Unsupervised Face Detection in the Dark

Wenjing Wang, *Graduate Student Member, IEEE*, Xinhao Wang, Wenhan Yang, *Member, IEEE*, and Jiaying Liu, *Senior Member, IEEE*

Abstract—Low-light face detection is challenging but critical for real-world applications, such as nighttime autonomous driving and city surveillance. Current face detection models rely on extensive annotations and lack generality and flexibility. In this paper, we explore how to learn face detectors without low-light annotations. Fully exploiting existing normal light data, we propose adapting face detectors from normal light to low light. This task is difficult because the gap between brightness and darkness is too large and complicated at the object level and pixel level. Accordingly, the performance of current low-light enhancement or adaptation methods is unsatisfactory. To solve this problem, we propose a joint High-Low Adaptation (HLA) framework. We design bidirectional low-level adaptation and multitask high-level adaptation. For low-level, we enhance the dark images and degrade the normal-light images, making both domains move toward each other. For high-level, we combine context-based and contrastive learning to comprehensively close the features on different domains. Experiments show that our HLA-Face v2 model obtains superior low-light face detection performance even without the use of low-light annotations. Moreover, our adaptation scheme can be extended to a wide range of applications, such as improving supervised learning and generic object detection. Project publicly available at: https://daooshee.github.io/HLA-Face-v2-Website/.

Index Terms—Low-light, domain adaptation, illumination enhancement, high-level, low-level, face detection

1 Introduction

Face detection is one of the fundamental computer vision areas. It can facilitate various real-world applications, including but not limited to identity authentication, portrait beautification, and security monitoring. In recent decades, face detection technology [1], [2], [3], [4] has experienced rapid development and achieved remarkable results. However, finding and localizing faces under adverse illumination scenarios remains challenging. Underexposure leads to a series of visual degradations, including but not limited to unclear details, information loss, and color casting. These degradations not only degrade subjective visual quality but also deteriorate machine vision usability and robustness, posing potential risks to surveillance video analytics and auxiliary driving at night. For example, although dual shot face detector (DSFD) [4] achieves over 90% face detection precision on WIDER FACE [5], it can hardly detect faces covered in darkness, as shown in Fig. 1a.

Recent works have paid attention to the construction of face detection benchmarks under degraded conditions, e.g., DARK FACE [6] and the unconstrained face detection dataset (UFDD) [7]. DARK FACE is a pioneering large-scale low-light face detection dataset with both training/validation/testing sets, which provided rich resources for many low-light face detection studies [8] in recent years. The UFDD collects a new testing set of face images, including weather-based degradations, motion blur, focus blur and several others. Although these works fill in the blank in some sense, they still neglect some issues. First, existing datasets rely on extensive manual annotations. However, data collection and annotation can be difficult under extreme conditions, such as low light. Second, for real applications, there are still inevitable domain gaps due to different data collection environments and devices. Therefore, these datasets and methods can only support a limited scope of applications.

Different from existing research, in this paper, we explore how to adapt face detectors from normal light to low light without the help of low-light annotations. Since only low-light images in the given conditions need to be collected, compared with existing works [4], [9], our method can support more application scenarios. Our task is nontrivial. The challenges lie in the gaps at low and high levels, namely, pixel and feature levels. At the low level, the two domains have different pixel appearances in terms of illumination, color, and noise intensity, while at the high level, the two domains have different object semantics due to the difference between daytime and nighttime scenarios. Typical low-light enhancement methods [9], [10] aim at improving human visual quality and thus might fail to close the semantic gap, as shown in Fig. 1b. In contrast, traditional adaptation methods are mainly designed for tasks in which the scenes of source and target domains are similar, such as adapting from Cityscapes [11] to Foggy Cityscapes [12]. However, in our task, the domain gap is much more complex. Experimental results show that traditional adaptation methods are not effective enough for our task.

• *The authors are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China. E-mail: {daooshee, wxh0510, yangwenhan, liujiaying}@pku.edu.cn.*

Manuscript received 19 July 2021; revised 18 Dec. 2021; accepted 15 Feb. 2022. Date of publication 18 Feb. 2022; date of current version 5 Dec. 2022.
This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102702, in part by the National Natural Science Foundation of China under Grant 62172020, in part by State Key Laboratory of Media Convergence Production Technology and Systems, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).
(Corresponding author: Jiaying Liu.)
Recommended for acceptance by T. Hassner.
Digital Object Identifier no. 10.1109/TPAMI.2022.3152562

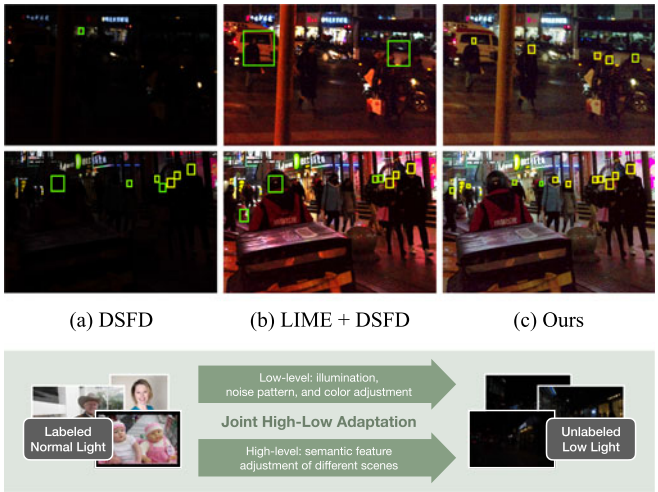


Fig. 1. Qualitative results and our adaptation paradigm for face detection in the dark. In comparison to DSFD [4] on original low light and the LIME [9]-enhanced versions, our model recognizes faces more accurately. Here, the color of the bounding boxes represents the confidence of recognition, with yellow indicating higher confidence.

To comprehensively narrow high-level and low-level gaps, we propose joint High-Low Adaptation (HLA). For low-level adaptation, existing methods apply unidirectional pixel-level brightening or darkening transformations, which are not powerful enough to solve our challenging task. We instead bidirectionally bring the two domains closer to each other. By enhancing low-light images and degrading normal light images with added noise and color casting, we build intermediate states, which serve as stepping-stones to cross the wide gap. Moreover, existing low-light enhancement methods mainly consider human vision rather than machine vision. In comparison, we design a low-light enhancement curve family, which is lightweight but more beneficial to downstream high-level tasks. For high-level adaptation, we jointly reduce the feature distance between the states constructed by low-level adaptation. We apply multitask self-supervised learning, consisting of inter- and intra-domain, context-based and contrastive learning. These learning strategies not only close high-level gaps but also further enhance the representation. With the proposed HLA scheme, our face detector achieves superior performance even though we do not use any annotation of low-light faces. Our contribution is threefold:

- Aiming at the problem of low-light face detection without low-light annotation, we propose a joint low-level and high-level adaptation scheme. Our model successfully transfers knowledge from normal light to the dark and surpasses state-of-the-art face detection and adaptation methods by a large margin.
- To fill the low-level gap, we propose bidirectional adaptation. By brightening dark images and degrading normal-light images, we make the two domains each take a step toward each other. Moreover, we design a new illumination adjustment deep curve targeting downstream high-level tasks.
- To fill the high-level gap, we design cross- and intra-domain self-supervised learning. We introduce a multitask scheme based on context-based pretext tasks and contrastive learning. Our adaptation not only narrows the gaps among multiple domains but also further strengthens the visual representation.

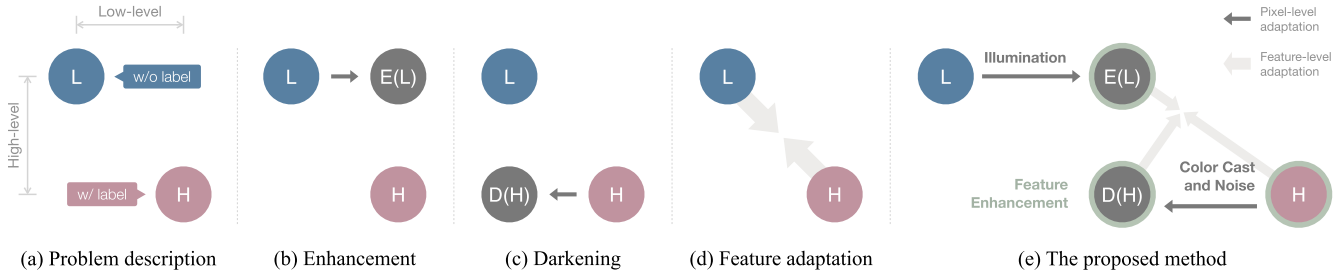
This paper is an extension of our earlier publication [13]. Our changes lie in the methodology and experiments. First, we design a new family of deep curves along with a new low-light enhancement network in Section 3.2. Our new low-light enhancement model not only achieves better detection performance but is also extremely lightweight and fast. Second, we replace the synthesis-based intermediate domain generation pipeline with augmentation transformation on the input images in Section 3.4. This modification simplifies the original complex training process and supports end-to-end training. To distinguish our model from HLA-Face in [13], we refer to our improved version as HLA-Face v2. With the new techniques, HLA-Face v2 outperforms HLA-Face by a large margin and narrows the gap between unsupervised learning and the ideal supervised learning upper bound to only 0.001 mAP. In addition to methodology improvement, the experiment has also been enriched. We provide more comparison results in Section 4.2, ablation studies in Section 4.3, and a variety of applications in Section 4.5.

The remainder of this paper is organized as follows. In Section 2, we review existing works in relevant areas. Then, in Section 3, we introduce the motivation and detailed designs of our method. Next, in Section 4, we demonstrate the effectiveness of our methods by experiments. Finally, in Section 5, we summarize the paper and discuss future directions.

2 Related Works

*Low-Light Enhancement.*Insufficient brightness is a common problem in image capture. Much research has been conducted on illumination enhancement. Early methods manually model illumination and design adjustment strategies. Histogram equalization and its variants [14] redistribute the intensities on the histogram. Regarding low light as dark haze, dehazing-based methods [15] first invert images and then apply dehazing techniques. The retinex algorithm decomposes images into two signals, a reflectance field and a shading field representing the scene lighting. On this basis, some methods [9], [16] first decompose reflectance and lighting and then process the two components separately or simultaneously. With the rapid development of deep learning in recent years, neural networks have been widely used for low-light enhancement. Some models process images in an end-to-end way [17], and some also introduce retinex theory [10], [18], [19]. A series of works further consider RAW images [20]. A systematic review and benchmarking can be found in [21].

Although existing low-light enhancement methods can improve human visual quality, these methods do not fully consider machine vision and downstream machine learning tasks. In this paper, we analyze the underlying mechanism between pixel-level restoration and high-level perception and propose corresponding solutions.



(a) Problem description (b) Enhancement (c) Darkening (d) Feature adaptation (e) The proposed method

Fig. 2. Principles of different adaptive low-light detection methods. L: low-light data. H: normal light data. Compared with existing methods (b)–(d), our method narrows low-level and high-level gaps by comprehensive learning constraints and therefore is more effective and powerful.

Face Detection. Traditional detectors rely on hand-crafted mechanisms, while recent models mainly learn features in a data-driven way. According to the detection framework, deep-based face detectors can be categorized into two types: one-stage and two-stage. One-stage models [22] directly predict the positions and recognition. Two-stage models [23], [24] instead separate the process of proposal generation and refinement. Compared with generic object detection, the challenges of face detection lie in the fact that the scale is often very diverse and various. To address this problem, many studies propose multiscale pyramids [25], [26] and various anchor sampling and matching strategies [27], [28], [29].

Although face detection is a popular high-profile research topic, existing studies seldom consider the scenario of insufficient illumination. In this paper, we not only design an effective dark face detector but also remove the dependence of low-light annotation.

Dark Object Detection. Despite the rapid development of normal light object detection, low-light objects have not received enough attention. For RAW short-exposure low-light images, Sasagawa *et al.* [30] proposed merging pretrained models in different domains with glue layers and generative knowledge distillation. Targeting RGB images, Loh *et al.* constructed the dataset ExDark [31], consisting of ten types of low-light images. Recently, a large-scale dataset DARK FACE [6] was built, providing an effective platform for the low-light face detection research field. Thereafter, a series of detectors came out in the UG^2 Prize Challenge.[1] However, these methods highly rely on annotations and therefore are not robust and flexible enough.

Unsupervised Domain Adaptation. Unsupervised domain adaptation (UDA) can be a direct solution to remove the dependency on labels. Typical methods are based on feature alignment [32], [33], [34], [35], adversarial learning [36], [37], [38], [39], pixel adaptation [40], [41], [42], and pseudo-labeling [43], [44]. For object detection, models usually consider not only the global domain gap but also the local domain gap [45], [46]. Although UDA achieves good performance for many applications, it is less effective in low-light face detection because of the large low-normal light domain gap. In this paper, we decompose the complex problem and design a superior solution.

Bidirectional adaptation is a common design in UDA. Existing methods [47], [48], [49] mainly introduce a pair of source-to-target and target-to-source mappings. They predict fake source and fake target data, which can work as intermediate domains to bridge the domain gap. Some methods [50] also use "bidirectional" to express that the pixel-level translation and the recognition model promote each other. Our bidirectional low-level adaptation differs from these methods. Instead of adopting source-to-target and target-to-source mappings, we decompose the low-light degradation into three components and introduce two new mapping directions of illumination restoration and noise & color distortion.

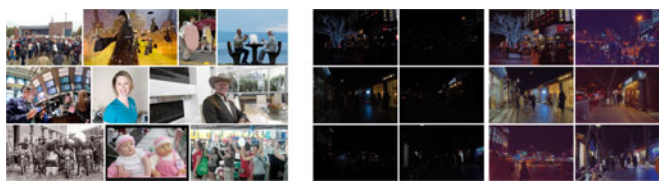
3 Joint High-Low Adaptation

This section first presents the motivation of our joint high-low adaptation and then introduces the detailed design.

3.1 Motivation

We hope to adapt face detection models from the annotated normal light domain H to the unannotated low-light domain L. Depending on the transfer route, we roughly categorize existing methods into three types: image enhancement, image darkening, and feature adaptation, as illustrated in Fig. 2. Image *enhancement*-based methods [51] first brighten the test images and then directly apply the pretrained normal-light face detector on the brightened images. Image *darkening*-based methods [52], [53], [54] first darken the normal light data and then retrain the face detector on the darkened images that have human annotation. Typical *feature adaptation* methods adopt alignment [35], adversarial learning [36], or pseudo labeling [42] to adapt the representation.

The challenge of low-light face detection lies in the complex gaps between H and L. As illustrated in Fig. 3, WIDER FACE [5] and DARK FACE [6] not only have different illumination and noise levels but also contain different content and semantic scenes (e.g., daytime scenes versus nighttime street views). Nevertheless, enhancement and darkening approaches can only handle pixel/signal-level gaps, while



(a) WIDER FACE (b) DARK FACE

Fig. 3. Representative samples from WIDER FACE and DARK FACE. We enhance DARK FACE for better visibility.

1. http://cvpr2020.ug2challenge.org/

feature adaptation tries to bridge the whole gap in one step, ignoring the usable distribution correspondence of low-level vision features.

To address these issues, we propose a two-step route to bridge the feature-level and pixel-level gaps jointly, namely, joint high-low adaptation (HLA). Our learning paradigm is shown in Fig. 2e. We construct low-level intermediate states between L and H and push the high-level features toward each other by comprehensive learning constraints. Specifically, we handle the low-level gap by enhancing L into $E(L)$ and darkening H into $D(H)$. In contrast to existing unidirectional translation, our bidirectional translation not only facilitates low-level adaptation but also supports high-level adaptation. Then, we bring the representations of the three domains, H, $E(L)$, and $D(H)$, closer by comprehensive learning constraints. To further improve the detection performance, we enhance the feature by representation learning. Our overall framework is shown in Fig. 6. In the following, we will introduce each of the proposed modules.

3.2 Enhancement Curve Family for High-Level Vision

Most of the popular deep-based low-light enhancement methods use neural networks to directly generate the pixel output. However, these approaches usually requires a powerful model with large parameters, leading to the risk of overfitting. Moreover, deep low-light enhancement models are often difficult to train. Existing methods either use paired data [55], adversarial learning [56], or retinex theory [19], which are of limited robustness and may result in visual artifacts. Recently, Li *et al*. proposed the curve-based model Zero-DCE [17], which is lightweight and does not rely on reference during training. However, it remains unexplored in [17] to explain the curve's properties and decide the optimal curve form. Moreover, not only Zero-DCE but also almost all existing low-light enhancement models aim at human vision rather than machine vision. Many methods keep noisy regions dark, enhance contrast, or apply local illumination adjustment, which improves subjective visual quality but might damage the high-level detection performance. In this paper, we will solve the abovementioned issues and propose a new enhancement model especially suitable for machine vision. Next, we will introduce a new form of curve family and the corresponding training strategy.

Curve Forms. Denote $f(\cdot,\alpha)$ as a family of enhancement functions, where $y = f(x,\alpha)$ represents enhancing $x \in [0,1]$ into $y \in [0,1]$ with a hyperparameter α. We require f to meet the following constraints:

I. $f(0,\alpha) = 0, f(1,\alpha) = 1, \forall\alpha$.
II. $f(\cdot,\alpha)$: $[0,1] \rightarrow [0,1]$ is monotonic and differentiable.
III. $\forall x_0 \in [0,1], \forall y_0 \in [0,1], \exists\alpha_0$ s.t. $f(x_0,\alpha_0) = y_0$.
IV. $\forall\alpha, \dfrac{\partial f(x,\alpha)}{\partial x} \neq 0$ holds for $\forall x \in (0,1)$.

Conditions I and II help improve visibility and maintain contrast. The first condition limits black and white to remain the same after enhancement, while the second keeps pixels in order. Condition III ensures that f can brighten the image. Intuitively, we want f to cover the entire $[0,1] \times [0,1]$ space as x and α change so that f can be flexible enough to handle various inputs. Condition IV prevents f from degenerating into a horizontal line; otherwise, the model can be difficult to train and tune.

Many functions can meet the above requirements. For simplicity, we only consider elementary functions, which have simple forms and are naturally differentiable. Considering Conditions I and II, for some elementary functions, given $0 = f(0,\alpha)$, $1 = f(1,\alpha)$, and the monotonic constraint, the formula can be directly determined. Take a quadratic function as an example. Defining $y = ax^2 + bx + c$, we have

$$\begin{cases} 0 = a \cdot 0 + b \cdot 0 + c, \\ 1 = a \cdot 1 + b \cdot 1 + c, \\ 2ax + b > 0, x \in [0,1]. \end{cases} \tag{1}$$

The solution is as follows:

$$\begin{cases} a \in [-1,1], \\ b = a - 1, \\ c = 0. \end{cases} \tag{2}$$

Therefore, the form of the quadratic enhancement curve is as follows:

$$y = x + ax(1-x), \quad a \in [-1,1]. \tag{3}$$

Note that Eq. (3) is exactly the form of the nonlinear curve mapping in Zero-DCE, indicating that Zero-DCE is only a special case of our proposed curve family.

For more complex functions, since the analytical solution can be hard to obtain, we instead first select a monotonic interval $[x_1, x_2]$ and then move and stretch the function to pass (0, 0) and (1, 1) by the following affine transformation:

$$\mathrm{AT}(f(x), x_1, x_2) = \frac{f((x_2 - x_1)x + x_1) - f(x_1)}{f(x_2) - f(x_1)}. \tag{4}$$

Finally, we obtain a collection of curves. Their formulas are given in Table 1.

Not all of these curves satisfy Condition III. As shown in Figs. 4b and 4d, 4e, 4f, 4g, quadratic, cubic, circle, sine, and arcsine curves cannot cover the entire $[0,1] \times [0,1]$ space. Accordingly, in Figs. 5c, 5d, 5e, 5f, and 5g, their enhancement results are still dim. Curves satisfying all the requirements, reciprocal, exponential, power, and arctangent functions, instead can substantially enhance the illumination, as shown in Figs. 5i, 5j, 5k, and 5l. Logarithm only covers half of the $[0,1] \times [0,1]$ space. Although the logarithmic curve may not be able to handle overexposure, it works well for restoring underexposure, as shown in Fig. 5h.

Zero-DCE [17] proposes to solve the limited representation scope of the quadratic curve by multiple iterations

$$f_n(x,\alpha) = f(f_{n-1}(x,\alpha),\alpha). \tag{5}$$

As shown in Fig. 4c, with the increase in iteration number n, the representation range expands. This feature explains why a larger n produces brighter results in [17]. However, more iterations lead to increased computation, more parameters, and difficulty in training and tuning. In comparison, the newly proposed logarithmic, reciprocal, exponential,

TABLE 1
The Forms of Our Low-Light Enhancement Curves

Name	Formula	Value Ranges	Req. 3	Req. 4
Tangent (Cotangent)	$y = \mathrm{AT}(\tan(x), x_1, x_2)$	$-\pi/2 < x_1 < x_2 < \pi/2$	✓	
Quadratic	$y = \alpha x^2 + (1-\alpha)x$	$\alpha \in [-1, 1]$		✓
Cubic	$y = \mathrm{AT}(x^3, x_1, x_2)$	$x_1 < x_2$		✓
Circle	$y = \mathrm{AT}(\sqrt{1-x^2}, x_1, x_2)$	$x_1, x_2 \in [0, 1]$		✓
Sine (Cosine)	$y = \mathrm{AT}(\sin(x), x_1, x_2)$	$-\pi/2 < x_1 < x_2 < \pi/2$		✓
Arcsine (Arccosine)	$y = \mathrm{AT}(\arcsin(x), x_1, x_2)$	$-1 < x_1 < x_2 < 1$		✓
Logarithmic	$y = \dfrac{\log(\alpha x + 1)}{\log(\alpha + 1)}$	$\alpha > 0$	Half	✓
Reciprocal	$y = \dfrac{(\alpha+1)x}{x+\alpha}$	$\alpha \in (-\infty, -1) \cup (0, +\infty)$	✓	✓
Exponential	$y = \dfrac{\alpha^x - 1}{\alpha - 1}$	$\alpha \in [0, 1) \cup (1, +\infty)$	✓	✓
Power	$y = \mathrm{AT}(x^\alpha, x_1, x_2)$	$0 < x_1 < x_2, \alpha > 0$	✓	✓
Arctangent (Arccotangent)	$y = \mathrm{AT}(\arctan(x), x_1, x_2)$	$x_1 < x_2$	✓	✓

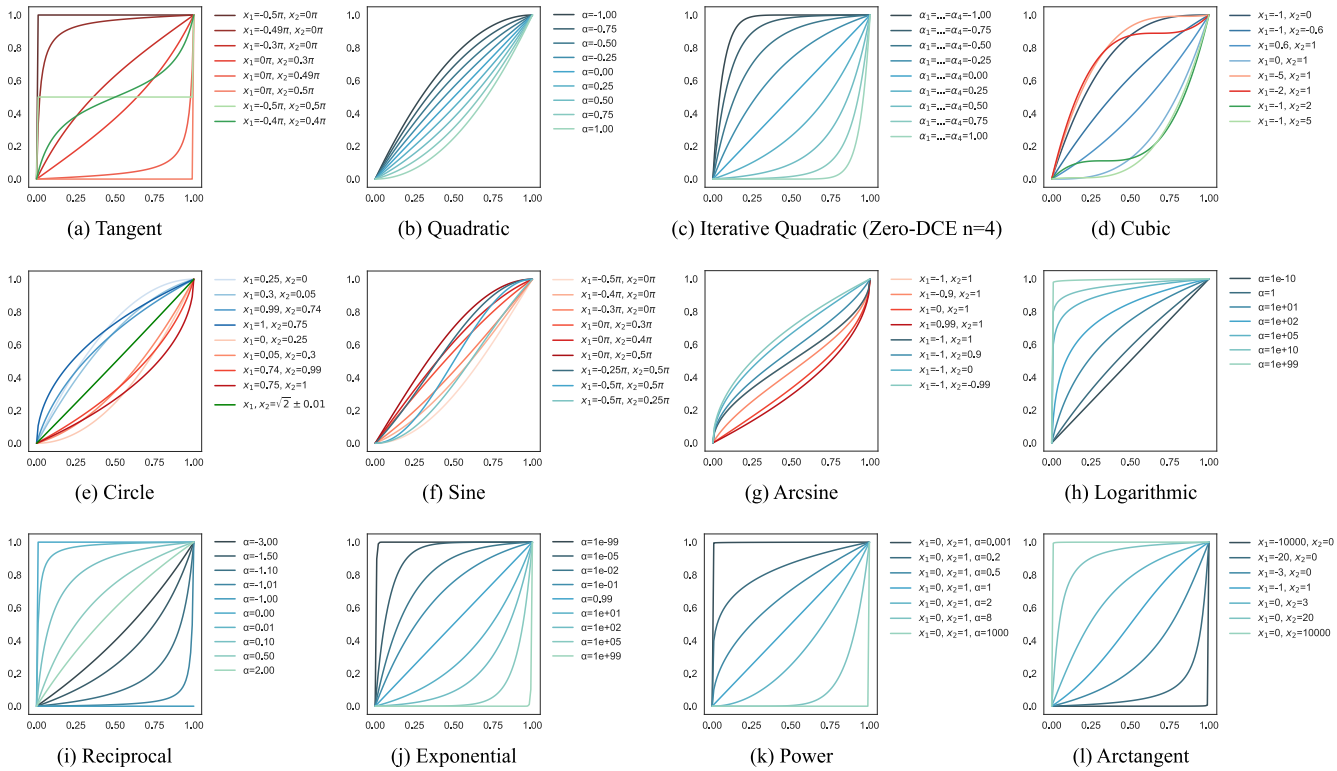


Fig. 4. Curves under different parameters. For each function, we use a variety of parameters to show its representation range. Among them, tangent has the risk of becoming a horizontal line; quadratic, cubic, circle, sine, and arcsine cannot cover the entire $[0, 1] \times [0, 1]$ space.

power, and arctangent curves can process images without iteration and thus are lightweight and robust.

Finally, we select the reciprocal function for simplicity. Logarithmic, exponential, power, and arctangent curves also have good low-light enhancement performance, which will be shown and further analyzed in Section 4.3.

Training Strategy. Given the low-light input L, we use a deep neural network to estimate α and generate the normal light prediction by $E(L) = f(L, \alpha)$. Following [17], the model is trained based on zero references. In comparison to common end-to-end or retinex-based low-light enhancement deep models, curved-based zero-reference learning is more capable of preserving the detailed structure of the input image, as it only needs to adjust the pixel signals instead of regenerating them.

Training involves three losses proposed in [17]. First, an exposure control loss teaches models to adjust the illumination

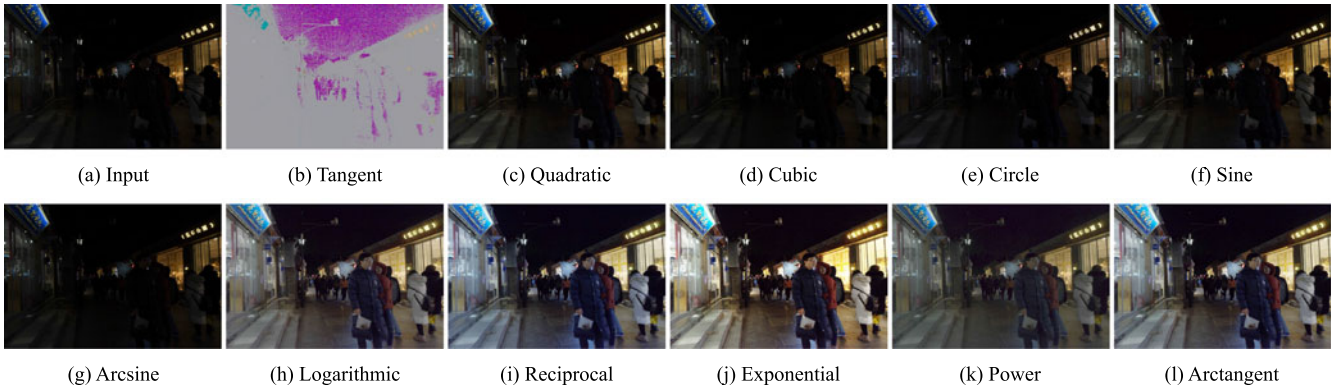


Fig. 5. Enhancement results by different curves. (b) Tangent curve suffers from severe distortion. (c)–(g) Quadratic, cubic, circle, sine, and arcsine curves fail to lighten the image. (h)–(l) Logarithmic, reciprocal, exponential, power, and arctangent curves successfully brighten the low-light image.

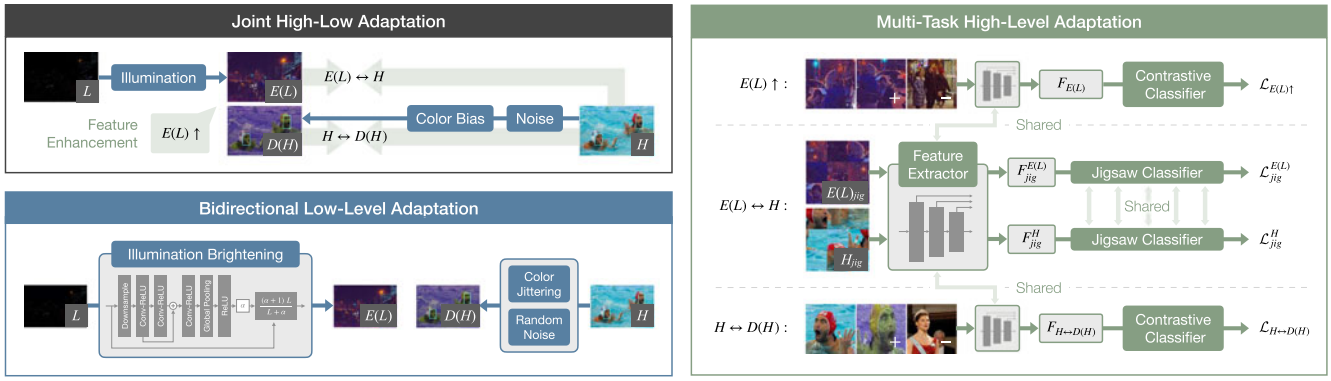


Fig. 6. Our joint high-low adaptation (HLA) for face detection under low-light conditions. We bidirectionally fill the low-level gap by building intermediate states and fill the high-level gap comprehensively through multitask cross-domain self-supervised learning.

$$\mathcal{L}_{exp} = |E(L) - e|, \tag{6}$$

where e controls the strength of enhancement. Then, a spatial consistency loss preserves the spatial continuity of neighboring regions

$$\mathcal{L}_{spa} = \sum_{j \in \Omega} (|E(L) - E(L)_j| - |L - L_j|)^2, \tag{7}$$

where Ω denotes four neighboring regions. Finally, a color constancy loss balances RGB color channels

$$\begin{aligned} \mathcal{L}_{color} = &(E(L)^R - E(L)^G)^2 + (E(L)^R - E(L)^B)^2 \\ &+ (E(L)^G - E(L)^B)^2. \end{aligned} \tag{8}$$

where R, G, and B represent the red, green, and blue color channels. For example, $E(L)^R$ is the red channel of $E(L)$. The final training objective is their combination

$$\mathcal{L}_{enh} = \lambda_{exp}\mathcal{L}_{exp} + \lambda_{spa}\mathcal{L}_{spa} + \lambda_{color}\mathcal{L}_{color}. \tag{9}$$

Our Condition IV comes from the fact that exposure control loss encourages curves to be horizontal. As shown in Fig. 4a, the tangent curve can be a horizontal line with $x_1 \to -1/2\pi$ and $x_2 \to 1/2\pi$. Accordingly, during training, it usually converges into $L(x) = 1/2$, leading to severe distortion, as shown in Fig. 5b.

Zero-DCE [17] differs from our model by considering nonuniform illumination and estimating pixelwise α. Although local illumination adjustment can improve subjective visual quality, it can also cause distortions and harm semantic information, degrading the performance of high-level vision models. We instead propose to make α spatially shared, namely, one α for the whole image. The benefits are twofold. First, our model can generate fewer artifacts, preserve more semantic information, and better benefit high-level tasks. Moreover, since we need to estimate fewer parameters, we can design a more lightweight model.

Network Architecture. Our model consists of 3 depthwise 32-channel separable convolutional layers. The output of the first layer is fed into the last layer by skip connection and fused by elementwise addition. After feature extraction, we use global average pooling to estimate α. The domain of α in the reciprocal function is not continuous. For training convenience, we abandon the negative $(-\infty, -1)$ part and add a rectified linear unit (ReLU) to clip α. Since we do not adjust illumination locally, we can downsample the image for acceleration. When estimating α, we resize the image to 16×16, which does not affect the performance. The overall network architecture is shown in Fig. 6.

Our model is extremely tiny and lightweight. Containing only 1.95 k trainable parameters, our model is more than 280-fold smaller than RetinexNet [19] (555.2 k). For computational complexity, our floating-point operations per second

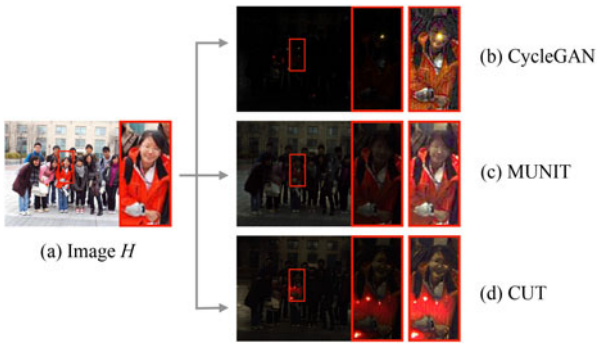


Fig. 7. Results of transferring from WIDER FACE to DARK FACE. The partial regions are enhanced for better visibility.

Fig. 8. Results of our bidirectionally brightening DARK FACE images and degrading WIDER FACE images.

(FLOPs) are only 0.0005 GMac for $1200 \times 900 \times 3$ images, which is 10^{-6} of RetinexNet (359 GMac).

3.3 Bidirectional Low-Level Adaptation

Combining the proposed low-light enhancement model, in this section, we explore how to narrow the low-level gap. The challenge is twofold. First, the coexistence of the semantic gap increases the difficulty of pixel-level transfer. We show some representative results for translating H into L in Fig. 7. Since DARK FACE contains many car lights and streetlamps, CycleGAN [57] generates yellow headlights on faces, while CUT [58] generates taillight-like artifacts on human bodies. MUNIT [59] instead fails to adjust the illumination of the image and generates results visually far from L. The other challenge lies in the difficulty of image restoration. Low-light brings heavy noise and color bias. Nevertheless, existing image restoration methods are not powerful enough to handle this heavy degradation.

Noting that low-light degradation can be approximated to be multifactorial and decomposable, we propose a bidirectional adaptation scheme. We roughly decompose the degradation of underexposure into three aspects: illumination, color, and noise. Although denoising and color restoration are difficult, reversely applying noise and color bias is easy. Therefore, we propose to brighten L into $E(L)$ and degrade H with added noise and color bias into $D(H)$, as shown in Fig. 8. By making the two domains each take a step toward each other, we fill the pixel-level gap at a lower cost. Meanwhile, by determining the specific forms of low-light degradation, the transfer model can be less affected by the semantics on the nighttime street view. Next, we will in turn introduce the detailed design of each process.

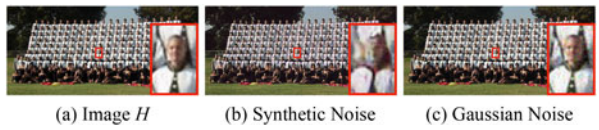


Fig. 9. Comparison of different strategies for adding noise. (b) Our previous publication [13]. (c) Our new strategy.

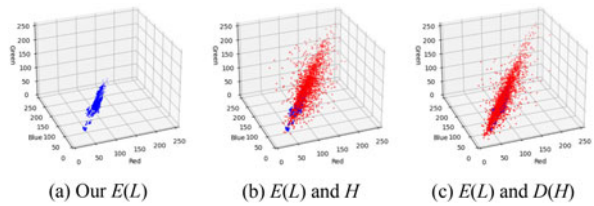


Fig. 10. The color distribution of different domains. Each spot represents the average pixel value of an image patch. We use blue to represent $E(L)$ samples, and red to represent H and $D(H)$ samples.

Brightening. Unlike most low-light enhancement models, we intend to only brighten the illumination and do not consider color bias and noise. We use the reciprocal curve proposed in Section 3.2, which adjusts illumination globally. The remaining pixel-level gap related to noise and color is modeled by the following $H \rightarrow D(H)$ process. Our enhancement may not have superior human visual effects but can better assist high-level tasks.

Noise Synthesis and Color Jittering. We divide the remaining pixel-level gap into color-dependent and color-independent. The former is simulated by random color jittering, while the latter is simulated by Gaussian noise.

In our previous publication [13], we try to exactly simulate low-light noise thus using a synthesis-based pipeline. Although adversarial learning is powerful for learning signal distribution. It introduces new artifacts and severely distorts small faces as shown in Fig. 9. The key problem is that $D(\cdot)$ does not need to strictly follow low-light degradations, it only needs to be the superset of low light. Our new $D(\cdot)$ not only covers the low-light degradation more comprehensively at a lower cost but is also easier to implement and supports end-to-end training.

We expect $D(H)$ to contain $E(L)$. Based on statistics, we set the mean and standard of Gaussian noise to $\mu = 0$ and $\sigma \sim \mathcal{U}[0, 16]$, respectively, and the jittering operation to hue: (0.8, 1.2), saturation & contrast: (0.6, 1.4), and brightness: (0.4, 1.2). The pixel value distribution of $E(L)$, H, and $D(H)$ is shown in Fig. 10. Compared with H, our $D(H)$ successfully covers the distribution of $E(L)$.

3.4 Multitask High-Level Adaptation

Based on the established intermediate domains, we conduct feature adaptation. Most existing feature adaptations are not sufficiently robust. Adversarial learning is not stable. Feature alignment [44], pixel adaptation [60], and pseudo labeling [42] cannot effectively address complex gaps. We instead use self-supervised learning, which is more stable and effective without the introduction of additional human supervision. By sharing self-supervised learning classifiers across different domains, the features are pushed into the same high-dimensional subspace; therefore, the high-level gap can be filled.

Specifically, for $E(L)$, H and $D(H)$, we narrow $E(L)$-H and H-$D(H)$, respectively. $E(L)$-H is closed by context-based self-supervised learning, while the gap H-$D(H)$ is filled by contrastive learning, which also improves the effectiveness of features. We also apply contrastive learning on the single $E(L)$ domain to learn more effective features. The whole adaptation fuses these learning methods in a multi-tasking way as shown in Fig. 6. Next, we will introduce each objective.

Filling E(L) and H. Self-supervised learning sets up pretext tasks with the labels generated by the images themselves. Through pretext tasks, the models can learn about shapes and semantics. Here, we use the jigsaw puzzles [61]. Rotation [62] or the combination of rotation and jigsaw are also adopted but are less effective than using jigsaw alone empirically. The reason may be that some photos in WIDER FACE are advertisements or paintings in which the faces have rare angles. Therefore, the rotation pretext task has ambiguous objectives on WIDER FACE. On the other hand, jigsaw puzzling can better teach models the concept of position, which is beneficial for detection.

Different from [61], we put 3×3 patches together into a single image and reduce the permutation number to 30. According to [63], this approach eases training and is sufficient when jigsaw plays the role of an auxiliary task. Denote p_{jig} as the ground truth permutation, $\mathcal{L}_c$ as cross-entropy loss, and F_{jig} as the extracted feature, the training objective for each domain is

$$\mathcal{L}_{jig}^{E(L)} = \mathcal{L}_c(F_{jig}^{E(L)}, p_{jig}^{E(L)}), \tag{10}$$

$$\mathcal{L}_{jig}^{H} = \mathcal{L}_c(F_{jig}^{H}, p_{jig}^{H}), \tag{11}$$

and the final training objective is

$$\mathcal{L}_{E(L)\leftrightarrow H} = \mathcal{L}_{jig}^{E(L)} + \mathcal{L}_{jig}^{H}. \tag{12}$$

To fill the high-level gaps, we share the permutation classification heads between $E(L)$ and H so that representations can be forcefully mapped into a shared space.

Filling H and D(H). The principle of contrastive learning is to distinguish between "positive" and "negative" samples. Denote the query as v, its positive pair as v^+ and negatives pairs as $v^- = \{v_1^-, v_2^- \ldots, v_N^-\}$; the contrastive loss is

$$\mathcal{L}_q(v, v^+, v^-) = -\log\left[\frac{\sigma(v, v^+)}{\sum_{n=1}^{N}\sigma(v, v_n^-) + \sigma(v, v^+)}\right]. \tag{13}$$

The similarity $\sigma(\cdot,\cdot)$ is often measured by taking the dot product

$$\sigma(x, y) = \exp(x \cdot y/\tau), \tag{14}$$

where τ is the temperature parameter [64]. Intuitively, contrastive learning is equivalent to a classification problem in which each image belongs to its own class. In our implementation, samples v and v^+ are one image under different views, while v^- are other images under some views. Following [65], the views are generated by a random data augmentation family. We use a combination of random resizing, cropping, color jittering, desaturation, Gaussian blur, and horizontal flip.

Contrastive learning can reduce the distance between positive samples. Exploiting this capability, we embed the distortion $D(\cdot)$ into contrastive learning

$$\begin{aligned}\tilde{\mathcal{L}}_{H\leftrightarrow D(H)}^{\text{cross}} = {} & \mathcal{L}_q(D(H), H^+, D(H)^-) \\ & + \mathcal{L}_q(H, D(H)^+, H^-),\end{aligned} \tag{15}$$

where H^+ and H^- represent the positive and negative pairs of H, $D(H)^+$ and $D(H)^-$ represent the positive and negative pairs of $D(H)$. In this way, the features of H and $D(H)$ can be pushed toward each other, and the high-level gap can be filled. Additionally, compared with traditional feature alignment strategies, contrastive learning can further enhance the representation.

We additionally apply single-domain contrastive learning on H and $D(H)$ to learn better representations

$$\begin{aligned}\tilde{\mathcal{L}}_{H\leftrightarrow D(H)}^{\text{single}} = {} & \mathcal{L}_q(D(H), D(H)^+, D(H)^-) \\ & + \mathcal{L}_q(H, H^+, H^-).\end{aligned} \tag{16}$$

To simplify the training process, we merge the above two losses by implementing $D(\cdot)$ in augmentation. Use $D^*(H)$ to represent a 50% probability of being H and the other 50% probability of being $D(H)$; the final objective is

$$\mathcal{L}_{H\leftrightarrow D(H)} = \mathcal{L}_q(D^*(H), D^*(H)^+, D^*(H)^-). \tag{17}$$

Moreover, we apply momentum contrast (MoCo) and follow [66] for other settings.

Improving E(L). To better adapt the face detector, we apply contrastive learning on the single $E(L)$ domain

$$\mathcal{L}_{E(L)\uparrow} = \mathcal{L}_q(E(L), E(L)^+, E(L)^-). \tag{18}$$

Final Objective. Finally, all the above learning objectives are combined in a multitask learning way. Denote $\mathcal{L}_{det}$ as the face detection objective on annotated normal light data, and the final training loss for our joint adaptation is

$$\begin{aligned}\mathcal{L} = {} & \lambda_{E(L)\leftrightarrow H}\mathcal{L}_{E(L)\leftrightarrow H} + \lambda_{H\leftrightarrow D(H)}\mathcal{L}_{H\leftrightarrow D(H)} \\ & + \lambda_{E(L)\uparrow}\mathcal{L}_{E(L)\uparrow} + \lambda_{det}\mathcal{L}_{det},\end{aligned} \tag{19}$$

where λ balances different learning objectives.

4 Experiments

In this section, we first describe the details of our implementation and then provide the experimental results.

4.1 Implementation Details

Network Architecture. Our framework is built based on DSFD [4]. DSFD adopts an extended VGG16 [67] backbone and extracts a 6-layer multiscale feature. Our self-supervised learning headers are placed after all six layers. The headers share the same architecture of Conv-Conv-FC. Please refer to the supplementary material for more details, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3152562.

Datasets. Experiments are mainly conducted on the normal light face detection dataset WIDER FACE [5] and the low-light face detection dataset DARK FACE [6]. We use their official splittings.

TABLE 2
Low-Light Face Detection Comparison Results

Category	Method	mAP (%)
Face Detection	Faster-RCNN [68]	1.7
	SSH [69]	6.9
	RetinaFace [3]	8.6
	SRN [70]	9.0
	SFA [2]	9.3
	PyramidBox [1]	14.0
	Small Hard Face [71]	16.1
	DSFD [4]	16.1
Enhancement (with DSFD)	SICE [18]	4.7
	RetinexNet [19]	12.0
	KinD [10]	15.8
	EnlightenGAN † [56]	20.8
	EnlightenGAN [56]	31.3
	Zero-DCE † [17]	37.3
	LIME [9]	40.7
	Zero-DCE++ [72]	40.9
	Zero-DCE [17]	41.3
	MF [16]	41.4
Darkening (with DSFD)	MUNIT [59]	29.7
	CycleGAN [57]	31.9
	CUT [58]	32.7
Unsupervised DA (with DSFD)	OSHOT [44]	25.4
	Progressive DA [60]	28.5
	Bidirectional DA [47]	33.7
	Pseudo Labeling [42]	35.1
Fully Supervised	Fine-tuned DSFD [4]	46.0
Ours	HLA-Face [13]	44.4
	HLA-Face v2	**45.9**

† *denotes retraining the deep-based methods with DARK FACE and WIDER FACE.*

Training Settings. First, we individually train the enhancement submodel E on the dataset in [17] and remain fixed during the remainder of the training.

We pretrain our complete framework on WIDER FACE using only the detection loss $\mathcal{L}_{det}$ following [4]. Then, we fine-tune it on WIDER FACE (images and labels) and DARK FACE (images only) with all the proposed loss functions. We use a batch size of 8 and an SGD optimizer. Training lasts for 70 k iterations. The learning rate is set to 1e-4 at first and then reduced to 1e-5 after 20 k iterations.

Evaluation Settings. For inference, we first enhance the image by the proposed deep reciprocal curve and then apply adapted DSFD following its original evaluation implementation. For performance measurement, we calculate the mean average precision (mAP) as the measure using the official evaluation tool.[2]

4.2 Comparison Results

We benchmark 24 state-of-the-art methods. For a comprehensive evaluation, the compared methods cover five categories: face detection, image enhancement, image darkening, feature adaptation, and fully supervised learning. The results are shown in Table 2 and Fig. 11.

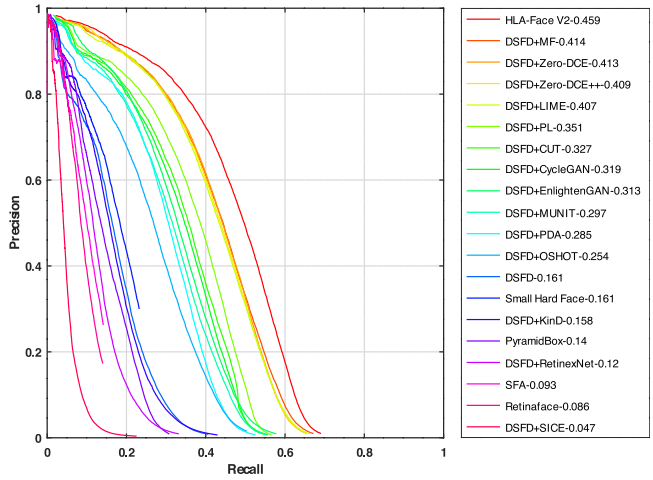


Fig. 11. Low-light face detection precision-recall (PR) curves.

TABLE 3
Benchmarking the Combination Performance of Different Face Detection and Low-Light Enhancement Methods

	PyramidBox [1]	Small Hard Face [71]	DSFD [4]	Average
Original	14.0	16.1	16.1	15.4
KinD [10]	15.6	16.2	15.8	15.9
EnlightenGAN [56]	28.5	29.3	31.3	29.7
LIME [9]	35.7	37.2	40.7	37.9
MF [16]	37.5	38.3	41.4	39.1
Zero-DCE [17]	35.9	37.7	41.3	38.3
Our $E(\cdot)$	**39.4**	**39.6**	**44.5**	**41.2**

Face Detection. We consider eight detection methods: one is for generic objects, and seven are designed specifically for faces. Affected by the low-light condition, the performance of these detectors is unsatisfactory, as shown in Table 2. Among them, Faster-RCNN[3] [68] (retrained on WIDER FACE) performs the worst. Other face detectors perform much better, but their mAPs are still lower than 20%. By the proposed adaptation scheme, our HLA-Face v2 greatly surpasses these approaches.

Enhancement. Here, we examine the scheme of first brightening then detection, i.e., Fig. 2b. As shown in Table 3, although Small Hard Faces [71] and DSFD [4] are comparable on low-light images, combining illumination enhancement, DSFD generally outperforms small hard faces. This result indicates that the DSFD has better robustness and generalization. Therefore, we choose DSFD as the face detection baseline in the following experiments.

Back to Table 2, we benchmark eight illumination adjustment methods. Although most of them can greatly improve the detection performance, some methods harm it. This is because these methods introduce too much visual distortion. As shown in Fig. 12, the result of SICE [18] is still dark. The results of RetinexNet [9] include a weird green color bias. KinD [10] instead overdenoises the images, resulting in dull color and blurry edges. Although the images become slightly brighter, the gap between the nighttime and

2. https://github.com/Ir1d/DARKFACE_eval_tools

3. https://github.com/playerkk/face-py-faster-rcnn

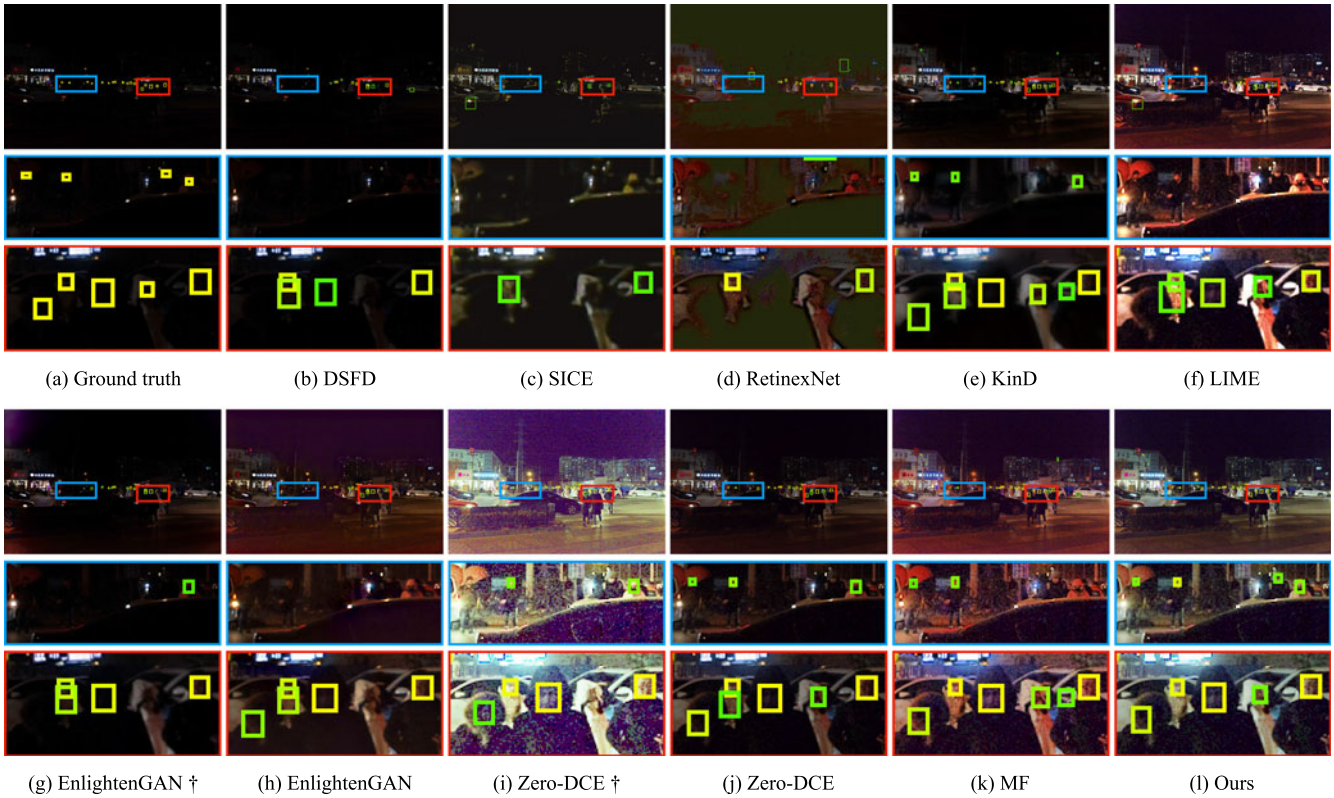


Fig. 12. Comparison against enhancement-based methods. (a) Input low-light image and ground truth bounding boxes. (b)-(k) Results of low-light enhancement methods with DSFD. (l) Our result. † denotes retraining the deep-based enhancement methods. The color of the bounding boxes represents the confidence of recognition, with yellow indicating higher confidence.

daytime photographs remains large, especially considering visual distortion. Therefore, these three low-light enhancement methods perform poorly.

LIME [9], EnlightenGAN [56], Zero-DCE [17] and MF [16] can help DSFD better recognize faces, as shown in Table 2. They also achieve better subjective visual quality. As shown in Figs. 12f, 12g, 12h, 12i, 12j, and 12k, faces originally buried in darkness can be seen more clearly. However, their performance is still relatively undesirable compared to our model. This is because when simply combining face detection with low-light enhancement, the semantic gap remains unsolved.

For deep-based EnlightenGAN and Zero-DCE, we also retrain them on the face detection datasets. Specifically, Zero-DCE is retrained with DARK FACE, while EnlightenGAN is retrained with DARK FACE and WIDER FACE. The training settings are the same as the original implementation. As shown in Table 2, the performance of face detection degrades after retraining. This is because compared with the original training datasets used by EnlightenGAN and Zero-DCE, DARK FACE suffers from more severe darkness and distortion, which disturbs the training process and misleads the enhancement model. We instead train our enhancement submodel E on clean images [17] and thus are not affected by noise or distortion.

Darkening. Next, we explore the effects of darkening-based schemes, i.e., Fig. 2c. We first translate WIDER FACE to DARK FACE and then use the synthetic dark version WIDER FACE to retrain the face detection model. Existing darkening-based adaptation methods [53], [73] mainly use CycleGAN [57]. We also evaluate more powerful image-to-image translation methods MUNIT [59] and CUT [58].

As shown in Table 2, all mAP scores of darkening-based adaptation are lower than 35%. This is because existing pixel-level translation models cannot fully simulate the characteristics of low-light, as shown in Fig. 7. The semantic gap between normal light and low-light cannot be narrowed by darkening and then retraining. We instead bridge the high-level and low-level gaps separately, therefore achieving better adaptation performance

Unsupervised Domain Adaptation. In the following, we evaluate UDA methods, i.e., Fig. 2d. To make the comparison fairly and avoid the impact of weak baseline methods, we reimplement all Faster-RCNN-based models with DSFD. OSHOT [44] applies the rotation angle prediction self-supervised learning scheme for one-shot domain adaptation. We extend it into training with the whole DARK FACE. As shown in Table 2, OSHOT does not deal well with helping the detection model transfer the knowledge from the bright condition to the dark one. The two-step Pseudo Labeling [42] first synthesizes artificial training data by CycleGAN and then fine-tunes the detector with the pseudo labels. In comparison to training on CycleGAN-synthetic dark WIDER FACE, the mAP improves from 31.9% to 35.1%. Nevertheless, the performance is still undesirable. Progressive DA [60] proposes joint adversarial appearance translation and knowledge transfer. However, the combination of low-level and high-level adversarial learning cannot fully narrow the domain gap. Based on CycleGAN, Bidirectional DA [47] jointly trains models on original and darkened normal light as well as original and

TABLE 4
Ablation Study Experimental Results

$E(\cdot)$	$E(L) \leftrightarrow H$	$H \leftrightarrow D(H)$	$E(L) \uparrow$	mAP (%)
-	-	-	-	15.3
✓	-	-	-	41.4
-	Rotation	-	-	22.7
-	Jigsaw	-	-	26.9
-	Rot + Jig	-	-	25.3
-	-	H only	-	18.6
-	-	✓	-	19.1
-	-	-	✓	20.2
-	-	✓	✓	20.4
-	Jigsaw	✓	✓	25.1
✓	Jigsaw	-	-	42.1
✓	Jigsaw	✓	✓	42.6
✓	Jigsaw	✓	✓	**45.9** †

† *denotes building image pyramids for multiscale testing.*

TABLE 5
Comparison of Different Curve Forms

Curve Form	mAP (%)
Original Zero-DCE [17]	38.3
Exponential	39.1
Power	38.8
Iterative Quadratic (n=8)	40.8
Logarithmic	40.9
Arctangent	41.1
Reciprocal	**41.6**

For convenience, we do not use the multiscale testing scheme in DSFD.

enhanced low light data. Feature-level adversarial loss and cross-domain consistency loss further assist the adaptation. This method is originally designed for dehazing. To implement it for face detection, we replace the dehazing-related unsupervised losses with jigsaw permutation, which has been proved to be effective for face detection in our paper. The mAP score of Bidirectional DA is 33.7%, which is better than training on CycleGAN-synthetic dark WIDER FACE. However, due to the instability of multi-level adversarial learning and the complexity of the framework, the training is not stable and the performance is not satisfactory. In comparison, we design a more comprehensive and tailored adaptation scheme for low-light conditions, thus achieving better results.

With Low-Light Annotations. Compared with fine-tuning DSFD using DARK FACE labels in the training, our model is only 0.001 mAP lower than the supervised learning method, demonstrating the effectiveness of our adaptation.

4.3 Ablation Studies

In this section, we validate and discuss each of our technical designs. The results are summarized in Table 4.

Effectiveness of E(·). Brightening the target image with our low-light enhancement submodel can improve the performance from 15.3% to 41.4% in mAP. As shown in Table 3, compared with other low-light enhancement methods, our $E(\cdot)$ best improves the performance of all three face detectors. This demonstrates the robustness and generalization of our deep reciprocal curve.

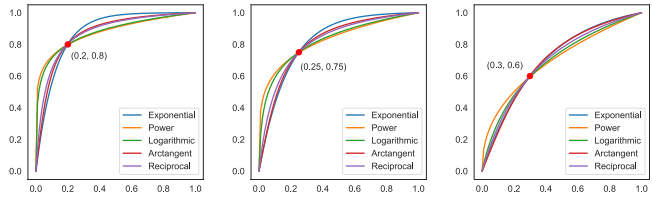


Fig. 13. Comparison of curve shapes. Given a point, we give possible solutions of different enhancement curves.

TABLE 6
Comparison of Full and Light Version Enhancement Curves

Method	mAP	FLOPs	Params	Time
Full version	41.6%	73 GMac	67.3 k	52 ms
Light version	41.4%	517 k Mac	1.95 k	0.9 ms

For convenience, we do not use the multiscale testing scheme in DSFD.

Next, we verify the effectiveness of each design in $E(\cdot)$. We first benchmark different curve forms without network simplification and acceleration. Here, we use a standard 7-layer CNN with symmetrical skip connections and estimate α without downsampling, which is the same as the backbone of Zero-DCE. The results are shown in Table 5.

As analyzed in Section 3.2, Zero-DCE is a special case of ours. Compared with Zero-DCE, i.e., iterative quadratic curve (n=8) with elementwise α, our iterative quadratic curve has better detection performance, which is consistent with our motivation to use spatially uniform α, i.e., a consistent α for all pixels of an image. The quadratic curve relies on iterative processing, which introduces more parameters and increases the risk of overfitting. Accordingly, in Table 5, its performance is slightly undesirable.

Exponential, power, logarithmic, arctangent, and reciprocal curves restore the illumination in one step. Among these curves, the exponential and power curves perform slightly worse, possibly because their growth rate is unstable and unbalanced. As shown in Fig. 13, the exponential curve tends to generate excessively high y-values for large x-values. The power curve grows too fast when x is small and becomes too flat when x is large. Accordingly, in Fig. 5j, the exponential curve overenhances the contrast, and in Fig. 5k, the result of the power curve has flat color and dull contrast. Arctangent and reciprocal forms instead have the most stable and balanced curve shapes in Fig. 13 and achieve the best performances in Table 5. The reciprocal curve slightly outperforms the arctangent curve, possibly because the reciprocal curve has a simpler form and thus is easier to train and tune.

The effect of model acceleration and simplification is shown in Table 6. We report the performance of first enhancing then detecting with DSFD (mAP), computational complexity (FLOPs), network parameters, and running time analysis for images of resolution $1200 \times 900 \times 3$. The experiment is conducted with an Intel i7-9700 K @3.60 GHz and GeForce GTX TITAN X. With a performance degradation of only 0.2%, we decrease the FLOPs to 1e-5 x, the number of parameters to $1/30$ x, and running time to $1/60$ x.

A more comprehensive comparison between our lightweight version curve and other deep enhancement methods can be found in Table 7. Our submodel not only achieves

TABLE 7
Comparison of Deep-Based Enhancement Methods on Detection Performance (mAP), Computational Complexity (FLOPs), Network Parameters, and Running Time Analysis

Method	mAP	FLOPs	Params	Time
RetinexNet [19]	12.0%	359 GMac	555.2 k	248 ms
EnlightenGAN [56]	31.3%	275 GMac	8.64 M	120 ms
Zero-DCE++ [72]	40.9%	0.08 GMac	10.56 k	6.3 ms
Zero-DCE [17]	41.3%	86 GMac	79.42 k	61 ms
Our submodel E	44.5%	517 k Mac	1.95 k	0.9 ms

better detection performance but is also smaller and runs faster.

Effectiveness of E(L) ↔ H. The results of different strategies for closing $E(L)$ and H are shown in Table 4. Jigsaw works better than rotation angle prediction and even their combination. This may be because the jigsaw pretext task better helps the detection model recognize positions.

Effectiveness of H ↔ D(H). Contrastive learning on the single H domain (H only) can improve the mAP from 15.3% to 18.6%. By introducing $D(H)$, the performance further improves by 0.5%, demonstrating the effectiveness of the proposed cross-domain contrastive learning scheme.

Effectiveness of E(L) ↑. Improving the feature on $E(L)$ increases the mAP score by 4.9%, showing that learning a good visual representation is vital for adaptive detection. The combination of $H \leftrightarrow D(H)$ and $E(L) \uparrow$ can increase the performance to 20.4% in mAP. Further introducing $E(L) \leftrightarrow H$ improves the mAP score to 25.1%. However, this score is slightly lower than using $E(L) \leftrightarrow H$ only. It may be that three learning tasks are too complicated for a single network. In contrast, with low-light enhancement $E(\cdot)$, introducing $H \leftrightarrow D(H)$ and $E(L) \uparrow$ to $E(L) \leftrightarrow H$ increases the mAP score from 42.1% to 42.6%. By narrowing the illumination gap between the source and target domains, our low light enhancement can assist different learning tasks to cooperate with each other.

Combination Effect. With all proposed modules, finally, the full version achieves the best result. DSFD applies a pyramid multiscale testing scheme, which leads to better detection results but also extensively increases the inference time from 1.25 hours to 10 hours on the DARK FACE test set. Even without multiscale testing, our HLA-Face v2 (42.6% in mAP) can still surpass all the compared methods in Table 2.

Failure Cases. Although our model greatly surpasses existing methods, it still has several limitations. As shown in Fig. 14, when people are facing away from the camera, our model may predict false positive faces on the back of their heads. These wrong faces may be deduced from human body contours. Our model may also recognize other bright circular objects as faces, such as car wheels. In addition, our model is not robust to window reflection and extremely small faces (less than 5×5 pixel), which is due to the limitation of the detection backbone. Another interesting failure case is shown in the last column of Fig. 14. Our model recognizes a face logo. However, this prediction is wrong because DARK FACE does not consider fake faces.

Fig. 14. Failure cases of false positive prediction, window reflection, extremely small faces, and fake faces. The color of the bounding boxes represents the confidence of recognition, with yellow indicating higher confidence.

(a) Input and GT (b) HLA-Face (c) HLA-Face v2

Fig. 15. Comparison against our earlier publication [13]. The color of the bounding boxes represents the confidence of recognition, with yellow indicating higher confidence.

4.4 Comparison Against the Earlier Publication

Compared with our earlier publication HLA-Face [13], our HLA-Face v2 further improves the unsupervised detection performance from 44.4% to 45.9%. As shown in Fig. 15, HLA-Face overbrightens the images and brings severe noise. HLA-Face v2 is less affected by illumination and noise and therefore has more accurate face prediction.

Moreover, our newly proposed reciprocal curve is more powerful and lightweight than the enhancement submodel in HLA-Face. With our new designs, the performance of first brightening and then detection improves from 39.1% to 41.4% without a pyramid multiscale testing scheme in DSFD. For model size and computational complexity, the enhancement submodel in [13] has 214.13k parameters, and the FLOPs for $1200 \times 900 \times 3$ images are 231.67 GMac. In

TABLE 8
Normal Light Face Detection Comparison Results on WIDER FACE

	Easy	Medium	Hard
DSFD [4]	94.6%	93.7%	88.0%
HLA-Face [13]	95.0%	93.9%	88.3%
HLA-Face v2	**95.2%**	**94.1%**	**88.8%**

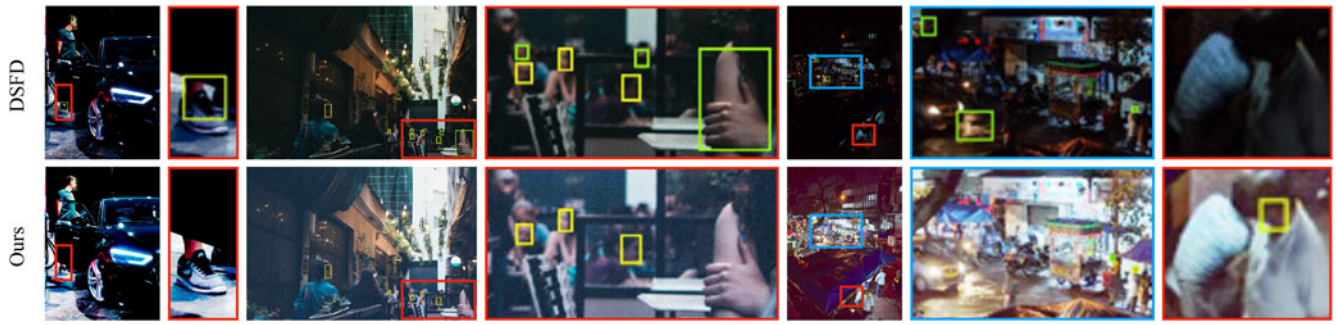


Fig. 16. Comparison on real-world cases. First row: input images and DSFD detection results. Second row: our enhancement and detection results. The color of the bounding boxes represents the confidence of recognition, with yellow indicating higher confidence.

TABLE 9
Comparison on Generic Object Detection, Classification, and Semantic Segmentation

$E(\cdot)$	$E(L) \leftrightarrow H$	$H \leftrightarrow D(H)$	$E(L) \uparrow$	ExDark			CODaN	Dark Zurich
				AP	AP_{50}	AP_{75}	Top-1 Acc.	mIoU
-	-	-	-	29.3%	59.8%	24.5%	46.9%	17.1%
✓	-	-	-	29.7%	59.5%	24.9%	56.2%	20.6%
-	✓	-	-	29.8%	60.7%	25.6%	55.6%	19.0%
-	-	✓	-	29.4%	59.5%	25.2%	52.3%	20.5%
-	-	-	✓	29.5%	59.3%	25.7%	54.7%	19.7%
✓	✓	✓	✓	30.4%	61.1%	26.2%	60.7%	23.1%

We adapt Faster-RCNN from COCO to ExDark, ResNet from daytime to nighttime on CODaN, and RefineNet from Cityscapes to Dark Zurich.

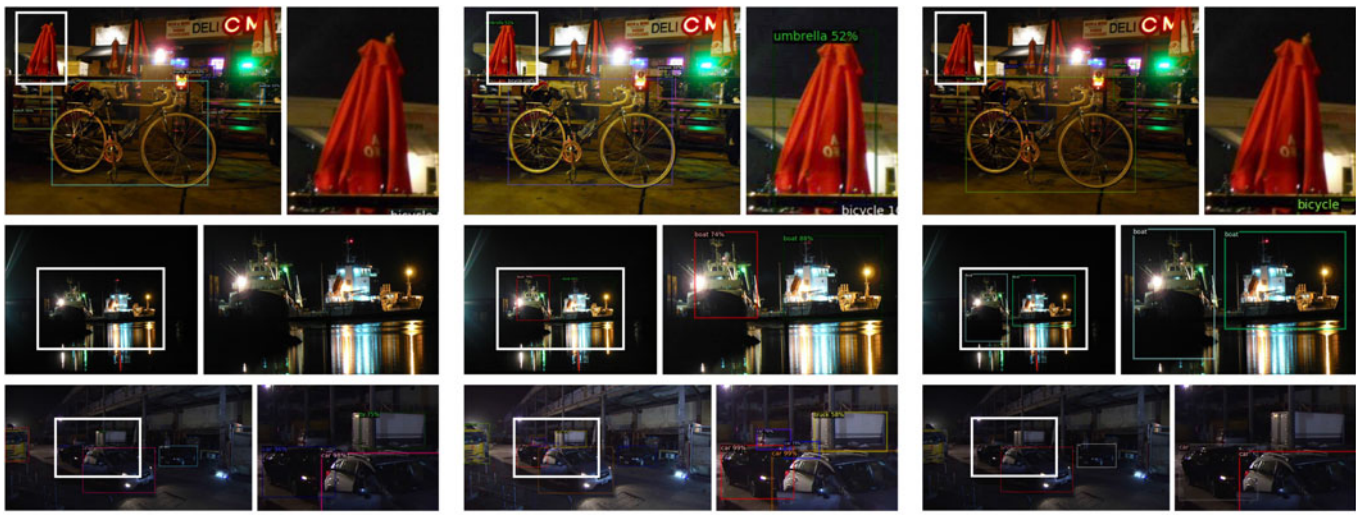
comparison, the newly proposed reciprocal curve only has 1.95 k parameters and 517 k Mac FLOPs.

4.5 Applications

In this section, we show our generalization by four applications: detecting normal light images, handling real-world cases, improving supervised learning, and adapting generic object detection models.

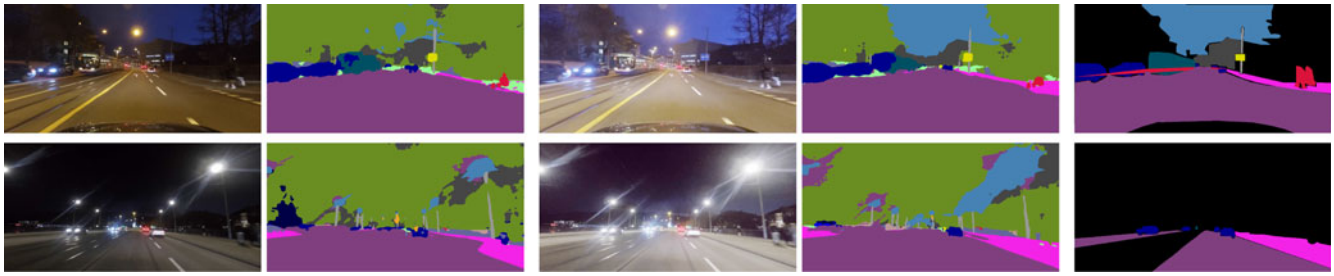
Performance on WIDER FACE. Our model can detect normal light faces as well. To avoid overexposure, we skip $E(\cdot)$ when the average pixel value of the image is higher than 45, which does not affect our original detection on DARK FACE. As shown in Table 8, compared with DSFD [4], HLA-Face v2 performs better on WIDER FACE for easy, medium, and hard faces, indicating that our joint high-low adaptation can improve the robustness and generalization of the model. Our model also outperforms HLA-Face [13], verifying the effectiveness of our newly proposed low-level enhancement model and high-level adaptation schemes.

Real World Cases. The results on in-the-wild images are shown in Fig. 16. The original DSFD [4] can be easily affected by illumination and incorrectly recognizes shoes, arms, car lights, and signboards as faces. Our HLA-Face v2 detects faces more accurately.



(a) Original COCO-pretrained Faster-RCNN (b) With our joint High-Low Adaptation (c) Ground Truth

Fig. 17. Comparison results of assisting generic object detection on low-light images without low-light annotation. Images are from the testing set of the ExDark dataset. Our adaptation can improve the detection accuracy and sometimes even find objects missed in the ground truth.



(a) Original Cityscapes-pretrained RefineNet (b) With our joint High-Low Adaptation (c) Ground Truth

Fig. 18. Adapting street scene semantic segmentation to nighttime without low-light annotation. With our adaptation, the pixels are better categorized and the predicted contours are clearer.

Improving Supervised Learning. Our unsupervised adaptation scheme can also help with supervised learning. For fine-tuning DSFD with labels, combining our adaptation schemes, the mAP is improved from 46.0% to 48.1%.

Generic Object Detection. The proposed joint high-low adaptation can also be extended to other tasks. For example, the AP of COCO-pretrained [74] Faster-RCNN [68] on ExDark [31] is 29.3%, as shown in Table 9. With our adaptation designs introduced, the performance improves, and the final full version achieves the highest AP of 30.4%, demonstrating the generalization of our method.

Moreover, with our adaptation, Faster-RCNN can detect objects missed in the ground truth, such as the umbrella in the first row and the cars and trucks in the third row in Fig. 17. This capability demonstrates that our adaptation can help reduce the burden of creating low-light annotations.

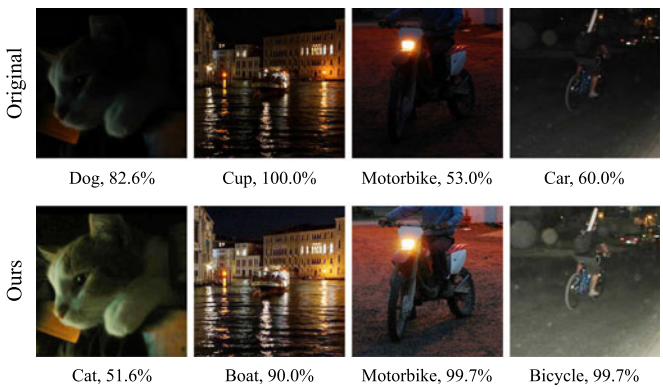


Fig. 19. Results of unsupervised low-light image adaptive classification. Below each image, we show the predicted category and its confidence. Our model can correct wrong predictions and increase the confidence of correct predictions.

(a) DSFD (b) HLA-Face v2 (c) w/ One-shot

Fig. 20. Result of adapting HLA-Face v2 to the given test image by one-shot fine-tuning.

Classification. In the above experiments, we explore various detection tasks. Now we further extend our framework to more vision tasks. First, we analyze how our framework performs for day-to-night classification adaptation based on a new dataset CODaN [75]. As shown in Table 9, the top-1 accuracy of daytime-pretrained ResNet-18 [76] is only 46.9%, which can be increased to 60.7% with our adaptation. Ablation study demonstrates the effectiveness of each of our technical designs.

As shown in Fig. 19, our adaptation can greatly improve visibility, correct misclassified results, and increase the confidence of correct results.

Semantic Segmentation. Nighttime street scene segmentation is a vital technique for autonomous driving. Our framework can also adapt RefineNet [77] from normal light Cityscapes [11] to low-light Dark Zurich [78]. As shown in Table 9, the mean intersection over union (mIoU) score is improved from 17.1% to 23.1%. More subjective results are shown in Fig. 18.

One interesting result is that, for object detection and image classification, the jigsaw permutation task is often more powerful than cross-domain and single-domain contrastive learning. However, for semantic segmentation, contrastive learning works better in day-to-night adaptation. It may be that segmentation is a pixel-level classification task, and contrastive learning can better guide the model to extract finegrained features.

5 Conclusion and Future Work

To fully exploit normal light annotation and reduce the burden of obtaining extra low-light annotation, we design a joint high-low adaptation (HLA) framework. Specifically, we introduce reciprocal-curve-based illumination adjustment, bidirectional pixel-level translation, and representation adaptation based on multitask self-supervised learning. Qualitative and quantitative experiments support our designs, show the superiority of our model, and demonstrate the potential of joint high- and low-level adaptation. Our work can inspire related new research on low-light enhancement, high-level vision under low-level visual quality, and feature adaptation.

Although we remove the dependency on low-light annotations, the proposed HLA still requires many low-light images. However, in real applications, it is sometimes even quite difficult to collect many images. There are 100 degradation types for 100 images on the internet. A suitable and

large dataset with thousands of images cannot be collected for each degradation. A naive solution would be one-shot fine-tuning. In Fig. 20, we show a case in which DSFD [4] and our HLA-Face v2 mistakenly recognize the rightmost man's hands or clothes as faces. By mixing the given test image into the L domain and fine-tuning for 100 iterations, our one-shot HLA-Face v2 produces more accurate results. In the future, we will explore faster and more reliable adaptation strategies.

Another direction for future work is extending HLA to other high-level tasks. We have demonstrated our generalization on generic object detection in Section 4.5. In the future, we will explore more tasks, such as low-light action recognition [79] and nighttime surveillance analytics [80].

REFERENCES

[1] 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.

[2] S. Luo, X. Li, R. Zhu, and X. Zhang, "SFA: Small faces attention face detector," *IEEE Access*, vol. 7, pp. 171 609–171 620, 2019.

[3] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," 2019, *arXiv: 1905.00641*.

[4] J. Li *et al.*, "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5055–5064.

[5] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.

[6] 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.

[7] 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 9th Int. Conf. Biometrics Theory, Appl. Syst.*, 2018, pp. 1–10.

[8] J. Liang *et al.*, "Recurrent exposure generation for low-light face detection," *IEEE Trans. Multimedia*, early access, Mar. 25, 2021, doi: 10.1109/TMM.2021.3068840.

[9] 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.

[10] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1632–1640.

[11] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[12] C. Sakaridis, D. Dai, and L. V. Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, 2018.

[13] W. Wang, W. Yang, and J. Liu, "HLA-face: Joint high-low adaptation for low light face detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16190–16199.

[14] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," in *Proc. 1st Conf. Vis. Biomed. Comput.*, 1990, pp. 337–345.

[15] X. Dong *et al.*, "Fast efficient algorithm for enhancement of low lighting video," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2011, pp. 1–6.

[16] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. W. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Process.*, vol. 129, pp. 82–96, 2016.

[17] C. Guo *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1777–1786.

[18] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[19] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 155.

[20] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3291–3300.

[21] J. Liu, D. Xu, W. Yang, M. Fan, and H. Huang, "Benchmarking low-light image enhancement and beyond," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1153–1184, 2021.

[22] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[23] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[24] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

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

[26] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[27] C. Zhu, R. Tao, K. Luu, and M. Savvides, "Seeing small faces from robust anchor's perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5127–5136.

[28] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3FD: Single shot scale-invariant face detector," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 192–201.

[29] X. Ming, F. Wei, T. Zhang, D. Chen, and F. Wen, "Group sampling for scale invariant face detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3441–3451.

[30] Y. Sasagawa and H. Nagahara, "YOLO in the dark - domain adaptation method for merging multiple models," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 345–359.

[31] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. Image Understanding*, vol. 178, pp. 30–42, 2019.

[32] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[33] S. Ben-David , J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1/2, pp. 151–175, 2010.

[34] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[35] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.

[36] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 59:1–59:35, 2016.

[37] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2962–2971.

[38] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12448–12457.

[39] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6091–6100.

[40] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: Symmetric bi-directional adaptive GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8099–8108.

[41] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[42] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5001–5009.
[43] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 480–490.
[44] A. D'Innocente, F. C. Borlino, S. Bucci, B. Caputo, and T. Tommasi, "One-shot unsupervised cross-domain detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 732–748.
[45] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3339–3348.
[46] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6949–6958.
[47] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain adaptation for image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2805–2814.
[48] G. Yang, H. Xia, M. Ding, and Z. Ding, "Bi-directional generation for unsupervised domain adaptation," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 6615–6622.
[49] P. Jiang, A. Wu, Y. Han, Y. Shao, M. Qi, and B. Li, "Bidirectional adversarial training for semi-supervised domain adaptation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, Art. no. 130.
[50] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6929–6938.
[51] S. W. Cho, N. R. Baek, J. H. Koo, M. Arsalan, and K. R. Park, "Semantic segmentation with low light images by modified CycleGAN-based image enhancement," *IEEE Access*, vol. 8, pp. 93 561–93 585, 2020.
[52] H. Lee, M. Ra, and W. Kim, "Nighttime data augmentation using GAN for improving blind-spot detection," *IEEE Access*, vol. 8, pp. 48 049–48 059, 2020.
[53] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7373–7382.
[54] V. F. Arruda *et al.*, "Cross-domain car detection using unsupervised image-to-image translation: From day to night," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
[55] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, 2017.
[56] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.
[57] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.
[58] T. Park, A. A. Efros, R. Zhang, and J. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 319–345.
[59] X. Huang, M. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 179–196.
[60] H. Hsu *et al.*, "Progressive domain adaptation for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 738–746.
[61] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.
[62] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018.
[63] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2224–2233.
[64] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[65] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 1392.
[66] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv: 2003.04297*.
[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[68] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[69] 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. 4885–4894.
[70] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, Art. no. 1009.
[71] Z. Zhang, W. Shen, S. Qiao, Y. Wang, B. Wang, and A. L. Yuille, "Robust face detection via learning small faces on hard images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1350–1359.
[72] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 03, 2021, doi: 10.1109/TPAMI.2021.3063604.
[73] T. Liu, Z. Chen, Y. Yang, Z. Wu, and H. Li, "Lane detection in low-light conditions using an efficient data enhancement: Light conditions style transfer," in *Proc. IEEE Intell. Veh. Symp.*, 2020, pp. 1394–1399.
[74] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[75] A. Lengyel, S. Garg, M. Milford, and J. C. van Gemert, "Zero-shot domain adaptation with a physics prior," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021.
[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[77] G. Lin, A. Milan, C. Shen, and I. D. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.
[78] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7373–7382.
[79] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "ARID: A new dataset for recognizing action in the dark," 2020, *arXiv: 2006.03876*.
[80] X. Wang *et al.*, "When pedestrian detection meets nighttime surveillance: A new benchmark," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 509–515.

Wenjing Wang (Graduate Student Member, IEEE) received the BS degree in data science in 2019 from Peking University, Beijing, China, where she is currently working toward the PhD degree with the Wangxuan Institute of Computer Technology. Her current research interests include image enhancement, image synthesis, and deep learning.

Xinhao Wang is currently working toward the BS degree in intelligence science from Peking University, Beijing, China. His current research interests include style transfer and deep learning.

Wenhan Yang (Member, IEEE) received the BS and PhD degrees (Hons.) in computer science from Peking University, Beijing, China, in 2012 and 2018, respectively. He is currently a postdoctoral research fellow with the School of EEE, Nanyang Technological University, Singapore. He has authored more than 80 technical articles in refereed journals and proceedings, and holds nine granted patents. His current research interests include image/video processing/restoration, bad weather restoration, and human –machine collaborative coding. He was the recipient of the IEEE ICME-2020 Best Paper Award, IEEE CVPR-2018 UG2 Challenge First Runner-up Award, and CSIG Best Doctoral Dissertation Award, in 2019. He was an area chair of IEEE ICME-2021 and the organizer of IEEE CVPR-2019/2020/2021 UG2+ Challenge and Workshop.

Jiaying Liu (Senior Member, IEEE) received the PhD degree (Hons.) in computer science from Peking University, Beijing, China, 2010. She is currently an associate professor, 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, California. She was a visiting researcher with Microsoft Research Asia, in 2015 supported by the Star Track Young Faculties Award. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 60 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a senior member of CSIG and CCF. She was a member of Multimedia Systems and Applications Technical Committee (MSA TC), and Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She is also a member of Image, Video, and Multimedia (IVM) and Signal and Information Processing Theory and Methods (SIPTM) Technical Committee in APSIPA. She was the recipient of the IEEE ICME 2020 Best Paper Award and IEEE MMSP 2015 Top10% Paper Award. She was also an associate editor for *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *Journal of Visual Communication and Image Representation*, the technical program chair of IEEE ICME-2021/ACM ICMR-2021, the area chair of CVPR-2021/ECCV-2020/ICCV-2019, and the CAS Representative at the ICME Steering Committee. During 2016–2017, she was the APSIPA distinguished lecturer.