This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3325357



# Watch You Under Low-resolution and Low-illumination: Face Enhancement via Bi-factor Degradation Decoupling

Xin Ding, *Member, IEEE,* Zheng Wang, *Member, IEEE,* Jing Fang, Zhenyu Shu, Ruimin Hu, *Member, IEEE,* and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*—Face enhancement aims to improve low-quality face images to a higher-quality level. However, in real-world nighttime scenes, complex degradation factors often affect these images, making it challenging to preserve important facial details. Existing image enhancement algorithms typically focus on independently conducting image super-resolution and brightness enhancement, assuming a fixed degradation level based on simulated training datasets. Nonetheless, real nighttime scenes involve complex degradation processes, where degradation factors dynamically and variably manifest. Therefore, achieving effective face enhancement in such scenarios is particularly daunting. This work analyzes and unveils the multiple factors of low resolution and low illumination during degradation. Based on this analysis, we propose a Bi-factor Degradation Decoupling network. Our method leverages a decoupling network to generate qualitative and quantitative features corresponding to each factor's degradation degree in the low-quality environment. These features are then combined with robust facial feature constraints to recover the details of low-quality faces. Extensive experiments demonstrate that our method surpasses state-of-the-art approaches in both enhancement and face super-resolution.**

***Index Terms*—Degradation Analysis, Super-resolution, Low-illumination Enhancement, Deep Learning, VAE.**

## I. Introduction

### A. Problems

IMAGE capturing systems often encounter challenges in variable shooting distances and difficult lighting conditions, resulting in various quality issues such as low resolution (LR) and inadequate illumination. To address these challenges, face enhancement algorithms aim to reconstruct high-quality face images from low-quality inputs. The term 'low-quality' encompasses images with LR, poor illumination, or multiple degradation factors.

Xin Ding and Zhenyu Shu are with School of Computer and Data Engineering, NingboTech University, Ningbo 315100, China. (e-mail: xding07@163.com; shuzhenyu@nit.zju.edu.cn)

Zheng Wang, Jing Fang, and Ruimin Hu are with the National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan 430072, China. (e-mail: wangzheng@whu.edu.cn; jingfang@whu.edu.cn; hrm@whu.edu.cn).

Chia-Wen Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

This work was supported by the National Natural Science Foundation of China under Grants (62172356, 61872321), Zhejiang Provincial Natural Science Foundation of China (LY22F020026). Corresponding author: Zheng Wang.

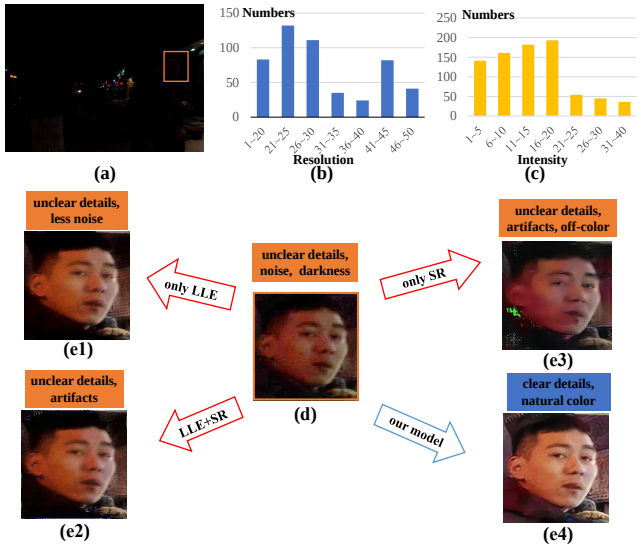


Fig. 1. Illustration of a low-quality image captured in a real nighttime scene and the face's recovery results. (a) The image of the nighttime scene. It suffers from the low resolution and low brightness. In particular, the low-quality faces have different degrees of resolution and brightness. To illustrate the issue, the relative data are counted. (b) The statistically different resolutions of the real faces selected from the low-quality images. (c) The statistically different intensities of the real dark faces selected from the low-quality images. The comparison methods suffer from the difficult degradation in the real night scene. (d) The input low illumination and LR face by linear amplification and interpolation cropped from the low-quality image (a), (e1) the LLE face by HDR-net [1], (e2) the cascaded LLE and SR face by HDR-net+Super-FAN [2], (e3) the SR face by Super-FAN, (e4) our results.

The technique of face SR was pioneered by Baker and Kanade [3], which proposed an approach that formulates high-resolution (HR) face reconstruction as a prior factor. The method uses spatial distributions, such as Gaussian pyramids, to predict training patches for frontal face images.

Recently, deep learning-based methods have shown great success in image SR [4], [5], including algorithms such as SRCNN, VDSR, and LapSRNet [6], [7], [8]. However, applying SR techniques to low-illumination face images may not be suitable due to the potential introduction of unwanted noise and artifacts. Low-quality images with complex degradation factors, especially in dark conditions, can pose challenges for SR algorithms.

In real surveillance scenarios, face images are often captured in low-light conditions, resulting in complex degradation phe-

nomena caused by multiple factors and random degradation. Existing algorithms typically focus on individual aspects, such as face SR or brightness enhancement [9], [10]. However, they usually ignore the combined effects of these degradation factors in real scenes, leading to artifacts, blurred details, and other deficiencies in the results. For example, Fig. 1(a) illustrates a blurred low-light face captured in a night scene, exhibiting uneven illumination, poor face details, and high image noise due to linear zooming. Moreover, real low-quality face images often exhibit variations in resolution and brightness, as shown in Figs. 1(b) and (c), which illustrate examples in DarkFace [11]. This dataset demonstrates a wide range of resolution and illumination levels in low-quality faces. We enhance the input low-quality images using the HDR-net algorithm [1], which improves details and reduces noise, as depicted in Fig. 1(e1). However, when low-light and LR images are employed as inputs for face enhancement algorithms, such as Super-FAN [2], significant artifacts, blurred details, and color biases are introduced, resulting in poor subjective outcomes, as shown in Fig. 1(e3). The combined effects of low illumination and low resolution further degrade the reconstruction quality, leading to off-color and unclear details, as indicated in Fig. 1(e2). To address these challenges and improve facial details, we propose an approach that analyzes the combined effects of low illumination and low resolution in order to enhance low-quality faces.

The process of enhancing low-quality faces captured in nighttime scenes can be conceptualized as an information tracing procedure. Its objective is to recover missing valid pixels from the degraded inputs and reconstruct them into high-quality images. However, existing face enhancement algorithms primarily focus on recovering clear face images and do not explicitly address retrieving missing pixels or reconstructing valid face information from the degradation process. Consequently, the performance of these algorithms heavily relies on the ability to accurately analyze the degradation process and trace back to the potential state prior to information loss. Treating face enhancement solely as a down-sampling process without precise knowledge of the specific degradation state of the image is insufficient for effectively recovering actual low-quality face images. Instead, by incorporating degradation analysis as a constraint, we can reconstruct face details that align with the likely state preceding information loss. This approach goes beyond the prediction of unknown information by existing algorithms, thereby enhancing the overall performance of face enhancement in terms of reconstruction quality.

### B. Main Idea and Contributions

To tackle the aforementioned challenges, we devise a robust face enhancement algorithm grounded in degradation decoupling analysis.

**Main Idea**: Specifically, we propose an algorithm designed to reconstruct low-quality face images in nighttime environments with various resolutions and illumination levels. First, the network can retrospectively identify and evaluate the factors contributing to image degradation in the given environment, and quantify the corresponding degree of degradation. These identified factors and degradation levels are subsequently incorporated into the face reconstruction process. Additionally, the algorithm extracts resilient facial features that effectively counter environmental interference, thereby constraining the face reconstruction procedure.

To this end, we adopt the $\beta$-vae decoupling network [12] to separate the low-quality face images into interpretable latent factors [13]. Next, we design encoders to simultaneously address the challenges of illumination and resolution degradation. Additionally, we employ an encoder to estimate robust facial feature factors. As a result, the resolved degradation factors are generated and assigned with corresponding labels and intensities. These factors are then utilized by a decoder to generate enhanced SR images under low-light conditions and images that highlight the extracted facial features. Finally, these generations are fed into the reconstruction network for accurate facial results.

In this paper, **our contributions** are summarized as follows:

- We propose a novel approach that simultaneously tackles the enhancement of face images captured in LR and low-light conditions within real-world scenarios. Our technique effectively addresses the challenges commonly encountered when dealing with a wide range of random scales of resolutions and illumination levels.
- In this study, we present a comprehensive analysis that decouples the bi-factor degradation observed during the face enhancement process. This analysis provides valuable insights and effective methodologies that can be applied to real-world scenarios, enabling practical applications.
- We conduct a comprehensive evaluation on publicly available and real-world low-quality face datasets, demonstrating the superior performance of our proposed method compared to several state-of-the-art methods.

The remainder of the paper is structured as follows. Sec. II provides an overview of related works. Our proposed method is described in Sec. III. Experimental results on simulated and real-world low-quality images are presented in Sec. IV. Finally, we conclude our work in Sec. V.

## II. Related Work

In this section, we provide a comprehensive review of the relevant literature pertaining to our proposed method. The literature review covers various aspects, including image degradation analysis, face SR, and brightness enhancement.

### A. Image Degradation Analysis

Image degradation analysis plays a crucial role in image enhancement tasks such as image SR, restoration, and deblurring. Many research works have incorporated this analysis into their methodologies. For example, Efrat *et al.* [14] emphasized the significance of accurately estimating the blur kernel for effective image deblurring. Tracking the degradation process has become a key objective for various image processing algorithms. To address the challenge of handling diverse blur kernels encountered in real LR images, Zhang *et al.* [15] proposed a plug-and-play framework specifically designed for degraded SR. Their approach focused on effectively addressing arbitrary blur kernels and relies on additional kernel estimation

methods to accurately estimate the blur kernels in real low-quality images.

In order to enhance the algorithm's adaptability to real-world processes, some researchers have explored training augmented networks by estimating various factors, such as blur kernels [16], [17] and noise [18], present in real low-quality images. This approach involves generating training datasets of real low-quality images or constructing datasets with both high and low-quality images. Bulat *et al.* [19] introduced the High-to-Low+Low-to-High model, which analyzes the degradation process of a low-quality image so as to reconstruct it accordingly. The KernelGAN model [20] treats SR in real images as an estimation of unknown downsampling kernels for reconstruction, thus aligning with the motivation behind High-to-Low+Low-to-High to make training data closely resemble the given inputs. The DASR model in [21] employs a degradation encoder to estimate features related to the degradation level, which are then used as constraints in the reconstruction process through the degradation-aware block.

In contrast, real-world low-quality images exhibit a wide range of degradation factors, making the actual degradation process complex and intractable. Additionally, approaches like that proposed in [19] rely on generating low-quality datasets for enhanced models but do not address the fundamental issue of the dependency on the training data. Without effective constraints on quality reduction, the resulting representation models may still struggle to adequately reconstruct low-quality images in real scenes with varying and dynamic conditions.

### B. Face Super-resolution

Face SR has witnessed remarkable advancements thanks to the powerful deep learning techniques [9], [22], [23], [24], [25], [26], [27]. However, preserving manifold consistency between the LR and HR spaces in real-world scenarios remains challenging due to the complexity of degradation processes. To capture both the global topology information and local texture details of human faces, Huang *et al.* introduced Wavelet-SRNet [28], a method based on wavelet transform. This approach incorporates three types of loss functions: wavelet prediction loss, texture loss, and full-image loss. Similarly, Bulat *et al.* introduced Super-FAN [2], an end-to-end framework addressing face SR and facial landmark detection simultaneously, achieving improved face resolution and robust facial landmark detection. Ma *et al.* [24] proposed a face SR approach that employs two recurrent networks in an iterative collaboration framework. These networks focuse on facial image recovery and landmark estimation, respectively, enhancing face SR performance. In a related study, Mei *et al.* [29] proposed a Non-Local Sparse Attention (NLSA) method for single image SR. This method utilizes dynamic sparse attention patterns to effectively address the challenges associated with SR tasks.

In order to address the challenge of achieving large-scale face SR, Wang *et al.* [30] proposed a method that incorporates the learning of facial prior knowledge during the training process to enhance the level of detail. However, this approach introduced a significant issue whereby the training model becomes excessively dependent on the simulated datasets, leading to unsatisfactory results when generating very LR faces in real-world scenarios. To break such limitations and improve the quality of faces at different scales in practical settings, the multi-scale recurrent scalable network (MRS-Net+) was proposed by Liu *et al.* [31]. This method aims to effectively enhance the quality of faces at varying resolutions, providing more accurate and visually appealing results. By considering the multi-scale nature of face SR, MRS-Net+ offers a promising solution for addressing the challenges associated with face SR across different scales in real-world applications.

However, the aforementioned studies consider face enhancement as the SR of simulated data sampled at a fixed scale, which substantially differs from real-world degradation scenarios. As revealed in [32], models trained using these approaches have demonstrated unsatisfactory reconstruction results when applied to real-world images. Thus, the facial reconstruction task lacks an effective constraint on the environmental conditions, leading to suboptimal reconstruction performance.

### C. Brightness Enhancement

Brightness enhancement approaches [33], [34], [35] aim to enhance the illumination and visibility of dark images. These approaches can be categorized into two main types: Retinex decomposition-based and deep learning-based methods. Retinex-based methods such as single-scale Retinex [36] and multi-scale Retinex [37] utilize Gaussian or bilateral filters to remove halo artifacts and improve image quality. Other methods manipulate both the illumination and reflectance layers to achieve enhanced results.

Deep learning methods for enhancing low-light images have been extensively studied. Lore *et al.* [38] utilized a deep auto-encoder called Low-Light Net (LLNet) for contrast enhancement and denoising. Gharbi *et al.* introduced HDR-net [1], a neural network architecture inspired by bilateral grid processing and local affine color transforms. Wang *et al.* proposed Retinex-Net [39], which includes Decom-Net and Enhance-Net for image decomposition and illumination adjustment, respectively. Chen *et al.* introduced the SID model [23], enhancing low-light images using corresponding long-exposure reference images.

In a recent study by Yang *et al.* [40], an attempt was made to explore semi-supervised learning techniques for low-light image enhancement. The work proposed a deep recursive band (DRBN) representation that serves as a connection between fully supervised and unsupervised learning frameworks, thereby leveraging the advantages of both approaches. To enhance images and suppress noise in the reflectance map, a Low-Rank Regularized Retinex Model (LR3M) was proposed in [41], incorporating a low-rank prior into the Retinex decomposition process. Another representative model, called Retinex-inspired Unrolling with Architecture Search (RUAS) was proposed in [42], aiming to construct a lightweight yet effective enhancement network for low-light images in real-world scenarios. By exploring the principles of Retinex, RUAS

achieves superior performance in enhancing low-light images while considering the computational efficiency of the network architecture.

Zhao *et al.* [43] proposed a unified deep framework for Retinex decomposition and low-light image enhancement. However, existing methods often struggle with adjusting exposure effectively, resulting in uneven exposure or partial overexposure. To overcome these limitations, Fan *et al.* [44] introduced the multiscale low-light image enhancement network with illumination constraint (MLLEN-IC). This end-to-end model incorporates an illumination constraint into the network architecture, aiming to achieve superior generalization ability and stable performance. By leveraging this constraint, MLLEN-IC effectively addresses exposure-related issues and produces desirable enhancement results.

However, real-world nighttime images are often affected by various complex degradation factors, including noise, resolution degradation, and interference from low light conditions. Additionally, the image brightness can vary significantly with different low light intensities, posing challenges for recovering low-quality face images in various degradation conditions [19]. Consequently, achieving satisfactory results in restoring real-world low-quality face images under such multi-dimensional degradation conditions remains a significant challenge.

In real-world scenarios, low-quality faces frequently exhibit various complex and diverse degradation processes. Consequently, a considerable gap exists in the formation and effective treatment of the degradation processes associated with low-quality faces in real scenes. Further research and development are needed to address this challenge and improve the performance of face enhancement and SR algorithms in handling complex degradation scenarios.

## III. Proposed Method

### A. Motivation

In this section, we provide a comprehensive explanation of the motivation behind our proposed method.

Fig. 1(a) illustrates the common challenges faced when capturing low-quality faces at night, including issues related to illumination and shooting distance. These factors often result in missing object information and inaccurate recognition. Furthermore, the wide range of illumination conditions and various shooting distances introduce additional complexity, leading to diverse and random degradation in face images. Our objective is to address these challenges and restore the quality of faces in real-world scenarios.

In Fig. 2(s1), the object is captured under natural lighting conditions $I_1$. By maintaining an appropriate capture distance $D_1$, the face image $y$ retains as much high resolution and detailed information of the original object as possible. In contrast, Fig. 2(s2) illustrates the impact of low illumination $I_2$, which diminishes the reflected light from the object. This reduction in information intensity leads to lower overall illumination in the final imaging result $x_{rt}$, thus resulting in image degradation. This degradation affects all pixels in the image, where the illumination degradation can be expressed as $x_{rt} = I_l y$. Fig. 2(s3) demonstrates the formation of degradation

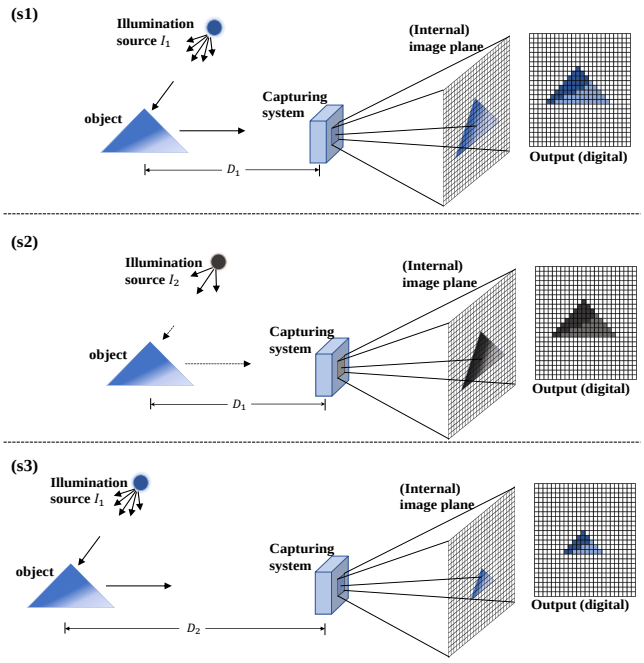


Fig. 2. Conceptual diagram of image degradation process from a high-quality image to a low-quality image, involving two distinct and independent components: low illumination and long-distance factors. (s1) The capturing system with the natural illumination $I_1$ and short-distance $D_1$. (s2) The capturing system with the low illumination $I_2$ and short-distance $D_1$. (s3) The capturing system with the natural illumination $I_1$ and long-distance $D_2$.

as the object is captured with a smaller resolution $x_{it}$ due to an increased capture distance from $D_1$ to $D_2$. This resolution degradation process [45] can be viewed as a downscaling model applied to the HR image $y$. It involves reducing the image resolution $x_{it} = R_l y$ by combining multiple pixels in the neighborhood through mean or weight fusion techniques. The downsampling operation converts the HR image $y$ to a lower-resolution image $x_{it}$ using $R_l$. While illumination degradation refers to an overall reduction in image pixel values, resolution degradation involves the fusion of neighboring pixels. Considering the degradation process and the patterns affecting image information, the degradation of illumination and resolution are essentially independent of each other. As a result, the low-quality image $x$ can be expressed by

$$x = R_l(I_l y). \tag{1}$$

The bi-factor degradation decoupling significantly impacts the rendering of face images and the representation of crucial information. As shown in Figs. 1(b) and (c), the combination of a long distance (resulting in a lower resolution of the target face) and low exposure (leading to a dark image) contributes to a more intricate degradation of the image.

Since the degradation mechanisms of illumination and resolution are independent, we can represent the degraded image $x$ in terms of low illumination and low resolution as follows:

$$p(x) = \int\int\int p(x \mid (z_r, z_i, z_f))p(z_i)p(z_r) \\ p(z_f)dz_r dz_i dz_f, \tag{2}$$

where $z_i$, $z_r$, and $z_f$ represent the degradation factors of illumination, resolution, and facial features, respectively. We aim to separate and quantify these two independent degradation factors by generating labels for the related facial features in the image.

As illustrated in Fig. 3, we aim to recover a high-quality face $y$ from an input degraded face $x$ obtained from a nighttime environment. The low-quality image $x$ undergoes random intensity degradation, including low illumination and low resolution. To address this issue, we propose a model that decouples input $x$ and extracts multiple latent codes representing the latent feature spaces [46], involving illumination, resolution, and facial feature factors. By utilizing the latent code of the facial feature factor and independently analyzing the latent codes of the illumination and resolution factors, we generate the corresponding factor intensity labels. These labels are used to constrain the low-light enhancement and SR techniques. Finally, the low-light enhancement, image SR, and robust facial features are fused to the reconstruction network to generate the hallucinated results.

### B. Extraction of Latent Codes

In this subsection, we focus on analyzing low-quality images from three aspects: illumination, resolution, and robust facial features. To accomplish this, we employ three encoders to implement the corresponding functions. The encoders generate latent codes that represent illumination and resolution, respectively. These latent codes enable us to track the degradation process and provide a detailed analysis of the complex factors in an actual environment. Thus, we use $E_r$ to describe illumination and $E_i$ to describe resolution as

$$z_r = E_r(x), z_i = E_i(x), \tag{3}$$

where the latent codes $z_r$ and $z_i$ capture the degradation characteristics of the environmental factors and provide insights into the effects of respective degraded factors.

As our focus is on reconstructing face images, it is important to extract the representation of facial features that can adapt to various environmental conditions. By emphasizing robust facial features during the recovery process, we ensure clear and accurate depictions of faces regardless of complex environmental conditions. To this end, we employ the face feature encoder $E_f$ to generate the latent code specifically related to facial characteristics

$$z_f = E_f(x), \tag{4}$$

thus obtaining the facial characteristics $z_f$.

### C. Factor Analysis

We extract the latent codes representing distinct factors: illumination, resolution, and facial characteristics. By considering these factors using the extracted codes, we can recover the desired information.

**Facial feature extraction.** Our primary objective is to reconstruct facial information in low-quality images. Facial features play a crucial role in these low-quality images as they contain significant semantic information that can be utilized for face reconstruction. Using the existing face semantic latent code $z_f$, we employ the facial decoder $x_m$ to generate the facial feature map $D_f$:

$$x_m = D_f(z_f). \tag{5}$$

The feature map $x_m$ primarily acts as a skin mask, accurately capturing the contour and position of the face. To impose additional constraints on the facial details, $x_m$ is used as the input for two CNN modules $F_m$, which in turn generate mask features $x_{ms}$ that contain more fine-grained facial details:

$$x_{ms} = F_m(x_m). \tag{6}$$

As a result, the facial features can be thoroughly analyzed and accurately represented in fine detail.

**Illumination analysis and enhancement.** In a degraded environment, illumination analysis is primarily conducted using the latent code $z_i$ extracted by the illumination encoder $E_i$. The generated $z_i$ is then used as the input to the illumination decoder, which generates the input image $x_i = x$ to impose constraints on the illumination factor.

$$p(x_i) = \int p(x_i \mid z_i)p(z_i)dz_i. \tag{7}$$

The illumination decoder is used for constraining the illumination factor $D_i$:

$$x_i = D_i(z_i). \tag{8}$$

Meanwhile, the latent code $z_i$ is utilized in the fully connected layers $Fc_i$ for labeling. However, representing illumination poses challenges due to the various sensitivities of different objects in each scene of the actual image, making convergence difficult. To address this, we focus on accurately expressing illumination by specifically targeting facial skin with a similar light sensitivity range. To this end, we utilize the Face Extraction Unit (FEU) to represent the image as face illumination without the background:

$$x_f = x \cdot x_m + \max(x \cdot x_m)(1 - x_m). \tag{9}$$

The face information $x_f$ is fed into another illumination encoder $E_{fi}$ to generate the latent code for face illumination, denoted as

$$z_{fi} = E_{fi}(x_f). \tag{10}$$

The latent code is subsequently processed by the fully connected network $Fc_i$ to generate the illumination label $l_i$ (aka the illumination factor) expressed as $l_i = Fc_i(z_{fi})$.

To achieve low-light enhancement, we determine the optimal value for the illumination factor, which serves as a boundary condition to precisely express the resulting illumination enhancement. The illumination parameters and face factors are

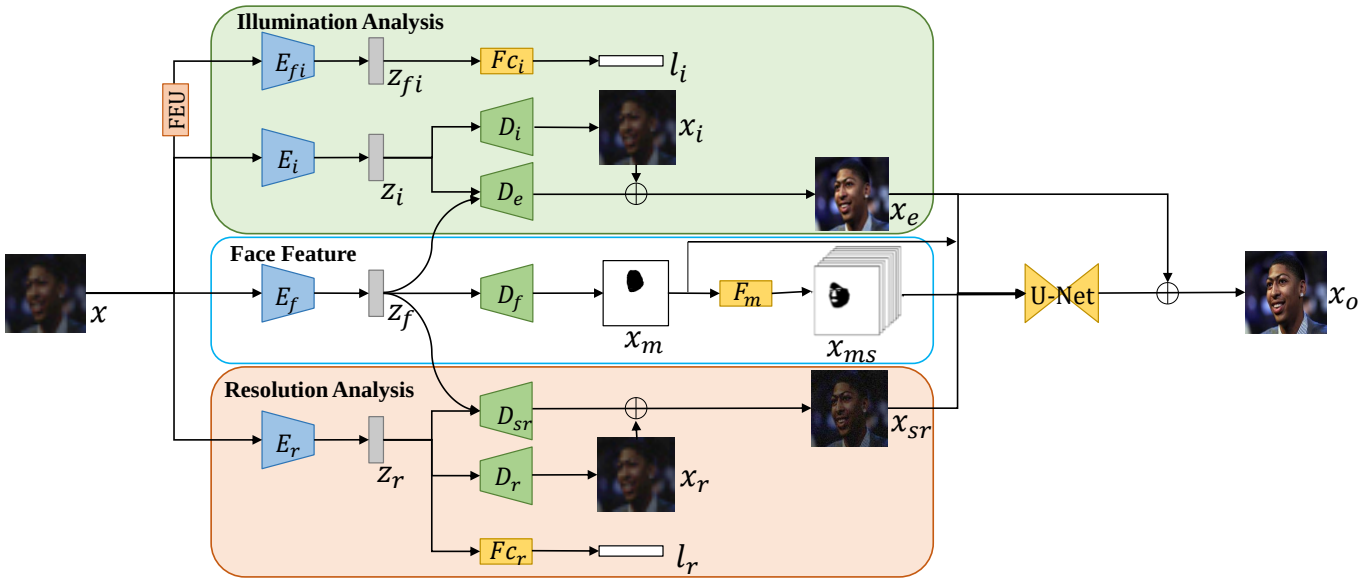


Fig. 3. Decoupling analysis framework of Bi-factor Degradation and robust facial features for face enhancement. Given a low-quality image $x$, our approach involves three modules: LLE, SR, and facial features. The input image $x$ goes through illumination analysis and is then encoded by $E_i$ to extract the latent code $z_i$, which is then refined by $D_i$. We introduce the FEU module for face illumination analysis that focuses on the face skin, which is encoded by $E_{fi}$ to obtain the illumination latent code $z_{fi}$. An illumination label $l_i$ is generated using the fully connected layer $Fc_i$. Secondly, in resolution analysis, $x$ is encoded by $E_r$ to obtain the latent code $z_r$, which is then refined by $D_r$. The fully connected layer $Fc_r$ uses $z_r$ to generate the resolution label $l_r$. Consequently, we use the VAE module with $E_f$ and $D_f$ for face feature analysis to derive the face feature map $x_m$ and the latent code $z_f$. The network $F_m$ extracts mask features $x_{ms}$ from $x_m$. Using the latent codes $z_f$, $z_i$, $z_r$, and the decoders $D_e$, $D_{sr}$, we generate respective residual features for low-light enhancement ($x_e$) and SR ($x_{sr}$). The intermediate outputs $x_e$, $x_{sr}$, $x_{ms}$, and $x_m$ are then fed into the U-Net for reconstructing the residual result, which contains more face details. Finally, the intermediate enhanced image $x_e$ is added to obtain the final output $x_o$.

thoroughly constrained, ensuring a well-balanced and highly effective enhancement. The enhancement result is denoted as $x_e$:

$$p(x_e) = \int\int p(x_e \mid (z_i, z_f))p(z_i)p(z_f)dz_i dz_f, \quad (11)$$

where the illumination factor is treated as an environmental condition, while the face factor represents a robust object representation independent of illumination for the enhanced images $x_e$.

To reconstruct the corresponding image, both the facial and illumination features are utilized as inputs to the illumination-enhanced decoder $D_e$:

$$x_e = x_i + D_e(z_i, z_f). \quad (12)$$

Through this network, we are able to perform degradation analysis and feature enhancement of the illumination factor.

**Resolution analysis and enhancement.** We also analyze to decompose the factors responsible for the random degradation in image resolution. The latent code $z_r$ is derived from the resolution encoder $E_r$ and utilized to reconstruct as the input image $x_r = x$:

$$p(x_r) = \int p(x_r \mid z_r)p(z_r)dz_r. \quad (13)$$

The constrained resolution factor is represented by the resolution decoder $D_r$:

$$x_r = D_r(z_r). \quad (14)$$

The latent code $z_r$ is passed through the fully connected network $Fc_r$ to generate the corresponding label denoted as the resolution factor $lb_r = Fc_r(z_r)$.

When addressing the resolution degradation, the scale parameters are utilized as inputs for the resolution enhancement process, specifically in image SR. The resolution and facial factors are employed as boundary conditions to express the resulting SR image $x_{sr}$:

$$p(x_{sr}) = \int\int p(x_{sr} \mid (z_r, z_f))p(z_s)p(z_f)dz_s dz_f, \quad (15)$$

where the resolution factor serves as the environmental condition, while the face factor is utilized as the semantic feature. These features are mutually independent, These features are mutually independent, allowing for effective constraints to be applied in the reconstruction of the SR image $x_{sr}$. The face and resolution features are fed into the image SR encoder $D_{sr}$ to reconstruct the corresponding images:

$$x_{sr} = x_r + D_{sr}(z_r, z_f). \quad (16)$$

As a result, this network allows us to utilize the corresponding factors of both resolution and face features for enhancement.

### *D. Image Reconstruction*

By performing decoupling analysis, we decompose the input image $x$ into multiple factor features: $x_{sr}$, $x_e$, $x_{ms}$, and $x_m$. These features effectively offer a comprehensive representation

of the degraded conditions and facial characteristics. Hence, they are reconstructed via $F_u$ to generate improved HR images $x_o$ by incorporating more detailed residuals from the intermediate enhanced image $x_e$:

$$x_o = x_e + F_u(x_{sr}, x_e, x_{ms}, x_m), \tag{17}$$

where the reconstructed network is grouped by a U-Net [47].

### E. Loss Functions

The loss functions are defined in two aspects: feature decoupling and image reconstruction.

*1) Feature Decoupling:* Feature decoupling primarily represents essential attributes, while image enhancement is accomplished through degradation analysis and robust feature extraction. This process encompasses the analysis of multiple factors, including facial features, illumination degradation, and resolution degradation.

Face feature analysis involves the representation of robust facial features in the image, which is achieved through the use of the face mask maps, and the loss function is

$$\mathcal{L}_f = ||x_{ms} - x_{mst}||_2^2 + ||x_m - x_{mt}||_2^2, \tag{18}$$

where $x_{mst}$ and $x_{mt}$ refer to the face mask maps corresponding to the target faces.

The illumination degradation analysis process involves illumination feature representation, illumination enhancement, and label generation. The loss function $\mathcal{L}_i$ for this process comprises these three components, with feature representation and illumination enhancement primarily expressed through the Euclidean distance. The generated labels are represented using the cross-entropy loss function as follows:

$$\begin{aligned}\mathcal{L}_i = ||x_i - x||_2^2 + ||x_e - x_{it}||_2^2 + \sigma_1(label_i * \log l_i \\ + (1 - label_i)\log(1 - l_i)),\end{aligned} \tag{19}$$

where $label_i$ represents the target illumination label, $\sigma_1$ denotes the illumination constraint component, and $x_{it}$ refers to the LR face.

The resolution degradation analysis process involves resolution feature representation, resolution enhancement, and resolution label generation. The loss function $\mathcal{L}_r$ for this process also consists of these three components:

$$\begin{aligned}\mathcal{L}_r = ||x_r - x||_2^2 + ||x_{sr} - x_{rt}||_2^2 + \sigma_2(label_r * \log l_r \\ + (1 - label_r)\log(1 - l_r)),\end{aligned} \tag{20}$$

where $label_r$ represents the target resolution label, $\sigma_2$ denotes the constraint component in terms of resolution and $x_{rt}$ refers to the HR low-illumination image.

*2) Image Reconstruction:* The image reconstruction process primarily involves fusing three deconstructed feature components: facial features, brightness factors, and resolution factors. The resulting output is the enhanced face image $x_o$, and the loss function for image reconstruction is denoted as $\mathcal{L}_o$:

$$\mathcal{L}_o = ||x_o - y||_2^2. \tag{21}$$

After analyzing the environmental characteristics and facial features, a clear face image can be reconstructed.

*3) Overall Loss Function:* Consequently, the overall loss function $\mathcal{L}_{to}$ encompasses both feature decoupling and face reconstruction:

$$\mathcal{L}_{to} = \mathcal{L}_f + \mathcal{L}_i + \mathcal{L}_r + \alpha\mathcal{L}_o, \tag{22}$$

where the constraint weight $\alpha$ determines the importance of face reconstruction, ultimately improving feature extraction and face reconstruction.

### F. Implementation Details

**Training setup**: The network architecture is depicted in Fig. 3. The model is trained using the ADAM optimizer [48] with constraint parameters set to $\alpha = 1$, $\sigma_1 = 0.2$, $\sigma_2 = 0.2$. The learning rate is set to 0.0001, and it is halved every 100 training epochs. The experiments are implemented using PyTorch [49] and trained on an NVIDIA RTX 1080ti GPU.

## IV. Experiments

We conduct extensive experiments to assess the effectiveness of our algorithm on low-quality face images captured in nighttime scenes. The term "low-quality" here encompasses both low resolution and low brightness in a random fashion. It is worth noting that enhancing low-quality face images at night presents a unique challenge, as existing algorithms have not directly addressed this specific scenario. To address this, we treat the compared algorithms for face enhancement as a combination of low-illumination enhancement and face SR. In terms of face SR, we compare our approach against state-of-the-art (SOTA) algorithms such as Super-FAN [2], Wavelet-SRNet [28], and NLSA [29]. For reference, we also include Bicubic interpolation as a baseline comparison method. Prior to the face SR process, we perform pre-processing for low-illumination enhancement. This includes linear enhancement, LR3M [41], HDR-net [1], SID [23], Retinex-Net [39], RUAS [42], and DRBN [40]. All experimental validations are conducted on the CelebA face dataset [50], in which the luminance degradation is simulated using a realistic low-illumination model [51], [52], whereas the resolution degradation is implemented as a random degradation parameter.

### A. Realistic Low-illumination Model

Following the Camera Response Function (CRF) approach proposed in [51], [52], we synthesize low-light images by randomly setting illumination conditions and introducing noise to the original face images.

We set the realistic low-illumination model as

$$y = f(1/\Gamma(DM(L + n(L)))), \tag{23}$$

where $f$ refers to CRF sampled from the set of 201 CRFs mentioned in [53], $DM$ stands for the demosaicing function, $\Gamma$ represents the degree of the low-illumination, $L$ is the

irradiance image of raw pixels, and $n(L)$ denotes the adding random noise.

During the experiments, we employ the low-illumination model on the simulated face images, where the image noise variance $\sigma_n$ for $n(L)$ ranges from 0 to 0.06, and $\Gamma$ is a random value selected from the range of 1 to 80, with 10 consecutive numbers assigned as a level (*e.g.*, values from 1 to 10 represent one level).

### *B. Datasets*

**CelebA**: We conduct extensive experiments on the Large-scale CelebFaces Attributes (CelebA) dataset [50]. For our experiments, we utilize 12,000 images for training and approximately 3,500 images for testing. The process for generating these dark images is described in Sec. IV-A. We apply the model trained by CelebAMask-HQ [54] on the original (well-lit) images to obtain the targeted parsing maps. They are all resized to $256\times256$, and then randomly downscaled to one of the sizes $32\times32$ and $64\times64$. Finally, they are super-resolved to $256\times256$.

**Helen**: We also conduct our experiments on the Helen dataset [55]. The methods for synthesizing dark images are described in Sec. IV-A. For testing purposes, we utilize approximately 300 images, which are all resized to $256\times256$, and then randomly downscaled to one of the sizes $32\times32$ and $64\times64$. Finally, they are super-resolved to $256\times256$.

**D-faces+**: We obtain face images from the publicly available real-world nighttime face dataset DarkFace [11]. These realistic low-quality face images are captured in low-light conditions. We collect approximately 150 face images as a realistic low-quality face testing set, which involves various factors such as random LR and realistic nighttime low-illumination conditions. The primary objective of this data collection is to validate the effectiveness of our model and compare its performance with other algorithms using subjective metrics on real-world images.

**Face Detection**: In the training process, incorporating facial features can enhance the performance of the enhancement model. In our experiment, we extract face masks from CelebA and Helen using a pre-trained model for parsing maps [56] as facial features. The resulting face masks exhibit similar characteristics to those obtained from CelebAMask-HQ.

### *C. Performance Evaluation*

In this section, we compare the performances of our method and several state-of-the-art (SOTA) methods. As our work primarily focuses on restoring low-light and LR face images, we evaluate the performances on the simulated datasets derived from Helen [55] and CelebA [50] mentioned above. The evaluation involves three scenarios: restoration from low-light images with 4× down-sampling (resolution: 64×64), 8× down-sampling (resolution: $32\times32$), and randomly down-scaled to a resolution ranging from $32\times32$ to $64\times64$. In the training stage, our model and the compared models are all trained using randomly down-scaled images of the simulated CelebA. In the inference stage, the above three kinds of down-scaled images from both CelebA and Helen are used.

**Low-light LR images**. For a fair comparison, all LR test images from CelebA and Helen are upscaled to the size of $256\times256$ by using the various LLE and face SR models trained on the simulated Celeb. In our approach, we utilize the SID algorithm to recover the original image. The model initially starts with four channels, but in our validation, it was adjusted to three channels. The resulting images are then pre-processed with low-light enhancement methods before being upscaled using SR algorithms. Additionally, we train a separate model using only SR methods to enhance low-quality images. Fig. 4 compares the enhancement results for low-light images downscaled by a factor of 4, showing that while images with larger-sized faces are adequately represented, our results particularly exhibit finer details. The results show that although Wavelet-SRNet can reconstruct fine details, it leads to unnatural colors and artifacts, significantly degrading subjective performance. The third and fourth rows of Fig. 4 depict the results for enhancing low-quality faces of different sizes, showing the compared algorithms yield poor reconstructions for low-quality face images with 8× down-sampling. In contrast, our model excels in preserving fine face details under the same image conditions. The sizes of faces in the subsequent four rows are smaller, posing a greater challenge for the compared algorithms to recover image textures at 8× down-sampling. Overall, our algorithm produces relatively clear predictions of facial details in these scenarios.

To further validate the performance of our training model, we perform face SR and enhancement on randomly down-sampled images. Similar to the aforementioned results, the LR images are also upscaled to $256\times256$. The last four lines of Fig. 4 illustrate the different sizes of the reconstructed faces. The comparison algorithm exhibits minimal improvement in performance, with the faces not being adequately recovered. In contrast, our approach leverages degradation analysis and robust face features, resulting in our results exhibiting precise details that facilitate identification. This highlights the superiority of our approach. Fig. 4 also shows the results with our model for randomly down-sampled low-light images.

Table I compares the results of various algorithms along with their corresponding experimental setups. Considering the specific nature of our experimental setup, the compared algorithms utilize cascades of pre-trained models for LLE and face SR. The experimental setups involve the down-sampling factor (*i.e.*, 4×, 8×, or random downsampling) with random low-lighting. As demonstrated in Table I, the results obtained from LLE can be viewed as a pre-processing step for image SR. With Wavelet-SRNet, most performance metrics improve compared to the LLE pre-processing methods. Additionally, we conducted tests on the reconstruction performance of Wavelet-SRNet without luminance pre-processing, which yields significantly improved objective metrics compared to the pre-processing approach. This alternative setup can also be applied to super-FAN. It is worth noting that the NLSA approach, the most recent work, outperforms Wavelet-SRNet and is very close to our method in terms of PSNR and SSIM. Furthermore, we conducted performance verification specifically on the enhancements achieved solely by super-FAN, Wavelet-SRNet, and NLSA. Low-light conditions and the lack of fine details negatively impact the objective results of these individual approaches. Interestingly, the results obtained from


Fig. 4. Qualitative performance comparison of high-resolution faces recovered from simulated low-light and down-scaled faces sampled from the CelebA and Helen datasets. All these images are upscaled to the size of $256\times256$. In addition to this, linear amplification images are also upscaled to the same size, while keeping the rest of the results consistent with this setting. The input faces are of various resolutions. The first four rows correspond to images downscaled by a factor of 4, while the next two rows show images downscaled by a factor of 8.

super-FAN and Wavelet-SRNet are even worse than those achieved through LLE pre-processing methods. In contrast, our method outperforms all the compared algorithms, including the combined methods.

**Influences on input sizes and low-light conditions**. In this subsection, we present the influences of different degradation factors involving input sizes and low-light conditions. Fig. 5 visualizes the reconstructed results of our method and HDR-net and Super-FAN on two test face images ($Face_1$ and $Face_2$) with sizes of $56\times56$ and $40\times40$, respectively. The first two rows, the middle two rows, and the last two rows correspond to the experimental results for low-light images with different parameter settings, namely $\Gamma$ = 20, 40, and 80, respectively. The results with HDR-net and Super-FAN are influenced by various degrees of degradation, resulting in different levels of detail in the reconstructed faces. In contrast, our method leads to stably good details across different cases. This indicates that for complex degradation factors, our method can perform well on a wide range of low-quality images. Furthermore, Table IV compares the objective metrics for different degradation factors, including resolution ($56\times56$, $48\times48$, and $40\times40$) and luminance level ($\Gamma$ = 20, 40, and 80), showing that our algorithm consistently outperforms the compared algorithms in terms of the objective metrics.

**Influences on face recognition performance.** We further evaluate the face recognition performances on the reconstructed faces with various methods measured by the cosine similarity based on ArcFace embedding [57]. Our evaluation includes generations of random degraded images obtained from CelebA and Helen. We compare our method with several SOTA methods, including Super-FAN [2], Wavelet-SRNet [28], NLSA [29], HDR-net [1], SID [23], Retinex-Net [39], RUAS [42], and DRBN [40]. Table V demonstrates that our method achieves the highest face recognition accuracy

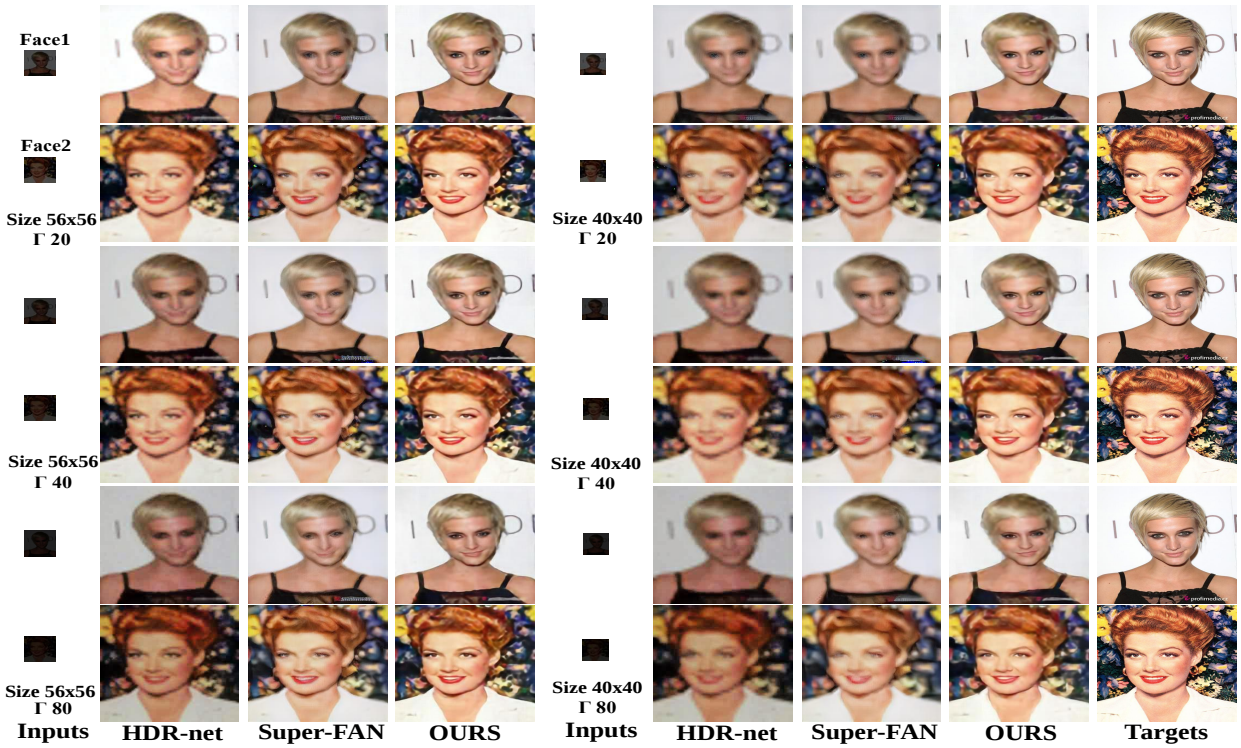


Fig. 5. Qualitative performance comparison of recovered faces in different sizes due to various degradation situations.

TABLE I
QUANTITATIVE PERFORMANCE (PSNR AND SSIM) COMPARISON OF OUR METHOD AND THE COMPARED METHODS ON SIMULATED IMAGES DERIVED FROM CELEBA AND HELEN UNDER VARIOUS DEGRADATION SCENARIOS

| | | CelebA | | | Helen | | |
|---|---|---|---|---|---|---|---|
| **Algorithms** | | 4× | 8× | **random degrees** | 4× | 8× | **random degrees** |
| **Metrics** | | PSNR(db)/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM |
| **Linear amplication** | | 18.53/0.6340 | 17.78/0.5584 | 18.13/0.5929 | 19.06/0.5300 | 17.95/0.4500 | 18.47/0.4867 |
| **LR3M [41]** | | 8.86/0.3449 | 8.73/0.3168 | 8.89/0.3396 | 10.90/0.2989 | 10.70/0.2705 | 10.93/0.2955 |
| **Retinex-Net [39]** | **+Bicubic** | 20.08/0.6262 | 18.74/0.5353 | 19.36/0.5772 | 19.51/0.5398 | 17.65/0.4389 | 18.47/0.4837 |
| **SID [23]** | | 23.03/0.7126 | 20.80/0.6011 | 21.78/0.6501 | 20.54/0.5706 | 18.49/0.4577 | 19.37/0.5072 |
| **HDR-net [1]** | | 22.81/0.6992 | 20.93/0.5995 | 21.78/0.6441 | 22.49/0.6419 | 20.20/0.5374 | 21.25/0.5845 |
| **RUAS [42]** | | 19.22/0.5918 | 18.16/0.5242 | 18.61/0.5282 | 18.38/0.4794 | 17.10/0.4070 | 17.68/0.4191 |
| **DRBN [40]** | | 22.31/0.6788 | 21.02/0.6016 | 21.65/0.6328 | 22.49/0.6419 | 20.20/0.5374 | 21.25/0.5845 |
| **Linear amplication** | | 16.37/0.5892 | 16.06/0.5234 | 16.20/0.5522 | 17.53/0.5011 | 16.87/0.4296 | 17.17/0.4609 |
| **LR3M [41]** | | 18.47/0.6109 | 17.48/0.5132 | 15.99/0.5407 | 16.72/0.4628 | 15.67/0.3687 | 13.66/0.3953 |
| **Retinex-Net [39]** | **+Super-FAN [2]** | 19.43/0.6316 | 18.35/0.5451 | 19.36/0.5772 | 18.83/0.5832 | 17.38/0.4763 | 16.80/0.4361 |
| **SID [23]** | | 22.79/0.7117 | 20.80/0.6024 | 21.63/0.6483 | 20.58/0.5734 | 18.62/0.4615 | 19.44/0.5087 |
| **HDR-net [1]** | | 21.90/0.7001 | 20.27/0.5972 | 20.94/0.6398 | 19.72/0.6364 | 18.27/0.5274 | 18.86/0.5734 |
| **RUAS [42]** | | 18.53/0.5988 | 17.44/0.5171 | 17.86/0.5304 | 17.83/0.4820 | 16.43/0.3978 | 16.99/0.4167 |
| **DRBN [40]** | | 21.37/0.6697 | 20.09/0.5840 | 20.71/0.6163 | 20.54/0.5434 | 18.93/0.4701 | 19.71/0.4977 |
| **Super-FAN [2]** | | 21.50/0.7326 | 19.82/0.5930 | 20.57/0.6386 | 20.42/0.5854 | 17.09/0.4501 | 18.47/0.4875 |
| **Linear amplication** | | 18.70/0.6463 | 17.86/0.5563 | 18.22/0.5936 | 19.38/0.5447 | 18.01/0.4487 | 18.62/0.4889 |
| **LR3M [41]** | | 16.51/0.5875 | 15.89/0.5113 | 17.65/0.5699 | 17.41/0.4707 | 16.43/0.3896 | 17.43/0.4555 |
| **Retinex-Net [39]** | **+Wavelet-SRNet[28]** | 20.24/0.6508 | 18.61/0.5413 | 19.31/0.5869 | 19.68/0.5591 | 17.41/0.4410 | 18.35/0.4898 |
| **SID [23]** | | 22.87/0.7159 | 20.47/0.5966 | 21.46/0.6467 | 20.58/0.5818 | 18.23/0.4588 | 19.14/0.5099 |
| **HDR-net [1]** | | 23.21/0.7086 | 20.89/0.5956 | 21.84/0.6421 | 22.71/0.6648 | 19.91/0.5386 | 21.06/0.5908 |
| **RUAS [42]** | | 18.85/0.6018 | 17.64/0.5269 | 18.18/0.5425 | 18.14/0.4861 | 16.80/0.4102 | 17.40/0.4309 |
| **DRBN [40]** | | 21.95/0.6725 | 20.88/0.5992 | 21.43/0.6286 | 21.52/0.5636 | 19.95/0.4934 | 19.71/0.4977 |
| **Wavelet-SRNet[28]** | | 20.45/0.6914 | 19.48/0.6084 | 19.93/0.6401 | 17.74/0.5463 | 17.29/0.4727 | 17.66/0.5027 |
| **Linear amplication** | | 18.22/0.6032 | 17.65/0.5394 | 17.93/0.5680 | 18.79/0.5227 | 18.24/0.4791 | 18.24/0.4791 |
| **LR3M [41]** | | 8.85/0.3622 | 8.71/0.3254 | 8.88/0.3525 | 10.93/0.3235 | 8.71/0.3254 | 10.95/0.3146 |
| **Retinex-Net [39]** | **+NLSA [29]** | 19.65/0.5909 | 18.57/0.5139 | 19.09/0.5486 | 19.43/0.5299 | 17.67/0.4356 | 18.41/0.4750 |
| **SID [23]** | | 23.44/0.7296 | 20.81/0.5990 | 21.86/0.6533 | 21.42/0.6059 | 19.81/0.5262 | 19.81/0.5262 |
| **HDR-net [1]** | | 23.08/0.7078 | 20.86/0.5928 | 21.78/0.6411 | 23.14/0.6748 | 20.15/0.5489 | 21.32/0.5875 |
| **RUAS [42]** | | 19.19/0.5867 | 18.31/0.5369 | 18.62/0.5216 | 18.66/0.4841 | 17.36/0.4231 | 17.88/0.4194 |
| **DRBN [40]** | | 21.21/0.6559 | 21.27/0.6171 | 21.49/0.6299 | 20.68/0.5429 | 20.27/0.5107 | 20.75/0.5207 |
| **NLSA[29]** | | 24.68/0.7581 | 22.63/0.6749 | 22.78/0.6863 | 23.27/0.6338 | 17.62/0.4659 | 21.50/0.5641 |
| **Ours** | | **24.85/0.7694** | **22.95/0.6908** | **23.44/0.7039** | **23.69/0.6479** | **22.07/0.5830** | **22.25/0.5878** |

TABLE II
ABLATION STUDY SHOWING THE QUANTITATIVE PERFORMANCES (PSNR AND SSIM) WITH INDIVIDUAL MODULES UNDER DIFFERENT SETTINGS: "W/O ANALYSIS," "W/O DEGREES," "W/O FACES," "W/O LL" AND "W/O LR"

| | PSNR | SSIM |
|---|---|---|
| **w/o analysis** | 21.97 | 0.6731 |
| **w/o degrees** | 23.34 | 0.7024 |
| **w/o faces** | 22.70 | 0.6915 |
| **w/o LL** | 23.03 | 0.6960 |
| **w/o LR** | 22.87 | 0.6995 |
| **ours** | **23.44** | **0.7039** |

TABLE III
THE OBJECTIVE METRICS (EA AND CS (1)) FOR DIFFERENT SCALES OF LOW-ILLUMINATION AND LOW-RESOLUTION.

| | EA | CS(1) |
|---|---|---|
| **Low-illumination** | 13.41% | 36.21% |
| **Low-resolution** | 74.08% | 100% |

in terms of measured by the cosine similarity score.

### D. Computational Complexity

Table VI compares the computational complexity of our method with that of other models in terms of parameter size and run-time. Our algorithm consumes longer run-time and more model parameters than the other models employed in the comparison. However, the run-time increase is still reasonable.

### E. Ablation Study

**Settings of Ablation Study**. In this section, we conduct an ablation study to examine the effectiveness of individual modules, focusing on the analysis of low-light, LR, and facial degradation. The test images are randomly down-sampled to different resolutions. The ablation study settings consist of the following: 1) our method with all modules ("Ours"), 2) our method without analysis of facial degradation and features ("w/o analysis"), 3) our method without degradation degree setting ("w/o degrees"), 4) our method without face feature estimation ("w/o faces"), 5) our method without low-light decoupling ("w/o LL"), and 6) our method without LR decoupling analysis ("w/o LR").

Fig. 6 illustrates the results of different module settings. The results without LL and LR exhibit unclear details and artifacts. Their corresponding PSNR and SSIM scores, as shown in Table II, are poor. Analyzing image quality solely based on a single degradation factor is inaccurate in environments with complex types of degradation, leading to subpar reconstruction results. The absence of facial features adversely affects the results of the setting "w/o faces", resulting in imperfect face details. The poor PSNR and SSIM performances highlight the importance of extracting robust facial features in reconstruction. Comparing the quantitative performances of the setting "w/o degrees" in Table II, our method achieves better reconstruction of details degraded by low-light and low-resolution. Moreover, Fig. 6 shows that our method achieves evident subjective performance improvement over the compared models, particularly in recovering facial details. Therefore, introducing greater detail degradation as a training constraint enhances face reconstruction.

TABLE IV
QUANTITATIVE PERFORMANCE (PSNR AND SSIM) COMPARISON UNDER DIFFERENT LR AND LOW-LIGHT CONDITIONS

| | sizes | **HDR-net** | **Super-FAN** | **OURS** |
|---|---|---|---|---|
| $Face_1$ | | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM |
| $\Gamma = 20$ | $56 \times 56$ | 24.50/0.7930 | 20.83/0.7867 | **25.80/0.8500** |
| | $48 \times 48$ | 23.79/0.7682 | 20.08/0.7571 | **25.41/0.8395** |
| | $40 \times 40$ | 23.01/0.7386 | 19.26/0.7254 | **24.75/0.8237** |
| $\Gamma = 40$ | $56 \times 56$ | 16.33/0.7557 | 20.48/0.7890 | **25.45/0.8474** |
| | $48 \times 48$ | 16.09/0.7315 | 20.04/0.7586 | **25.23/0.8380** |
| | $40 \times 40$ | 15.45/0.6991 | 19.06/0.7248 | **24.57/0.8171** |
| $\Gamma = 60$ | $56 \times 56$ | 14.91/0.7366 | 20.96/0.7862 | **24.72/0.8416** |
| | $48 \times 48$ | 14.65/0.7127 | 20.59/0.7587 | **24.17/0.8269** |
| | $40 \times 40$ | 14.14/0.6796 | 20.05/0.7263 | **23.96/0.8062** |
| $\Gamma = 80$ | $56 \times 56$ | 13.71/0.7141 | 20.84/0.7849 | **23.03/0.8298** |
| | $48 \times 48$ | 13.57/0.6924 | 20.52/0.7571 | **22.90/0.8169** |
| | $40 \times 40$ | 12.94/0.6609 | 19.92/0.7242 | **22.53/0.7895** |
| $Face_2$ | | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM |
| $\Gamma = 20$ | $56 \times 56$ | 19.22/0.6368 | 18.61/0.6318 | **21.25/0.7084** |
| | $48 \times 48$ | 18.78/0.5989 | 18.34/0.6036 | **20.80/0.6830** |
| | $40 \times 40$ | 18.63/0.5565 | 17.55/0.5547 | **20.13/0.6418** |
| $\Gamma = 40$ | $56 \times 56$ | 17.38/0.6052 | 19.15/0.6287 | **20.90/0.6908** |
| | $48 \times 48$ | 16.89/0.5693 | 18.77/0.6020 | **20.53/0.6674** |
| | $40 \times 40$ | 16.84/0.5299 | 18.08/0.5558 | **19.88/0.6272** |
| $\Gamma = 60$ | $56 \times 56$ | 14.69/0.5507 | 19.01/0.6180 | **20.38/0.6701** |
| | $48 \times 48$ | 14.73/0.5263 | 18.63/0.5894 | **20.00/0.6448** |
| | $40 \times 40$ | 14.75/0.4946 | 18.12/0.5499 | **19.54/0.6094** |
| $\Gamma = 80$ | $56 \times 56$ | 13.03/0.5106 | 18.75/0.6094 | **19.92/0.6555** |
| | $48 \times 48$ | 12.97/0.4863 | 18.42/0.5817 | **19.59/0.6325** |
| | $40 \times 40$ | 12.85/0.4548 | 17.80/0.5387 | **19.16/0.5913** |

**Degradation Parameter Validation**. We also validate the estimated degree of low-light and LR degradation. We conduct the corresponding test on CelebA, using the same random low-quality dataset as described in Sec. IV-B. The test results involve different levels of random low-light and LR degradation. To assess the accuracy of these estimates, we use objective metrics such as Exact Accuracy (EA) and Cumulative Score (CS) [58]. Specifically, EA measures the correctness of illumination and resolution labels within a specified tolerance level $n$, while CS calculates the percentage of test labels with absolute errors less than or equal to $n$. In Table III, we present the scale accuracy of resolution and illumination, showing that the resolution scale accuracy is highly precise, with CS(1) even approaching 100%. However, the illumination scale accuracy is relatively lower, as obtaining accurate values that express the extent of illumination reduction for different pixels in the images is challenging. Nonetheless, the illumination enhancement decoder compensates for the inaccuracies in illumination analysis to some extent, leading to promising experimental results for low-light enhancement.

### F. Performances on Real-World Dark Face Images

*1) Validation on D-faces+:* We also conduct evaluations of our method on real-world images, which are recovered using models trained on synthetic low-light images from CelebA. Fig. 7 shows a collection of frontal and non-frontal face images captured under various illumination conditions, including surveillance camera footage and other real low-lighting facial datasets. The Size and $M\_I$ values correspond to the resolution and mean intensity of the images. The two bottom images depict the same person under different lighting conditions. The input low-light images are interpolated to the target size using the BICUBIC method. It is evident

TABLE V
COMPARISON OF COSINE SIMILARITY SCORES BASED ON ARCFACE EMBEDDING FOR EVALUATING THE FACE RECOGNITION PERFORMANCE ON RECONSTRUCTED FACES WITH VARIOUS RESTORATION MODELS, WHERE A HIGHER VALUE INDICATES A BETTER PERFORMANCE

| | SID | HDR-net | DRBN | Wavelet-SRNet | Super-FAN | NLSA | Ours |
|---|---|---|---|---|---|---|---|
| **CelebA** | 0.4926 | 0.4883 | 0.4868 | 0.4986 | 0.4542 | 0.5017 | **0.5251** |
| **Helen** | 0.4906 | 0.4823 | 0.4666 | 0.4566 | 0.4870 | 0.4054 | **0.5238** |

TABLE VI
COMPUTATIONAL COMPLEXITY COMPARISON AMONG THE METHODS IN TERMS OF MODEL PARAMETER SIZE (M) AND RUN-TIME (MS)

| | SID | HDR-net | DRBN | RUAS | Wavelet-SRNet | Super-FAN | NLSA | Ours |
|---|---|---|---|---|---|---|---|---|
| **Parameter size (M)** | 7.76 | 0.43 | 0.56 | 0.003 | 7499.93 | 1.30 | 1.81 | 19.90 |
| **Run-time (ms)** | 14.0 | 10.0 | 56.4 | 18.4 | 38.2 | 21.0 | 18.0 | 63.8 |


Fig. 6. Ablation study examining the effectiveness of individual modules, where six variants are compared: "our method," "w/o analysis," "w/o degrees," "w/o faces," "w/o LL," and "w/o LR."


Fig. 7. Qualitative performance comparison of faces with different sizes captured in real nighttime scenes.

TABLE VII
SUBJECTIVE EVALUATION SCORES OF VARIOUS RESTORATION MODELS ON FACE IMAGES CAPTURED AT NIGHTTIME

| | **Subjective Scores** |
|---|---|
| HDR-net | 3.2 |
| HDR-net+Super-FAN | 3.2 |
| HDR-net+Wavelet-SRNet | 3.3 |
| Super-FAN | 5.5 |
| Wavelet-SRNet | 3.7 |
| OURS | **6.8** |

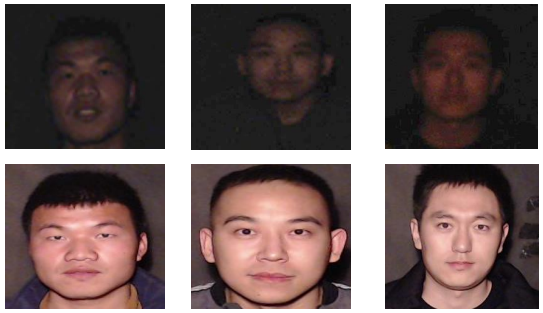


Fig. 8. Three samples of low-quality (top) and high-quality (bottom) face pairs.

from the linearly zoomed images that the inputs suffer from low-light degradation and blurs. The results from Super-FAN and Wavelet-SRNet exhibit artifacts and lack clear details. While the results from SID achieve excellent enhancement in low light, they lack clear facial texture, especially for low-resolution images. In comparison, our method not only recovers the most detailed information without artifacts but also ensures clear and recognizable faces. Our approach effectively removes noise while preserving natural lighting.

We further evaluate the performances based on the non-reference metrics CEIQ [59] and NIQE [60], and the cosine similarity of ArcFace embedding for face recognition on real faces. These non-reference metrics provide insights into image quality. Additionally, we utilize cosine similarity to measure the improvement achieved by our method. To compare the results, we capture images of the same individuals under both


Fig. 9. Illustration of failure cases of face restoration in real nighttime scenes.

TABLE VIII
COMPARISON OF NON-REFERENCE QUALITY METRICS (CEIQ [59] AND NIQE [60]) AND THE COSINE SIMILARITY OF ARCFACE EMBEDDING FOR FACE RECOGNITION ON REAL FACES

| | **CEIQ** | **NIQE** | **cosine similarity** | | |
|---|---|---|---|---|---|
| | | | C_Fs1 | C_Fs2 | C_Fs3 |
| HDR-net | 2.71 | **29.11** | 0.2784 | 0.2261 | 0.5335 |
| HDR-net+Super-FAN | 2.79 | 38.65 | 0.2764 | 0.2213 | 0.5138 |
| HDR-net+Wavelet-SRNet | 2.87 | 29.85 | 0.3139 | 0.2358 | 0.4950 |
| Super-FAN | **3.42** | 41.79 | 0.3883 | 0.3761 | 0.4403 |
| Wavelet-SRNet | 3.40 | 36.81 | 0.4043 | 0.3399 | 0.4753 |
| OURS | 2.97 | 36.28 | **0.4115** | **0.4573** | **0.5387** |

low-quality and high-quality conditions, as shown in Fig. 8. Subsequently, we enhance the low-quality images and measure the similarity between these enhanced images and the high-quality images of different individuals.

We also evaluate the subjective quality of the restored faces. Since the corresponding ground-truth images are not available, we cannot measure PSNR and SSIM. To assess the subjective quality of the reconstructed images, we randomly select image groups from our algorithm and a comparison algorithm. Each group consists of 6 cases of reconstructed images, which are then ranked by subjects based on their subjective scores ranging from 1 to 10, where a higher score indicates a better subjective quality. The results, obtained through a significant number of statistics, are presented in Table VIII, demonstrating that our subjective scores are higher overall.

*2) Failure Cases:* Our algorithm does have limitations when it comes to the resolution of real images. Fig. 9 depicts some failure cases, particularly in extreme cases such as very low resolution (the first row), extremely low light (the second row), and complex illumination conditions (the third row). An insufficient resolution leads to blurry results, while inadequate illumination causes color distortion. Moreover, our method may fail to enhance images in complex illumination scenes. These limitations arise partly due to expressiveness constraints used in our method and partly due to the limited recoverable information from the data source. Real-world degradation involves various factors, including resolution, illumination, and other complex scenarios that are challenging to address with a generic decoupling model. Although Super-FAN performs better in low-quality faces with complex illumination, it still suffers from artifacts. Overall, our model applies to most real degradation scenes. However, in our future work, we aim to develop more comprehensive representation models to handle complex degradation in real scenes, making them applicable in various scenarios.

## V. CONCLUSION

Our work addresses the issue of enhancing low-quality, low-light face images affected by complex degradation in real scenes. The presence of random and complex degradation factors poses challenges for existing face restoration methods. We approach low-quality face enhancement in a complex environment by parsing and feature extraction. We proposed to learn robust facial features by decoupling the degradation process to achieve precise reconstruction and representation of the face.From our experiments, we have identified certain limitations in our method. In specific real-world scenarios, such as

extremely low resolution, extremely low-light conditions, and non-uniform illumination, our algorithm struggles to achieve satisfactory recovery of face images. Additionally, there are some flaws in the details of our results, particularly in the eyes. In future work, we aim to analyze the degradation process and intricate details of face images more comprehensively. We anticipate improved details recovery in these challenging situations by incorporating specific constraints.

## REFERENCES


[1] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graphics*, vol. 36, no. 4, pp. 1–12, 2017.

[2] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 109–117.

[3] S. Baker and Kanade, "Hallucinating faces," in *Proc. IEEE Int. Conf.n Autom. Face Gesture Recognit.*, 2000, pp. 83–88.

[4] S. Zhang, S. Chang, and Y. Lin, "End-to-end light field spatial super-resolution network using multiple epipolar geometry," *IEEE Trans. Image Process.*, vol. 30, pp. 5956–5968, 2021.

[5] T. Guo, H. Seyed Mousavi, and V. Monga, "Adaptive transform domain image super-resolution via orthogonally regularized deep networks," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4685–4700, 2019.

[6] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2015.

[7] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[8] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 624–632.

[9] C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung, "SiGAN: Siamese generative adversarial network for identity-preserving face hallucination," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6225–6236, Dec. 2019.

[10] H.-C. Shao, K.-Y. Liu, W.-T. Su, C.-W. Lin, and J. Lu, "DotFAN: A domain-transferred face augmentation net," *IEEE Trans. Image Process.*, vol. 30, pp. 8759–8772, Oct. 2021.

[11] W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, and Z. Wang, "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020.

[12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[13] I. Higgins, L. Chang, V. Langston, D. Hassabis, C. Summerfield, D. Tsao, and M. Botvinick, "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons," *Nature commun.*, vol. 12, no. 1, pp. 1–14, 2021.

[14] N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin, "Accurate blur models vs. image priors in single image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2832–2839.

[15] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1671–1681.

[16] R. Zhou and S. Susstrunk, "Kernel modeling super-resolution on real low-resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis.*, 2019, pp. 2433–2443.

[17] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," vol. 32, 2019.

[18] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, "Real-world super-resolution via kernel estimation and noise injection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 466–467.

[19] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proc. European Conf. Comput. Vis.*, 2018, pp. 185–200.

[20] B.-K. Sefi, S. Assaf, and I. Michal, "Blind super-resolution kernel estimation using an internal-gan," in *Proc. Conf. Neural Inf. Process. Syst.*, 2019, pp. 284–293.

[21] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," *arXiv preprint arXiv:2104.00416*, 2021.

[22] W.-Z. Shao, J.-J. Xu, L. Chen, Q. Ge, L.-Q. Wang, B.-K. Bao, and H.-B. Li, "On potentials of regularized wasserstein generative adversarial networks for realistic hallucination of tiny faces," *Neurocomputing*, vol. 364, pp. 1–15, 2019.

[23] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 666–682.

[24] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5569–5578.

[25] J. Chen, J. Chen, Z. Wang, C. Liang, and C.-W. Lin, "Identity-aware face super-resolution for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 27, pp. 645–649, Apr. 2020.

[26] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, "Atmfn: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2734–2747, 2019.

[27] R. Wang, M. Jian, H. Yu, L. Wang, and B. Yang, "Face hallucination using multisource references and cross-scale dual residual fusion mechanism," *In. J. Intell. Syst.*, vol. 37, no. 11, pp. 9982–10 000, 2022.

[28] H. Huang, H. Ran, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1689–1697.

[29] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3517–3526.

[30] C. Wang, J. Jiang, Z. Zhong, and X. Liu, "Propagating facial prior knowledge for multitask learning in face super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7317–7331, 2022.

[31] T. Liu, M. Xu, S. Li, R. Ding, and H. Liu, "Mrs-net+ for enhancing face quality of compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2881–2894, 2022.

[32] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, "Real-world super-resolution of face-images from surveillance cameras," *IET Image Process.*, vol. 16, no. 2, pp. 442–452, 2022.

[33] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3063–3072.

[34] X. Han, H. Yang, G. Xing, and Y. Liu, "Asymmetric joint gans for normalizing face illumination from a single image," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1619–1633, 2019.

[35] Y. Gao, H.-M. Hu, B. Li, and Q. Guo, "Naturalness preserved nonuniform illumination estimation for image enhancement based on retinex," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 335–344, 2017.

[36] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 451–462, 1997.

[37] 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, 1997.

[38] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, 2017.

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

[40] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3060–3069.

[41] X. Ren, W. Yang, W.-H. Cheng, and J. Liu, "Lr3m: Robust low-light enhancement via low-rank regularized retinex model," *IEEE Trans. Image Process.*, vol. 29, pp. 5862–5876, 2020.

[42] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10 561–10 570.

[43] Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, and F. Kuang, "Retinexdip: A unified deep framework for low-light image enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1076–1088, 2021.

[44] G.-D. Fan, B. Fan, M. Gan, G.-Y. Chen, and C. P. Chen, "Multiscale low-light image enhancement network with illumination constraint," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7403–7417, 2022.
[45] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[46] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "Gan inversion: A survey," *arXiv preprint arXiv:2101.05278*, 2021.
[47] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Comput.d Comput.-Assist. Intervention*, 2015, pp. 234–241.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv: Learning*, 2014.
[49] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *Proc. Conf. Neural Inf. Process. Syst.*, 2017.
[50] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.
[51] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1712–1722.
[52] Y. Ren, Z. Ying, T. H. Li, and G. Li, "Lecarm: Low-light image enhancement using the camera response model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 968–981, 2018.
[53] M. D. Grossberg and S. K. Nayar, "What is the space of camera response functions?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2003, pp. II–602.
[54] C. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5549–5558.
[55] V. Le, J. Brandt, Z. Lin, L. D. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. European Conf. Comput. Vis.*, 2012, pp. 679–692.
[56] Zll, "face-parsing.pytorch," https://github.com/zllrunning/face-parsing.PyTorch.
[57] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.
[58] K. Lim, N.-H. Shin, Y.-Y. Lee, and C.-S. Kim, "Order learning and its application to age estimation," in *Proc. Int. Conf. Learn. Represent.*, 2019.
[59] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 838–842, 2014.
[60] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, 1995.