# Unsupervised Illumination Adaptation for Low-Light Vision

Wenjing Wang, *Student Member, IEEE*, Rundong Luo, *Student Member, IEEE*, Wenhan Yang, *Member, IEEE*, and Jiaying Liu, *Senior Member, IEEE*

***Abstract*—Insufficient lighting poses challenges to both human and machine visual analytics. While existing low-light enhancement methods prioritize human visual perception, they often neglect machine vision and high-level semantics. In this paper, we make pioneering efforts to build an illumination enhancement model for high-level vision. Drawing inspiration from camera response functions, our model could enhance images from the machine vision perspective despite being lightweight in architecture and simple in formulation. We also introduce two approaches that leverage knowledge from base enhancement curves and self-supervised pretext tasks to train for different downstream normal-to-low-light adaptation scenarios. Our proposed framework overcomes the limitations of existing algorithms without requiring access to labeled data in low-light conditions. It facilitates more effective illumination restoration and feature alignment, significantly improving the performance of downstream tasks in a plug-and-play manner. This research advances the field of low-light machine analytics and broadly applies to various high-level vision tasks, including classification, face detection, optical flow estimation, and video action recognition.**

***Index Terms*—Domain adaptation, high-level vision, illumination enhancement, low-light, self-supervised learning.**

## I. Introduction

INSUFFICIENT lighting is a prevalent image degradation resulting from adverse environments, defective equipment, or improper shooting settings. It can impair images' visual quality, leading to detail loss, decreased visibility, and aesthetic distortion. Moreover, with the advance of deep learning, visual analytics is becoming increasingly crucial in numerous applications. Low-light conditions can also challenge machine analytics, raising difficulties in high-level vision tasks, such as nighttime surveillance video analysis and autonomous driving.

The restoration of low-light images has received extensive attention since the birth of digital imaging. Many works have effectively improved the human visual quality of low-light images, from early manually designed algorithms [1] to recent learning-based models [2]. However, most existing low-light enhancement methods aim to improve the images' visual quality but ignore machine vision demand, misleading downstream high-level vision models. Some methods try to embed semantic perception [3] for visual reconstruction but still cannot guarantee performance in downstream high-level vision tasks.

To further improve machine vision performance in the dark, an intuitive idea is to directly train the model on annotated low-light data [4]. Despite performing well on certain tasks, annotation requirements severely limit their application scope. Therefore, unsupervised normal-to-low-light domain adaptation has emerged as a promising research direction that eliminates the annotation needs. Among this field, some methods propose to synthesize target domain annotation through image translation [5], [6], while others adopt self-supervised learning [7] or utilize handcraft operators [8]. However, existing algorithms either rely on multiple source domains [9], adopt troublesome multi-stage and multi-level processes [7], or fail in darker cases [8]. Moreover, most adaptive methods concentrate on the high-dimensional features of machine analytics models while neglecting the characteristics of input images themselves.

In contrast to the aforementioned methods, we fully utilize the potential of illumination adjustment. We propose a curve-based enhancement model and two self-supervised training strategies to enhance images from the machine vision perspective, thus benefiting the model's performance on downstream high-level tasks. First, inspired by the Weber-Fechner law and camera statistics, we constrain our enhancement function by "concavity", which enjoys an efficient implementation by predicting a non-positive second-order derivative and then applying discrete integration. This design enables the enhancement model to produce natural-looking images and improves its adaptability to multiple downstream tasks. Then, we design two self-supervised strategies to train this model towards unsupervised adaptation. When task-relevant information is available, we propose assembling the knowledge of a pre-defined set of base enhancement curves. This process is implemented by congregating the model's prediction results on images enhanced by these curves into pseudo labels. Although simple in formulation, assembling the base curves could bring reliable supervision for subsequent self-training. Meanwhile, we draw support from pretext tasks

Manuscript received 7 June 2023; revised 17 December 2023; accepted 10 March 2024. Date of publication 27 March 2024; date of current version 6 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, and in part by the AI Joint Lab of Future Urban Infrastructure sponsored by Fuzhou Chengtou New Infrastructure Group and Boyun Vision Company Ltd. Recommended for acceptance by J. Gu. *(Wenjing Wang and Rundong Luo contributed equally to this work.) (Corresponding author: Jiaying Liu.)*

Wenjing Wang, Rundong Luo, and Jiaying Liu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: daooshee@pku.edu.cn; rundong_luo@stu.pku.edu.cn; liujiaying@pku.edu.cn).

Wenhan Yang is with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: yangwh@pcl.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3382108, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3382108

when task-relevant information is difficult to use. Specifically, we train a pretext task head to guide the model by our proposed rotated jigsaw task. The network architecture and training strategies complement each other and form a framework with strong adaptability despite being easy to train.

Our self-aligned concave curve (SACC) framework could be a strong baseline for unsupervised normal-to-low-light adaptation. Note that although our method leverages task-relevant information, it does not rely on any labeled data under low-light conditions, and the obtained models are plug-and-play for a wide range of downstream tasks. Besides, we suggest using pseudo-labeling to address the noise and semantic discrepancies between the enhanced low-light and the natural normal-light images. We call this SACC+, which is straightforward to implement but could further boost the performance and surpass existing low-light enhancement and domain adaptation techniques.

In summary, our contributions are threefold:

- We are the first to propose an illumination enhancement model for low-light high-level vision. Our model could enhance images from the machine vision perspective despite being lightweight in architecture and simple in formulation.
- We train the enhancement model with base enhancement curves or pretext tasks to satisfy different downstream scenarios. Our training strategy can narrow the normal/low-light domain gap and improve the model's performance without annotated data. Besides, our framework serves as a plug-and-play remedy for multiple downstream tasks.
- We evaluate our method on various high-level vision tasks, including classification, face detection, optical flow estimation, and video action recognition. Extensive experiments across multiple benchmarks demonstrate the superiority of our approach over state-of-the-art low-light enhancement and domain adaptation methods.

This manuscript is an extension of our prior publication [10] while exhibiting improvements in the following aspects: (1) We introduce a novel training strategy for our deep concave curve in Section IV-A. This strategy enables us to fully leverage task-relevant information, which could substantially improve performance even in low-light conditions where labeled data is unavailable. We denote this novel technique by SACC-CE and refer to the original method as SACC-PT. (2) We incorporate asymmetric data augmentation for unsupervised adaptation to simulate the diverse low-light conditions while keeping the normal-light distribution simple, thus further boosting the performance, as detailed in Section IV-C. (3) We enrich the ablation analyses in Section IV, provide more qualitative results in Section VI-B–VI-E, and discuss a broader scope of applications in Section VI-H. Extensive experiments on various datasets demonstrate the superiority of our method, as shown in Fig. 1.

The structure of the remaining sections is as follows: In Section II, we review previous literature in relevant areas. Subsequently, we introduce the motivation and design of our deep concave curve in Section III, elaborate on its training strategy in Section IV, and empirically analyze both the curve architecture and training strategy in Section V. Thereafter, we validate the efficacy of our framework in Section VI and summarize the paper in Section VII.

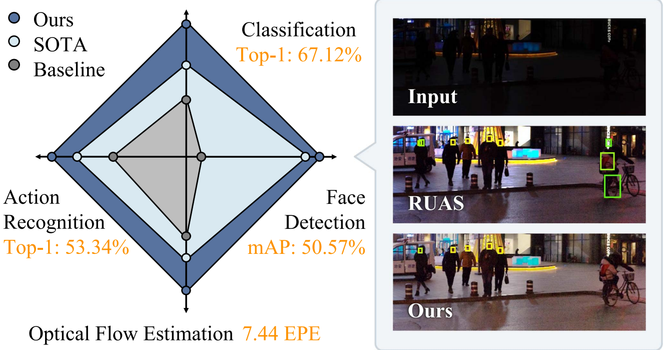


Fig. 1. Left: Comparison with the baseline model (trained with normal light data only) and previous state-of-the-art on multiple downstream tasks. Right: Example nighttime face detection results. Our approach better enhances faces hidden in darkness, resulting in more accurate detection.

## II. Related Works

*Low-Light Restoration:* Low-light enhancement aims to improve the human visual experience towards images captured under insufficient lighting conditions. Conventional approaches exploit non-learning techniques such as histogram equalization [11], gamma correction, and image formation theories such as the Retinex Theory [12], which decomposes images into albedo and reflectance components. The advent of deep learning has facilitated the development of more effective approaches. Some works [13], [14], [15] suggested simulating the Retinex decomposition process using paired training data, while others, like EnlightenGAN [16], employ the adversarial learning paradigm to eliminate the need for paired data. Besides, DRBN [17] introduced a semi-supervised framework that combines the benefits of supervised and unsupervised methods; RUAS [2] unrolled the optimization process of Retinex-inspired models and used neural architecture search to find better network architectures; Zero-DCE [18] introduced quadratic curves for enhancement, whose parameters could be learned without normal-light references; Wu et al. [19] introduced a semantic-aware knowledge-guided framework that assists a low-light enhancement model in learning semantic priors from segmentation models; Ren et al. [20] proposed a hybrid network to capture the global content and salient structures of images in a unified network. Advanced techniques and frameworks, including frequency decomposition [21], feature pyramids [22], [23], flow models [24], and transformers [25], are also adopted in recent papers. In addition to these approaches for general RGB images, several works focus specifically on restoring backlit images [26], multi-stereo images [27], ultra-high-definition images [28], [29], RAW images [30], and videos [31], [32].

*Unsupervised Domain Adaptation:* Unsupervised domain adaptation aims to adapt the model trained on a labeled source domain to an unlabeled target domain. Existing methods can be generally categorized into feature alignment [33], adversarial learning [34], domain translation [35], and self-training [36].

Feature alignment methods [33], [37], [38] quantify the feature discrepancy between two domains by a certain statistical metric and minimize it. Adversarial learning methods [34], [39], [40] incorporate an adversarial loss term to distinguish domains. Domain translation methods [35], [41], [42] synthesize target domain samples by generative adversarial networks (GANs) for training. Self-training methods [36], [43], [44], [45] create pseudo-labels for the unlabeled target domain data and then re-train the network. Nevertheless, despite achieving good performance for many applications, existing domain adaptation approaches are less effective in low-light conditions due to the inherent complexity of the normal/low-light domain gap.

*High-Level Vision in Low-Light Conditions:* Recent years witnessed a proliferation of research on low-light high-level vision due to its increasing application demand. Apart from utilizing enhancement methods as a pre-processing step, adapting the model pre-trained on normal-light data to low-light is also a popular solution. For example, Sasagawa et al. [46] devised an approach for dark object detection by combining pre-trained models from different domains using glue layers. MAET [47] exploits the image signal processing (ISP) pipeline for nighttime image generation and uses both synthetic and real nighttime images for training. HLA-Face [7] uses a joint feature- and pixel-level framework for low-light face detection. DANNet [9] proposes a multi-source adversarial training framework for nighttime semantic segmentation to adapt models in a single stage. CIConv [8] presents a color-invariant convolutional network for learning illumination-invariant features that apply to various tasks. Other research has focused on low-light image retrieval [48], depth estimation [49], and matching [50].

Despite these progresses, most existing unsupervised adaptation approaches concentrate on feature migration while ignoring the significance of pixel-level adjustment. In contrast, this paper presents an enhancement-based adaptation method that outperforms existing methods by a large margin.

## III. Architecture: The Deep Concave Curve

This section introduces our motivation and the architecture of our illumination enhancement module.

### A. Motivation: From CRF to Concave Curve

The camera response function (CRF) defines the relationship between a scene's light irradiance and the pixel values (intensity) captured by a camera. Illumination is linearly related to the irradiance level but has a complex non-linear relationship with the intensity. On this basis, some low-light enhancement works [51], [52] exploit the linearity of irradiance. They first transform the image's intensity to irradiance, adjust its irradiance linearly, and then map irradiance back to intensity. However, CRFs vary across different cameras and even different ISO settings on the same camera, posing a great challenge for estimating irradiance. Moreover, this back-and-forth mapping is a pure low-level operation requiring prior knowledge of the target camera's ISP settings.

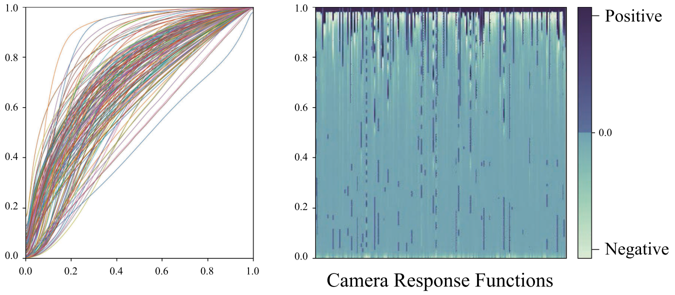


Fig. 2. Left: real camera response functions from the DoRF [54] dataset. Right: The heat map of second-order derivatives in DoRF, where each column represents a camera response function.

We propose simplifying this complex pipeline into a single-step adjustment at the intensity level, making it easier to incorporate high-level vision guidance during training. Specifically, our adjustment function can be defined as a mapping $g$ from the original intensity value to the enhanced intensity value. We first analyze the constraints $g$ should satisfy. Generally, CRFs can be considered identical for each pixel in an image, disregarding spatial variations such as vignetting, lens fall-off [53], or signal-dependent noise. Thus, we set $g$ as a global operation. Additionally, $g$ should be a monotonically increasing function that passes through (0,0) and (1,1) to maintain the intensity's monotonicity and numerical range. Moreover, though $g$ is defined on discrete values, it should be approximately continuous to prevent distortion on neighboring pixels.

Furthermore, we draw inspiration from real CRFs to determine the shape of $g$. Fig. 2 shows a collection of CRFs obtained from the DoRF [54] dataset. Statistically, we found that 89.5% of the CRFs have a negative second-order derivative (i.e., concavity), which is in line with the fact that a stimulus's perceived intensity is proportional to its physical intensity's logarithm, as described by the Weber-Fechner law. Consequently, a linear increase in irradiance leads to a concave transformation in intensities, indicating that $g$ should be concave.

The above constraints appropriately limit the solution space. On the one hand, the constraints are concise for implementation. On the other hand, they ensure that enhanced images follow the distribution of natural images and thus support downstream vision tasks. Next, we introduce how to embed the above constraints into neural networks.

### B. Formulation and Implementation

The above analysis suggests that our illumination enhancement function should adhere to two fundamental characteristics: *increasing monotonicity* and *concavity*. Accordingly, we call our enhancement model the "deep concave curve".

Now, we present its comprehensive design. Given an input low-light image $I_L$, we use a neural network $C$ to predict the adjustment function $g$. $g$ can be represented as a vector $g \in \mathbb{R}^P$ for a color space with $P$ numeric values (e.g., $P = 256$ for 8-bit images). Specifically, for a pixel of value $(p-1)/P$ ($p \in \{1, 2, \ldots, P\}$) from the input image, its new value will be the $p$-th element of $g$. We provide an exemplary case in the

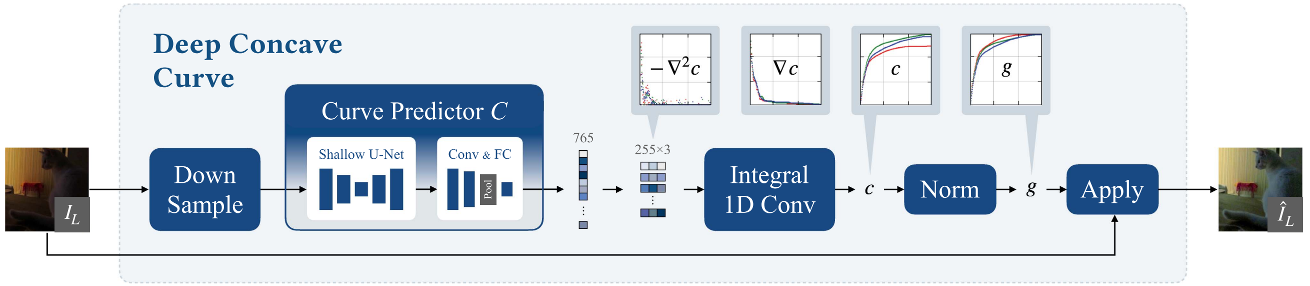


Fig. 3. The architecture of the proposed deep concave curve, which intends to enhance the illumination of the input low-light image. We first predict the minus second-order derivative $-\nabla^2 c$ and then integrate and normalize it into a concave curve $g$. Finally, we apply $g$ to the input image $I_L$ to obtain the enhanced image $\hat{I}_L$.

supplementary material to demonstrate the functionality of the curve.

Intuitively, concavity and monotonicity could be achieved by penalty terms on the discrete first-order or second-order derivatives. Nevertheless, this naive approach brings new loss functions and thus adds burdens to the balance between multiple learning objectives.

We instead propose to satisfy the two constraints by altering the model's prediction to avoid introducing new loss functions. Our model predicts the negative second derivative $-\nabla^2 c = C(I_L)$ instead of directly estimating $g$. The concavity of $g$ is ensured by $-\nabla^2 c \geq 0$, and we implement $-\nabla^2 c \geq 0$ by incorporating a ReLU activation after the final layer of network $C$. Then, we integrate $-\nabla^2 c$ into $c$. Finally, we normalize $c$ into $g = \mathrm{Norm}(c)$ to fit the mapping to the range of [0,1].

The discrete integration from $-\nabla^2 c$ to $\nabla c$ can be considered sub-sequence summation. Specifically, we represent $\nabla c$ as the matrix-vector multiplication $\nabla c = A \cdot (-\nabla^2 c)$ where $A = [a_{ij}]$ is the upper triangular matrix:

$$a_{ij} = \begin{cases} 1, & i \leq j, \\ 0, & i > j. \end{cases} \tag{1}$$

Since $-\nabla^2 c \geq 0$ and $a_{ij} \geq 0$, it follows that $\nabla c \geq 0$, guaranteeing the function $g$'s monotonicity. Similarly, $c = B \cdot (\nabla c)$, where $B = [b_{ij}]$ is a strictly lower triangular matrix:

$$b_{ij} = \begin{cases} 0, & i \leq j, \\ 1, & i > j. \end{cases} \tag{2}$$

For computational efficiency, we combine the two integrations into one step $c = D \cdot (-\nabla^2 c)$, where $D = B \cdot A$ is calculated in advance. Given that the first element of $g$ is zero, we only need to predict the remaining $P - 1$ values, i.e., $-\nabla^2 c \in \mathbb{R}^{P-1}$ and $c \in \mathbb{R}^P$. Finally, we divide $c$ by its maximum value: $g = \mathrm{Norm}(c) = c/||c||_\infty$ to normalize its entry to [0,1].

### C. Network Architecture

Given an input low-light image $I_L$, We predict an independent $g$ for each color channel, i.e., $g_R, g_G, g_B$ for RGB images. As for implementation, we place $g_R$, $g_G$, $g_B$ adjacently and carry out the integration as a one-dimensional convolution with an output channel of 256 and a kernel size of 1. The convolution weight is set to $D = B \cdot A$ shared across the three channels.

We depict the overall architecture in Fig. 3. During training and inference, we downsample the input image to a resolution of $16 \times 16$ to enhance the receptive field and efficiency. The curve predictor $C$ comprises a shallow U-Net [55], two convolutional layers, a global average pooling, and a fully connected layer. The output dimension is 765 for 8-bit RGB images. After obtaining the model's prediction, we reshape it to $3 \times 255$ and then acquire $c$ by integrating it through 1D convolution. Finally, we normalize $c$ to derive $g$ and apply it back to $I_L$.

## IV. Training Strategy: Self-Aligned Adaptation

Upon introducing the form of the deep concave curve, this section discusses how to train this curve for unsupervised illumination adaptation.

Our objective is to train the illumination enhancement curve capable of improving the model's performance on downstream tasks in low-light conditions. Contrary to conventional low-light enhancement techniques [2], [18] that solely focus on visual quality, we propose to utilize high-level machine vision as guidance. Specifically, given a downstream model pre-trained on normal light images, we freeze its backbone and want our deep concave curve to bridge the feature-level gap between low and normal light. Nonetheless, generating appropriate supervision becomes our greatest challenge without direct supervision.

Previous works focused on aligning global features, using techniques such as discrepancy metrics [38] or adversarial discriminators [34]. However, the domain gap between low and normal light includes both illumination-relevant and irrelevant aspects, and we aim to align only the illumination-relevant portions. The illumination-irrelevant distribution drift, such as differences in background or object appearance, is not our objective but abounds. As a result, inaccurate supervision is introduced when aligning global features, potentially misleading the model and complicating the training process.

To disentangle illumination-relevant and irrelevant features, we employ self-supervised learning and devise two strategies for scenarios where downstream task information can and cannot be used. The following presents the design of each strategy.

### A. Curve Ensemble Learning

To begin with, we discuss a circumstance where task information (i.e., the task head) is readily available. Utilizing

**Algorithm 1:** Pseudo-Labeling in Curve Ensemble.

```
input  : Unlabeled low-light dataset $\mathcal{D}_{ul}$, curve
         family $\mathcal{T}$, normal-light pre-trained model $F$
         and task head $h_T$, confidence threshold $t_1$,
         and candidate threshold $t_2$.
output : Pseudo-labeled low-light dataset $\mathcal{D}_{pl}$
initiate: $\mathcal{D}_{pl} = \emptyset$
forall $I_L \in \mathcal{D}_{ul}$ do
    // store the predictions
    Multiset $S = \emptyset$ ;
    for $f_i \in \mathcal{T}$ do
        Enhance $I_L$ with $f_i$ and obtain $I_{L,i}$ ;
        Use $F$ and $h_T$ to predict its label and
          confidence: $(y_i, c_i)$ ;
        if $c_i \geq t_1$ then
            Add $y_i$ to $S$ ;
    if $|S| \geq t_2$ and all $y_i \in S$ are identical then
        // denote the prediction by $y$
        Add $(I_L, y)$ to $\mathcal{D}_{pl}$ ;
return : $\mathcal{D}_{pl}$
```

task information allows us to optimize our enhancement model towards downstream tasks directly, thus avoiding being misled by illumination-irrelevant features that are unimportant for downstream tasks.

However, the crux remains the absence of labels, and the most straightforward idea is to draw support from pseudo labels, i.e., self-training. Self-training is a prevalent approach for semi-supervised learning and domain adaptation [43], [44], [45]. Specifically, it uses a model pre-trained on the source domain to generate pseudo labels for the unlabeled target domain data. Afterward, the model is re-trained on the target domain dataset with pseudo labels. Nevertheless, directly adopting this approach for normal-to-low-light adaptation will result in ill performance due to too many noisy labels, thus yielding unsatisfactory results.

To generate high-quality pseudo labels, we propose to assemble a set of simple curves, which serve as base enhancement models. Despite being simple in formulation and poor in performance, we expect that assembling those weak models could result in a more robust model and thus bring reliable supervision to our deep concave curve. Our method is called curve ensemble learning, and the following elaborates on its details.

Given a low-light image $I_L$ and the curve family $\mathcal{T}$, we generate an enhanced image $I_{L,i} = f_i(I_L)$ for each curve $f_i \in \mathcal{T}$ and then query the pre-trained daytime model to predict $I_{L,i}$'s label $y_i$ and confidence $c_i$. Predictions with confidence below a certain threshold $t_1$ will be discarded. Then, we grant $I_L$ a pseudo label if the remaining predictions are identical and the number of them exceeds another threshold $t_2$. The granted pseudo label $y$ is the prediction all remaining curves agree to. This process is repeatedly operated on all target domain images. Finally, we collect all images with a granted pseudo label and obtain a low-light dataset $\mathcal{D}_{pl}$. The pseudo-code of this algorithm is provided in Algorithm 1.

Then we define the form of curves in $\mathcal{T}$. For typical low-light enhancement mappings $f : [0, 1] \to [0, 1]$, it should satisfy $f(0) = 0, f(1) = 1$ as well as being monotonic increasing and concave. As discussed in [7], many curve forms (power, exponential, arc-tangent, reciprocal, etc.) perform reasonably well, while the reciprocal curve is the best. We follow their empirical findings and set $\mathcal{T}$ be a set of reciprocal curves:

$$\mathcal{T} = \bigcup_{\alpha \in \mathcal{A}} \{f(x; \alpha)\}, \tag{3}$$

where

$$f(x; \alpha) = \frac{(\alpha + 1)x}{x + \alpha} \tag{4}$$

and $\mathcal{A} = \{0.001, 0.01, 0.04, 0.1, 0.25, 0.6, 2, 10^5\}$. Note that $\alpha > 0$ ensures the concavity of the curve. The $\mathcal{T}$ curves are illustrated in Fig. 5.

Finally, we train the network with the cross-entropy loss $\ell_{ce}$ on the pseudo labeled set $\mathcal{D}_{pl}$:

$$\mathcal{L}_L^{CE} = \sum_{(I_L, y) \in \mathcal{D}_{pl}} \ell_{ce}(h_T(F(\hat{I}_L)), y), \tag{5}$$

where $F$ is the feature extractor, $h_T$ the task-specific head, and $\hat{I}_L$ the enhanced result of $I_L$ by our deep concave curve. During training, both $F$ and $h_T$ are frozen.

### B. Pretext Task Learning

When task information is available, we can leverage pseudo labels to reduce the effect of illumination irrelevant factors. However, for many downstream tasks, task heads involve non-differentiable modules such as non-maximum suppression. In such scenarios, directly using the model's prediction for the subsequent self-training process is inappropriate, as pointed out in [56]. To address this issue, we propose an alternate approach that draws support from self-supervised pretext tasks and does not require task information.

We architect our approach on the following observation. Despite having varied illumination conditions, both low-light and normal-light domains share a prior distribution of natural images. Therefore, a model trained on normal-light data should also be able to do pretext tasks on the enhanced low-light images. Leveraging this point, we first append a trainable MLP head after the fixed feature extractor and train it on normal-light images by the pretext task. We then add our deep concave curve before the backbone and train it on low-light images by the pretext task with the backbone and MLP head fixed. Through optimizing the pretext task loss, our deep concave curve could learn to enhance low-light images for the downstream models, thus implicitly improving the downstream task performance in underlit conditions.

Then, we discuss the choice of the pretext task. Contrastive learning [57], [58] is a self-supervised learning paradigm that contrasts positive and negative image pairs. However, contrastive learning perplexes the model's perception of illumination as it requires fierce signal-related augmentations (e.g., color jittering). Moreover, contrastive learning maximizes

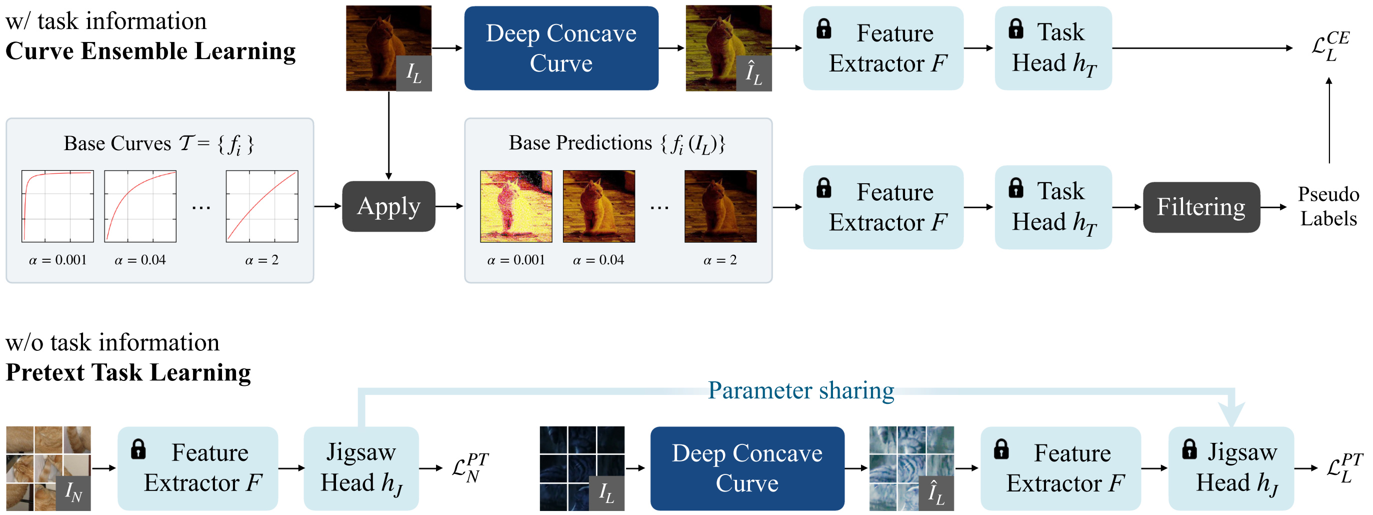
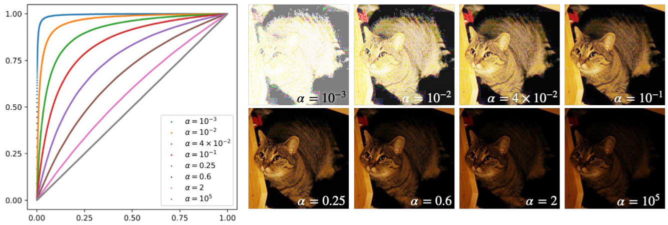


Fig. 4. The framework of self-aligned concave curve (SACC). When task information is available, we generate high-quality pseudo labels by assembling knowledge from a pre-defined curve family. When task information is unavailable, we first train a pretext head on normal-light data and then learn the deep concave curve on dark data with a fixed pretext head.



(a) Enhancement curves (b) Enhanced results

Fig. 5. Illustration of reciprocal curve family $\mathcal{T}$ (a), and corresponding enhanced results (b).

the feature distance between different samples, which will mislead the enhancement model into generating diverse but abnormal results. Meanwhile, although conventional techniques such as rotation prediction [59] or jigsaw puzzling [60] do not alter the images' visual appearance, they are not powerful enough to provide enough supervision for illumination adaptation.

To effectively guide the enhancement model, we propose the rotated jigsaw puzzle. The rotated jigsaw puzzle involves randomly rotating the input image at various angles and applying a 3×3 jigsaw shuffling process. The network is then trained to recognize the permutation. This approach is more challenging than traditional jigsaw puzzling and enhances the MLP heads' understanding of semantics, thus introducing more supervision for adaptation.

The training pipeline involves two steps. We first fix the feature extractor $F$ and train the jigsaw head $h_J$ on normal-light dataset $\mathcal{D}_N$ consisting of normal-light images and their ground-truth rotated jigsaw permutation $p$:

$$\mathcal{L}_N^{PT} = \sum_{I_N \in \mathcal{D}_N} \ell_{ce}(h_J(F(I_N)), p). \tag{6}$$

Then we fix both $F$ and $h_J$, and train our deep concave curve on the low-light dataset $\mathcal{D}_L$:

$$\mathcal{L}_L^{PT} = \sum_{I_L \in \mathcal{D}_L} \ell_{ce}(h_J(F(\hat{I}_L)), p), \tag{7}$$

where $\hat{I}_L$ is the enhanced result of $I_L$ by our deep concave curve. We illustrate the two training strategies in Fig. 4.

### C. Unsupervised Normal-to-Low-Light Adaptation

The above designs form our novel unsupervised low-light adaptation framework, dubbed the self-aligned concave curve (SACC). Given a downstream model pre-trained on normal-light data, we use its feature extractor to train the deep concave curve via two distinct strategies with and without task information: curve ensemble (SACC-CE) and pretext task (SACC-PT). In the test phase, input images are first enhanced before applying the downstream model.

Contrary to existing low-light adaptation techniques [7], [8], [61], SACC eliminates the need for normal and low-light data annotations and does not alter the downstream model. Despite only focusing on illumination enhancement, SACC achieves superior results by concurrently addressing low-level characteristics and high-level features.

However, low-light conditions present challenges beyond insufficient lighting, such as noise and semantics. While SACC effectively manages illumination shifts, other distribution gaps persist between enhanced low-light images and real-world normal-light images, proving difficult to address via curve-based enhancement. On this basis, we again leverage self-training to finetune the downstream model. Consistent with previous self-training approaches, our method involves two steps: pseudo-label generation and finetuning.

*Pseudo-Label Generation:* A prevalent issue in self-training is overfitting due to noisy labels. To mitigate this issue, we employ a simple threshold rejection mechanism. Specifically,

TABLE I
DESCRIPTION OF OUR METHODS

| Method | SACC | | SACC+ | |
|---|---|---|---|---|
| | SACC-CE | SACC-PT | SACC-CE+ | SACC-PT+ |
| Curve Ensemble<br>(Sec. 4.1) | ✓ | | ✓ | |
| Pretext Task<br>(Sec. 4.2) | | ✓ | | ✓ |
| Self-Training<br>(Sec. 4.3) | | | ✓ | ✓ |
| Enhance Model | Trained | Trained | Trained | Trained |
| Downstream Model | Fixed | Fixed | Trained | Trained |

SACC-CE and SACCPT: we train the enhancement model only by Curve Ensemble and Pretext task, respectively, with the downstream model fixed. SACC-CE+ and SACC-PT+: we incorporate an additional stage that trains the downstream model.

TABLE II
EFFECTS OF DIFFERENT LOW-LIGHT ENHANCEMENT BACKBONES UNDER OUR ASYMMETRIC SELF-SUPERVISED STRATEGY

| Network Architecture | Classification (SACC-CE) | Detection (SACC-PT) |
|---|---|---|
| Baseline | 55.04 | 16.09 |
| EnlightenGAN [16] | 57.92 | 18.42 |
| Zero-DCE [18] | 61.84 | 25.36 |
| Gamma Correction $x^{\gamma}$ | 61.12 | 38.25 |
| No Constraint | 27.20 | 12.44 |
| $\nabla g \geq 0$ | 61.28 | 34.40 |
| $\nabla g \geq 0$ and $\nabla^2 g \leq 0$ (proposed) | **62.24** | **41.31** |
| $\nabla g \geq 0$, $\nabla^2 g \leq 0$, and $\nabla^3 g \geq 0$ | 60.88 | 40.74 |

We report top-1 accuracy (%) for classification and mean average precision (%) for detection. The best performance in each column is bolded.

low-confidence pseudo-labels are discarded during the finetuning stage. Although seemingly simplistic, we empirically found that this method can achieve comparable results to those self-paced techniques.

*Finetuning with Asymmetric Augmentation:* Additionally, we propose to inject asymmetric augmentations into low-light and normal-light images. Concretely, we exert weak data augmentations (e.g., horizontal flip and mild resize & crop) to the normal light images and strong data augmentations (e.g., fierce resize & crop and color jittering) to low-light images to simulate various low-illumination environments and make the task more difficult.

This self-training strategy on the downstream model can further mitigate the domain gap between enhanced low-light and normal-light images. We call this advanced version **SACC+**. For disambiguation, we briefly explain our proposed methods in Table I.

## V. Analysis of Method Design

### A. Justification for Model Architecture

In the following, we provide empirical justifications for the proposed illumination enhancement model, i.e., the deep concave curve. The analysis is based on the classification dataset CODaN [8] and face detection dataset DARK FACE [62] and WIDER FACE [62]. Our goal is to improve the performance of the normal-light pre-trained model in low-light conditions. The normal-light pre-trained models are ResNet-18 [63] for classification and DSFD [64] for detection. More details can be found in Section VI. As for training strategy, we adopt SACC-CE for the former and SACC-PT for the latter if not explicitly specified. For efficiency, we adopt the fast inference mode of DSFD [64] during evaluation in this section.

To begin with, we compare our approach with other network architectures for low-light enhancement. We explore two representative enhancement models: EnlightenGAN [16], which directly performs image-to-image translation through a U-net [55], and Zero-DCE [18], which performs pixel-wise adjustment by an iterative quadratic curve. All enhancement models are trained using the same strategy as ours so that we can solely evaluate the effect of network architecture.

As the subjective results in Table II demonstrate, both EnlightenGAN and Zero-DCE perform poorer than our proposed deep concave curve.

As shown in Fig. 6, we observe weird artifacts on enhanced images generated by EnlightenGAN and Zero-DCE. In Fig. 6(a), both methods cause discontinuous color variation. In Fig. 6(b), the edges have abnormal textures. In comparison, our deep concave curve enjoys better visual results and downstream performance in both image classification and face detection. These results demonstrate our spatially sharing, monotonically increasing, and concavity constraints could faithfully regularize the model from carving cheating symbols or hints (i.e., artifacts) on images.

Additionally, we test gamma correction $x^{\gamma}(0 < \gamma < 1)$ as an alternate curve-based enhancement approach by using a shallow convolutional network to predict $\gamma$. As shown in Table II, its performance is limited. We ascribe it to our curve's high degree of freedom as we predict an independent target value for each input pixel value. At the same time, gamma correction only has the global adjustment parameter $\gamma$.

Finally, we explore the appropriate intensity of constraint imposed on the curve prediction network. As shown in Fig. 7 and Table II, setting the curve unrestricted will result in abnormal enhancement results. Meanwhile, the curve seems discontinuous with only $\nabla g \geq 0$. Our SACC achieves the best result when we require the curve to satisfy $\nabla g \geq 0$ and $\nabla^{2\tilde{}} g \leq 0$. Note that in Fig. 7, the red, blue, and green curves have different shapes, indicating that our model can perform channel-aware enhancement in accordance with the input image, i.e., correct the color bias. We also try $\nabla^{3\tilde{}} g \geq 0$ and discover it will degrade the overall performance and generate partially over-exposed images. This is because three successive iterative integration exponentially increases the value of the weight matrix's entries, thus incurring gradient vanishing and complicating the training process.

### B. Justification for Training Strategy

Next, we discuss the proper training strategy for our deep concave curve. We compare our proposed training strategy (SACC-CE and SACC-PT) with other prevailing paradigms, including discrepancy metrics [38], [65], [66], adversarial learning [67],

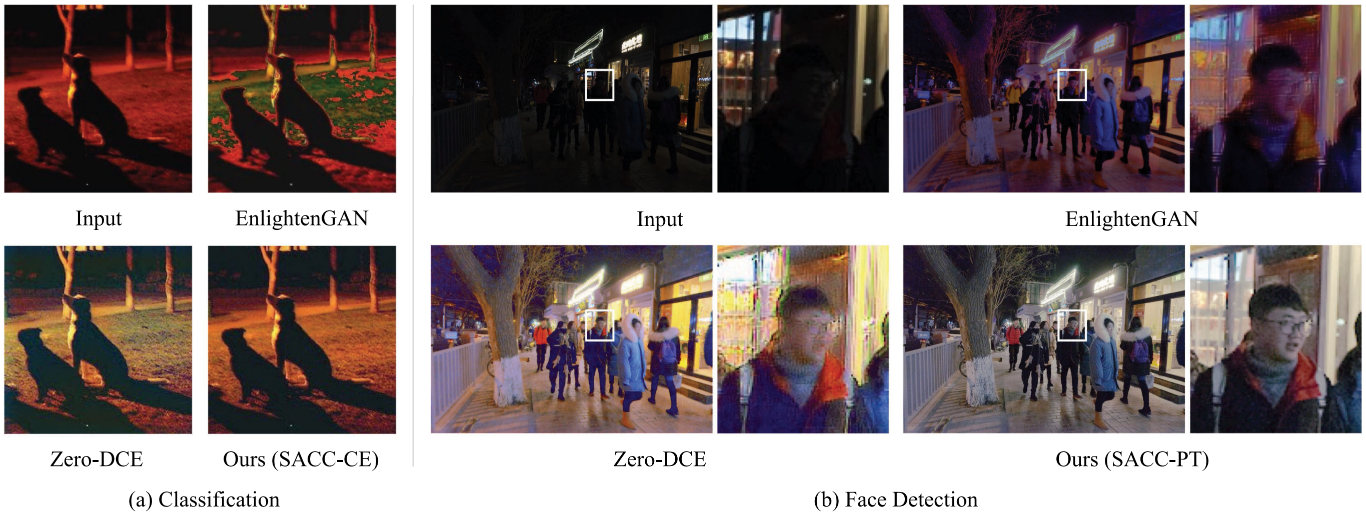


Fig. 6. Effects of asymmetric self-supervised learning with different low-light enhancement model backbones.

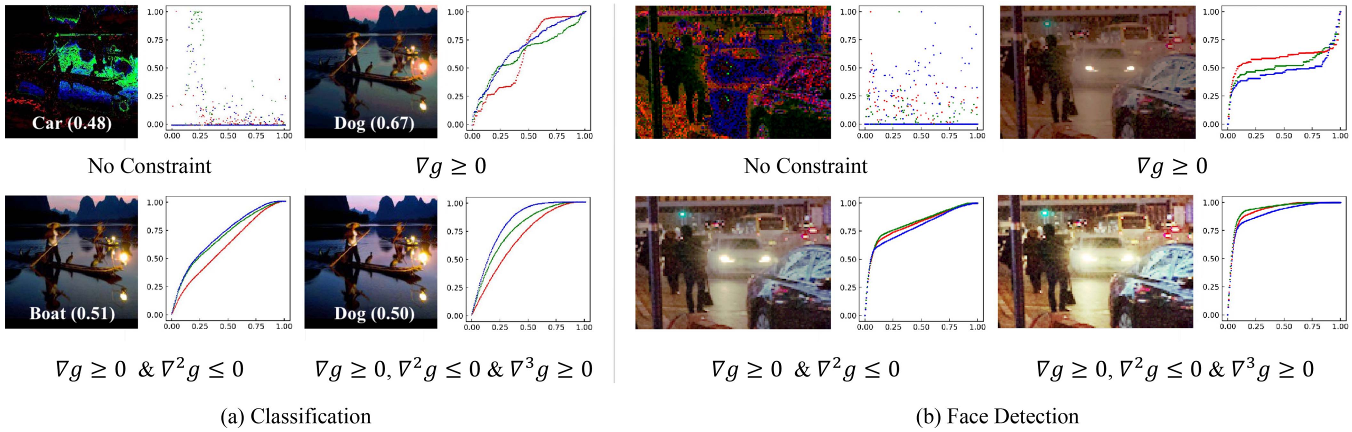


Fig. 7. Empirical results on (a) CODaN classification dataset and (b) DARK FACE face detection dataset. We show each group's low-light enhancement results (left) and curve shapes (right) under different curve form constraints. The curve shapes demonstrate the correspondence between the original (x-axis) and new pixel values (y-axis).

and self-supervised learning [58], [59], [60]. Quantitative results are provided in Table III.

Discrepancy metrics, Central Moment Discrepancy (CMD) [65], Maximum Mean Discrepancy (MMD) [66], and Deep CORAL [38], are originally proposed to deal with general domain adaptation and are ineffective in bridging the normal/low-light domain gap. Adversarial learning [67] also brings unsatisfactory results since it is easy to discriminate between normal/low-light modalities, which breaks the balance between the feature extractor and discriminator. Besides, adversarial learning also suffers from unstable training dynamics due to its complex architecture.

For self-supervised learning, we consider both contrastive learning-based [58] and pretext task-based approaches [59], [60]. Despite the superiority of contrastive learning on model pre-training, it considers global features and thus is unsuitable for training our enhancement curve. Conventional pretext tasks, including rotation prediction and jigsaw permutation, perform unsatisfactorily due to their limited expressiveness.

As for proposed SACC-CE and SACC-PT, the former performs better on the classification task while poorer on face detection. We owe this to the discrete and easily comparable nature of the output in classification, which improves the accuracy of the curve ensemble. In contrast, the ensemble predictions exhibit reduced accuracy when applied to face detection. Due to the need for a specially designed merge operation for bounding box predictions, directly combining them like classification would lead to inaccurate or too few pseudo-labels, resulting in sub-optimal training results. Designing a proper label merging strategy might be a solution; however, developing a strategy for each downstream task is impractical. Therefore, we propose pretext task-based SACC-PT that operates directly on features. SACC-PT performs worse than SACC-CE on classification since it cannot leverage task information but is more general

TABLE III
COMPARISON BETWEEN DIFFERENT TRAINING STRATEGIES OF OUR DEEP CONCAVE CURVE

| Category | Method | Classification | Detection |
|---|---|---|---|
| Baseline | - | 55.04 | 16.09 |
| Discrepancy Metrics | MMD [66]<br>CMD [65]<br>CORAL [38] | 60.80<br>61.04<br>61.28 | 33.76<br>28.90<br>21.53 |
| Adv. Learning | LSGAN [67] | 60.32 | 38.05 |
| Self-supervised Learning | MoCo [58]<br>Rotation [59]<br>Jigsaw [60] | 60.00<br>61.76<br>60.48 | 38.46<br>39.59<br>41.01 |
| **Proposed** | SACC-PT<br>SACC-CE | 61.44<br>**62.24** | **41.31**<br>38.62 |

We report top-1 accuracy (%) for classification and mean average precision (%) for detection. The best performance in each column is bolded.

TABLE IV
COMPARISON BETWEEN DENOISING THE ENHANCED LOW-LIGHT IMAGES AND OUR PROPOSED SELF-TRAINING APPROACH (SACC+)

| Category | Method | Classification | Detection |
|---|---|---|---|
| Baseline | - | 62.24 | 41.31 |
| Denoising | N2N [68]<br>BM3D [69] | 54.40<br>55.28 | 40.54<br>24.51 |
| Self-Training (Proposed) | w/o asymmetric aug.<br>w/ asymmetric aug. | 63.36<br>**67.12** | 45.51<br>**48.73** |

We evaluate top-1 accuracy (%) for classification and mean average precision (%) for detection. We adopt SACC-CE and SACC-PT for classification and face detection, respectively. The best performance in each column is bolded.

to non-classification downstream tasks than SACC-CE. Overall, for applications that require the aggregation of full image information for decision-making, such as image classification, SACC-CE can be used; for those requiring finer-grained local information for inference, SACC-PT can be used.

### C. SACC+: Image Denoising or Model Finetuning?

Finally, we verify our design of SACC+. Since our deep concave curve is a global operation, it leaves the noise untouched, which is one primary characteristic of low-light images. Therefore, denoising the enhanced images may further improve the performance. To test this assumption, we consider two approaches: the non-learning method BM3D [69] and the learning-based Neighbor2Neighbor [68].

Despite their success in human visual experience, denoising blurs details crucial to high-level semantics. As a result, the model's classification and face detection performance degrade notably, as shown in Table IV. Contrarily, our finetuning approach, SACC+, inherently addresses noise issues through representation learning without compromising image information. Furthermore, employing asymmetric augmentation during finetuning helps bridge the gap between the normal-light domain and the enhanced low-light domain, effectively boosting the model's overall performance.

## VI. EXPERIMENTS

This section provides the implementation details, benchmarking results, empirical analysis, and applications of our proposed method.

### A. Implementation Details

Our proposed framework applies to various models and vision tasks. To justify its effectiveness, we evaluate it on several representative low-light vision tasks, including classification, face detection, optical flow estimation, and video action recognition.

Our framework functions on a pre-trained model using normal-light data. For SACC-CE/SACC-PT, we freeze the pre-trained model and train our deep concave curve. Afterward, for SACC-CE+/SACC-PT+, we freeze the enhancement model and finetune the downstream model. Both stages require no labeled low-light data. In the following, We adopt SACC-CE for classification and video action recognition, while SACC-PT for face detection and optical flow estimation. More implementation details can be found in the supplementary material.

### B. Low-Light Image Classification

To begin with, we evaluate our method by the image classification task. CODaN [8] is a 10-class dataset gathered for low-light adaptation, which comprises 12,500 normal-light images and 2,500 low-light images. We use their official normal light settings while splitting 1,250 low-light images for training/validation and the remaining for testing. We aim to adapt the ResNet-18 pre-trained on normal light data to the low light domain. We compare our approach with eleven low-light enhancement methods and four unsupervised domain adaptation methods. Note that enhancement methods only modify input images, while adaptation methods also alter the model. We also provide the fully supervised learning result (i.e., trained with low-light labels) as a reference.

Table V shows the comparison results. Low-light enhancement methods improve illumination from a human vision perspective but neglect machine vision, resulting in a limited performance gain. Unsupervised domain adaptation methods, CMD [65], AdaBN [75], and DANN [34], are designed for general domain adaptation scenarios and fail to handle the significant normal/low-light domain gap effectively. CIConv [8] introduces a color-invariant convolutional layer to acquire illumination-invariant features. However, its handcrafted operator is not robust enough to handle the diverse illumination conditions of low-light environments. In contrast, our SACC-CE adjusts illumination according to machine vision guidance, yielding the best outcome. Furthermore, our SACC-CE+ achieves even higher results through pseudo-label finetuning. We also compare our results with vanilla supervised training, which could be an ideal upper bound for unsupervised methods to evaluate their potential. However, it is worth noting that supervised methods are also constrained by data volume, collection domain, etc., making the upper bound not strict. Due to incorporating our adaptive enhancement curve and generalizable training strategy, our method has achieved comparable results (less than

TABLE V
COMPARISON RESULTS OF LOW-LIGHT IMAGE CLASSIFICATION

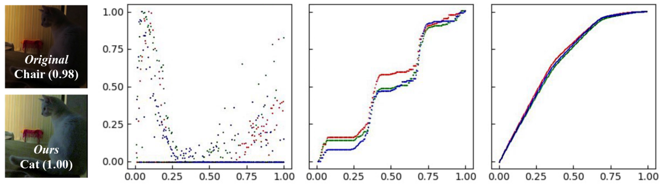
| Category | Method | Top-1 (%) |
|---|---|---|
| Baseline | ResNet-18 [63] | 55.04 |
| Supervised | Finetuned ResNet-18 [63] | 71.52 |
| Low-Light Enhancement | RetinexNet [13] | 44.72 |
| | LLFlow-SKF [19] | 49.60 |
| | Zero-DCE† [18] | 50.80 |
| | EnlightenGAN [16] | 57.76 |
| | Zero-DCE++ [70] | 58.56 |
| | SCI [71] | 59.12 |
| | MF [72] | 59.20 |
| | KinD [14] | 59.28 |
| | LIME [73] | 59.44 |
| | Zero-DCE [18] | 59.44 |
| | URetinexNet [74] | 59.52 |
| | RUAS† [2] | 59.84 |
| | EnlightenGAN† [16] | 60.24 |
| | LLFlow [24] | 60.72 |
| | **SACC-CE** (Ours) | **62.24** |
| Unsupervised Domain Adaptation | CMD [65] | 55.92 |
| | AdaBN [75] | 59.68 |
| | DANN [34] | 59.76 |
| | CIConv [8] | 60.96 |
| | **SACC-CE+** (Ours) | **67.12** |

† denotes that we have re-trained the enhancement model.
The best performance for each category is bolded.

TABLE VI
COMPARISON RESULTS FOR DARK FACE DETECTION

| Category | Method | mAP (%) |
|---|---|---|
| Baseline | DSFD [64] | 16.09 |
| Supervised | Finetuned DSFD [64] | 45.99 |
| Low-Light Enhancement | RetinexNet [13] | 12.04 |
| | LLFlow-SKF [19] | 13.94 |
| | KinD [14] | 15.84 |
| | EnlightenGAN† [16] | 20.77 |
| | EnlightenGAN [16] | 31.31 |
| | URetinexNet [74] | 31.39 |
| | SCI [71] | 34.45 |
| | Zero-DCE † [18] | 37.30 |
| | LLFlow [24] | 37.41 |
| | RUAS† [2] | 38.36 |
| | LIME [73] | 40.71 |
| | Zero-DCE++ [70] | 40.90 |
| | Zero-DCE [18] | 41.27 |
| | MF [72] | 41.43 |
| | **SACC-PT** (Ours) | **44.57** |
| Unsupervised Domain Adaptation | CIConv [8] | 4.40 |
| | OSHOT [76] | 25.38 |
| | Progressive DA [77] | 28.47 |
| | Pseudo Labeling [78] | 35.07 |
| | HLA-Face [7] | 44.44 |
| | HLA-Face v2 [79] | 45.91 |
| | **SACC-PT+** (Ours) | **50.57** |

† denotes that we have re-trained the enhancement model on DARKFACE [62]. The best performance for each category is bolded.



(a) Results (b) $-\nabla^2 c$ (c) $\nabla c$ (d) $g$

Fig. 8. An example result of low-light image classification. (a) The original image is classified as "Chair" with confidence 0.98. After enhancement, the model recognizes it correctly as "Cat". (b)-(d) Different curve shapes.

5% performance drop on classification) or even surpasses the performance of supervised methods (on face detection in the following section). These results demonstrate the great potential of our method.

We show an example of the predicted enhancement curves in Fig. 8 to justify they have satisfied our requirements. Despite $-\nabla^2 c$ appearing discrete, its value remains positive, resulting in a monotonically increasing integration $\nabla c$ and a concave final enhancement curve $g$, which aligns with the constraint we introduced in Section III-B. Satisfying these constraints ensures that the predicted curves obey the natural image priors, thus providing good generalization for low-light domain adaptation.

## C. Dark Face Detection

We further benchmark our approach on dark face detection. The WIDER FACE dataset [84] comprises 32,000 images from various events and scenes, while the DARK FACE dataset [62] consists of 10,000 nighttime street scene images. We employ their official splits and use DSFD [64] as the baseline.

Table VI presents the results. Although RetinexNet [13] and KinD [14] focus on detail reconstruction and noise reduction, they introduce additional artifacts that negatively affect machine vision and face detection performance. Other low-light enhancement methods also bring limited improvement. We attribute these results to their ignorance of machine vision, which induces undesirable but human-imperceptible noise to the enhanced images.

On the other hand, unsupervised domain adaptation methods, such as OSHOT [76], Progressive DA [77], and Pseudo Labeling [78] only bring limited performance gain. Meanwhile, despite being effective on CODaN, the color invariant layer of CIConv [8] fails on the DARK FACE because of the extremely low illumination of nighttime street scenes. As shown in Fig. 9(l), CIConv's representation suffers from terrible noise. HLA-Face [79] and HLA-Face v2 [79] adopt complex joint high-low-level adaptation strategies. In contrast, our SACC-PT surpasses HLA-Face while only adjusting the illumination. Besides, our advanced version, SACC-PT+, could even outperform the fully supervised result, demonstrating our method's superiority.

We then generalize the DSFD-trained SACC-PT to other face detectors, including PyramidBox [80] and MogFace [81]. Table VII demonstrates our method performs consistently better even on unseen downstream frameworks, indicating that knowledge distilled from high-level vision is applicable across various machine models.

## D. Optical Flow Estimation in the Dark

Next, we investigate optical flow estimation to show the wide-ranging applicability of our approach. The VBOF dataset [87]

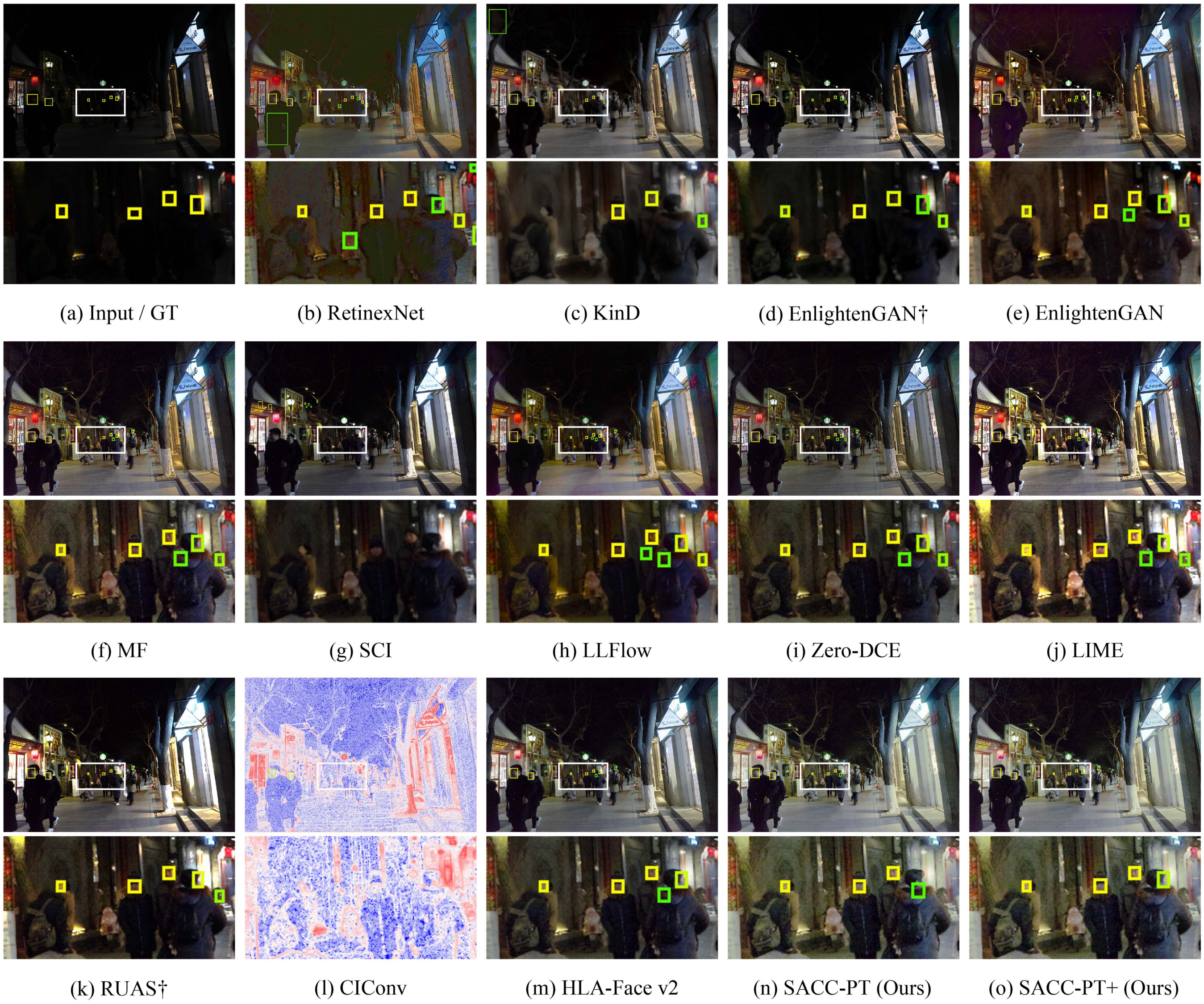
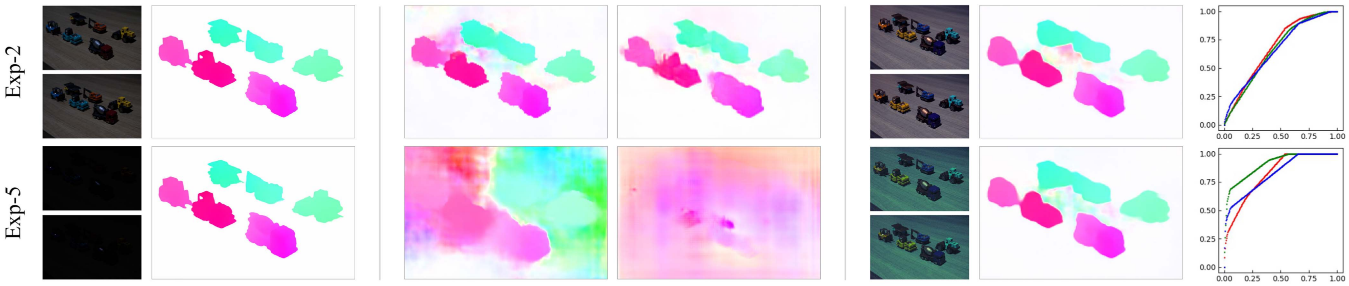


Fig. 9. Qualitative comparison results of dark face detection. † denotes that we have re-trained the enhancement model on DARKFACE [62]. The bounding boxes' color signals the model's confidence in detection, with yellow representing a higher degree of confidence.

TABLE VII
More Results on DARK FACE [62]

| | Detection mAP (%) | | | | NR-IQA | |
|---|---|---|---|---|---|---|
| | PyramidBox [80] | DSFD [64] | MogFace [81] | Average | NIQE [82] | SSEQ [83] |
| Original | 13.99 | 16.09 | 16.36 | 15.48 | 8.5 | 24.1 |
| RetinexNet [13] | 11.42 | 12.04 | 14.56 | 12.67 | 7.7 | 21.9 |
| KinD [14] | 15.61 | 15.84 | 21.27 | 17.57 | 9.8 | 42.2 |
| EnlightenGAN† [16] | 19.54 | 20.77 | 24.02 | 21.44 | 9.0 | 22.6 |
| EnlightenGAN [16] | 28.45 | 31.31 | 35.79 | 31.85 | 9.7 | 14.1 |
| Zero-DCE† [18] | 33.41 | 37.30 | 37.75 | 36.15 | 6.7 | 18.4 |
| LLFlow [24] | 32.84 | 37.41 | 41.08 | 37.11 | 8.4 | 10.3 |
| RUAS† [2] | 32.77 | 38.36 | 40.71 | 37.28 | 6.2 | 4.4 |
| LIME [73] | 35.69 | 40.71 | 42.82 | 39.74 | 6.5 | 5.7 |
| Zero-DCE++ [70] | 35.56 | 40.90 | 43.45 | 39.97 | 6.4 | 6.7 |
| Zero-DCE [18] | 35.95 | 41.27 | 43.62 | 40.28 | 6.4 | 7.9 |
| MF [72] | 37.49 | 41.43 | 43.87 | 40.93 | 6.5 | 8.2 |
| **SACC-PT** (Ours) | 39.20 | 44.57 | 46.45 | 43.41 | 6.3 | 3.2 |

Red denotes the best and blue denotes the second best performance.



(a) Input (b) Ground Truth (c) Original (d) Zheng *et al.* (e) SACC-PT (Ours) (f) Our $g$

Fig. 10. Optical flow estimation results of the same scene under different illumination levels.

TABLE VIII
COMPARISON RESULTS OF LOW-LIGHT OPTICAL FLOW ESTIMATION RESULTS

| | Exp-2 | Exp-3 | Exp-4 | Exp-5 | AVG |
|---|---|---|---|---|---|
| Original | 10.21 | 11.90 | 14.36 | 17.82 | 13.57 |
| Zheng *et al.* [87] | 9.04 | 8.81 | 8.78 | 9.31 | 8.99 |
| **SACC-PT** (Ours) | **6.70** | **7.03** | **7.57** | **8.47** | **7.44** |

The best performance in each column is bolded.

contains 10,000 image pairs with varying brightness levels. We choose Exp-2 to Exp-5 subsets as the target domain for adaptation, where Exp-2 is brighter and Exp-5 is darker. The ground truth flow fields are estimated by the state-of-the-art method GMA [88] on the Exp-1 subset captured in normal-light conditions. We adopt the PWC-Net [89] as the baseline and measure the performance by end-point error (EPE).

Zheng et al. [87] proposed to adapt the optical flow models by simulating dark image noise and generating a low-light training dataset. Nevertheless, this synthesis process only considers the signal distribution while neglecting machine vision, resulting in a performance gain that is inferior to ours, as shown in Table VIII. Moreover, Fig. 10 demonstrates our SACC's robustness to varying input illumination levels.

### E. Low-Light Action Recognition

Although originally designed for images, our approach also applies to videos. This section further evaluates our framework by low-light video action recognition. We collect approximately 800 low-light video clips from the ARID dataset [92]. Normal light training data consists of 2,600 normal light video clips from HMDB51 [93], UCF101 [94], Kinetics-600 [90], and Moments in Time [95]. We use the 3D-ResNet [96]-based I3D [85] as our primary classifier.

When predicting enhancement curves, we combine all frames in a video clip into a large image and globally apply the predicted curve $g$ to all frames. This uniformity of $g$ ensures the temporal consistency of the enhanced video. We report the results as top-1 accuracy.

As shown in Table IX, video enhancement methods StableLLVE [86], SMOID [32], and SGZ [91] improve the baseline by around 3%. In contrast, our SACC-CE substantially improves performance by 6.44% without additional complex operations, demonstrating our method's superiority for videos. Subjective results can be found in Fig. 11.

TABLE IX
COMPARISON RESULTS OF LOW-LIGHT VIDEO ACTION RECOGNITION

| Category | Method | Top-1 (%) |
|---|---|---|
| Baseline | I3D [90] | 46.90 |
| Low-Light Enhancement | StableLLVE [86] | 48.97 |
| | SMOID [32] | 49.57 |
| | SGZ [91] | 49.94 |
| | **SACC-CE** (Ours) | **53.34** |

The best performance is bolded.

TABLE X
COMPARISON BETWEEN THIS MANUSCRIPT AND OUR PREVIOUS WORK [10]

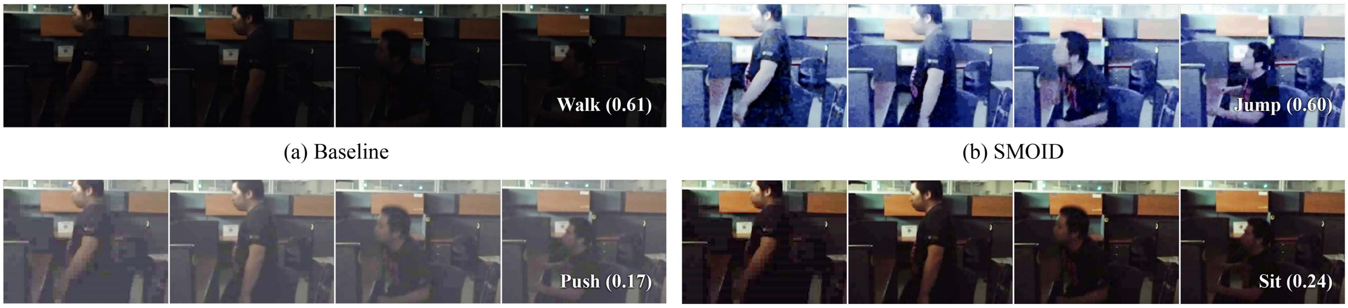
| Task | Conference Version | Journal Version | Improvement |
|---|---|---|---|
| Classification | 63.92% | 67.12% | +3.20% |
| Face Detection | 46.91% | 50.57% | +3.66% |
| Video Action Recognition | 52.13% | 53.34% | +1.21% |

### F. Comparison With Earlier Publication

Compared with our earlier publication [10], we present a novel curve ensemble technique to train our deep concave curve for classification tasks and an asymmetric augmentation strategy for SACC+. These methodological advancements significantly improve our model's performance, as evidenced in Table X.

### G. Running Time Analysis

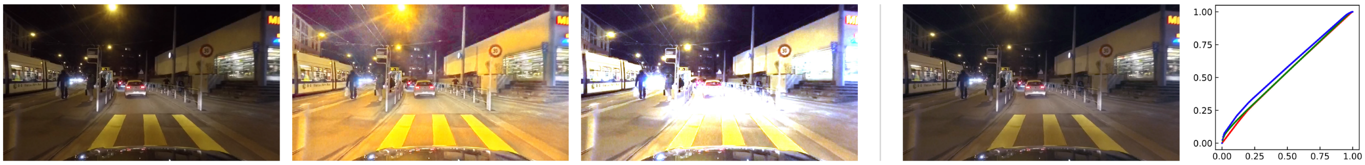
We provide the computation complexity (multiple-accumulate operations, MACs), network parameters, and running time for input images of resolution $1200 \times 900 \times 3$ in Table XI. The image is processed on a GeForce GTX TITAN X GPU with an Intel i7-9700 K @3.60 GHz CPU. Our method achieves significantly better performance while using lower computation than previous methods.



(a) Baseline (b) SMOID

(c) StableLLVE (d) SACC-CE (Ours)

Fig. 11. Qualitative low-light video action recognition results. Comparison methods include the baseline model I3D [85], and low-light video enhancement methods SMOID [32], StableLLVE [86].



(a) Input (b) EnlightenGAN (c) RUAS (d) Our result and predicted $g$

Fig. 12. Enhancement results on Nighttime Driving [97]. Our model automatically finds that the input is bright enough and thus decides not to adjust the illumination, which prevents over-exposure.

TABLE XI
Comparison Between Learning-Based Enhancement Methods on Computation Complexity (MACs), Network Parameters, and Running Time Analysis

| Method | MACs | Params | Running time |
|---|---|---|---|
| RetinexNet [13] | 359G | 555.20k | 248ms |
| EnlightenGAN [16] | 275G | 8.64M | 120ms |
| Zero-DCE [18] | 86G | 79.42k | 61ms |
| Zero-DCE++ [70] | 0.08G | 10.56k | 6.3ms |
| **SACC-CE/SACC-PT** (Ours) | 0.01G | 158.06K | 6.7ms |

### H. Broader Applications

*Generalize to Supervised Learning Scenarios:* Although primarily developed for domain adaptation, our low-light enhancement approach also offers advantages in supervised learning situations. In particular, we utilize the deep concave curve pre-trained with SACC, freeze its parameters, and finetune the normal-light model using low-light labeled training data. Compared to the supervised learning baseline, incorporating our deep concave curve enhances the model's low-light classification top-1 accuracy from 71.52% to 72.64% on CODaN.

*In-the-wild Evaluation:* We evaluate our proposed SACC for dark face detection on samples outside the DARK FACE [62] dataset and show the results in Fig 13. While the baseline detector [64] is sensitive to changes in lighting conditions and produces incorrect predictions, our approach could deliver accurate predictions consistently, regardless of the lighting conditions.

*Hard Cases:* Scenarios with sufficient artificial lighting are hard cases for low-light enhancement. We use nighttime semantic segmentation as an example, where our goal is to adapt RefineNet [98] pre-trained on the Cityscapes [99] dataset to the Nighttime Driving [97] dataset. Although the background night sky is dark in nighttime street views, artificial lights sufficiently illuminate the foreground. However, existing low-light enhancement methods cannot distinguish whether the target objects are bright.

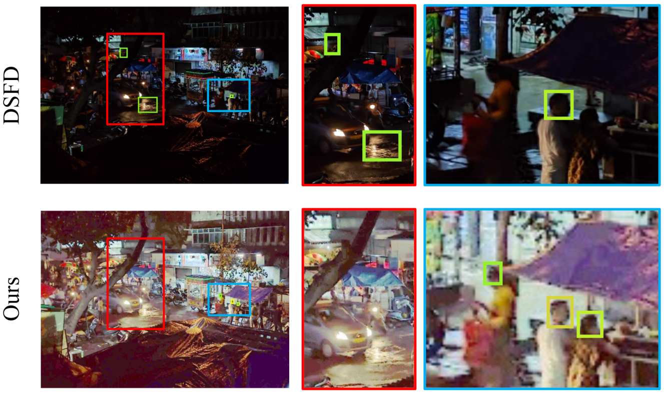


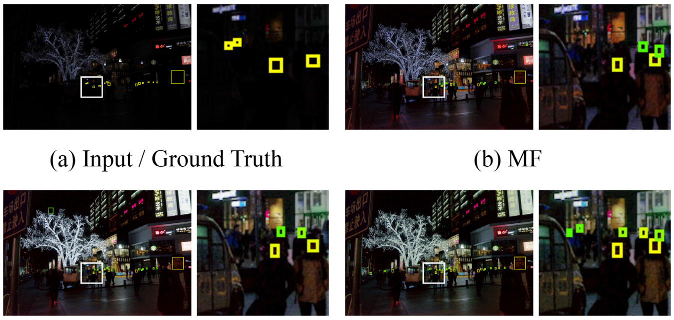
Fig. 13. In-the-wild evaluation results. First row: Input images and detection results by DSFD [64]. Second row: The corresponding enhanced image and detection results of our method. The bounding boxes' color signals the model's confidence in detection, with yellow representing a higher degree of confidence.

In comparison, our deep concave curve determines there is no need for further foreground enhancement, as illustrated in Fig. 12. In Table XII, our SACC-PT outperforms low-light enhancement methods even though they are re-trained on target datasets, showing that our framework has superior adaptability.

TABLE XII
NIGHTTIME SEMANTIC SEGMENTATION RESULTS

| Method | mIoU (%) |
|---|---|
| Baseline [98] | 33.54 |
| EnlightenGAN [16] | 21.03 |
| RUAS† [2] | 22.60 |
| LIME [73] | 25.87 |
| MF [72] | 28.26 |
| Zero-DCE [18] | 29.78 |
| Zero-DCE † [18] | 31.19 |
| EnlightenGAN† [16] | 34.73 |
| **SACC-PT** (Ours) | **35.24** |

† denotes that the low-light enhancement model is re-trained on DARKFACE [62]. The best performance is bolded.



(a) Input / Ground Truth

(b) MF

(c) HLA-Face v2

(d) Ours

Fig. 14. Failure case on extremely dark and small faces. All models have false positive predictions, but our model does not miss faces.

### I. Failure Case Study

Our model may yield incorrect predictions for extremely dark and tiny faces, which presents a challenging case for all methods. As shown in Fig. 14, MF [72] and HLA-Face v2 [79] also struggle to predict all faces correctly. This difficulty arises when face detectors rely on contextual inference to determine facial positions based on body shapes when faced with tiny faces. Consequently, this approach cannot always differentiate between a face and the back of the head, resulting in false positives. While MF and HLA-Face missed some faces, our model successfully identified all of them. This highlights our model's enhanced robustness in handling challenging scenarios involving small faces.

## VII. CONCLUSION

This paper presents a novel methodology for unsupervised normal-to-low-light domain adaptation, referred to as the Self-Aligned Concave Curve (SACC). Contrary to conventional enhancement methods that concentrate on the human visual experience, we propose to employ high-level machine vision as guidance. Our approach utilizes the deep concave curve for illumination enhancement in conjunction with two self-aligned techniques for effectively training such a curve. Extensive experiments on multiple high-level vision tasks demonstrate the superiority of our framework. Existing and future works may incorporate our proposed enhancement curve for better performance.

## REFERENCES

[1] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

[2] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10561–10570.

[3] Y. Liu et al., "Single-image HDR reconstruction by learning to reverse the camera pipeline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1648–1657.

[4] R. G. VidalMata et al., "Bridging the gap between computational photography and visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4272–4290, Dec. 2021.

[5] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7373–7382.

[6] 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.

[7] W. Wang, W. Yang, and J. Liu, "HLA-Face: Joint high-low adaptation for low light face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16195–16204.

[8] 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, pp. 4379–4389.

[9] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "DANNet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15769–15778.

[10] W. Wang, Z. Xu, H. Huang, and J. Liu, "Self-aligned concave curve: Illumination enhancement for unsupervised adaptation," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 2617–2626.

[11] 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.

[12] E. H. Land, "The retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–28, 1977.

[13] 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.

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

[15] Q. Jiang, Y. Mao, R. Cong, W. Ren, C. Huang, and F. Shao, "Unsupervised decomposition and correction network for low-light image enhancement," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19440–19455, Oct. 2022.

[16] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[17] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3060–3069.

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

[19] Y. Wu et al., "Learning semantic-aware knowledge guidance for low-light image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1662–1671.

[20] W. Ren et al., "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4364–4375, Sep. 2019.

[21] Y. Jin, W. Yang, and R. T. Tan, "Unsupervised night image enhancement: When layer decomposition meets light-effects suppression," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 404–421.

[22] Y. Zhao, Y. Xu, Q. Yan, D. Yang, X. Wang, and L.-M. Po, "D2HNet: Joint denoising and deblurring with hierarchical network for robust night image restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 91–110.

[23] S. Zhou, C. Li, and C. Change Loy, "LEDNet: Joint low-light enhancement and deblurring in the dark," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 573–589.

[24] Y. Wang, R. Wan, W. Yang, H. Li, L. Chau, and A. C. Kot, "Low-light image enhancement with normalizing flow," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2604–2612.

[25] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 12470–12479.

[26] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, “Iterative prompt learning for unsupervised backlit image enhancement,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 8060–8069.

[27] J. Huang, X. Fu, Z. Xiao, F. Zhao, and Z. Xiong, “Low-light stereo image enhancement,” *IEEE Trans. Multimedia*, vol. 25, pp. 2978–2992, 2023.

[28] C. Li et al., “Embedding fourier for ultra-high-definition low-light image enhancement,” in *Proc. Int. Conf. Learn. Representations*, 2023.

[29] Z. Zheng, W. Ren, X. Cao, T. Wang, and X. Jia, “Ultra-high-definition image HDR reconstruction via collaborative bilateral learning,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4429–4438.

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

[31] C. Chen, Q. Chen, M. N. Do, and V. Koltun, “Seeing motion in the dark,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3184–3193.

[32] H. Jiang and Y. Zheng, “Learning to see moving objects in the dark,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7323–7332.

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

[34] Y. Ganin and V. S. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[35] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 95–104.

[36] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 297–313.

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

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

[39] 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.

[40] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1647–1657.

[41] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, “Learning from simulated and unsupervised images through adversarial training,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2242–2251.

[42] J. Hoffman et al., “CyCADA: Cycle-consistent adversarial domain adaptation,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1994–2003.

[43] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, “Confidence regularized self-training,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5981–5990.

[44] K. Mei, C. Zhu, J. Zou, and S. Zhang, “Instance adaptive self-training for unsupervised domain adaptation,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 415–430.

[45] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, “ST++: Make self-training work better for semi-supervised semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4258–4267.

[46] 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.

[47] Z. Cui, G.-J. Qi, L. Gu, S. You, Z. Zhang, and T. Harada, “Multitask AET with orthogonal tangent regularity for dark object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2533–2542.

[48] T. Jenícek and O. Chum, “No fear of the dark: Image retrieval under varying illumination conditions,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9695–9703.

[49] K. Wang et al., “Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 16035–16044.

[50] W. Song, M. Suganuma, X. Liu, N. Shimobayashi, D. Maruta, and T. Okatani, “Matching in the dark: A dataset for matching image pairs of low-light scenes,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6009–6018.

[51] H. Huang, W. Yang, Y. Hu, and J. Liu, “Raw-guided enhancing reprocess of low-light image via deep exposure adjustment,” in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 118–133.

[52] H. Huang, W. Yang, Y. Hu, J. Liu, and L.-Y. Duan, “Towards low light enhancement with raw images,” *IEEE Trans. Image Process.*, vol. 31, pp. 1391–1405, 2022.

[53] N. Asada, A. Amano, and M. Baba, “Photometric calibration of zoom lens systems,” in *Proc. Int. Conf. Pattern Recognit.*, 1996, pp. 186–190.

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

[55] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, 2015, pp. 234–241.

[56] H. Li, Z. Wu, A. Shrivastava, and L. S. Davis, “Rethinking pseudo labels for semi-supervised object detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1314–1322.

[57] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[58] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[59] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in *Proc. Int. Conf. Learn. Representations*, 2018.

[60] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[61] R. Luo, W. Wang, W. Yang, and J. Liu, “Similarity min-max: Zero-shot day-night domain adaptation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 8070–8080.

[62] 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.

[63] 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.

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

[65] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, “Central moment discrepancy (CMD) for domain-invariant representation learning,” in *Proc. Int. Conf. Learn. Representations*, 2017.

[66] K. M. Borgwardt, A. Gretton, M. J. Rasch, H. Kriegel, B. Schölkopf, and A. J. Smola, “Integrating structured biological data by kernel maximum mean discrepancy,” in *Proc. 14th Int. Conf. Intell. Syst. Mol. Biol.*, 2006, pp. 49–57.

[67] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2813–2821.

[68] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, “Neighbor2neighbor: Self-supervised denoising from single noisy images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14781–14790.

[69] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[70] 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.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022.

[71] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, “Toward fast, flexible, and robust low-light image enhancement,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5627–5636.

[72] 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.

[73] 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.

[74] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, “URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5891–5900.

[75] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, “Revisiting batch normalization for practical domain adaptation,” in *Proc. Int. Conf. Learn. Representations Workshops*, 2017.

[76] A. D’Innocente, F. C. Borlino, S. Bucci, B. Caputo, and T. Tommasi, “One-shot unsupervised cross-domain detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 732–748.

[77] H. Hsu et al., “Progressive domain adaptation for object detection,” in *Proc. Winter Conf. Appl. Comput. Vis.*, 2020, pp. 738–746.

[78] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, “Cross-domain weakly-supervised object detection through progressive domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5001–5009.

[79] W. Wang, X. Wang, W. Yang, and J. Liu, "Unsupervised face detection in the dark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1250–1266, Jan. 2023.

[80] 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.

[81] Y. Liu, F. Wang, B. Sun, and H. Li, "MogFace: Rethinking scale augmentation on the face detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4083–4092.

[82] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[83] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process.: Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.

[84] 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.

[85] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[86] F. Zhang, Y. Li, S. You, and Y. Fu, "Learning temporal consistency for low light video enhancement from single images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4967–4976.

[87] Y. Zheng, M. Zhang, and F. Lu, "Optical flow in the dark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6748–6756.

[88] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. I. Hartley, "Learning to estimate hidden motions with global motion aggregation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9752–9761.

[89] D. Sun, X. Yang, M. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8934–8943.

[90] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv: 1705.06950*.

[91] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *Proc. Winter Conf. Appl. Comput. Vis. Workshops*, 2022, pp. 581–590.

[92] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "ARID: A new dataset for recognizing action in the dark," in *Proc. Deep Learn. Hum. Activity Recognit.: 2nd Int. Workshop*, 2021, pp. 70–84.

[93] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[94] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[95] M. Monfort et al., "Moments in time dataset: One million videos for event understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 502–508, Feb. 2020.

[96] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6201–6210.

[97] D. Dai and L. V. Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *Proc. 21st Int. Conf. Intell. Transport. Syst.*, 2018, pp. 3819–3824.

[98] 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.

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

**Wenjing Wang** (Student Member, IEEE) received the BS degree in data science from Peking University, Beijing, China, in 2019. She is currently working toward the doctoral degree with the Wangxuan Institute of Computer Technology, Peking University. She has authored more than 20 technical articles in refereed journals and proceedings, and she holds five granted patents. Her current research interests include image enhancement, image synthesis, and deep learning.

**Rundong Luo** (Student Member, IEEE) is working toward the graduate degree from Peking University, Beijing, China. His research spans multiple fields among computer vision, with a focus on 3D vision and generative models.

**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 an associate researcher with PengCheng Laboratory, Shenzhen, Guangdong, China. His current research interests include image/video processing/restoration, bad weather restoration, human-machine collaborative coding. He has authored more than 50 technical articles in refereed journals and proceedings, and holds 9 granted patents. He received the 2023 IEEE Multimedia Rising Star Runner-Up Award, the IEEE ICME-2020 Best Paper Award, the IFTC 2017 Best Paper Award, the IEEE CVPR-2018 UG2 Challenge First Runner-up Award, and the MSA-TC Best Paper Award of ISCAS 2022. He was the Candidate of CSIG Best Doctoral Dissertation Award in 2019. He served as the area chair of IEEE ICME-2021/2022/2023/2024, the session 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, in 2010. She is currently an associate professor and a Boya young fellow with the Wangxuan Institute of Computer Technology, Peking University. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 70 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a senior member of CSIG and a distinguished member of CCF. She was a visiting scholar with the University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. She was a visiting researcher with Microsoft Research Asia, in 2015, supported by the Star Track Young Faculties Award. She has served as a member of the Multimedia Systems and Applications Technical Committee (MSA TC) and the Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She received the IEEE ICME 2020 Best Paper Award and the IEEE MMSP 2015 Top10% Paper Award. She has also served as an associate editor for *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits Systems for Video Technology*, and *Journal of Visual Communication and Image Representation*, the technical program chair for ACM MM Asia-2023/IEEE ICME-2021/ACM ICMR-2021/IEEE VCIP-2019, the area chair for CVPR-2021/ECCV-2020/ICCV-2019, an ACM ICMR Steering Committee member, and a CAS representative for the ICME Steering Committee. She was an APSIPA distinguished lecturer from 2016 to 2017.