG and MED modules during learning. Intuitively, MED provides facial location information to guide REG such that the facial regions could be specially enhanced for the purpose of detection. An example detection result is shown in the rightmost column of Fig. 3, and it shows that REGDet successfully localizes far more faces than simply applying the base detector on different intermediate images. It is worth noting that MED is flexible in choosing the base detector. In our experiments, several state-of-the-art algorithms such as DSFD [28], PyramidBox [43] and S3FD [58] all demonstrate clear performance improvement when embedded in REGDet.

IV. Experiments

A. Setup

1) Dataset and Metric: We adopt the recently constructed DARK FACE dataset [56] as our testbed. 6000 real-world low-light images captured under extreme low-light environment. The resolution of the images is 1080×720. Totally 43 849 manually annotated faces are released. The annotated faces have large scale variance, ranging from 1×2 to 335×296. There are usually 1 to 20 annotated faces in an image. Since the original test split [56] is withheld, we randomly leave 1000 images as our test set. Figure 5 shows the distribution of face number and face resolution in the train/test splits. Following prior work [28], [43], [58], face detection performance is measured by mean Average Precision (mAP), which is calculated as the area under precision-recall curve.

2) Network Architecture: To benefit from the publicly available pre-trained models, we build up REGDet on the *base detectors* pre-trained on the existing largest dataset for face detection in the wild, *i.e.*, WIDER FACE [55] dataset. DSFD [28], PyramidBox [43] and S3FD [58], the state-of-the-art methods that achieve remarkable performance on WIDER FACE, are chosen as the base detectors. The weights of REGDet are initialized and bootstrapped as described in Section III-B and III-C. For reproducibility, we adopt public implementation of the base detectors with VGG-16 backbone network, which are all implemented with the PyTorch library. For scalability, the configurations of anchor design, sample matching, optimization and inference for different base detectors are set as suggested in the original papers [28], [32], [43] unless otherwise specified. For the proposed REG module, we set the number of stages as $T = 4$ and the number of REGU blocks as $L = 4$.

3) Data Augmentation: During training, for all methods we randomly crop image patches with random scales and then resize them to 640×640. To construct a model more robust to commonly seen variations, we adopt data augmentation schemes such as random patch sampling and random flipping following [32]. For our proposed REGDet, the random photometric distortion in data augmentation is removed as it has already involved an enhancement module. Note that we keep the photometric augmentation for the baselines following [28], [43], [58] for fair comparison.

4) Anchor Design: The anchor scales are the same for all the three base detectors at the inference stage, *i.e.*, 16, 32, 64, 128, 256, and 512. Following the baselines, we set the anchor ratio as 1:1 for S3FD and PyramidBox, and 1.5:1 for DSFD. The designed anchors cover a wide range of face scales, specifically, from faces with around 16×16 pixels to faces with around 512×512 pixels.

5) Hard Negative Mining: After the anchor matching step, a large number of negative anchors are produced, which causes significant imbalance between the positive and negative training samples and poor convergence performance. To address this issue, following [32], hard negative mining is adopted to select the negatives with highest cost in the training phase and make the ratio between the negative and positive anchors below 3:1.

6) Optimization: The models are trained with a batch size of 16 for 120 epochs. We adopt SGD with momentum of 0.9 to train the MED module. Annealing learning rate initialized with 0.001 and decay factor of 0.1 (decayed at the 64-th and 96-th epoch) are used for training the MED module following the common practice. The adaptive moments [24] (Adam) with default parameter setting is adopted for training the REG module, since it has shown promising results for training NNs with recurrent architecture.

7) Inference: During inference, the image is first rescaled to make $\sqrt{H \times W} = 2000$, where H and W denote the height and width of the test image respectively. The boxes output by the proposed method are firstly filtered out by a confidence threshold of 0.01 and keep the top 5000 boxes before applying non-maximum suppression (NMS). Then NMS is applied with Jaccard overlap of 0.3 and the top 750 bounding boxes are kept.

8) Compared Methods: We compare REGDet against various face detectors with illumination pre-processing using the state-of-the-art low-light image enhancement approaches including **MF** [11], **SRIE** [12], **LIME** [16], **BIMEF** [57], **GLAD-Net** [49], **RetinexNet** [51], **RRM** [29], **DeepUPE** [46], and **KinD** [59] to preprocess the images. **Baseline** denotes the plain detector fed by the original low-light images as input. We evaluate all the aforementioned approaches with both pre-trained and finetuned version. The pre-trained version directly uses the pre-trained weights on WIDER FACE and performs inference on pre-processed DARK FACE images using the aforementioned methods. The finetuned version further finetunes the model using pre-processed DARK FACE images as input. As the performances reported in [56] are for the withheld test data split with only pre-trained version, *we re-train the aforementioned methods on our train split and fairly compare them on our 1000-image test split*.

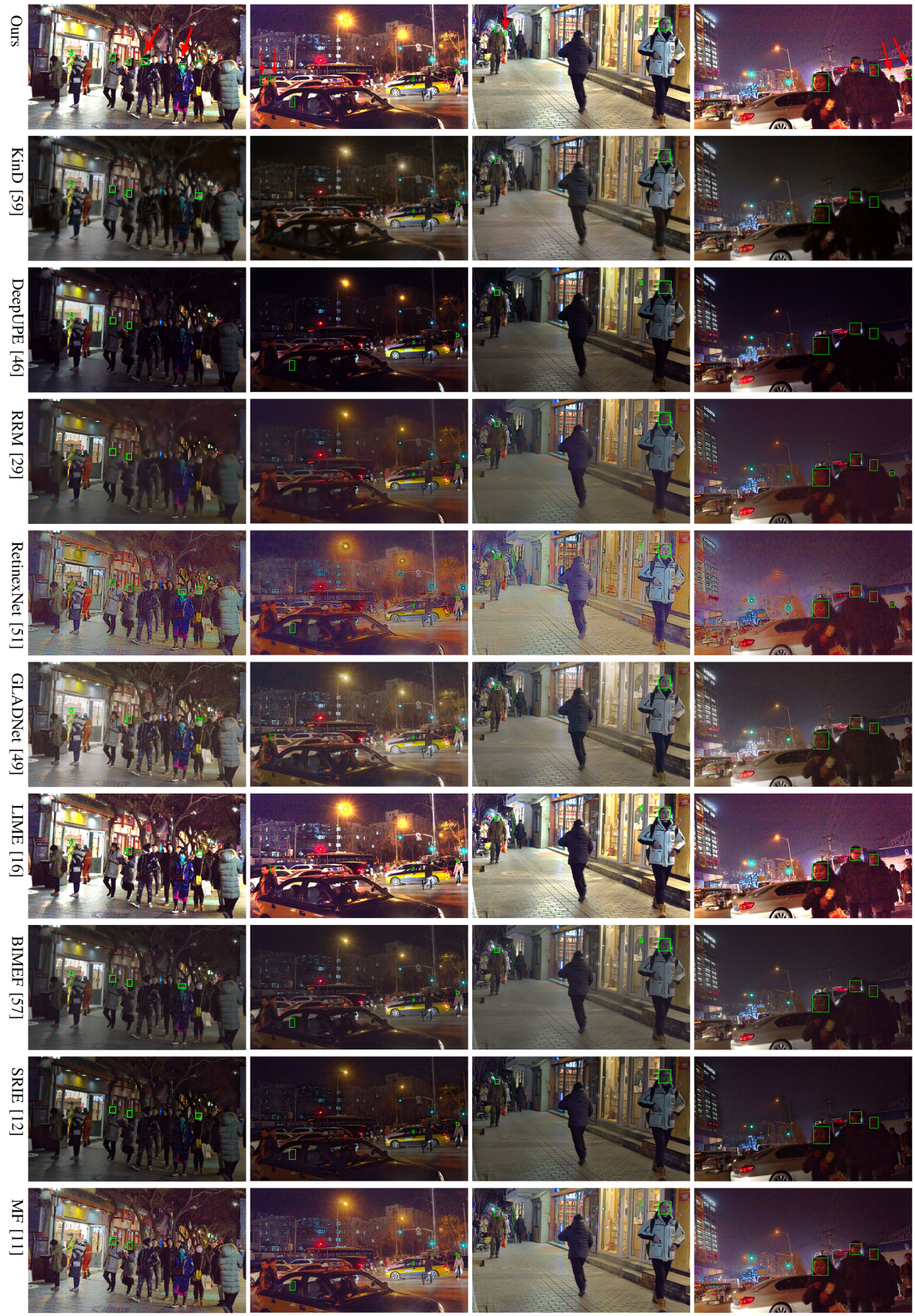
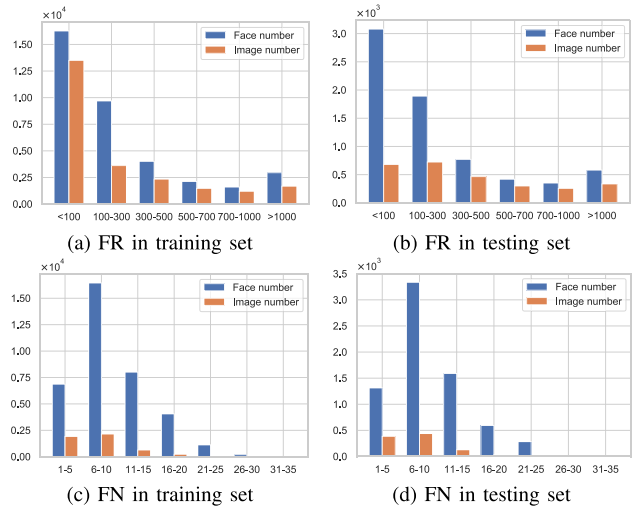


Fig. 4. Qualitative comparison of different methods. For better visualization, we draw the results of REGDet on images enhanced by LIME [16]. Red arrows indicate those faces that are challenging to be detected by the other methods. *Please Zoom in to see better*.



(a) FR in training set (b) FR in testing set

(c) FN in training set (d) FN in testing set

Fig. 5. Face resolution (FR) and face number (FN) distribution in train and test splits.

B. Result Analysis

The quantitative comparison of different approaches is shown in Fig. 6. The three pre-trained baseline detectors achieve results of 32.69%, 31.00%, and 26.58% mAP respectively. The relative performance disparity among the three detectors are consistent with their performance on WIDER FACE. The former two detectors perform much better as they apply modern context aggregation techniques such as feature enhancement using two shots [28] or context assisted pyramid anchors [43]. Compared with the pre-trained detectors, all finetuned ones achieve much higher performance, indicating that the existing large-scale dataset WIDER FACE dominated by normal-light images carries very different lighting distribution compared to DARK FACE dataset. Compared with original image input, many of the image enhancement approaches improve the face detection performance. Specifically, the pre-trained detectors equipped with pre-processing using MF, LIME, BIMEF, DeepUPE, GLAD-Net, and SRIE outperform the baseline with respectively 4.87%, 5.08%, 5.33%, 4.60%, and 0.45% performance gain when using DSFD as the base detector. In the finetuned setting, MF, LIME, BIMEF, and DeepUPE improve the baseline with respectively 1.12%, 0.94%, 1.75%, and 1.05% performance gain when using DSFD as the base detector. While these image enhancement methods show clear advantages over the baseline with the pre-trained setting, they achieve less performance gain in the finetuned setting, as finetuning already greatly reduces the data distribution discrepancy between normal-light and low-light images. However, it is noticeable that KinD, RetinexNet, and RRM cause performance degeneration to different extents due probably to the severe over-smoothness (KinD, RRM) or artifacts (RetinexNet) on regions containing faces (also evidenced by Fig. 4. Among them, the multi-exposure fusion method BIMEF performs best. The relatively good performance of BIMEF may also imply that it is promising to adaptively generate pseudo exposures with different light conditions, which is consistent with what we explored in this paper. In particular, compared with the finetuned baseline on original images equipped with photometric data augmentation [19], the proposed REGDet shows much higher detection mAP with respectively about **5.5%**, **5.2%**, and **3.0%** performance gain using the three base detectors, with negligible extra parameters (as shown in Table II). The overwhelmingly high detection rates of REGDet demonstrates its superiority over existing state-of-the-arts.

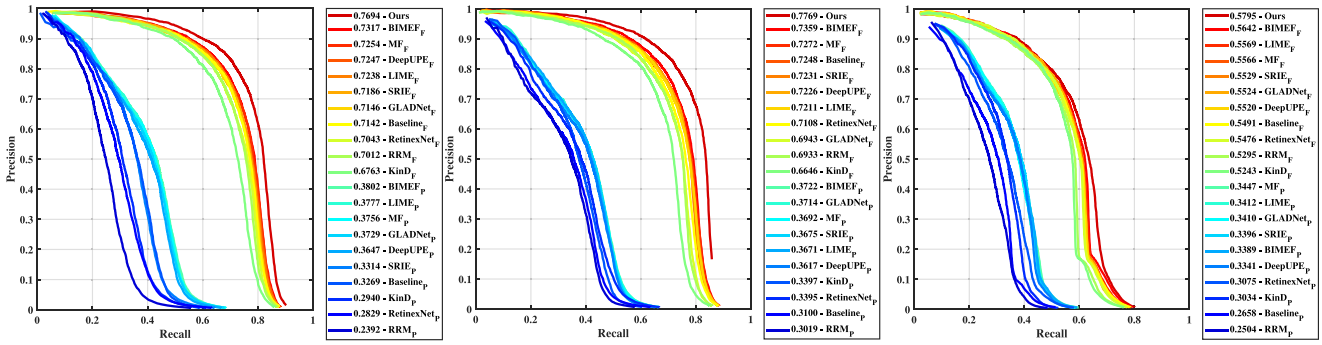
The qualitative results of different approaches on sampled images from DARK FACE are shown in Fig. 4. While those large and clear faces can also be detected by other methods, our method has successfully found much more dark and tiny faces, as pointed out by the red arrows in the presented images. Although it is hard to detect those faces even by human eyes, the proposed method is able to localize most of them and clearly outperforms other approaches.

C. Ablation Studies

1) Model Design of the Recurrent Architecture: To examine the effectiveness of the proposed recurrent component, variant generation modules are designed as illustrated in Fig. 7, which includes

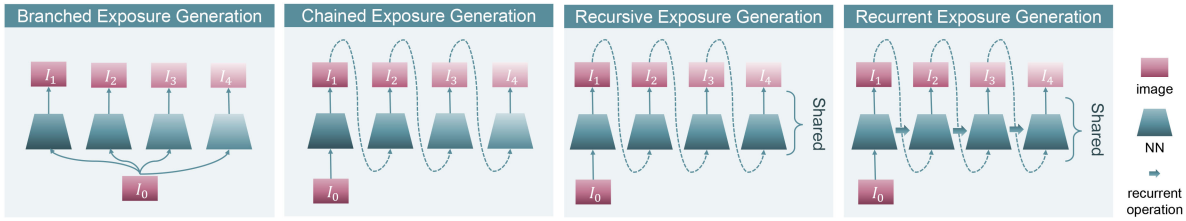
- **Branched Exposure Generation (BEG)** This module generates different exposures I_t parallelly from the original image I_0 by a module with T branches,
- **Chained Exposure Generation (CEG)** The t-th image is generated at the t-th stage of the module with non-shared weights conditioned on the image I_{t-1} generated at the $(t-1)$-th stage,
- **RecurSive Exposure Generation (SEG)** Similar with CEG, except that the module shares parameters at different stages,
- **Recurrent Exposure Generation (REG)** The module used in our proposed method. Different from the aforementioned modules, REG encodes historical feature maps in order to alleviate the probable unrecoverable information loss caused by the over-exposure and over-smoothness at the middle stages. The detailed description of the REG module is provided in Sec. III-A.

We replace REG with BEG, CEG, SEG respectively and conduct experiments on DARK FACE. As shown in Table I, all the designed lightweight modules introduce merely a few extra parameters while they almost all achieve improved detection results. BEG constructs multiple branches from the original image I_0 to generate different pseudo-exposures in parallel, and clearly boosts performance, indicating that the MED module does provide important guidance to the enhancement module for generating complementary information in different pseudo-exposures, as illustrated in Sec. III-C. In contrast, CEG and SEG that generate I_t conditioned on I_{t-1} with non-shared and shared weight, respectively, produce not so stable performance gain, due probably to unrecoverable information loss caused by the over-exposure and over-smoothness at the middle stages. This suggests that a proper modeling of the multi-exposure generation is the key to achieve good face detection performance. For the performance of using S3FD as base detector, Ours-CEG and Ours-SEG only



(a) DSFD [28] (b) PyramidBox [43] (c) S3FD [58]

Fig. 6. Quantitative results of different approaches are shown. All the other approaches have both pre-trained version (marked with subscript 'P') and finetuned version (marked with subscript 'F') excepting for ours.



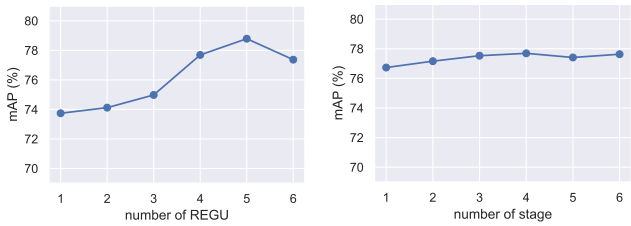
(a) BEG (b) CEG (c) SEG (d) REG

Fig. 7. Alternative pseudo-exposure generation modules.

achieve comparable or even decreased detection rates. We conjecture that the reason of the inferior performance is that S3FD has much less parameters and consequently much smaller model capacity compared with DSFD and PyramidBox, resulting in insufficient guidance effects for the generation modules. By encoding historical feature maps, the proposed REG alleviates the issue and performs the best. It indicates that the relationship between adjacent pseudo-exposures could be well modeled by maintained memory in the recurrent structure of REG. The consistent performance boost also demonstrates the scalability of REG across different base face detectors.

2) Pseudo-Supervised Pre-Training: The REG module is supervised and guided to generate images corresponding to diversified exposures with the designed pseudo-supervised pre-training. We provide experimental comparison on whether applying the proposed pseudo-supervised pre-training on the REG module or not. The performance of the resulted REGDet using PyramidBox as base detector are shown in Table II. When randomly initializing the REG module (w/o pre-training), the proposed REGDet remains good performance with an mAP of 76.36%. Equipped with the proposed pseudo-supervised pre-training technique, our method achieves the best performance with 1.33% absolute performance gain.

3) Joint Training With MED: The ability of generating images with diverse levels of exposure is not enough. For images captured under different lighting conditions, the accordingly proper level of exposure is also different. Moreover, it is not clear what characteristics of images can help face detection more. A direct guidance signal coming from the face detector could be helpful, which can be implemented by jointly training with MED. To verify its effectiveness, we freeze the weights of the pre-trained REG module, *i.e.*, without jointly training with MED. The corresponding result is reported in the third row of Table II. There is a dramatic performance degeneration without jointly training, specifically, 70.63% *vs.*77.69%.

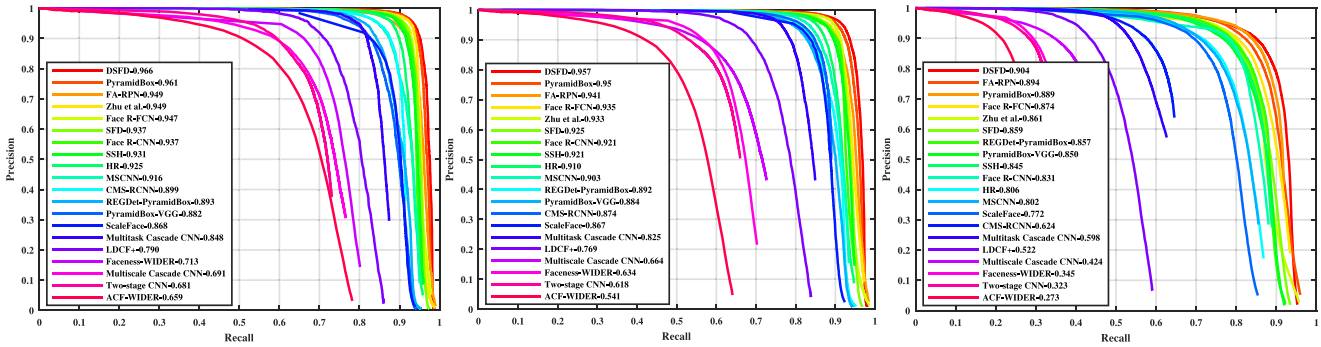


(a) mAP *vs.* number of REGU (b) mAP *vs.* number of stage

Fig. 8. Sensitivity studies on the hyper-parameters.

4) Channel-Wise Attention: In the proposed REGU, channel-wise attention enables appropriate cross-channel interaction inside a feature map. As shown in Table II, the channel-wise attention leads to performance gain of about 1% mAP.

5) Filter Inflation: We tailor the first convolutional layer of the detector using filter inflation technique [4] in the channel dimension so that the detector can simultaneously process multiple images and perform adaptive integration. The weights of the T convolutional layers are bootstrapped from the first layer in the pre-trained base detector, by duplicating and normalizing the pre-trained filter weights T times. The corresponding ablation is shown in Table II. Applying filter inflation results into 0.54% mAP gain.



(a) Val: Easy (b) Val: Medium (c) Val: Hard

Fig. 9. Precision-recall curves on the WIDER FACE validation set.

D. Hyper-Parameter Analysis

1) Numbers of REGU Blocks: There are $L = 4$ REGU blocks in the REG module. Generally, increasing the number of REGU blocks increases the capacity of the model as well as the computational costs associated with the model. Moreover, overfitting might occur when L is too large. To study the effect of the hyper-parameter L, we conduct several experiments using PyramidBox as base detector. The results are shown in Figure 8(a). We find that increasing L consistently improves the results when $L < 5$, and achieve a best mAP of 78.79% when $L = 5$. To tradeoff between effectiveness and efficiency, we set $L = 4$ in all other experiments.

2) Numbers of Stages: We conduct experimental comparison of different numbers of stages T for the REG module using PyramidBox as base detector. The results are shown in Fig 8(b). Setting $T = 1$ is equivalent to a special case of REGDet, namely, a single-exposure 'detection-with-enhancement' model. It achieves much higher detection performance (mAP) than the finetuned baseline (72.48%), but achieves inferior result than the multi-exposure frameworks ($T > 1$). On one hand, it supports the claim that jointly performing enhancement and detection is superior compared to plain detection for low-light face detection. On the other hand, it verifies the superiority of the proposed multi-exposure framework over single-exposure framework. Setting $T = 4$ achieves the best performance, indicating that it is a good practice.

E. More Analysis

1) Results on WIDER FACE: In this paper, we aim at face detection in low-light conditions, which might be the most commonly seen one among various poor visibility environments. However, the proposed method is also applicable for more general cases, *e.g.*, a model with robustness to large illumination variation. To evaluate the performance of the proposed method for real-world scenarios covering more general lighting conditions, we use a mixture of the WIDER FACE dataset (normal-light) and the DARK FACE dataset (low-light) to train our REGDet with PyramidBox as base detector.

The results on the WIDER FACE dataset are shown in Figure 9. Our method is denoted as 'REGDet-PyramidBox'. The corresponding baseline model is denoted as 'PyramidBox-VGG,' which is re-implemented based on the same VGG-16 backbone and test protocol as our method for fair comparison. Intuitively, REGDet cannot be expected to outperform the latest methods that are built upon backbone with larger model complexity, *e.g.*, ResNet-152, or multi-scale testing, and trained on the pure WIDER FACE train split that have much smaller distribution discrepancy with the test data. Still, our proposed model achieves comparable or even better performance compared to the baselines 'PyramidBox-VGG'. Specifically, compared to the baseline, the proposed method bring a performance gain of 1.1%, 0.8%, and 0.7% mAP respectively on the easy/medium/hard subsets of WIDER FACE despite of the discrepancy of data distribution. On the low-light dataset DARK FACE, REGDet-PyramidBox achieves an mAP of 73.86%. The empirical performances indicate that our proposed REGDet has good robustness to large illumination variation.

TABLE III
TRAINING AND TESTING COMPUTATIONAL COMPLEXITY

Base Detector	Training time (h)	Test time (ms)	FLOPS (G)
DSFD [28]	22.59	341	520.92
PyramidBox [43]	21.89	338	715.04
S3FD [58]	13.24	301	435.32

2) Training and Testing Computational Complexity: For training, it takes about 22 hours on a server with 8 Tesla V100 GPUs when using a batch size of 16 for 120 epochs. For testing, it takes about 0.3 s to process an input image of VGA-resolution (640×480). The computational complexity based on different base detectors is summarized in Table III. The test time is obtained by averaging from 10 runs on a single Titan RTX GPU.

V. CONCLUSION

In this work we proposed an end-to-end face detection framework, named REGDet, for dealing with low-light input images. The key component in REGDet is a novel recurrent exposure generation (REG) module that extends ConvGRU to mimic the multi-exposure technique used in photography. The REG module is then sequentially connected with a multi-exposure detection (MED) module for detecting faces from images under poor

lighting conditions. The proposed method significantly outperforms previous algorithms on a public low-light face dataset, with detailed ablation study further validating the effectiveness of the proposed learning component.

REFERENCES

[1] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 721–732, Jul. 1997.

[2] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1921–1935, Sep. 2009.

[3] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *Proc. Int. Conf. Learn. Representations*, Mar. 2016.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[5] W. Chen, M. J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Trans. Syst. Man Cybern. B. Cybern.*, vol. 36, no. 2, pp. 458–466, Apr. 2006.

[6] C. Chi *et al.*, "Selective refinement network for high performance face detection," in *Proc. AAAI Conf. Artif. Intell.*, Sep. 2019, pp. 8231–8238.

[7] K. Cho *et al.*, "Learning phrase representations using RNN encoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[8] G. Ding *et al.*, "Feature affinity-based pseudo labeling for semi-supervised person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2891–2902, Nov. 2019.

[9] H. Farid, "Blind inverse gamma correction," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1428–1433, Oct. 2001.

[10] J. Feng *et al.*, " 3D-Aided deep pose-invariant face recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1184–1190.

[11] X. Fu *et al.*, "A. fusion-based enhancing method for weakly illuminated images," *Signal Process.*, vol. 129, pp. 82–96, Dec. 2016.

[12] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding,, "A Weighted variational model for simultaneous reflectance and illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2782–2790.

[13] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[14] K. Gong *et al.*, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 770–785.

[15] M. Grossberg and S. Nayar, "Modeling the space of camera response functions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1272–1282, Oct. 2004.

[16] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[17] H. Han, S. Shan, X. Chen, and W. Gao, "A. comparative study on illumination preprocessing in face recognition," *Pattern Recognit.*, vol. 46, no. 6, pp. 1691–1699, Jun. 2013.

[18] Z. Hao *et al.*, "Scale-aware face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6186–6195.

[19] A. G. Howard, "Some improvements on deep convolutional neural network based image classification," Dec. 2013, *arXiv:1312.5402 [cs]*.

[20] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1522–1530.

[21] Y. Huang *et al.* "Multi-pseudo regularized label for generated data in person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1391–1403, Mar. 2019.

[22] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Tech. Rep. UM-CS-2010-009, Univ. Massachusetts, Amherst, 2010.

[23] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2017, pp. 650–657,.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[25] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1931–1939.

[26] E. H. Land, "The retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, 1977.

[27] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: The importance of good features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2004, pp. II–II.

[28] J. Li *et al.*, "DSFD: Dual shot face detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5060–5069.

[29] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.

[30] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 262–277.

[31] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[32] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[33] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, vol. 30, 2013, p. 3.

[34] S. Mann, "Comparametric equations with practical applications in quantigraphic image processing," *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1389–1406, Aug. 2000.

[35] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel, "Pushing the limits of unconstrained face detection: A challenge dataset and baseline results," in *Proc. IEEE Int. Conf. Biometrics Theory Appl. Syst.*, Oct. 2018, pp. 1–10.

[36] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4875–4884.

[37] M. Najibi, B. Singh, and L. S. Davis, "FA-RPN: Floating region proposals for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7723–7732.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, 2015, pp. 91–99.

[39] Y. Ren, Z. Ying, T. H. Li, and G. Li, "LECARM: Low-light image enhancement using the camera response model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 968–981, Apr. 2019.

[40] S. Shan, W. Gao, B. Cao, and D. Zhao, "Illumination normalization for robust face recognition against varying lighting conditions," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, USA, 2003, pp. 157–164.

[41] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2295–2303.

[42] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018.

[43] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 812–828.

[44] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[45] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2020, pp. 11534–11542.

[46] R. Wang *et al.*, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6849–6857.

[47] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.

[48] T.-H. Wang *et al.*, "Pseudo-multiple-Exposure-Based tone fusion with local region adjustment," *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 470–484, Apr. 2015.

[49] W. Wang, C. Wei, W. Yang, and J. Liu, "GLADNet: Low-light enhancement network with global awareness," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 751–755.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

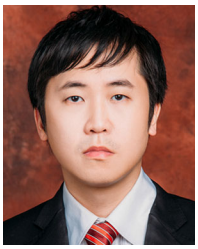
[51] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. BMVC*, Aug. 2018.

[52] S. Yan, S. Shan, X. Chen, and W. Gao, "Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[53] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.
[54] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3676–3684.
[55] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5525–5533.
[56] W. Yang *et al.*, "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020. doi: 10.1109/TIP.2020.2981922.
[57] Z. Ying, G. Li, and W. Gao, "A Bio-inspired multi-exposure fusion framework for low-light image enhancement," Nov. 2017, *arXiv:1711.00591 [cs]*.
[58] S. Zhang *et al.*, "S3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 192–201.
[59] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, May 2019, pp. 1632–1640.
[60] J. Zhao, "Deep learning for human-centric image analysis: From face recognition to human parsing." Ph.D. dissertation, National Univ. Singapore, Nov. 2018.
[61] J. Zhao *et al.*, "Towards pose invariant face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2207–2216.
[62] J. Zhao, J. Xing, L. Xiong, S. Yan, and J. Feng, "Recognizing profile faces by imagining frontal view," *Int. J. Comput Vis*, vol. 128, no. 2, pp. 460–478, Feb. 2020.
[63] J. Zhao *et al.*, "Dual-agent GANs for photorealistic and identity preserving profile face synthesis," *Proc. NeurIPS*, vol. 30, 2017, pp. 66–76.
[64] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-Aided dual-agent GANs for unconstrained face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, Oct. 2019.
[65] Y. Zhou, D. Liu, and T. Huang, "Survey of face detection on low-quality images," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, pp. 769–773, May 2018.

Jinxiu Liang received the B.E. degree, in 2016 in computer science from the South China University of Technology, Guangzhou, China, where she is currently working toward the Ph.D. degree in computer science. Her research interests include computer vision, image processing, and machine learning.

Jingwen Wang received the Ph.D. degree in computer science and technology from the South China University of Technology, Guangzhou, China, in 2018. He is currently a Senior Researcher with Tencent AI Lab, Shenzhen, China. His research interests include deep learning and computer vision, with respect to action classification, action detection, vision, and language.



Yuhui Quan received the Ph.D. degree in computer science from the South China University of Technology, Guangzhou, China, in 2013. From 2013 to 2016, he was the Postdoctoral Research Fellow of mathematics with the National University of Singapore, Singapore. He is currently an Associate Professor with the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing, and sparse representation.

Tianyi Chen is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include image processing and computer vision.

Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing China, 2010. She is currently an Associate Professor, Peking University Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University. From 2007 to 2008, she was a Visiting Scholar with the University of Southern California, Los Angeles, Los Angeles, CA, USA. In 2015, she was a Visiting Researcher with the Microsoft Research Asia supported by the Star Track Young Faculties Award. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 50 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. Dr. Liu is a Senior Member of theCSIG and CCF. She was a member of Multimedia Systems and Applications Technical Committee, and Visual Signal Processing and Communications Technical Committee in IEEE Circuits and Systems Society. She was the recipient of the IEEE ICME-2020 Best Paper Award and the IEEE MMSP-2015 Top10% Paper Award. She was also an Associate Editor for the IEEE TRANSACTION ON IMAGE PROCESSING, IEEE TRANSACTION ON CIRCUIT SYSTEM FOR VIDEO TECHNOLOGY, and Elsevier JVCI, the Technical Program Chair of the IEEE ICME-2021/ACM ICMR-2021, the Publicity Chair of IEEE ICME-2020/ICIP-2019, and the Area Chair of CVPR-2021/ECCV-2020/ICCV-2019. From 2016 to 2017, she was the APSIPA Distinguished Lecturer.

Haibin Ling received the B.S. and M.S. degrees from Peking University, Beijing China, in 1997 and 2000, respectively, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2006. From 2000 to 2001, he was an Assistant Researcher with Microsoft Research Asia, from 2006 to 2007, he was a Postdoctoral Scientist with the University of California, Los Angeles, Los Angeles, CA, USA, from 2007 to 2008, he was with Siemens Corporate Research as a Research Scientist, and from 2008 to 2019, he was a Faculty Member with the Department of Computer Sciences, Temple University, Philadelphia, PA, USA. In fall 2019, he joined the Department of Computer Science, Stony Brook University, Stony Brook, NY, USA, where he is currently a SUNY Empire Innovation Professor. His research interests include computer vision, augmented reality, medical image analysis, visual privacy protection, and human computer interaction. He was the recipient of the Best Student Paper Award of ACM UIST in 2003 and the NSF CAREER Award in 2014. He is an Associate Editor for the IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, and *Computer Vision and Image Understanding*. He was the Area Chair various times for CVPR and ECCV.

Yong Xu (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mathematics from Nanjing University, Nanjing, China, in 1993, 1996, and 1999, respectively. From 1999 to 2001, he was a Postdoctoral Research Fellow of computer science with the South China University of Technology, Guangzhou, China, where he became a Faculty Member and is currently a Professor with the School of Computer Science and Engineering. He is currently the Dean of Guangdong Big Data Analysis and Processing Engineering and Technology Research Center. His current research interests include computer vision, pattern recognition, image processing, and big data. He is a Senior Member of the IEEE Computer Society and the ACM. He was the recipient of the New Century Excellent Talent Program of MOE Award.