

# Recurrent Exposure Generation for Low-Light Face Detection

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

arXiv:2007.10963v1 [cs.CV] 21 Jul 2020

**Abstract**—Face detection from low-light images is challenging due to limited photos 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**—Low-light face detection, multi-exposure, gated recurrent networks

## I. INTRODUCTION

As the cornerstone for many face-related systems, face detection has been attracting long-lasting research attention [21], [24], [43], [52], [54]. It has extensive applications in human-centric analysis such as person re-identification [8], [20] and human parsing [13]. 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

Jinxiu Liang, Yuhui Quan, Tianyi Chen and Yong Xu are with School of Computer Science and Engineering at South China University of Technology, Guangzhou, China. (Email: cssherryliang@mail.scut.edu.cn; csyuhuiquan@scut.edu.cn; csttychen@mail.scut.edu.cn; yxu@scut.edu.cn)

Jingwen Wang is with Tencent AI Lab, Shenzhen, China. (Email: jaywong-jaywong@gmail.com)

Jiaying Liu is with Wangxuan Institute of Computer Technology, Peking University, Beijing, China. (Email: liujjiaying@pku.edu.cn)

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

Asterisk indicates the corresponding author.

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 [55] is recently constructed, which shows clear performance degradation of state-of-the-art face detectors. For example, DSFD [27] produces an mAP of 15.3%, in a sharp contrast to above 90% on the *hard* subset of the popular WIDER FACE [54] 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 [55]. 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 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 nontrivial [47], 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

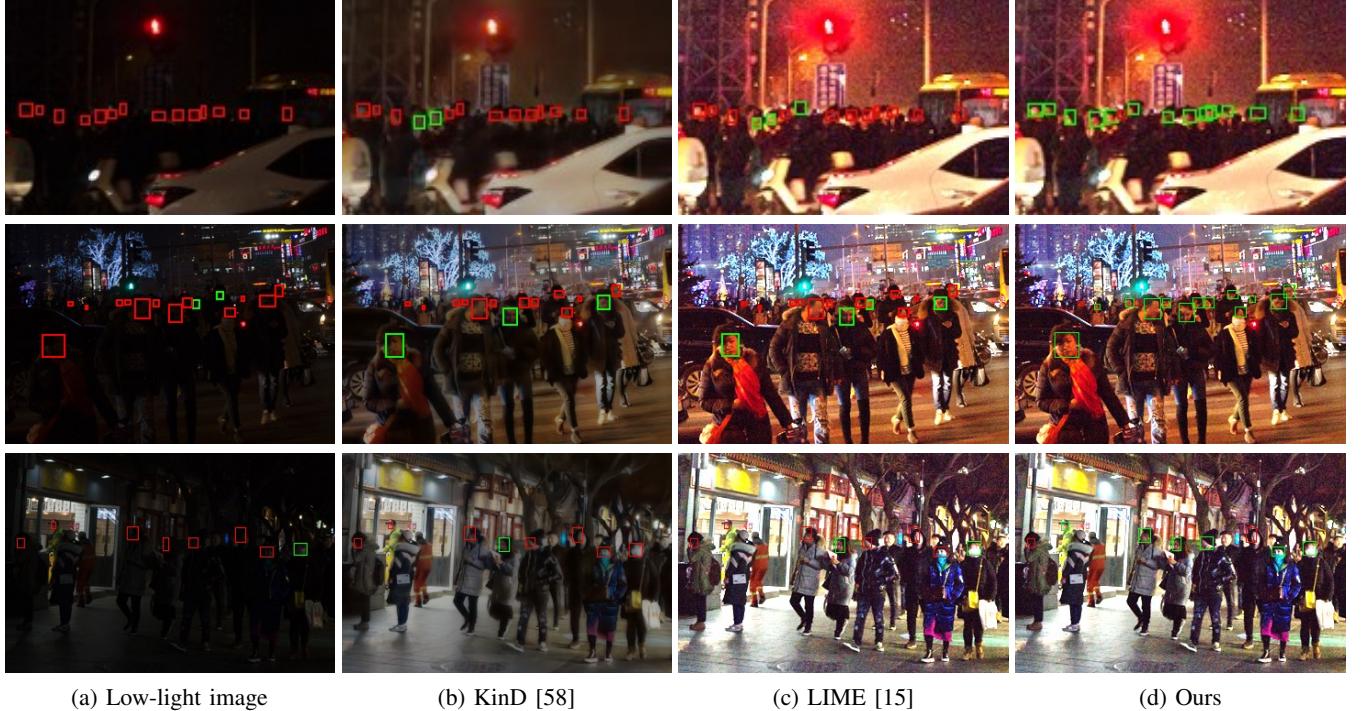


Fig. 1: Detection results of DSFD [27] on a low-light image (a) and its enhanced versions using KinD [58] (b) and LIME [15] (c). Green and red boxes indicate true positives and missed targets, respectively. It can be seen that the improvement brought by lighting enhancement is very limited. By contrast, our result in (d) (plotted on the same image of (c) for better visibility) show clear advantages.

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 [25] 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., [10], [11], [15], [28], [46]). 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., [45], [48], [50], [58]). However, the performance gain, when applied to low-light face detection, is still far from saturated [55]. 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 [56].

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 [21], [24], [54] and the proliferation of deep learning technologies [12], [30], [31], [37], face detection in unconstrained environments (a.k.a. ‘in the wild’) has made remarkable progress [17], [19], [27], [35], [36], [40], [42], [57]. Most recent technological developments have focused on robustness to geometric variance. Typical geometric distortion includes scale variation, deformation, occlusion and so on. For scale variation, researchers have proposed many effective strategies based on the idea of multi-scale analysis: designing image pyramids with different image scales [19], designing a pre-defined set of anchor boxes with different sizes and aspect ratios [?], [22], [36], detecting at different layers of the network [35], [57] and so on. Deformable part-based model improves deformation invariance by decomposing the task of face detection into detecting different facial parts [53]. The idea of face calibration is explored to obtain deformation invariance in [40]. 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 [41] and top-down feature fusion [27], [42].

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 [16], [26], [51]. 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 [34], [55], [59]. Illumination variation is known to be a major challenge for modern face detection algorithms [1], [59]. Pioneering approaches preprocess images by intensity mapping such as logarithmic transform [1] and gamma transform [39]. Photometric normalization is another commonly adopted method that counteracts the varying lighting conditions in hand-crafted feature [5], [51] and deep learning-based methods [31], [59]. Hand-crafted feature based methods derive the illumination invariance from various priors such as image differences or gradients [1], [16], while deep learning-based methods use random photometric distortions as augmentation to implicitly enhance the illumination invariance [27], [42], [57]. 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 [55]. 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 [55].

### 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 [29]. 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 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

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

where  $\mathcal{F}_\theta$  and  $\mathcal{G}_\omega$  denote the encoder and the decoder of the proposed module, respectively, with corresponding parameters  $\theta$  and  $\omega$ . 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.

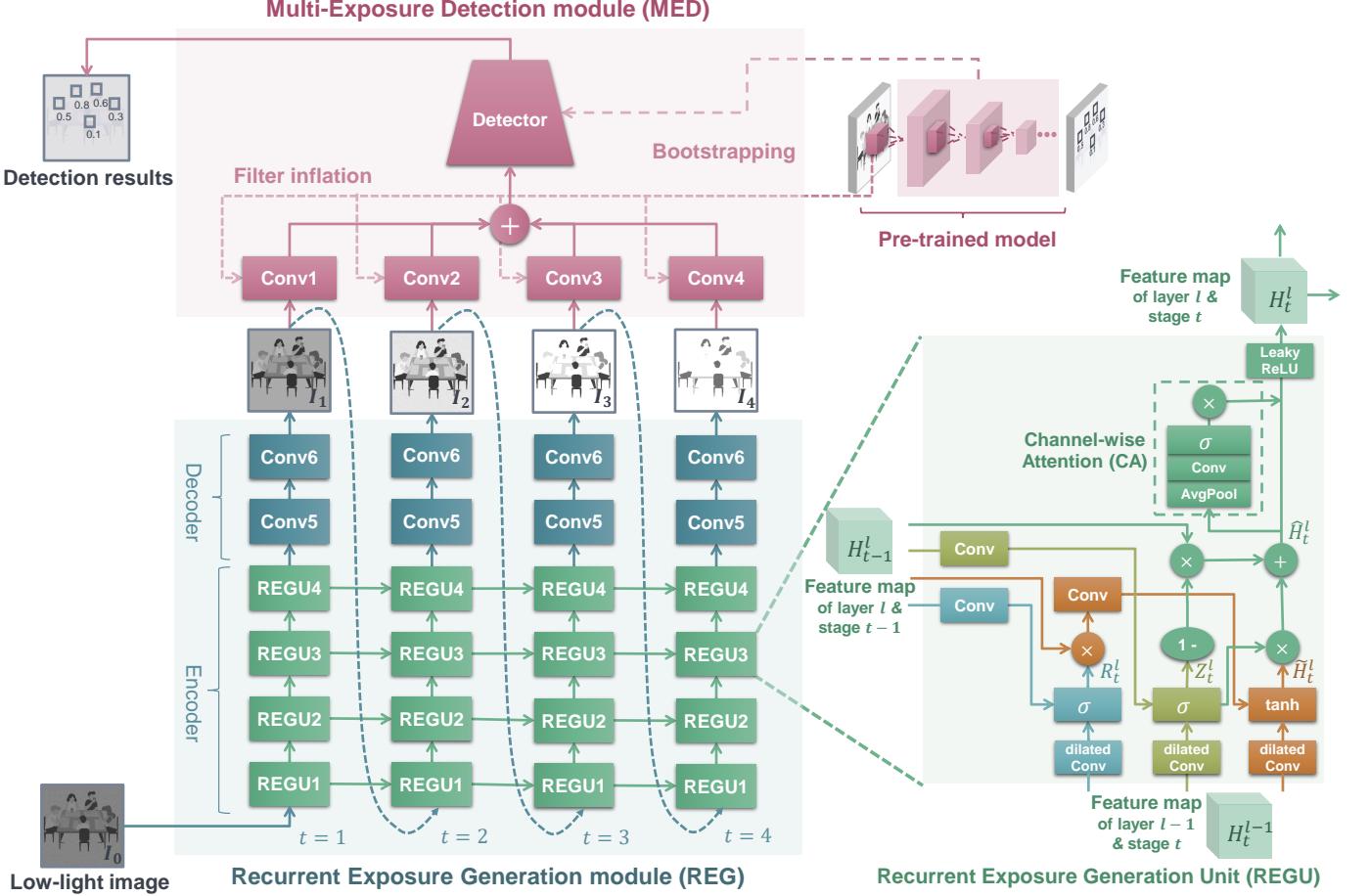


Fig. 2: The main framework of the proposed REGDet for low-light face detection.

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

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

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:

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

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

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

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

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

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.  $\xi$  denotes leaky ReLU [32] activation function

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

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) [44]  $\mathcal{A}^l$  can be computed as

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

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.

REGB 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 [44], 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 [32] 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 [56] that characterises 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 [38]. Once the parameters of CRF corresponding to a specific camera is known, the parameters of BTF can be estimated by solving the comparometric equation [33]. 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 [56] 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

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

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 [14]. Specifically, the exposure ratios are  $k, \dots, k^T$ , 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  $\ell_1$  norm and the Structure Similarity (SSIM) index [49] that reflects the difference on luminance and contrast, which is formulated as

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

and the SSIM meature is defined as

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

where means  $\mu$  and deviations  $\sigma$  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  $64 \times 64$  patches followed by random mirror, resize and rotation for data augmentation.

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 III.

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 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  $1, 2, \dots, N_a$  as

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

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 [31], we use weighted sum of the confidence loss and the localization loss:

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

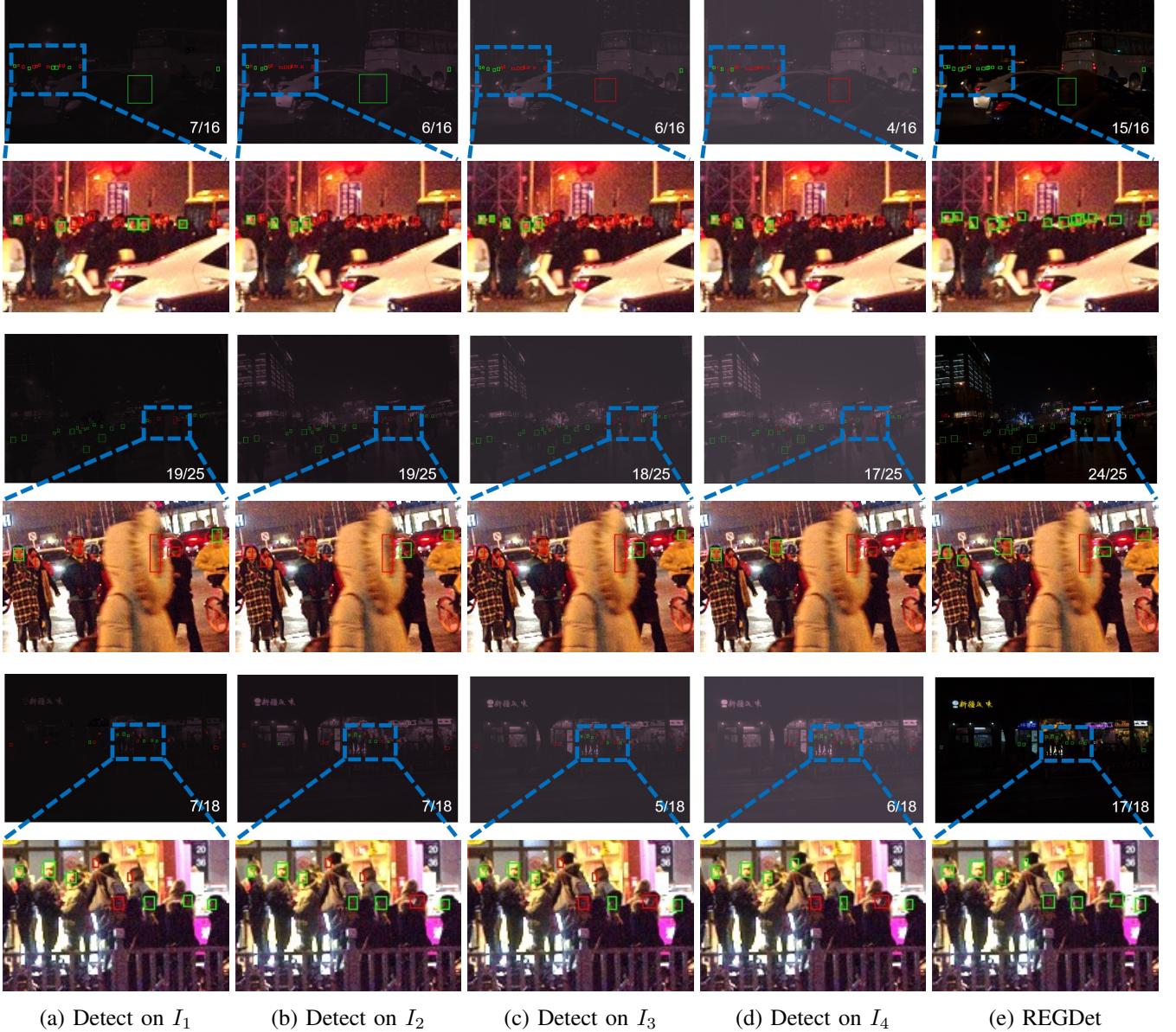


Fig. 3: The left four are detection results on intermediate  $I_1, I_2, I_3$  and  $I_4$  generated from the REG module, which show complementarity among the generated images, supporting the effectiveness of our proposed REG module. Note these ‘images’ are linearly normalized for visualization so that 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 [15] for better visibility.

where  $N_p$  denotes the number of positive anchors,  $\lambda$  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,

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

where the  $\hat{p}_i$  in the second term means that the localization loss is only calculated for those positive anchors. Following [12], the localization loss  $\mathcal{L}_{\text{loc}}(g_i, \hat{g}_i)$  is defined as the smooth  $\ell_1$  loss, *i.e.*, the distance between the predicted box  $g_i$  and the

ground-truth  $\hat{g}_i$  measured by Huber norm

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

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

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

The Huber norm is less sensitive to outliers than the  $\ell_2$  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 [27], PyramidBox [42] and S3FD [57] 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 [55] as our testbed. 6,000 real-world low-light images captured under extreme low-light environment. The resolution of the images is  $1080 \times 720$ . Totally 43,849 manually annotated faces are released. The annotated faces have large scale variance, ranging from  $1 \times 2$  to  $335 \times 296$ . There are usually 1 to 20 annotated faces in an image. Since the original test split [55] is withheld, we randomly leave 1000 images as our test set. Following prior work [27], [42], [57], face detection performance is measured by mean Average Precision (mAP), which is calculated as the area under precision-recall curve.

2) *Base detectors:* 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 [54] dataset. DSFD [27], PyramidBox [42] and S3FD [57], 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. As photometric augmentation has been a common practice in modern detectors, data augmentation related to exposure levels is used in the compared baselines.

3) *Implementation details:* Following [27], [31], [42], the batch size is 16 and multiple GPUs are used for speedup. The initial learning rate is 0.001, which is decreased by 0.1 at the 64-th and 96-th epoch. We adopt Adam [23] to train the REG module and SGD with momentum of 0.9 to train the MED module. Face anchor boxes that have over 0.35 IoU with the ground-truth annotated faces are labeled as positive anchors. The ratio between sampled negative anchors and positive anchors is fixed to 3 at each training iteration. For our proposed REGDet, we remove random photometric distortion in data augmentation as it has already involved an enhancement module. Note that we keep the photometric augmentation for the baselines following [27], [42], [57] for fair comparison. During inference, the image is first rescaled to make  $\sqrt{H} \times \sqrt{W} = 2000$ . Non-maximum suppression is applied with Jaccard overlap of 0.3 and the top 750 bounding boxes are kept.

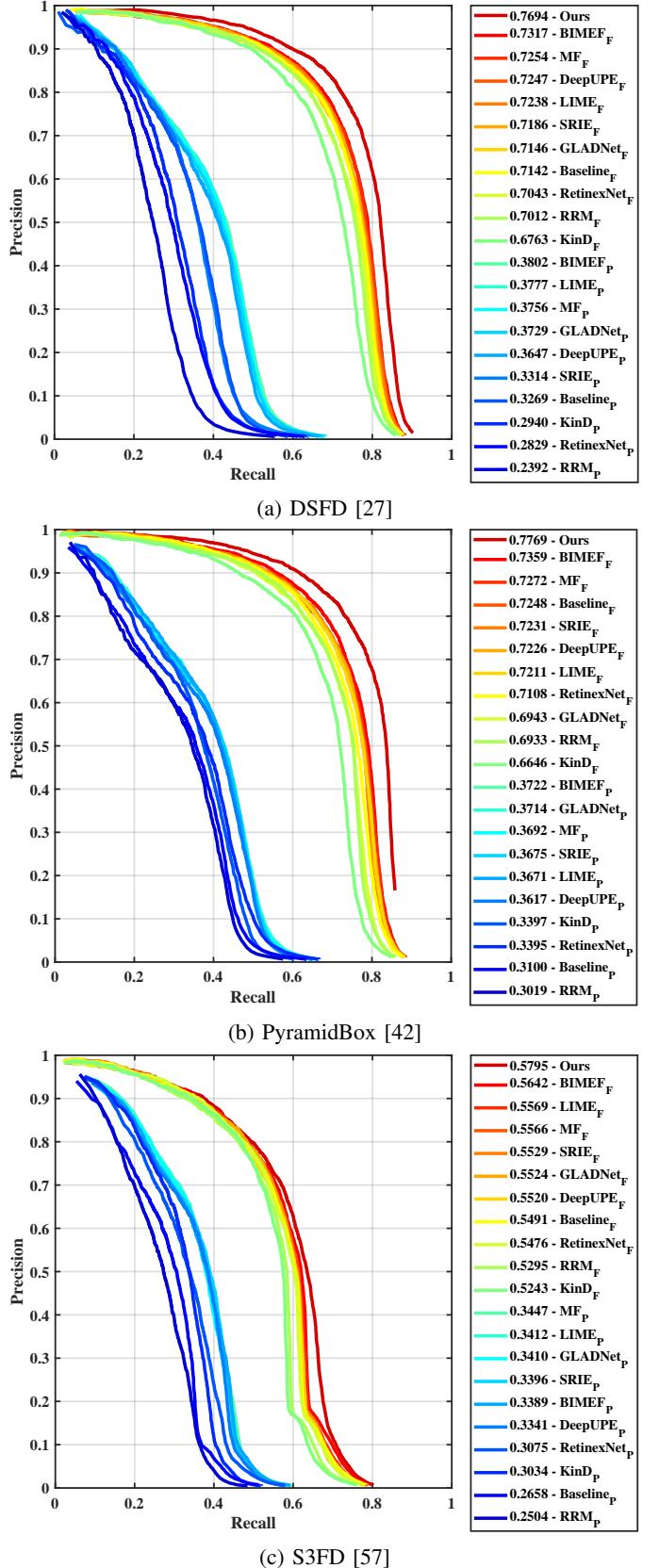


Fig. 4: Quantitative results of different approaches are shown. All the other approaches have both pre-trained version (marked with subscript ‘P’) and finetuned version (marked with subscript ‘F’) excepting for ours.

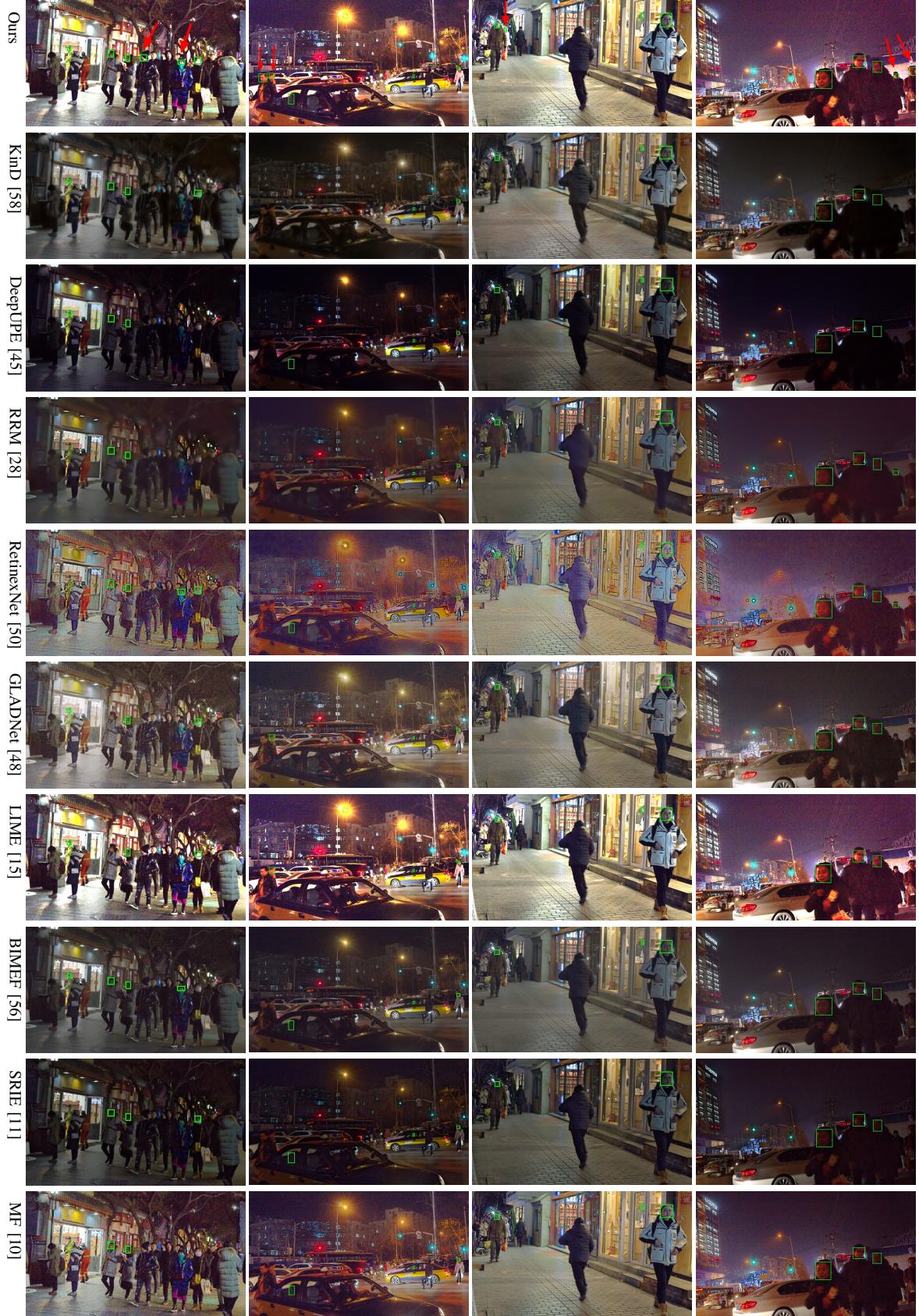


Fig. 5: Qualitative comparison of different methods. For better visualization, we draw the results of REGDet on images enhanced by LIME [15]. Red arrows indicate those faces that are challenging to be detected by the other methods. *Please Zoom in to see better.*

4) *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** [10], **SRIE** [11], **LIME** [15], **BIMEF** [56], **GLADNet** [48], **RetinexNet** [50], **RRM** [28], **DeepUPE** [45], and **KiND** [58] 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 [55] 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.*

## B. Result Analysis

The quantitative comparison of different approaches is shown in Fig. 4. 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 [27] or context assisted pyramid anchors [42]. 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, GLADNet, 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. 5). 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 [18], 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 I). 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. 5. 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) *Effectiveness of the recurrent architecture:* To examine the effectiveness of the proposed recurrent component, variant generation modules are designed as illustrated in Fig. 6, 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 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

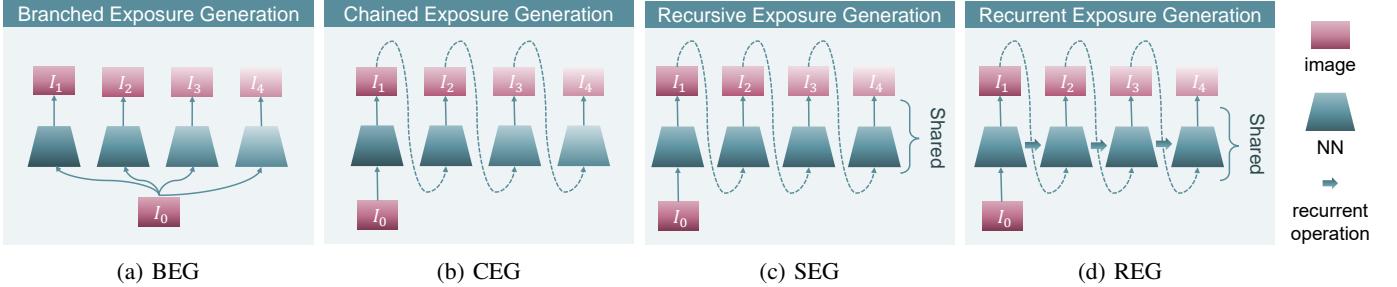


Fig. 6: Alternative pseudo-exposure generation modules.

TABLE I: Results of ablation study on the proposed REG module.

Method	DSFD [27]		PyramidBox [42]		S3FD [57]	
	#Params	mAP (%)	#Params	mAP (%)	#Params	mAP (%)
Finetuned 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	<b>76.94</b>	+ 0.12M	<b>77.69</b>	+ 0.12M	<b>57.95</b>

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) *Comparison of different numbers of stages*: We provide experimental comparison of different numbers of stages (denoted as  $T$ , Section III-A) using PyramidBox as base detector to support our choice of  $T = 4$ . The results corresponding to  $T = 1, 2, 4, 6$  are shown in Table II. 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 ( $T = 2, 4, 6$ ). 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. Setting  $T$  to a higher number (e.g., 6) does not bring more performance gain, meaning that generating too many pseudo exposures is unnecessary. Also, setting large  $T$  brings heavier computational cost.

TABLE II: Results of ablation study on the number of stages.

Numbers of stages	mAP (%)
$T = 1$	76.73
$T = 2$	77.16
$T = 4$	<b>77.69</b>
$T = 6$	77.63

3) *Effectiveness of the 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 compared in Table III. 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. As illustrated in Sec. III-B, the REG module is supervised and guided to generate images corresponding to diversified exposures with the designed pseudo-supervised pre-training. The collaborative and complementary information from different pseudo exposures can potentially be learnt from such a pre-training technique, which we believe is the key for better performance.

TABLE III: Ablation of the pseudo-supervised pre-training process for the REG module.

Setting	mAP (%)
Ours w/o pre-training	76.36
Ours w/ pre-training	<b>77.69</b>

## 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.



- 1–7, June 2008.
- [52] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(1):34–58, Jan. 2002.
  - [53] 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. (ICCV)*, pages 3676–3684, 2015.
  - [54] S. Yang, P. Luo, C.-C. Loy, and X. Tang. WIDER FACE: A face detection benchmark. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016.
  - [55] W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, and Z. Wang. UG2+ Track 2: A Collective Benchmark Effort for Evaluating and Advancing Image Understanding in Poor Visibility Environments. *arXiv:1904.04474 [cs]*, Apr. 2019.
  - [56] Z. Ying, G. Li, and W. Gao. A Bio-Inspired Multi-Exposure Fusion Framework for Low-light Image Enhancement. *arXiv:1711.00591 [cs]*, Nov. 2017.
  - [57] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3FD: Single Shot Scale-Invariant Face Detector. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 192–201, Oct. 2017.
  - [58] Y. Zhang, J. Zhang, and X. Guo. Kindling the Darkness: A Practical Low-light Image Enhancer. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*, May 2019.
  - [59] Y. Zhou, D. Liu, and T. Huang. Survey of Face Detection on Low-Quality Images. In *Proc. IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, pages 769–773, May 2018.