Pattern Recognition 154 (2024) 110558

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Collaborative brightening and amplification of low-light imagery via bi-level adversarial learning

Jiaxin Gao [a], Yaohua Liu [a], Ziyu Yue [b], Xin Fan [a], Risheng Liu [a,*]

[a] School of Software Technology, Dalian University of Technology, Dalian, 116024, China
[b] School of Mathematical Sciences, Dalian University of Technology, Dalian, China

ARTICLE INFO

Keywords:
Low-light image
Bi-level optimization
Image enhancement
Deep learning

ABSTRACT

Poor light conditions constrain the high pursuit of clarity and visible quality of photography especially smartphone devices. Admittedly, existing specific image processing methods, whether super-resolution methods or low-light enhancement methods, can hardly simultaneously enhance the resolution and brightness of low-light images at the same time. This paper dedicates a specialized enhancer with a dual-path modulated-interactive structure to recover high-quality sharp images in conditions of near absence of light, dubbed *CollaBA*, which learns the direct mapping from low-resolution dark-light images to their high-resolution normal sharp version. Specifically, we construct the generative modulation prior, serving as illuminance attention information, to regulate the exposure level of the neighborhood range. In addition, we construct an interactive degradation removal branch that progressively embeds the generated intrinsic prior to recover high-frequency detail and contrast at the feature level. We also introduce a multi-substrate up-scaler to integrate multi-scale sampling features, effectively addressing artifact-related problems. Rather than adopting the naive time-consuming learning strategy, we design a novel bi-level implicit adversarial learning mechanism as our fast training strategy. Extensive experiments on benchmark datasets — demonstrate our model's wide-ranging applicability in various ultra-low-light scenarios, across 8 key performance metrics with significant improvements, notably achieving a 35.8% improvement in LPIPS and a 23.1% increase in RMSE. The code will be available at https://github.com/moriyaya/CollaBA.

1. Introduction

Image Super-Resolution (SR) remains a fundamental challenge with in the realm of low-level computer vision, finding extensive applications across domains such as surveillance, high resolution imaging, and autonomous driving [1,2]. However, contemporary research primarily centers on super-resolving images captured under standard lighting conditions, leaving a noticeable void in the investigation of super-resolution techniques for images acquired in low-light settings [3,4]. Nevertheless, the enhancement of low-resolution images procured under extremely low illumination levels, with the objectives of luminance adjustment, detail magnification, and the ultimate generation of high-resolution, visually coherent images, holds paramount significance in practical applications [5,6]. This research endeavors to address this formidable challenge, with the overarching goal of elevating image quality in low-light environments.

In contrast to well-illuminated environments, the process of super-resolution when applied to low-light imagery may exacerbate a multitude of challenges, encompassing the amplification of artifacts, noise, uneven exposure, color aberrations, as well as the transformation of coarse and distorted textures [7,8]. Images captured in dimly lit scenarios are often influenced by highly erratic illumination distributions, which can potentially yield conspicuous shadows and highlight regions, thereby intensifying the intricacy of super-resolution tasks [9,10]. Furthermore, the act of capturing images in low-light conditions introduces additional complexities, such as diminished contrast, detail loss, and color deviations, collectively elevating the intricacy of super-resolution endeavors. Hence, it is imperative that, throughout the course of super-resolution processing, precise control over the exposure levels of input images is exercised to accommodate variations in illumination across different regions. This is paramount for ensuring that the resulting

* Corresponding author at: School of Software Technology, Dalian University of Technology, Dalian, 116024, China.
E-mail address: rsliu@dlut.edu.cn (R. Liu).

https://doi.org/10.1016/j.patcog.2024.110558
Received 9 October 2023; Received in revised form 3 April 2024; Accepted 29 April 2024
Available online 7 May 2024

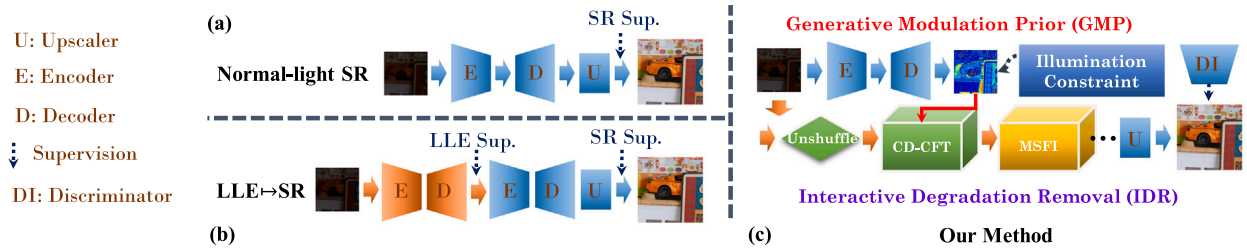

Fig. 1. Illustrating three different network architectures for low-light imagery super-resolution, including the (a) conventional normal-light Super-Resolution method (Normal-light SR), (b) cascaded Low-Light Enhancement (LLE) and SR method (LLE↦SR), and (c) our proposed dual-path modulated-interactive enhancer *CollaBA*.

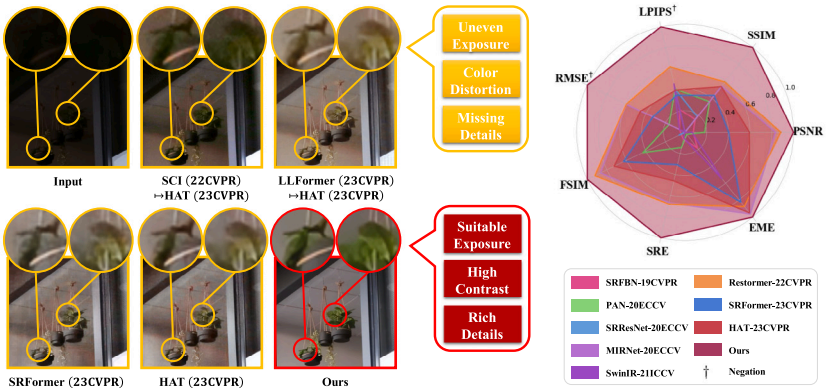

Fig. 2. Illustrating comparison results among normal-light SR, LLE↦SR, and our proposed method. Upon zooming in on specific regions, our enhanced images exhibit appropriate, realistic colors and rich textures compared to other methods. Simultaneously, we presented quantitative results for 9 representative state-of-the-art methods under 7 different metrics. Our method consistently demonstrated exceptional performance.

high-resolution imagery not only retains sharp details but also preserves a balanced perception of illumination. Moreover, given the propensity for missing details in images captured under low-light conditions, meticulous handling is warranted during the super-resolution process to avert the generation of erroneous artifacts and halos, which have the potential to compromise the quality and utility of the final output. Recently, several methods addressing this combined issue have been proposed, yet they fall short of achieving the desired enhancement outcomes [4,11]. Notably, these approaches often depend on synthetic datasets crafted by manually adjusting brightness levels, such as through gamma correction, aiming for luminance enhancement and super-resolution [4]. This leads to an ineffective generalization to real-world nighttime scenes, with issues like color distortions and artifacts.

Two direct approaches exist to tackle this complex task, as depicted in Fig. 1. The first method entails directly applying SR techniques intended for typical lighting conditions, denoted as normal-light SR. The second approach involves a cascaded combination of Low-Light Enhancement (LLE) and normal-light image SR methods, denoted as LLE↦SR, executed sequentially to achieve both brightness adjustment and resolution enhancement. Nevertheless, extensive experiments have shown that solely using these task-specific models results in noticeable constraints and a decline in performance. As illustrated in Fig. 2, we have chosen two state-of-the-art LLE methods, SCI [12] and LLFormer [2], along with the latest image SR methods, HAT [13] and SRFormer [14], as examples for visual result comparisons. Upon closer scrutiny, it becomes apparent that using HAT and SRFormer independently fails to recover fine details in low-light input images, leading to significant color discrepancies and the presence of artifacts. Similarly, the cascaded LLE↦SR methods (e.g., SCI↦HAT, LLFormer↦HAT) struggle to restore natural brightness and exhibit noticeable color distortions and texture loss. In contrast, our approach demonstrates a substantial performance advantage, yielding results with more natural and realistic colors, as well as improved structural details. Furthermore, in the right-hand side of Fig. 2, we present a numerical score comparison between our method and eight state-of-the-art methods across seven evaluation metrics. It can be observed that our method achieves a significant performance advantage compared to the second-ranked method (i.e., Restormer [15]), with a notable 35.8% improvement in terms of LPIPS and a 0.4 dB improvement in terms of PSNR. For additional comparative analysis results, please refer to the Experimental Section.

In more detail, we address the above challenges with a well-designed enhancer to simultaneously enhance brightness and resolution, dubbed *CollaBA*, which learns the direct mapping from low-resolution dark-light image to their high-resolution normal sharp version. Our core network consists of a dual-path modulated-interactive enhancer, including an interactive degradation removal module and a generative modulation prior module. The former aggregates multi-scale interactive feature flow for generating finer textures. The latter generates learnable illumination attention with the illumination prior loss, which is then embedded into subsequent degradation removal module to modulate the exposure level, through a channel-decoupling spatial transform as an intermediate bridge. Additionally, we propose a multi-substrate compositive up-sampling approach, effectively alleviating the issue of artifacts. From an algorithm-level perspective, we introduce the bi-level implicit adversarial learning strategy based on bi-level optimization to further improve the visualization performance while ensuring the stability of training. In summary, our contributions are fourfold:

- We present *CollaBA*, a specialized dual-path modulated-interactive enhancer, as a fresh approach to tackling the intricate challenge of collaborating amplification and brightening images taken in extremely low-light conditions. This marks our innovative exploration into the joint task's intricacies and the validation of practical solution strategies.

- Our *CollaBA* gains remarkable performance by imposing generative modulation priors to guide exposure regulation, progressively integrating them into the multi-scale degradation removal branch through spatial feature transformation.
- Instead of naive time-consuming adversarial learning strategy, a novel bi-level implicit adversarial learning mechanism inspired by hierarchical optimization is designed, combined with outer product approximation operation as a fast training strategy, effectively improving the stability of training and the quality of visual perception.
- Extensive quantitative and qualitative experiments were conducted to thoroughly validate that our method surpasses existing state-of-the-art approaches on real-world benchmark datasets, particularly in extremely low-light conditions, achieving a 35.8% improvement in LPIPS and a 23.1% increase in RMSE.

2. Related work

2.1. Low-light image enhancement

LLE strives to brighten dimly lit images. Early techniques, like the Retinex model [16,17], relied on handcrafted solutions, which often struggled with preserving details and colors, resulting in artifacts or loss of information [18]. In recent years, data-driven approaches have marked substantial progress, with current focus largely on both supervised [19,20] and unsupervised methods [3,12]. For instance, Yang et al. [21] proposed a recursive ribbon network and trained it using a semi-supervised strategy. Li et al. [22] focused on estimating curves specific to each low-image for brightness adjustment, integrating multiple prior-related losses (i.e., exposure, brightness, and color losses) to preserve the integrity of tones. A more lightweight and robust unsupervised method was introduced in [12], which employs a basic illuminance learning module during the inference phase to further enhance the generalization capability for low-light scenes. Most recently, there has been growing interest in leveraging simulated degradation to enhance image restoration capabilities. Jiang et al. [20] developed the degradation-to-refinement generation network, focusing on intrinsic degradation understanding and naturalness preservation. Wang et al. [19] utilized raw degradation priors to guide deep restoration models, building a prompting degradation perception modulator for adaptive image restoration learning. In practice, the most straightforward strategy entails integrating an upsampling operation within a low-light processing framework. Yet, this approach frequently results in complications such as blurriness, diminished brightness, and loss of texture clarity. Therefore, it becomes imperative to investigate techniques that elevate the resolution of low-light scenes by focusing on feature-level enhancements.

2.2. Image super-resolution

Super-resolution, in its general sense, entails generating high-resol ution imagery from their lower-resolution counterparts, usually conducted under normal lighting environments. Recent focus has primarily been on model-based approaches [23,24] and data-driven strategies [25,26]. For instance, Li et al. [27] leveraged a feedback mechanism to refine low-level representations with high-level information step by step, resulting in a strong early reconstruction ability. Jiang et al. [25] presented a hierarchical dense network for efficient image super-resolution, offering improved performance with reduced complexity. The work in [28] derived a top-k token selective transformer that eliminated irrelevant tokens, integrated multi-scale features, and utilized global context attention for high-frequency texture detail generation. Liang et al. [29] constructed transformer based network architecture for high-quality image reconstruction, where features are fused to preserve low-frequency information and enhance feature aggregation. Zamir et al. [30] proposed a multi-scale image restoration architecture with a non-local attention mechanism and attention-based multi-scale feature aggregation. Zhou et al. [14] introduced permuted self-attention to balance channel and spatial information, improving super-resolution model performance with less computational burden. Chen et al. [13] combined channel attention and window-based self-attention schemes to fully exploit the potential of Transformer networks in low-level vision tasks, achieving a balance between performance and efficiency. More recently, Cao et al. [24] presented CiaoSR, which uses implicit and scale-aware attention for super-resolution, effectively handling arbitrary-scale inputs. Li et al. [26] proposed leveraging dynamic local and global self-attention for image super-resolution, integrating transformer-inspired elements with minimal computational requirements. However, it is worth noting that conventional super-resolution methods designed for normal lighting conditions often fall short in addressing the unique challenges presented by low-light environments. Consequently, they tend to produce sub-optimal results characterized by low brightness, increased artifact presence, and blurred textures. Therefore, this paper employs specialized domain knowledge, especially in understanding ultra-dark environments, to devise a custom bi-level adversarial learning strategy aimed at generating high-frequency, fine-grained texture details, specifically designed for these joint tasks.

3. The proposed method

In this section, we will provide a comprehensive overview of our proposed *CollaBA*, including the designed network architecture, the objective functions, and the learning strategies.

3.1. Overview of CollaBA

Given a low-light, low-resolution input image $\mathbf{x}$ suffering from an unknown degradation, our joint task aims to estimate a high-quality normal-light, super-resolution image $\tilde{\mathbf{y}}$. The goal is to make $\tilde{\mathbf{y}}$ as close as possible to its ground-truth counterpart $\mathbf{y}$, which is of normal-light and high-resolution, in terms of realism and fidelity.

As depicted in Fig. 3, our *CollaBA* is meticulously designed as an enhancer utilizing a dual-path modulated-interactive structure. It primarily comprises a Generative Modulation Prior (GMP) module and a multi-scale Interactive Degradation Removal (IDR) module for refining context. These modules are connected through conditional illumination modulation using spatial feature transformations. The IDR module is specifically designed to eliminate complex degradations across multiple spatial scales, ensuring the preservation of intricate high-frequency details. By imposing a generative color prior as modulation information, we introduce the illuminance prior loss as prior constraint to regulate the exposure level. In this regard, the intensity of illumination map is obtained through the Unet-style network and its latent features are extracted. These features provide rich color and content information for interactive guidance, which are then aggregated into multi-scale degradation removal modules through spatial transformation operations to achieve realistic results while maintaining high fidelity. Also, a Unet based global-patch discriminator with spectral normalization is constructed to mitigate oversharp and artifacts. Additionally, we introduce a Multi-Substrate Merging Up-scaler (MSMU) module, which replaces the conventional single-layer naive sampling approach (e.g., bilinear, and bicubic interpolation), effectively mitigating artifact issues. Instead of relying on empirical naive adversarial learning based on alternate optimization, a novel Bi-level Implicit Adversarial (BIA) learning strategy based on a master–slave hierarchical optimization is introduced for efficient and stable training.

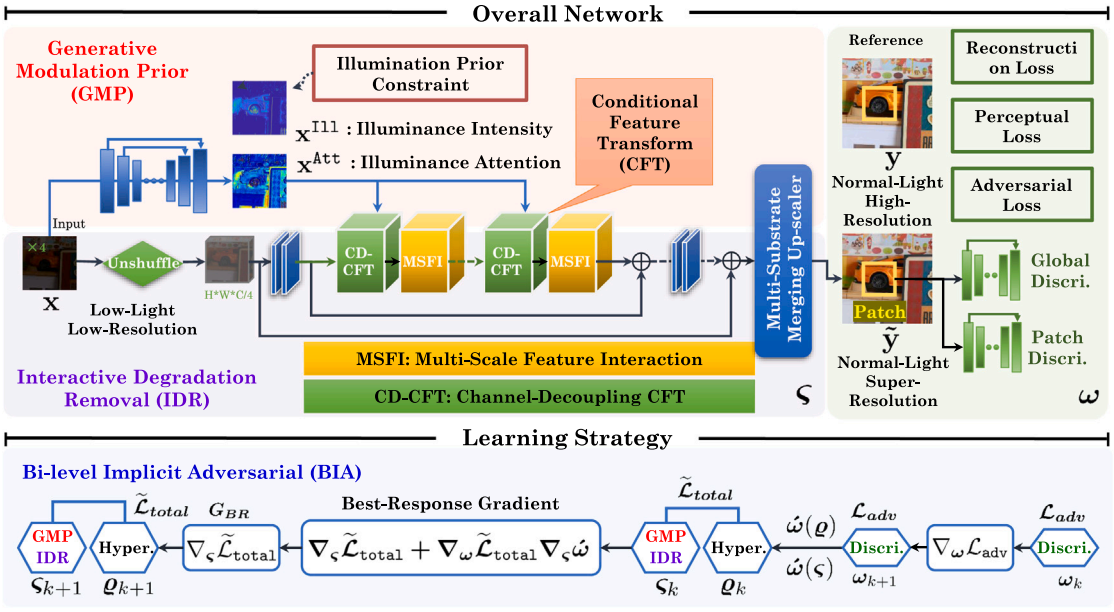

Fig. 3. Overview of *CollaBA*. The top panel illustrates the network architecture, which comprises a dual-path modulated-interactive enhancer that includes Generative Modulation Prior (GMP) and Degradation Removal modules based on multi-scale feature Interaction (abbreviate IDR). The bottom panel depicts the computational flow of the bi-level implicit adversarial strategy solution scheme.

3.2. Dual-path modulated-interactive enhancer

As illustrated in Fig. 3, we design a dual-path modulated-interactive enhancer as the base architecture because it can **(1)** extract parallel multi-resolution convolutional streams of multi-scale features and **(2)** interact with information across multi-resolution streams. Specifically, we first employ the unshuffle conversion (an inverse operation of pixelshuffle) to reduce the spatial size and enlarge the channel size before feeding the input into the main residual block architecture. Afterwards, the unshuffle features is transmitted to multi-scale residual block structure for feature extraction and aggregation, and this procedure will be completed with the aid of the following introduced channel-decoupling Conditional Feature Transform (CFT) as a bridge.

3.2.1. Channel-decoupling CFT

In order to better enhance interactive learning of contextual information, we further use the input attention features $\mathcal{A}_{tt}^{f}$ produced by several convolutional layers to modulate the multi-scale interactive feature $\mathcal{B}_{in}^{j}$. Inspired by a series of literatures related to spatial feature transformation and its variants [31,32], we first impose an affine transformation to attention features $\mathcal{A}_{tt}^{f}$ to generate modulating factors (i.e., w_{α}, w_{β}), and then execute the scaling and shifting operation for $\mathcal{B}_{in}^{j}$, formulated by:

$$\mathcal{B}_{CFT}^{j} = \mathtt{CFT}(\mathcal{B}_{in}^{j}|w_{\alpha}, w_{\beta}) = \mathtt{Conv}(\mathcal{B}_{in}^{j}) \odot w_{\alpha} + w_{\beta}, \ \text{where } w_{\alpha}, w_{\beta} = \mathtt{Conv}(\mathcal{A}_{tt}^{f}). \tag{1}$$

Here j represents the jth layer of the enhancer based on a residual structure. To better balance fidelity and transformability, we introduce channel-decoupling CFT transformation, denoted as $\mathtt{CD-CFT}$, with the aim of using learned prior features as conditional inputs to guide the exposure level of $\mathcal{B}_{in}^{j}$. As illustrated in Fig. 4(a), we first decouple features $\mathcal{B}_{in}^{j}$ by channel, splitting them into two halves: one half is retained, while the other undergoes affine transformation:

$$\begin{aligned}\mathcal{B}_{out}^{j} &= \mathtt{CD-CFT}(\mathcal{B}_{in}^{j}|w_{\alpha}, w_{\beta}) \\ &= \mathtt{Concat}\left[\mathcal{B}_{in,0}^{j}, \mathtt{Conv}(\mathcal{B}_{in,1}^{j}) \odot w_{\alpha} + w_{\beta}\right], \ \text{where } [\mathcal{B}_{in,0}^{j}, \mathcal{B}_{in,1}^{j}] = \mathtt{Split}(\mathcal{B}_{in}^{j}).\end{aligned} \tag{2}$$

Here $\odot$ denotes element-wise multiplication operation, $\mathtt{Concat}$ implies the concatenation operation, and $\mathtt{Split}$ indicates splitting the features along the channel dimension.

3.2.2. Multi-scale feature interaction

As shown in Fig. 4, we construct three scales of feature streams, for refined content reconstruction and aggregation. In each of these scales, content reconstruction blocks $\mathcal{T}_{crb}^{j}$ and selected feature aggregation modules Φ_{skff}^{j} are included, which are merged and interacted by the combination of parallel cascades. In order to focus on more significant context information, we introduce a self-attentive mechanism (i.e., Selective Kernel Feature Fusion, SKFF [30]) as Φ_{skff}^{j} to aggregate and select features at multiple resolutions. As for $\mathcal{T}_{crb}^{j}$, three-path feature transformation and translation are performed for the input features F_{in}^{j}, and then we obtain the refined features performed by $F_{out}^{j} = \mathcal{T}_{crb}^{j}(F_{in}^{j})$. Parameters are shared within the same group to reduce memory consumption. As shown in small right yellow box of Fig. 4, the implemented process are presented by

$$\mathcal{H}_{k}, \mathcal{H}_{o}, \mathcal{H}_{v} = \mathtt{Reshape}(F_{in}^{j}), \ \mathcal{T}_{crb}^{j} := \mathtt{Concat}\left[\mathcal{H}_{k} \otimes \mathtt{Softmax}(\mathcal{H}_{o}) + \mathcal{H}_{v}, F_{in}^{j}\right], \tag{3}$$

where j represents the jth layer and $\mathtt{Reshape}$ denotes the dimension transformation operation, and $\mathtt{Softmax}$ is the Softmax function. Note that $\otimes$ denotes the Kronecker product multiplication.

3.2.3. Generative modulation prior and auxiliary illuminance constraint

To guide images to enhance contrast according to exposure level, we construct a learnable illuminance prior module. The goal is to provide an adaptive guidance to properly enhance underexposed areas and avoid over-enhancement of normally exposed areas. Inspired by the physical model of Retinex theory[1], we design a self-guided heuristic learning process, aiming to use the illuminance maps that imply

[1] Retinex theory describes the phenomenon of human color vision. It can be expressed as $L = I \odot R$, where L is the low-light input, R is the reflectance layer, I is the illumination layer which is estimated by the maximum value of three color channels [16].

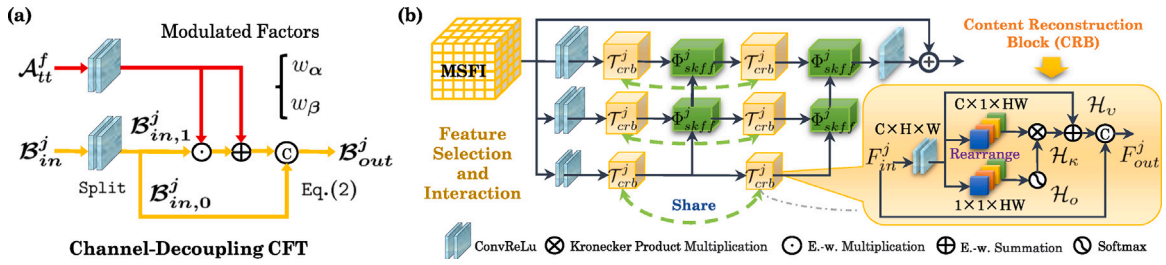

Fig. 4. Illustrating the detailed (a) CD – CFT module and (b) Multi-Scale Feature Interaction (MSFI) module. MSFT greatly ensures the selection and aggregation of features at different resolutions, with a key component (i.e., Content Reconstruction Block, CRB) maintaining parameter sharing across three scales.

exposure levels as prior information, determined by:

$$\mathbf{x}^{\mathtt{Ill}} = \frac{|\Gamma_M^{\xi}(\mathbf{x}) - \Gamma_M^{\xi}(\mathbf{y})|}{\Gamma_M^{\xi}(\mathbf{y})}, \ \Gamma_M^{\xi}(\cdot) = \max_{\xi \in \pi}(\cdot), \ \pi = \{r, g, b\}, \tag{4}$$

where $\Gamma_M^{\xi}(\cdot)$ returns the maximum value among three color channels $\{r, g, b\}$. $\mathbf{x}^{\mathtt{Ill}}$ is the desired illumination map as prior information to guide contrast enhancement. During training, we introduce the auxiliary constraint (i.e., illuminance prior loss), refer to Eq. (7), as an intermediate supervision for the expected illuminance attention map $\mathbf{x}^{\mathtt{Att}}$. We directly adopt U-Net's network as the infrastructure, and leverage such generative illumination prior to provide diverse and rich high-frequency details, i.e., $\mathbf{x}^{\mathtt{Att}} = \mathtt{UNet}(\mathbf{x})$. A typical way of deploying generative priors is to merge the latent attention feature $\mathcal{A}_{tt}^{f} = \mathtt{Conv}(\mathbf{x}^{\mathtt{Att}})$ and multi-scale feature through ψ_{CD-CFT}^{j} operation based on Eqs. (1)–(2). Such a illumination prior structure ensures all local patches of an enhanced images look like realistic normal-light ones, which proves to be critical in avoiding local over- or under-exposures as our experiments will reveal later.

3.2.4. Multi-substrate merging up-scaler

Traditional SR methods rely on interpolation upsampling techniques to enhance image resolution, but they suffer from the following drawbacks: *(1) Lack of Texture Richness:* They often fail to generate high-resolution images with natural textures and rich details, resulting in output images that appear smooth and lacking in realism. *(2) Artifacts and Halos:* Due to the ineffective modeling of scale-sensitive features, traditional methods often introduce unrealistic artifacts, halos, or aliasing artifacts. Motivated by this, we combine learnable sampling methods like pixel-shuffle with traditional bicubic sampling, employing a multi-scale integrated approach to increase resolution.

As shown in Fig. 5, MSMU utilizes pixel-shuffle $\mathcal{G}_{\mathcal{PS}}$ in parallel to generate three distinct scale features τ_i^U from the input feature C^f, where $i \in \{1, 2, 4\}$ represents three scale layers, formalized as follows:

$$\tau_i^U = \mathcal{G}_{PS}(C^f), \ \text{where } \mathcal{G}_{\mathcal{PS}} := \begin{cases} \mathtt{Identity}(C^f) & \text{If} \quad i = 1 \\ \mathtt{PS}(C^f) & \text{If} \quad i = 2 \\ \mathtt{PSoPS}(C^f) & \text{If} \quad i = 4 \end{cases} \tag{5}$$

Here Identity denotes the identity operation, which signifies no change in scale. PS is the pixel-shuffle operation for ×2 up-scaling. Subsequently, these three sets of features with different sizes enter the second stage called the dynamic aggregation block denoted as $\mathcal{P}_{DA}$ to aggregate features of different scales, producing corresponding outputs denoted as $\tilde{\tau_o^U}$, where o takes values from $\Omega = \{1, 2, 4\}$, representing different output resolution levels. Through $\mathcal{P}_{DA}$, essential features are dynamically selected for each $\tilde{\tau_{o\in\Omega}^U}$, which can be formalized as follows:

$$\tilde{\tau_{o\in\Omega}^U} = \left\langle \mathcal{P}_{DA}(\tau_i^U) \right\rangle_{i=1,2,4}, \ \text{where } \mathcal{P}_{DA} := \begin{cases} \mathtt{Identity} & \text{If} \quad i = o \\ \frac{o}{i}\uparrow_{\mathtt{BU}} & \text{If} \quad i < o \\ \frac{i}{o}\downarrow_{\mathtt{DS}} & \text{If} \quad i > o \end{cases} \tag{6}$$

Here $\uparrow_{\mathtt{BU}}$ is a bilinear operation for 2x up-scaling. $\downarrow_{\mathtt{DS}}$ is a transposed convolution for 2x down-sampling. Regarding the specific mathematical form of $\mathcal{P}_{DA}$ mentioned above, if $i = o$, an identity operation is applied, as indicated by the blue solid line in Fig. 5; if $o < i$, a transposed convolution down-sampling operation with a stride of 2 is used, as indicated by the purple solid line; if $o > i$, bilinear up-sampling operation is employed, as indicated by the green solid line. In a similar fashion, the third stage employs bicubic sampling in parallel to enhance the resolution of features at three different scales. This can be understood as a similar operation to Eq. (5), with the distinction being the use of Bicubic operations $\mathcal{G}_{\mathcal{BU}}$. Ultimately, the features at different scales are aggregated and undergo a 1 × 1 convolution transformation to yield the final upsampled image, formalized as $\tilde{\mathbf{y}} = \mathtt{Conv}_{1\times1}(\mathtt{Concat}(\langle \mathcal{G}_{\mathcal{BU}}(\tilde{\tau_o^U}) \rangle_{o\in\Omega}))$.

3.3. Training loss

The learning objective of training our *CollaBA* consists of: **(1)** illumination prior loss to provide an adaptive bootstrap for exposure level, **(2)** reconstruction loss and perceptual loss that constraints the outputs $\tilde{\mathbf{y}}$ close to the ground-truth $\mathbf{y}$ in both pixel space and perceptual space, **(3)** patch-based adversarial loss for restoring realistic textures.

Illumination Prior Loss. To obtain the expected illuminance attention map for modulating the degradation removal module, we use the $\mathcal{L}_2$ error metric to measure the prediction error as:

$$\mathcal{L}_{\mathtt{illu}}(\mathbf{x}^{\mathtt{Att}}, \mathbf{x}^{\mathtt{Ill}}) = \|\mathbf{x}^{\mathtt{Att}} - \mathbf{x}^{\mathtt{Ill}}\|_2. \tag{7}$$

Reconstruction Loss. We adopt the widely-used L_1 loss as our reconstruction loss $\mathcal{L}_{\mathtt{rec}}$:

$$\mathcal{L}_{\mathtt{rec}}(\tilde{\mathbf{y}}, \mathbf{y}) = \frac{1}{hwc} \sum_{i,j,k} |\tilde{\mathbf{y}}_{i,j,k} - \mathbf{y}i, j, k|, \tag{8}$$

where h, w and c denote the height, width and number of channels of $\tilde{\mathbf{y}}$, respectively.

Perceptual Loss. In order to ensure the consistency of image content in the feature space, we introduce the perceptual loss, defined as follows

$$\mathcal{L}_{per}(\tilde{\mathbf{y}}, \mathbf{y}) = \frac{1}{c_j h_j w_j} \|\phi_j(\tilde{\mathbf{y}}) - \phi_j(\mathbf{y})\|_2^2, \tag{9}$$

where ϕ is the pretrained VGG-19 network, c_j denotes the jth layer, $c_j h_j w_j$ denotes the size of the feature map at the jth layer.

Adversarial Loss. To support the generation of realistic textures in natural image manifolds, we introduce a spectral-normalized U-Net-style discriminator. The global-patch discriminator loss with *Binary Cross Entropy* with logistic function, i.e.,

$$\mathcal{L}_{\mathtt{adv}}(\boldsymbol{y}, \boldsymbol{z}) = -\mathtt{BCE}[D_{\mathtt{glo}}(\boldsymbol{y}), D_{\mathtt{glo}}(\boldsymbol{z})] - \mathtt{BCE}[D_{\mathtt{pat}}(\boldsymbol{y}), D_{\mathtt{pat}}(\boldsymbol{z})], \tag{10}$$

where $\mathtt{BCE}(D, \mathcal{G}) = \mathbb{E}_{\boldsymbol{z}}(\log D(\boldsymbol{z})) + \mathbb{E}_{\boldsymbol{y}}(\log(1 - D(\boldsymbol{y})))$. $D_{\mathtt{glo}}$ and $D_{\mathtt{pat}}$ denote the global and local discriminators, respectively. Such a global-patch discriminator setting facilitates a good balance of local detail enhancement and artifact suppression.

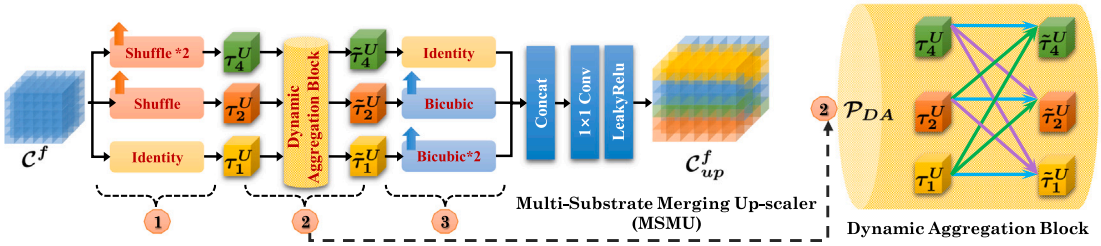

Fig. 5. Illustrating the proposed Multi-Substrate Merging Up-scaler (MSMU) module, which greatly ensures the selection and aggregation of features at different resolutions.

3.4. Bi-level implicit adversarial leaning

This section develops a novel BIA learning strategy based on the master–slave hierarchical optimization reformulation.

3.4.1. Master–slave bi-level reformulation

In order to be able to optimize learnable parameters $\{\boldsymbol{\omega}, \boldsymbol{\varsigma}\}$ for *CollaBA*, which in this case consists of the enhancer $\mathcal{E}$ with parameter $\boldsymbol{\varsigma}$ and discriminator D with parameter $\boldsymbol{\omega}$, the loss function $\mathcal{L}_u(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\omega}, \boldsymbol{\varsigma})$, comparing ground-truth labels $\mathbf{y}$ to predictions $\tilde{\mathbf{y}}$, has to be designed. The final loss is a combination of the single-modular losses $u \in \mathcal{U}$. Since each of the contributing single-modular loss function may behave differently, weighting each with a factor ϱ_u is essential. This yields a combined loss function $\mathcal{L}_{\text{total}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\omega}, \boldsymbol{\varsigma}) = \sum_{u \in \mathcal{U}} \mathcal{L}_u(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\omega}, \boldsymbol{\varsigma}) \cdot \varrho_u$. Instead of manually tuning $\boldsymbol{\varrho} = \{\varrho_u\}$ to account for the differing variances and offsets amongst the single-modular losses, the coefficients can be added to the learnable network parameters $\boldsymbol{\Theta} = \{\boldsymbol{\omega}, \boldsymbol{\varsigma}, \boldsymbol{\varrho}\}$. We introduce an augmenting $\mathcal{L}_{\text{total}}$ with a regularization term $\mathbf{R}(\varrho_u) = \ln(1 + \varrho_u^2)$ to avoid trivial solutions. The final combined loss function is expressed as

$$\mathcal{L}_{\text{total}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\Theta}) = \sum_{u \in \mathcal{U}} \frac{1}{2 \cdot \varrho_u^2} \mathcal{L}_u(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\omega}, \boldsymbol{\varsigma}) \cdot \varrho_u + \ln(1 + \varrho_u^2), \tag{11}$$

where ϱ_u denotes the weighting factor. $\mathcal{U}$ is the feasible region with respect to multiple losses.

Most existing training strategies focus on tedious hyper-parameter selection and empirical fine-tuning operations. In particular, the emerging adversarial learning strategy based on minimax formulation generally uses alternating gradient descent, but usually leads to training often tending to oscillation instability and gradient disappearance.[2] Motivated by the above, we propose a master–slave-induced hierarchical optimization framework, to explicitly characterize the dynamic coupling relationship between the three types of parameters in the gradient propagation process. Specifically, we consider the following even more abstract constrained optimization principle:

$$\begin{aligned} &\min_{\boldsymbol{\varsigma}, \boldsymbol{\varrho}} \widetilde{\mathcal{L}}_{\text{total}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\varsigma}, \boldsymbol{\varrho}, \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma}), \hat{\boldsymbol{\omega}}(\boldsymbol{\varrho})), \\ s.t.\ &\hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma}) \in S(\boldsymbol{\varsigma}) := \arg\max_{\boldsymbol{\omega}} \mathcal{L}_{\text{adv}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\varsigma}, \boldsymbol{\omega}, \boldsymbol{\varrho}), \\ &\hat{\boldsymbol{\omega}}(\boldsymbol{\varrho}) \in S(\boldsymbol{\varrho}) := \arg\max_{\boldsymbol{\omega}} \mathcal{L}_{\text{adv}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\varsigma}, \boldsymbol{\omega}, \boldsymbol{\varrho}), \end{aligned} \tag{12}$$

where the leader objective $\widetilde{\mathcal{L}}_{\text{total}}$ is the redefined upper-level objective function with best-response (i.e., $\hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma}), \hat{\boldsymbol{\omega}}(\boldsymbol{\varrho})$), and the follower objective $\widetilde{\mathcal{L}}_{\text{adv}}$ is the lower-level objective function w.r.t. $\boldsymbol{\omega}$. $S(\boldsymbol{\varsigma})$ and $S(\boldsymbol{\varrho})$ are the solution sets w.r.t. given $\boldsymbol{\varsigma}$ and $\boldsymbol{\varrho}$, respectively.[3] Intuitively, it is not symmetric in terms of two levels of learning process $\mathcal{L}_{\text{total}}$ and $\mathcal{L}_{\text{adv}}$, in which the latter is served as a constraint on the former for obtaining the optimal feedback transmitted to the leader objective.

Algorithm 1 Bi-level Implicit Adversarial (BIA).

Require: Initialization $\boldsymbol{\varsigma}^0, \boldsymbol{\omega}^0$, $\mathcal{E}$ parameterized by $\boldsymbol{\varsigma}$, discriminator D parameterized by $\boldsymbol{\omega}$, and necessary hyper-parameters ($\boldsymbol{\alpha}_l$ and $\boldsymbol{\alpha}_u$).

repeat
  for *not converged and* $1 \le t \le T$ **do**
    % *Lower-level variable probe.*
    Update $\boldsymbol{\omega}$ to obtain approximation $\hat{\boldsymbol{\omega}}$.
    $\hat{\boldsymbol{\omega}}(\boldsymbol{\omega}) \leftarrow \boldsymbol{\omega} - \boldsymbol{\alpha}_l \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\text{adv}}(\boldsymbol{\omega}, \boldsymbol{\varsigma}^t)$.
    Calculate$\nabla_{\boldsymbol{\varsigma}} \mathcal{L}_{\text{adv}}$, $\nabla_{\hat{\boldsymbol{\omega}}} \mathcal{L}_{\text{adv}}$, and $\nabla_{\hat{\boldsymbol{\omega}}} \widetilde{\mathcal{L}}_{\text{total}}$.
    Calculate $\boldsymbol{G}_{\boldsymbol{BR}}$ by Eq. (17) with $\hat{\boldsymbol{\omega}}$ and current $\boldsymbol{\varsigma}^t$.
    % *Upper-level variable probe.*
    Calculate $\nabla_{\boldsymbol{\varsigma}} \widetilde{\mathcal{L}}_{\text{total}}(\boldsymbol{\varsigma}^t)$ by Eq. (13) .
    Update $\boldsymbol{\varsigma}^{t+1} \leftarrow \boldsymbol{\varsigma}^t - \boldsymbol{\alpha}_u \nabla_{\boldsymbol{\varsigma}} \widetilde{\mathcal{L}}_{\text{total}}(\boldsymbol{\varsigma}^t)$.
    $t \leftarrow t + 1$.
  end for
until *training convergence.*
Output: *Optimal enhancer* $\mathcal{E}$ with $\boldsymbol{\varsigma}^*$.

3.4.2. Optimization algorithm

Based on a bi-level optimization reformulation, we introduce the best-response gradient algorithm [33,34]. The detailed algorithm flow regarding variable updates is illustrated in Fig. 3. We compute the gradient of the leader objective $\widetilde{\mathcal{L}}_{\text{total}}$ in terms of $\boldsymbol{\varsigma}$, formulated as[4]

$$\nabla_{\boldsymbol{\varsigma}} \widetilde{\mathcal{L}}(\boldsymbol{\varsigma}) = \overbrace{\nabla_{\boldsymbol{\varsigma}} \widetilde{\mathcal{L}}(\boldsymbol{\varsigma}, \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma}))}^{\text{direct gradient}} + \overbrace{\nabla_{\boldsymbol{\omega}} \widetilde{\mathcal{L}}(\boldsymbol{\varsigma}, \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma})) \nabla_{\boldsymbol{\varsigma}} \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma})}^{\boldsymbol{G}_{\boldsymbol{BR}}\text{: best-response gradient}}. \tag{13}$$

Considering that implicit function theorem can precisely estimate the best-response gradient, i.e., $\partial \mathcal{L}_{\text{adv}} / \partial \boldsymbol{\omega} = 0$, then $\nabla_{\boldsymbol{\varsigma}} \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma})$ is further derived as

$$\nabla_{\boldsymbol{\varsigma}} \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma}) = -\left[\nabla^2_{\boldsymbol{\omega}} \mathcal{L}_{\text{adv}}(\boldsymbol{\varsigma}, \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma}))\right]^{-1} \nabla^2_{\boldsymbol{\omega}\boldsymbol{\varsigma}} \mathcal{L}_{\text{adv}}(\boldsymbol{\varsigma}, \hat{\boldsymbol{\omega}}(\boldsymbol{\varsigma})). \tag{14}$$

To avoid directly calculating the products of various Hessians and their inversions, we further introduce a linear solver system $\mathbf{B}$ based on Eq. (14), reformulated as

$$\boldsymbol{G}_{\boldsymbol{BR}} = \left[\nabla^2_{\boldsymbol{\omega}\boldsymbol{\varsigma}} \mathcal{L}_{\text{adv}}\right]^{\top} \mathbf{B}, \text{ where } \left[\nabla^2_{\boldsymbol{\omega}\boldsymbol{\omega}} \mathcal{L}_{\text{adv}}\right] \mathbf{B} = -\nabla_{\boldsymbol{\omega}} \widetilde{\mathcal{L}}_{\text{total}}. \tag{15}$$

Here, $(\cdot)^{\top}$ denotes the transposition operation. To further suppress the complexity of two Hessian matrix $\nabla^2_{\boldsymbol{\omega}\boldsymbol{\omega}} \mathcal{L}_{\text{adv}}$ and $\nabla^2_{\boldsymbol{\omega}\boldsymbol{\varsigma}} \mathcal{L}_{\text{adv}}$, we consider replacing the original Hessian operation and introduce two approximations using corresponding outer products, as follows:

$$\nabla^2_{\boldsymbol{\omega}\boldsymbol{\omega}} \mathcal{L}_{\text{adv}} \approx \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\text{adv}} \nabla_{\boldsymbol{\omega}} \mathcal{L}^{\top}_{\text{adv}}, \ \nabla^2_{\boldsymbol{\omega}\boldsymbol{\varsigma}} \mathcal{L}_{\text{adv}} \approx \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\text{adv}} \nabla_{\boldsymbol{\varsigma}} \mathcal{L}^{\top}_{\text{adv}}. \tag{16}$$

By combining Eqs. (15)–(16), the response gradient $\boldsymbol{G}_{\boldsymbol{BR}}$ is approximately obtained, i.e.,

$$\boldsymbol{G}_{\boldsymbol{BR}} \approx -\nabla_{\boldsymbol{\varsigma}} \mathcal{L}_{\text{adv}} \left[\left(\nabla_{\boldsymbol{\omega}} \mathcal{L}^{\top}_{\text{adv}} \nabla_{\boldsymbol{\omega}} \widetilde{\mathcal{L}}_{\text{total}}\right) / \left(\nabla_{\boldsymbol{\omega}} \mathcal{L}^{\top}_{\text{adv}} \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\text{adv}}\right)\right]. \tag{17}$$

[2] Refer to Section 4.4 for experimental supports.

[3] Referring the bi-level optimization for the solution set $S(\boldsymbol{\varsigma})$ and $S(\boldsymbol{\varrho})$ [33, 34].

[4] For convenience, we do not distinguish the optimization of the two variables $\{\boldsymbol{\varsigma}, \boldsymbol{\varrho}\}$ in the leader objective, which is abbreviated as the variable $\boldsymbol{\varsigma}$.

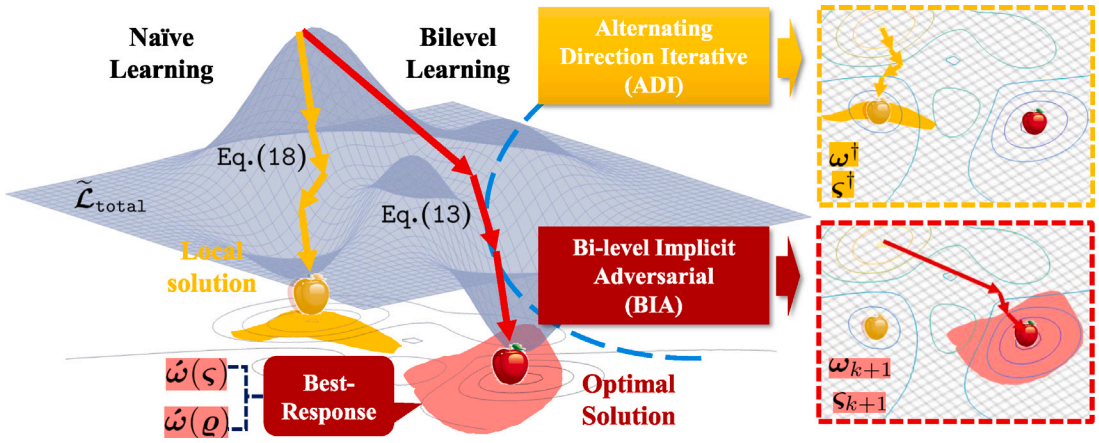

Fig. 6. Illustrating a comparison of two different learning strategies (i.e., ADI and BIA), while also showing the iterative trajectories of the variables corresponding to these two strategies. Dashed boxes depict the convergence of the two iterative programs to local solution and global optimal solution, respectively.

We summarize the BIA learning strategy in Alg. 1. In the training process, with the current parameters ς, we first optimize ω according to the objective $\mathcal{L}_{\text{adv}}$ in several steps to approximate the best-response, i.e., $\hat{\omega}(\varsigma) \approx \acute{\omega}(\varsigma)$. Then, $\acute{\omega}(\varsigma)$ is back-propagated to $\widetilde{\mathcal{L}}_{\text{total}}$ and calculate $\boldsymbol{G}_{\boldsymbol{BR}}$ is calculated based Eq. (17).

Remark 1. Naive Alternative Learning vs. Bi-level Implicit Adversarial Learning. As illustrated in Fig. 6, we compare the differences between two distinct training strategies from an optimization perspective. From the modeling aspect, the traditional alternating iteration scheme, based on the single-level minimax optimization model $\min_{\varsigma} \max_{\omega} \mathcal{L}_{\text{total}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \boldsymbol{\Theta})$, optimizes two parallel sub-problems in a disjointed manner. In each update, it fixes one set of parameters to update the other, following the iterative routine:

$$\omega^{\dagger} = \omega + \eta \nabla_{\omega} \mathcal{L}_{\text{total}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \omega, \varsigma, \varrho), \quad \varsigma^{\dagger} = \varsigma - \nu \nabla_{\varsigma} \mathcal{L}_{\text{total}}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{y}; \omega, \varsigma, \varrho), \tag{18}$$

where η and ν represent the learning rates for $\mathcal{E}$ and $\mathcal{D}$, respectively. This iterative optimization process only computes the direct gradients and overlooks the calculation of indirect gradients in Eq. (13), leading to training oscillations and inaccurate gradient feedback. In contrast, we have designed Alg. 1, which precisely computes the best-response gradients $\boldsymbol{G}_{\boldsymbol{BR}}$ using Eq. (17). The latter implies that the involved strategies accurately capture the impact of ω on ς in every step of the optimization. We have extensively validated the effectiveness of the BIA in the experimental section (refer to Section 4.4).

4. Experimental results

4.1. Experimental setting

Implementation Detail. All the experiments are conducted on a PC with an NVIDIA GeForce GTX 2080Titan GPU in the PyTorch 1.8.0 framework. We train our model with Adam optimizer for a total of 100k iterations. The learning rate was set to 2×10^{-3} and then decayed by a factor of 2 at the 70k-th, 75k-th iterations. The training mini-batch size is set to 16.

Benchmark and Metric. We evaluate the performance of our proposed method on two widely used datasets-RELLISUR.[5] and DarkFace[6] All these datasets have no overlap with our training dataset. We use the real-word RELLISUR dataset for training our *CollaBA* and other state-of-art methods. It contains data at ×1, ×2 and ×4 of different resolutions, the training set contains 3610 pairs of data at each resolution, and the test set contains 425 pairs of data at each resolution. During training, we resize the normal light and high resolution images to 160^2 as ground truth. The corresponding patches from low light low resolution pairs are resized to as 80 × 80 and 40 × 40 for ×2 and ×4 joint degradation removal tasks. As for testing, we construct two different real datasets (i.e., *RELLISUR-Test,* and *DarkFace-Test*) with distinct sources.

For the evaluation, we employ six widely-used full-reference metrics, including Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Root Mean Square Error (RMSE), Feature-based Similarity Index (FSIM) and Signal to Reconstruction Error ratio (SRE). We also introduce two no-reference quality assessment (i.e., EME and LOE) to evaluate the effectiveness.

Comparisons with State-of-the-Art. *1) Normal-light SR methods:* We compare our *CollaBA* with several state-of-the-art super-resolution methods, including ESRGAN [35], RDN [36], RCAN [37], SRFBN [27], PAN [38], SRResNet [39], MIRNet [30], SwinIR [29], Restormer [15], SRFormer [14] and HAT [13]. All these methods have been re-trained on RELLISUR for fair comparisons. *2) Cascaded LLE↦SR methods:* In addition, we conduct comparative experiments between cascaded SR and LLE methods. To be specific, we selected the most recent SR method, HAT [13], as an exemplar of super-resolution techniques and integrated it into a cascade behind three LLE methods: ZeroDCE [22], SCI [12], and LLFormer [2]. In this context, LLFormer represents the current state-of-the-art supervised LLE method. For the unsupervised LLE methods, ZeroDCE and SCI, we performed a dedicated retraining process exclusively on the ×1 low-light RELLISUR dataset. In contrast, HAT, serving as the subsequent super-resolution network, underwent retraining on both ×2 and ×4 normal-light datasets. We denote this modified version as $\text{HAT}_{\ddagger}$ to distinguish it from the standard low-light data configuration.

4.2. Experimental evaluation

Quantitative Evaluation.

Table 1 reports the numerical scores achieved in the ×2 and ×4 tasks using the authentic *RELLISUR-Test* dataset. It is evident that our approach surpasses the current state-of-the-art methods significantly, as demonstrated by various numerical metrics, securing the top position in all evaluation criteria for the ×2 task. For the ×2 task, our method achieves a remarkable PSNR improvement of 0.406 dB and a substantial LPIPS improvement of 35.8%. In the case of the ×4 task,

[5] https://vap.aau.dk/rellisur/
[6] https://flyywh.github.io/CVPRW2019LowLight/

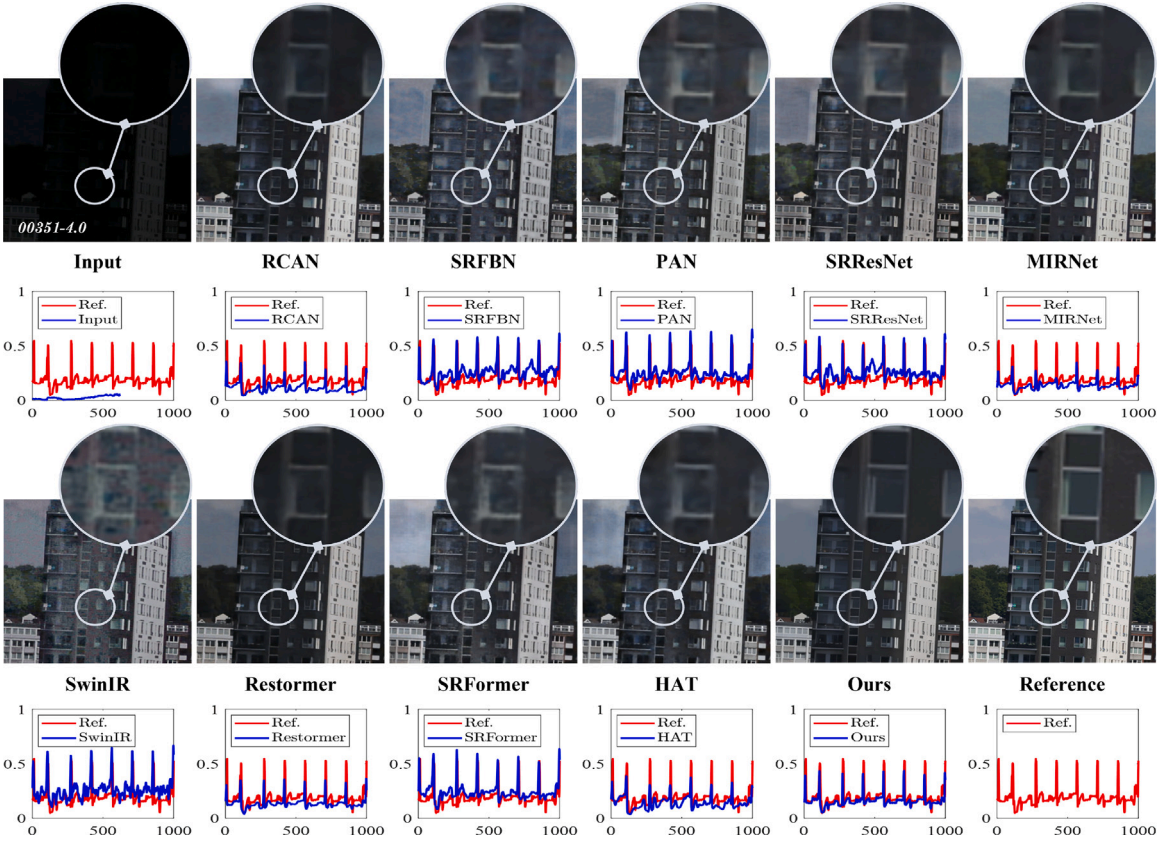

Fig. 7. Qualitative comparisons on *RELLISUR-Test* samples with scale factor of ×2 task. The image size scales from an input of 624 × 624 to an output of 1248 × 1248. The following signal maps illustrate the variations in pixel intensity between the generated images and the reference image along the randomly selected line segment. For ease of viewing, we display input images and the results of various methods at an equal size. *(Best viewed with zoom.)*

Table 1
Quantitative comparison on ***RELLISUR-Test*** dataset for ×2 and ×4 tasks (i.e., ×2 / ×4). Quantitative results in terms of six full-reference metrics including **PSNR**, **SSIM**, **LPIPS**, **RMSE**, **FSIM**, and **SRE**. Notice that all methods are retrained on *RELLISUR*. The top-ranked and the second-ranked method are highlighted in red bold and blue bold, respectively.

Metrics	ESRGAN[18ECCV]	RDN[18CVPR]	RCAN[18ECCV]	SRFBN[19CVPR]	PAN[20ECCV]	SRResNet[20ECCV]
PSNR↑	18.087/17.183	18.795/18.213	19.765/19.078	18.421/17.679	18.783/18.106	18.156/17.594
SSIM↑	0.655/0.647	0.701/0.703	0.712/0.713	0.662/0.665	0.693/0.700	0.667/0.684
LPIPS↓	0.300/0.471	0.455/0.584	0.426/0.550	0.510/0.640	0.450/0.559	0.451/0.581
RMSE↓	0.135/0.149	0.120/0.128	0.110/0.119	0.125/0.136	0.119/0.129	0.128/0.137
FSIM↑	0.873/0.858	0.874/0.866	0.881/0.874	0.847/0.836	0.867/0.859	0.848/0.841
SRE↑	55.523/58.306	55.865/58.776	56.372/59.213	55.678/58.514	55.852/58.713	55.524/58.457
EME↑	7.733/3.549	5.361/3.083	5.067/2.955	5.805/3.672	5.381/3.479	5.318/3.098
LOE↓	45.468/43.976	45.642/44.976	48.852/47.254	47.698/45.582	49.793/48.623	48.283/44.697
Metrics	MIRNet[20ECCV]	SwinIR[21ICCV]	Restormer[22CVPR]	SRFormer[23CVPR]	HAT[23CVPR]	CollaBA (Ours)
PSNR↑	21.053/19.783	18.386/17.534	**21.215/20.298**	19.556/18.723	20.213/19.751	**21.621/20.423**
SSIM↑	0.720/0.704	0.640/0.663	**0.727/0.720**	0.704/0.705	0.719/0.715	**0.787/0.734**
LPIPS↓	0.436/0.599	0.577/0.688	**0.385/0.492**	0.469/0.613	0.454/0.561	**0.247/0.371**
RMSE↓	0.095/0.109	0.125/0.139	**0.095/0.106**	0.110/0.121	0.103/0.110	**0.073/0.089**
FSIM↑	**0.889**/**0.878**	0.845/0.840	0.892/0.885	0.877/0.869	0.882/**0.873**	**0.896**/**0.873**
SRE↑	57.065/**59.635**	55.643/58.432	**57.096**/**59.824**	56.237/59.023	56.584/59.571	**57.831**/59.817
EME↑	**7.879**/**3.841**	6.773/3.346	7.645/**3.986**	7.513/3.462	7.847/3.543	**7.972**/3.794
LOE↓	30.968/33.966	46.042/42.581	**30.889**/**30.856**	50.411/45.606	42.265/35.284	**29.371**/**30.012**

Table 2
Quantitative comparison among cascaded LLE↦SR methods on *RELLISUR* dataset (i.e., ×2 and ×4 tasks (i.e., ×2/×4)). † indicates training on ×1 low-light RELLISUR dataset for LLE. ‡ denotes training on ×2 or ×4 *RELLISUR* for normal-light SR.

LLE↦SR Method	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	EME↑	LOE↓
ZeroDCE$^{21TPAMI}_{\dagger}$ ↦HAT$^{23CVPR}_{\ddagger}$	12.927/12.524	0.354/0.321	0.698/0.739	0.194/0.197	7.693/3.684	56.726/54.178
SCI$^{22CVPR}_{\dagger}$ ↦HAT$^{23CVPR}_{\ddagger}$	14.963/14.776	0.439/0.452	0.591/0.697	0.200/0.205	7.521/3.348	36.413/37.879
LLFormer$^{23AAAI}_{\dagger}$ ↦HAT$^{23CVPR}_{\ddagger}$	21.218/20.135	0.720/0.718	0.455/0.575	0.093/0.105	7.904/3.721	35.084/36.215

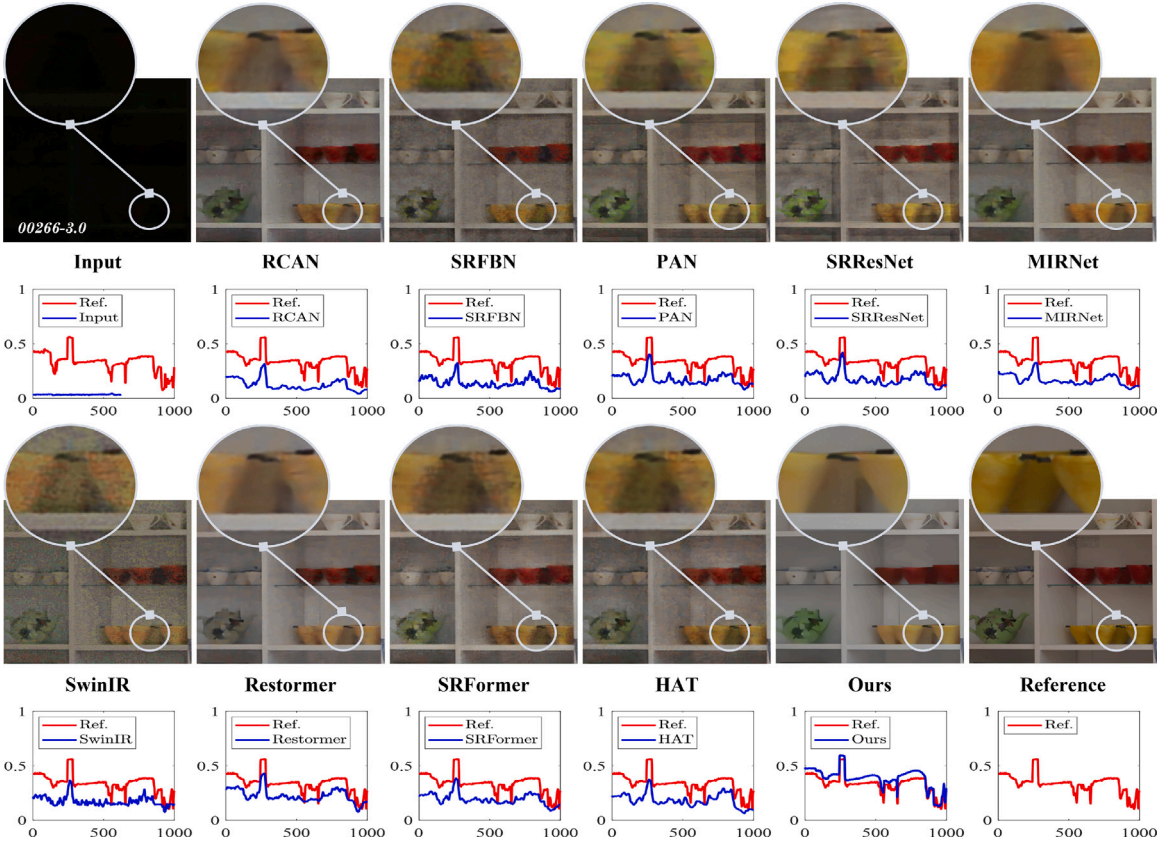

Fig. 8. Qualitative comparisons on *RELLISUR-Test* samples with scale factor of ×4 task. The image size scales from an input of 624 × 624 to an output of 2496 × 2496. The following signal maps illustrate the variations in pixel intensity between the generated images and the reference image along the randomly selected line segment. For ease of viewing, we display input images and the results of various methods at an equal size. *(Best viewed with zoom.)*

Table 3
Quantitative comparison on *DarkFace* dataset for ×2 and ×4 tasks (i.e., ×2/×4). Quantitative results in terms of two no-reference metrics including **EME** and **LOE**. Notice that all methods are retrained on *RELLISUR*. The top-ranked and the second-ranked method are highlighted in red bold and blue bold, respectively.

Metrics	ESRGAN[18 *ECCV*]	RDN[18 *CVPR*]	RCAN[18 *ECCV*]	SRFBN[19 *CVPR*]	PAN[20 *ECCV*]	SRResNet[20 *ECCV*]
EME↑	15.464/10.911	11.025/7.637	14.657/10.911	8.010/5.824	11.167/7.094	10.522/6.492
LOE↓	1.874/1.980	1.942/1.990	1.898/1.980	1.891/2.007	1.898/1.995	1.926/2.007
Metrics	MIRNet[20 *ECCV*]	SwinIR[21 *ICCV*]	Restormer[22 *CVPR*]	SRFormer[23 *CVPR*]	HAT[23 *CVPR*]	CollaBA (Ours)
EME↑	27.579/16.900	16.820/13.124	**30.453**/**21.156**	27.125/15.689	28.225/16.341	**32.710**/**21.086**
LOE↓	1.897/2.014	1.900/**1.975**	1.893/**1.975**	**1.865**/2.214	1.986/2.012	**1.846**/**1.965**

it still attains a PSNR improvement of 0.125 dB and a notable LPIPS improvement of 32.6%.

Furthermore, we present quantitative results for cascaded methods in Table 2. These scores clearly indicate that the cascade approach does not yield enhancements in joint task performance, particularly when employing unsupervised LLE methods (i.e., ZeroDCE and SCI), resulting in deteriorated enhancement outcomes after upscaling. Lastly, we provide quantitative comparisons using the genuine low-light human face dataset, *DarkFace-Test*, in Table 3. It is evident that our method outperforms existing approaches across nearly all no-reference metrics, reaffirming the robust generalization of our method in real-world, highly low-light scenarios.

Qualitative Evaluation. Figs. 7 and 8 present the visualization outcomes for the ×2 and ×4 tasks, respectively, using the authentic *RELLISUR-Test* dataset. Upon close observation of the locally magnified regions, it becomes apparent that a variety of normal-light SR methods fall short of attaining the anticipated visual quality. These methods exhibit pronounced noise, blurriness, and artifacts, as evident in approaches such as SwinIR and SRFormer, both afflicted with unsatisfactory blurry textures. In stark contrast, our approach excels in delivering the most natural and authentic visual quality, characterized by vibrant colors, pleasing brightness, and notably exceptional proficiency in the recovery of high-frequency structural details. Moreover, the signal maps depicting pixel intensity levels indicate that our approach consistently approaches the reference image.

Figs. 9 and 10 showcase the visualization results on the *DarkFace-Test* dataset. Likewise, it is apparent that our approach surpasses in the preservation of fine structures and natural color fidelity, presenting delightful luminosity while faithfully retaining realistic texture details.

Computational Efficiency. To examine model efficiency, we reported the *Parameters* and *FLOPs* of some recent methods in Table 4. We conduct the measures with images of size 128 × 128 on a single 2080Titan GPU. Note that, our proposed method could maintain a good balance of performance and computational efficiency, and is relatively competitive in terms of FLOPs.

4.3. Nighttime image semantic segmentation

Last but not least, we also validate the facilitation effects for the nighttime visual perception tasks. Specifically, we utilize the semantic segmentation model SAM [40] to investigate how different approaches

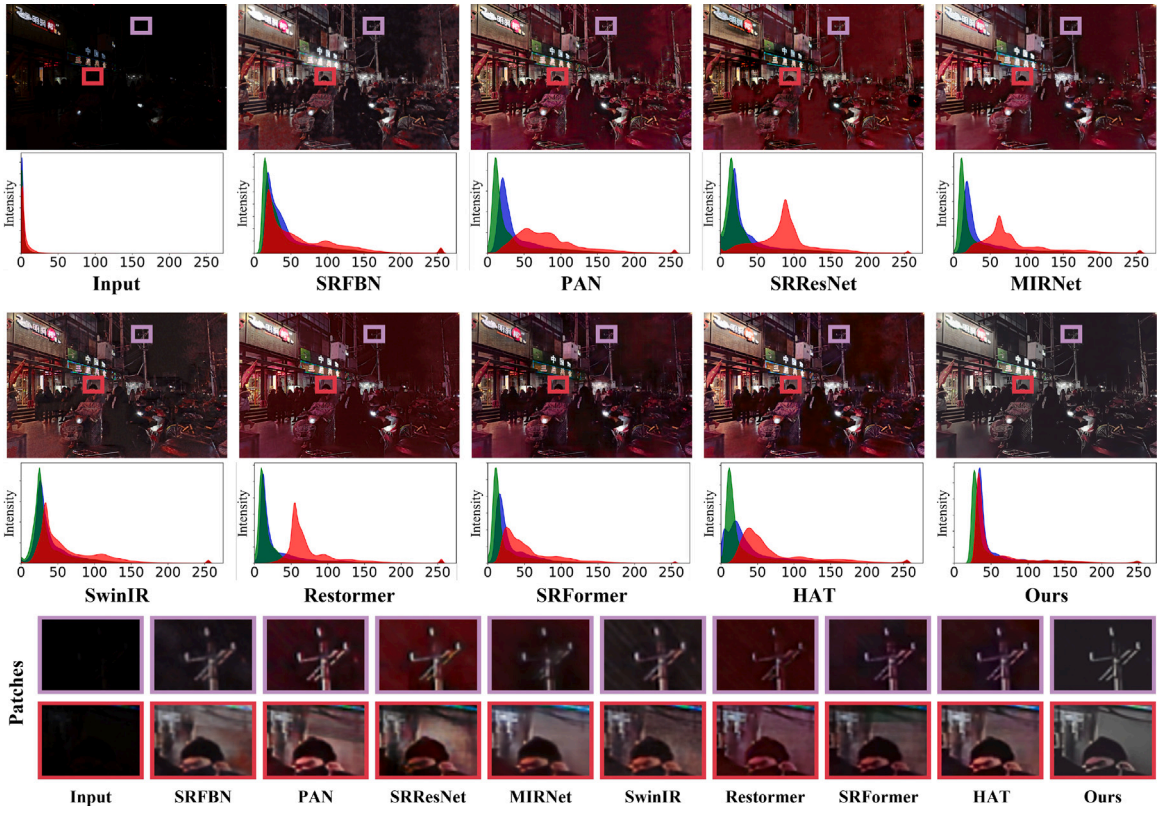

Fig. 9. Qualitative comparisons on real-world nighttime *DarkFace-Test* for ×2 super-resolution task. For ease of viewing, we display input images and the enhanced results of various methods at an equal size. *(Best viewed with zoom.)*

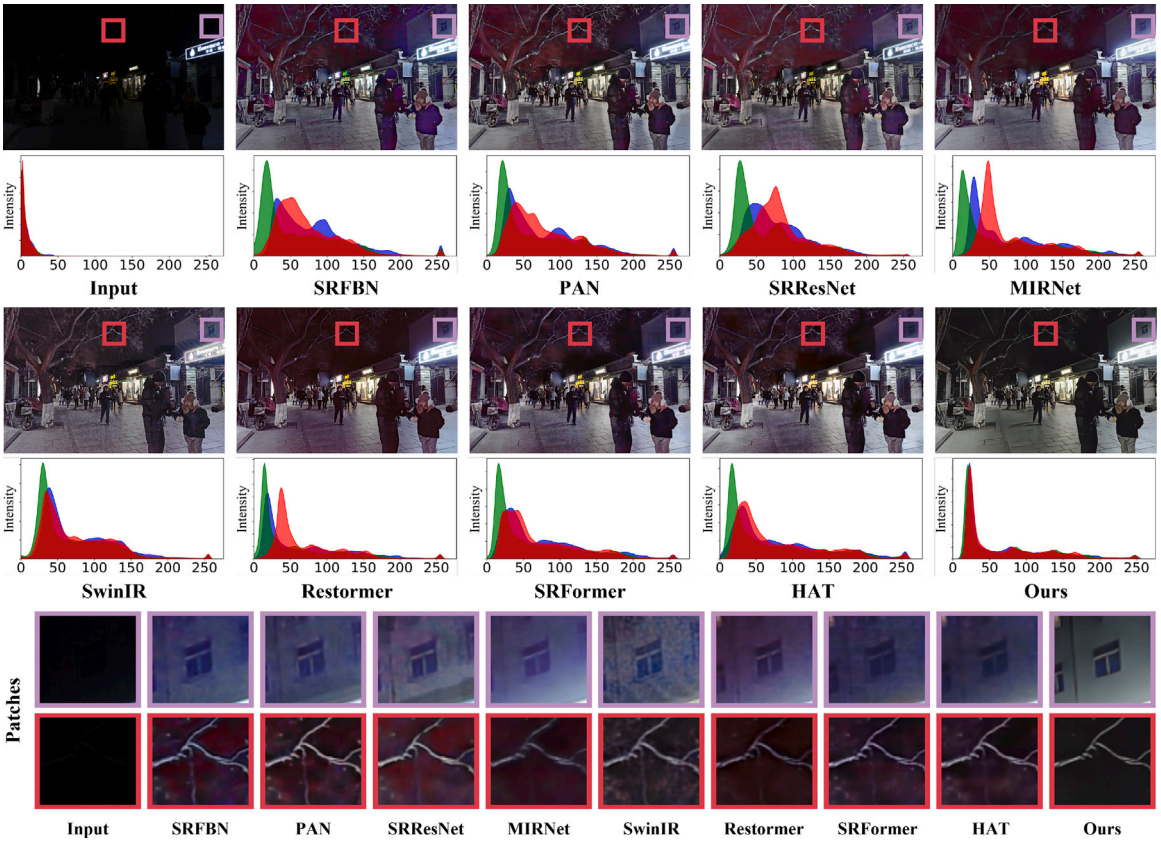

Fig. 10. Qualitative comparisons on real-world nighttime *DarkFace-Test* for ×4 super-resolution. For ease of viewing, we display input images and the enhanced results of various methods at an equal size. *(Best viewed with zoom.)*

Table 4
Quantitative comparison in terms of model size, FLOPs, and inference time among various methods. Notice that all methods are retrained on RELLISUR-Test. We conduct the measures for *RELLISUR-Test* dataset with images of size 128 × 128 on a single 2080Titan GPU. The best method are highlighted in bold.

Metrics	RCAN[18ECCV]	SwinIR[21ICCV]	Restormer[22CVPR]	SRFormer[23CVPR]	HAT[23CVPR]	CollaBA (Ours)
Params (MB)↓	15.444/15.592	11.683/11.825	26.126/26.209	10.162/10.220	**9.473/9.621**	13.453/13.625
FLOPs (G)↓	251.002/261.006	57.154/67.158	35.375/41.327	81.797/83.006	58.990/68.994	**22.059/66.872**
Inference (S)↓	0.047/0.049	3.879/4.226	0.033/0.035	0.218/0.221	0.184/0.186	**0.031/0.032**

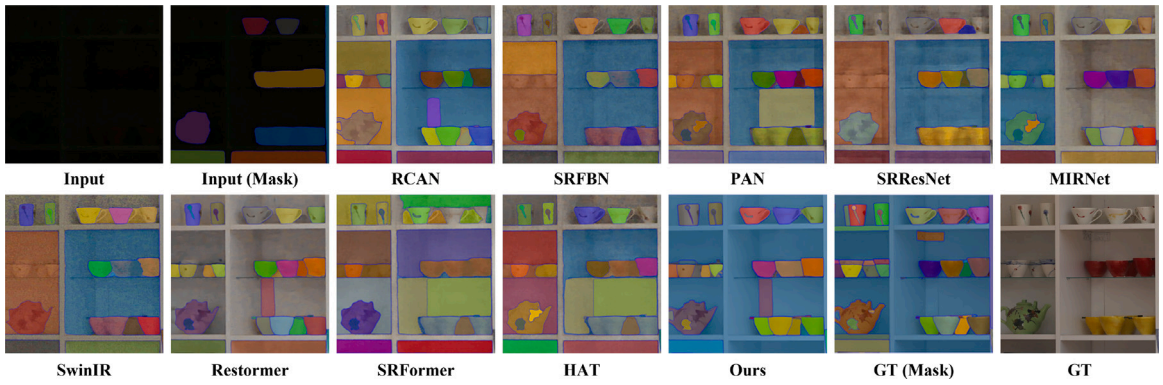

Fig. 11. Qualitative comparisons of various methods' visualized semantic segmentation results on the large-scale SAM model. The recovery results for various methods for the selected sample (i.e., 00266-3.0) are shown in Fig. 8. It can be observed that our segmentation results are the closest to the reference image.

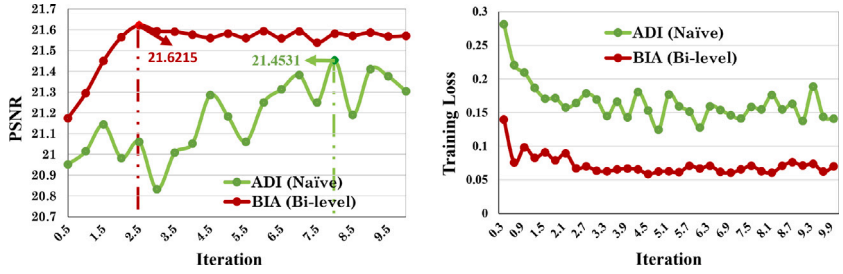

Fig. 12. Comparison of PSNR score (Left), and training loss (Right) in terms of the iteration ($\times 10^4$) under two different strategies, i.e., "Naive Learning Strategy (i.e., ADI)" and "Bi-level Learning Strategy (i.e., BIA)". Our strategy is to achieve peak performance quickly and steadily (i.e., *0.1684* dB PSNR improvements with *31.25%* iterations).

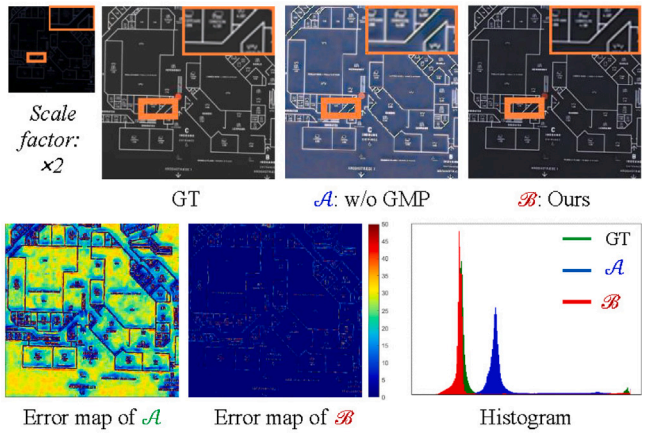

Fig. 13. Ablation study of the effect of GMP. The result w/o GMP tends to generate unnatural-looking results (e.g., over-exposure, distorted color, unpleasant artifacts).

are beneficial to the downstream high-level tasks. We display the qualitative comparison and a quantitative comparison in Fig. 11. In comparison, the proposed method obtains more accurate segmentation results.

4.4. Ablation study

We perform several ablation studies to demonstrate the effectiveness of each component (i.e., GMP, BIA and different losses) of *CollaBA* as follows.

Explorations on GMP and MSMU. Table 5 in [***No. 0***-***No. 2***] demonstrates the effectiveness of the GMP module in the absence of BIA. Comparing scenarios [***No. 0***] and [***No. 2***], it is observed that removing GMP results in a decrease across three metrics, with PSNR dropping by 0.468dB/0.317 dB for ×2 and ×4 tasks, respectively. In Fig. 13, when the GMP branch is discarded without adaptive exposure level, i.e., only keep the degradation removal module, the restored images fails to recover the normal color state. Severe color casts and overexposed whitening emerge between adjacent regions (i.e., $\mathcal{A}$: w/o GMP). A performance drop in [***No. 3***] is observed if we do not use the operation $\texttt{CD} - \texttt{CFT}$ but with operation $\texttt{Concat}$ (lower PSNR and higher LPIPS), demonstrating the positive impact of $\texttt{CD} - \texttt{CFT}$. Comparing cases [***No. 0***] and [***No. 4***], it is evident that removing the MSMU leads to a decrease in PSNR by 0.249dB/0.234 dB for ×2 and ×4 tasks, respectively, compared to a simple upsampling layer (e.g., Bilinear).

Effects of the BIA Strategy. We have investigated the proposed BIA learning mechanism as a replacement for the simplistic alternating learning approach. The corresponding quantitative scores are available in Table 5 in [***No. 1***], demonstrating noteworthy performance enhancements of 0.168 dB and 0.211 dB in terms of PSNR for the ×2 and ×4 tasks, respectively. Fig. 12 further illustrates that our BIA method maintained a more stable training process and achieved peak performance

Table 5
Ablation study of the proposed modules (i.e., GMP and MSMU) and BIA learning strategy for ×2 and ×4 tasks (i.e., ×2/×4). For convenience, it should be noted that our experiments from groups ***No. 2*** to ***No. 4*** were conducted without the implementation of the BIA strategy.

No.	*Configuration*	PSNR↑	SSIM↑	LPIPS↓
0)	CollaBA	**21.621/20.423**	**0.787/0.734**	**0.247/0.371**
1)	w/o BIA	$21.453_{0.168\downarrow}$/$20.212_{0.211\downarrow}$	$0.759_{0.028\downarrow}$/$0.718_{0.016\downarrow}$	$0.283_{0.036\uparrow}$/$0.389_{0.018\uparrow}$
2)	w/o GMP	$21.153_{0.468\downarrow}$/$20.106_{0.317\downarrow}$	$0.746_{0.041\downarrow}$/$0.705_{0.029\downarrow}$	$0.291_{0.044\uparrow}$/$0.411_{0.040\uparrow}$
3)	w/o CD – CFT (w/ Concat)	$21.342_{0.279\downarrow}$/$20.130_{0.293\downarrow}$	$0.761_{0.026\downarrow}$/$0.712_{0.022\downarrow}$	$0.281_{0.034\uparrow}$/$0.396_{0.025\uparrow}$
4)	w/o MSMU (w/ Bilinear)	$21.372_{0.249\downarrow}$/$20.189_{0.234\downarrow}$	$0.763_{0.024\downarrow}$/$0.715_{0.019\downarrow}$	$0.264_{0.017\uparrow}$/$0.385_{0.014\uparrow}$

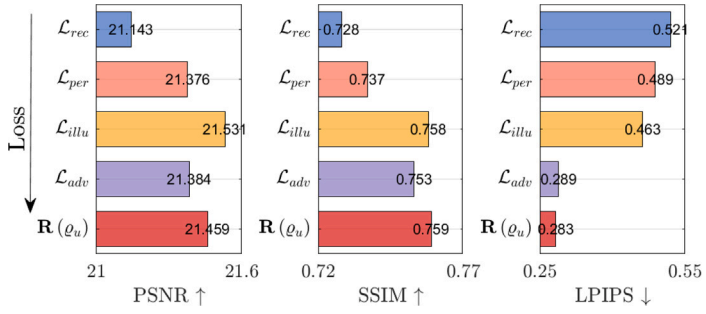

Fig. 14. Ablation study of different losses (w/o BIA). The arrow indicates losses are added in increments.

more rapidly. For instance, it attained the highest PSNR of 21.621 with only 25% of the iterations compared to the alternating strategy. Furthermore, the loss curve exhibited a smoother descent throughout the entire training process, in stark contrast to the oscillations observed in the alternating strategy.

Contribution of Each Loss. Fig. 14 presents the results of loss ablation experiments, with a specific focus on the ×2 task. The direction of the arrows signifies the gradual incorporation of various loss components. It is noteworthy that the table reveals score improvements achieved through $\mathcal{L}_{per}$, amounting to 0.24 dB for PSNR and 0.032 for the LPIPS metric, respectively. We observe a substantial enhancement in performance across all metrics following the inclusion of the $\mathcal{L}_{illu}$ loss. Particularly noteworthy is the profound impact of introducing the $\mathcal{L}_{adv}$ loss on the perceptual score LPIPS, resulting in a significant reduction of 0.232. Furthermore, the ablation experiments validate the positive influence of imposing constraints on hyper-parameters (i.e., $\mathbf{R}(\varrho_u)$) on the ultimate performance.

5. Conclusion

This research delves into the relatively uncharted area of super-resolution in low-lighting conditions. Utilizing a dual-path modulated-interactive enhancer, we have crafted a generative modulation prior for use as guidance information, coupled with an interactive degradation removal branch to enhance details and suppress artifacts. Furthermore, our novel learning mechanism acts as a training strategy to augment performance and foster stable training. This study represents significant progress in the field, adeptly addressing the complexities of super-resolution tasks in challenging low-lighting conditions. However, our method still faces limitations in terms of lightweight design and automated learning, particularly when addressing the more complex challenges of the real world, such as dynamically changing environmental conditions, diverse data distributions, and integration with downstream high-level perception tasks. In future work, we aim to develop more efficient learning strategies (i.e., multi-level optimization and meta-learning) and more lightweight network architectures (i.e., model pruning, and distillation) tailored for further advancements in various extreme and adverse scenarios.

CRediT authorship contribution statement

Jiaxin Gao: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Yaohua Liu:** Writing – review & editing, Visualization, Validation, Investigation, Formal analysis, Data curation. **Ziyu Yue:** Visualization, Validation, Investigation, Conceptualization. **Xin Fan:** Supervision, Resources, Project administration, Funding acquisition. **Risheng Liu:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (No. 2022YFA1004101), the National Natural Science Foundation of China (Nos. U22B2052, 62302078 and 61936002) and the Liaoning Revitalization Talents Program (No. 2022RG04).

References

[1] Z. Wang, J. Chen, S.C. Hoi, Deep learning for image super-resolution: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3365–3387.
[2] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, T. Lu, Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method, in: AAAI, vol. 37, (3) 2023, pp. 2654–2662.
[3] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, C.C. Loy, Low-light image and video enhancement using deep learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2021) 9396–9416.
[4] D. Cheng, L. Chen, C. Lv, L. Guo, Q. Kou, Light-guided and cross-fusion U-net for anti-illumination image super-resolution, IEEE Trans. Circuits Syst. Video Technol. 32 (12) (2022) 8436–8449.
[5] F. Zhou, X. Sun, J. Dong, X.X. Zhu, SurroundNet: Towards effective low-light image enhancement, Pattern Recognit. 141 (2023) 109602.
[6] R. Cai, Z. Chen, Brain-like retinex: A biologically plausible retinex algorithm for low light image enhancement, Pattern Recognit. 136 (2023) 109195.

[7] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, X. Fan, Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion, Int. J. Comput. Vis. (2023) 1–28.
[8] Z. Yue, J. Gao, Z. Su, Unveiling details in the dark: Simultaneous brightening and zooming for low-light image enhancement, in: AAAI, vol. 38, (7) 2024, pp. 6899–6907.
[9] J. Liu, J. Shang, R. Liu, X. Fan, Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion, IEEE Trans. Circuits Syst. Video Technol. 32 (8) (2022) 5026–5040.
[10] J. Gao, X. Liu, R. Liu, X. Fan, Learning adaptive hyper-guidance via proxy-based bilevel optimization for image enhancement, Vis. Comput. 39 (4) (2023) 1471–1484.
[11] A. Aakerberg, K. Nasrollahi, T.B. Moeslund, RELIEF: Joint low-light image enhancement and super-resolution with transformers, in: Scandinavian Conference on Image Analysis, 2023, pp. 157–173.
[12] L. Ma, T. Ma, R. Liu, X. Fan, Z. Luo, Toward fast, flexible, and robust low-light image enhancement, in: Computer Vision and Pattern Recognition, 2022, pp. 5637–5646.
[13] X. Chen, X. Wang, J. Zhou, Y. Qiao, C. Dong, Activating more pixels in image super-resolution transformer, in: Computer Vision and Pattern Recognition, 2023, pp. 22367–22377.
[14] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, Q. Hou, Srformer: Permuted self-attention for single image super-resolution, in: Computer Vision and Pattern Recognition, 2023.
[15] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.
[16] E.H. Land, J.J. McCann, Lightness and retinex theory, Josa 61 (1) (1971) 1–11.
[17] L. Ma, D. Jin, N. An, J. Liu, X. Fan, Z. Luo, R. Liu, Bilevel fast scene adaptation for low-light image enhancement, Int. J. Comput. Vis. (2023) 1–19.
[18] J.J. Jeon, J.Y. Park, I.K. Eom, Low-light image enhancement using Gamma correction prior in mixed color spaces, Pattern Recognit. (2023) 110001.
[19] C. Wang, J. Pan, W. Wang, J. Dong, M. Wang, Y. Ju, J. Chen, PromptRestorer: A prompting image restoration method with degradation perception, in: Neural Information Processing Systems, vol. 36, 2024.
[20] K. Jiang, Z. Wang, Z. Wang, C. Chen, P. Yi, T. Lu, C.-W. Lin, Degrade is upgrade: Learning degradation for low-light image enhancement, in: AAAI, vol. 36, (1) 2022, pp. 1078–1086.
[21] W. Yang, S. Wang, Y. Fang, Y. Wang, J. Liu, From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement, in: Computer Vision and Pattern Recognition, 2020, pp. 3063–3072.
[22] C. Li, C. Guo, C.C. Loy, Learning to enhance low-light image via zero-reference deep curve estimation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (8) (2021) 4225–4238.
[23] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, R. Timofte, Plug-and-play image restoration with deep denoiser prior, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 6360–6376.
[24] J. Cao, Q. Wang, Y. Xian, Y. Li, B. Ni, Z. Pi, K. Zhang, Y. Zhang, R. Timofte, L. Van Gool, Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution, in: Computer Vision and Pattern Recognition, 2023, pp. 1796–1807.
[25] K. Jiang, Z. Wang, P. Yi, J. Jiang, Hierarchical dense recursive network for image super-resolution, Pattern Recognit. 107 (2020) 107475.
[26] X. Li, J. Dong, J. Tang, J. Pan, DLGSANet: lightweight dynamic local and global self-attention networks for image super-resolution, in: Computer Vision and Pattern Recognition, 2023, pp. 12792–12801.
[27] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, W. Wu, Feedback network for image super-resolution, in: Computer Vision and Pattern Recognition, 2019, pp. 3867–3876.
[28] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, L. Zhang, TTST: A top-k token selective transformer for remote sensing image super-resolution, IEEE Trans. Image Process. (2024) 1–15.
[29] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: International Conference on Computer Vision, 2021, pp. 1833–1844.
[30] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, L. Shao, Learning enriched features for real image restoration and enhancement, in: European Conference on Computer Vision, 2020, pp. 492–511.
[31] X. Wang, K. Yu, C. Dong, C.C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in: Computer Vision and Pattern Recognition, 2018, pp. 606–615.
[32] E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, Film: Visual reasoning with a general conditioning layer, in: AAAI, vol. 32, (1) 2018.
[33] R. Liu, J. Gao, J. Zhang, D. Meng, Z. Lin, Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2021) 10045–10067.
[34] R. Liu, J. Gao, X. Liu, X. Fan, Learning with constraint learning: New perspective, solution strategy and various applications, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–18.
[35] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: European Conference on Computer Vision Workshops, 2018.
[36] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.
[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: European Conference on Computer Vision, 2018, pp. 286–301.
[38] H. Zhao, X. Kong, J. He, Y. Qiao, C. Dong, Efficient image super-resolution using pixel attention, in: European Conference on Computer Vision, 2020, pp. 56–72.
[39] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[40] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: International Conference on Computer Vision, 2023, pp. 4015–4026.

Jiaxin Gao received the bachelor's degree in school of mathematical science from Dalian University of Technology, Dalian, China, in 2018. She is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian. Her research interests include computer vision and bi-level optimization.

Yaohua Liu received the M.S. degree in software engineering from Dalian University of Technology, Dalian, China, in 2021. He is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian. His research interests include computer vision, bi-level optimization, attack and defense.

Ziyu Yue received the bachelor's degree in school of mathematical science from Dalian University of Technology, Dalian, China, in 2017. He is currently pursuing the Ph.D. degree in school of mathematical science with the Dalian University of Technology, Dalian. His research interests include computer vision, image enhancement, deep learning.

Xin Fan received the B.E. and Ph.D. degrees in information and communication engineering from Xian Jiaotong University, Xian, China, in 1998 and 2004, respectively. He joined the School of Software, Dalian University of Technology, Dalian, China, in 2009. His current research interests include computational geometry and machine learning.

Risheng Liu (Corresponding author) received the B.S. and Ph.D. degrees both in mathematics from the Dalian University of Technology in 2007 and 2012. He served as Hong Kong Scholar Research Fellow at the Hong Kong Polytechnic University from 2016 to 2017. He is currently a professor with DUT-RU International School of Information Science and Engineering, Dalian University of Technology. His research interests include machine learning, optimization and computer vision.