Low-FaceNet: Face Recognition-Driven Low-Light Image Enhancement

Yihua Fan, Yongzhen Wang, Dong Liang, *Member, IEEE*, Yiping Chen, *Senior Member, IEEE*, Haoran Xie, *Senior Member, IEEE*, Fu Lee Wang, *Senior Member, IEEE*, Jonathan Li, *Fellow, IEEE*, and Mingqiang Wei, *Senior Member, IEEE*

***Abstract*— Images captured in low-light conditions often induce the performance degradation of cutting-edge face recognition models. The missing and wrong face recognition inevitably makes vision-based systems operate poorly. In this article, we propose Low-FaceNet, a novel face recognition-driven network, to make low-light image enhancement (LLE) interact with high-level recognition for realizing mutual gain under a unified deep learning framework. Unlike existing methods, Low-FaceNet uniquely brightens real-world images by unsupervised contrastive learning and absorbs the wisdom of facial understanding. Low-FaceNet possesses an image enhancement network that is assembled by four key modules: a contrastive learning module, a feature extraction module, a semantic segmentation module, and a face recognition module. These modules enable Low-FaceNet to not only improve the brightness contrast and retain features but also increase the accuracy of recognizing faces in low-light conditions. Furthermore, we establish a new dataset of low-light face images called LaPa-Face. It includes detailed annotations with 11 categories of facial features and identity labels. Extensive experiments demonstrate our superiority against the state-of-the-art methods of both LLE and face recognition even without ground-truth image labels. Our code and dataset are available at https://github.com/fanyihua0309/Low-FaceNet.**

***Index Terms*— Contrastive learning, face recognition, Low-FaceNet, low-light image enhancement, semantic segmentation.**

Manuscript received 13 September 2023; revised 31 December 2023; accepted 2 February 2024. Date of publication 21 March 2024; date of current version 8 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant T2322012, Grant 62172218, and Grant 62032011; in part by the Shenzhen Science and Technology Program under Grant JCYJ20220818103401003 and Grant JCYJ20220530172403007; and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010170. The Associate Editor coordinating the review process was Dr. Chuang Sun. *(Corresponding author: Mingqiang Wei.)*

Yihua Fan, Dong Liang, and Mingqiang Wei are with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with the Shenzhen Institute of Research, Nanjing University of Aeronautics and Astronautics, Shenzhen 518038, China (e-mail: fanyihua@nuaa.edu.cn; liangdong@nuaa.edu.cn; mingqiang.wei@gmail.com).

Yongzhen Wang is with the School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243032, China (e-mail: wangyz@ahut.edu.cn).

Yiping Chen is with the School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai 528478, China (e-mail: chenyp79@mail.sysu.edu.cn).

Haoran Xie is with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China (e-mail: hrxie@ln.edu.hk).

Fu Lee Wang is with the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China (e-mail: pwang@hkmu.edu.hk).

Jonathan Li is with the Department of Geography and Environmental Management and the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

Digital Object Identifier 10.1109/TIM.2024.3372230

I. Introduction

FACE images captured in low-light conditions often suffer from unfavorable visibility and color bias [1], [2], which not only affect human visual quality but also drastically worsen the performance of face recognition networks [3]. Such degradation has a negative impact on the image-based measurement methods and advanced vision tasks [4], [5]. One promising solution to mitigate the performance drop in low-light conditions is to employ the supplementary lighting. However, not all scenarios support it, since practical constraints such as cost, power limitations, and the need for covert operations often make the supplementary lighting unfeasible. Low-light image enhancement (LLE) aims at brightening the illumination to make the information hidden in the dark visible and improve image quality. LLE is drawing much attention in multiple emerging computer vision areas, especially in the field of face recognition. As shown in Fig. 1, the unmanned patrol vehicle is often developed for surveillance and campus safety protections in university. When going on patrol at night, it may suffer from the inaccurate face recognition due to the low-light conditions. If equipping the vehicle with LLE techniques, it is able to enhance nighttime images captured by the onboard camera and simultaneously perform facial recognition. This enables improved nighttime surveillance and campus safety protections. However, we observe that even the best face recognition models struggle with low-light face images. While existing LLE models can improve these low-light images, the "re-lighted" faces may not serve face recognition successfully.

Why Does It Happen? There are mainly three reasons.

1) Existing LLE methods concentrate on pixel-level loss functions and fail to model the geometric and semantic information. They will cause uneven exposure and unrealistic details, damaging the recognition performance.
2) They exploit normal-light images to guide the optimization process while neglecting low-light images which are an effective training source.
3) LLE models are typically applied as an independent task, which cannot perceive the downstream application. Therefore, we attempt to investigate the benefit of exploiting facial semantics and low-light images as negative samples simultaneously for LLE and face recognition. To this end, we propose a task-driven paradigm to make high-level recognition interact with low-level image enhancement. The two tasks realize mutual gain under a unified deep learning framework.

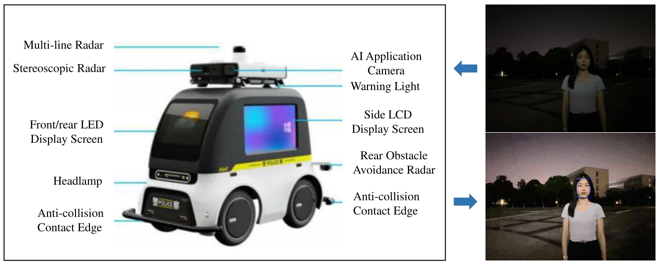

Multi-line Radar
Stereoscopic Radar
Front/rear LED Display Screen
Headlamp
Anti-collision Contact Edge
AI Application Camera
Warning Light
Side LCD Display Screen
Rear Obstacle Avoidance Radar
Anti-collision Contact Edge

Fig. 1. Diagram of an unmanned patrol vehicle at NUAA. The onboard camera is capable of capturing images and empowering intelligent decision–making. However, nighttime images inevitably suffer from poor visibility and unknown noise. Such degradation negatively impacts vision-based measurement systems and possibly even puts traffic safety at risk. To mitigate this, Low-FaceNet can potentially help achieve visually pleasing and realistic enhanced outputs to facilitate face recognition.

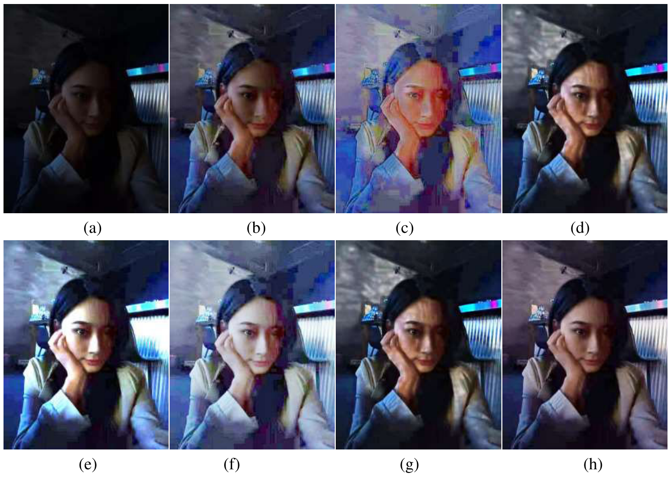


Fig. 2. Image enhancement results of a real-world low-light example. (a) Input low-light image, and the enhancement results of (b) LIME [7], (c) RetinexNet [10], (d) SSIENet [16], (e) RUAS [17], (f) Zero-DCE++ [18], (g) SCL-LLE [19], and (h) our Low-FaceNet. By leveraging facial semantics and low-light images, which are often ignored in previous approaches, our Low-FaceNet achieves results that are not only brighter but also more visually appealing.

A variety of conventional and deep learning-based techniques are developed to enhance the visual perception of low-light images. The conventional ones are often based on histogram equalization (HE) [6] or the Retinex theory [7], [8]. Although these methods are simple and easy to implement, hand-crafted priors do not always hold in them, leading to the unrealistic enhancement and a heavy computing burden.

The compelling performance of deep learning has shed light on the LLE field. Learning-based methods can fall into three categories: supervised [9], [10], [11], [12], unsupervised [13], [14], and semi-supervised methods [15]. Recent years have witnessed the impressive success of supervised methods based on synthetic data. However, collecting large-scale diverse paired data for supervised learning is often unrealistic, and training on synthetic datasets may cause overfitting and poor generality. Besides, how to achieve stable network training and establish connections between different domains remain intricate for semi-supervised and unsupervised methods. Furthermore, previous methods usually only utilize normal-light images as positive samples for training, ignoring the large amount of accessible underexposed images and semantic guidance that can facilitate the LLE task.

Most of existing LLE methods cannot fully connect LLE with face recognition. In contrast, we propose a face recognition-driven strategy that guides the model to work well for both improving image quality and enhancing recognition accuracy. Using unpaired images from normal-light and low-light conditions as positive and negative examples, we build a more general and discriminative network. The guidance of semantic information guarantees relatively even exposure of each part of the face. We further implement face recognition-driven embedding to promote both the low-level LLE task and the high-level recognition task and finally propose Low-FaceNet. Low-FaceNet contains an image enhancement network and four modules, with LLE acting as four specific constraints: learning contrasts, retaining features, smoothing brightness guided by semantics, and recognizing faces. Together, these constraints guarantee consistent lighting, color preservation, and improvements in recognition performance. Low-FaceNet achieves a visually better result, effectively supporting the face recognition task (see Fig. 2).

To facilitate the training of Low-FaceNet and encourage more works in this field, we build and release a new face-oriented LLE dataset. It contains paired normal-light/low-light face images with different race, age, expression, and pose. And each image has a corresponding semantic label map and identity label.

Experiments show clear improvements of Low-FaceNet over its competitors on LLE and face recognition. In summary, our contributions are threefold.

1) We propose Low-FaceNet, a task-driven paradigm to promote both low-level low-light image enhancement and high-level face recognition.
2) We exploit unpaired normal-light/low-light images as positive/negative samples. By making low-light image enhancement interact with facial understanding, we ensure the realistic restoration of face images.
3) We create LaPa-Face, a specialized benchmark for low-light image enhancement geared toward face recognition. It includes both low-light and normal-light images, with detailed semantic annotations and identity labels. LaPa-Face provides the vision community with a new dataset of low-light conditions.

II. Related Work

A. Low-Light Image Enhancement

LLE has been extensively studied to improve the visibility and reveal the hidden information. Existing solutions generally fall into prior-based and learning-based methods.

Traditional methods typically rely on hand-crafted priors, such as HE [6] and Retinex priors [7]. Retinex theory [20] which assumes that an image can be decomposed into reflectance and illumination, has gained significant attention. Mathematically, a given image can be expressed by

$$S = R \times I \tag{1}$$

where S denotes the source image, R and I denote the reflectance and illumination, respectively, and $\times$ denotes the pixel-wise product. Enhanced results are obtained by further adjusting the two components. Albeit these traditional methods enhance image brightness to some extent, they often exhibit a limited practical capacity under complex and diverse real-world scenes.

In recent years, deep learning-based methods have produced promising results in LLE. Supervised learning approaches utilize paired normal-light and low-light images for training. Retinex-based methods [10], [21] incorporate Retinex theory to decompose the images and refine the decomposed components to obtain the final enhanced images. In contrast, end-to-end methods [9], [22], [23], [24] directly learn the mapping from low-light input images to corresponding enhanced images. Despite their decent performance, the domain gap between real and synthetic data often results in poor generality and potential overfitting when using synthetic training data.

In practice, it is challenging or even impractical to obtain paired normal-light/low-light data of the same scene. To address this issue, RUAS [17] adopts a Retinex-inspired unrolling scheme with a network-searched architecture. Zero-DCE [14] formulates LLE as an image-specific curve estimation task and benefits from a set of well-designed nonreference loss functions. SCL-LLE [19] removes pixel-correspond paired training data through an effective semantically contrastive learning paradigm. However, achieving stable network training and creating cross-domain information relations remain nontrivial. In this work, inspired by [19], we design a simple Retinex-based enhancement network but gain superior performance by leveraging unpaired training data and semantic information.

B. Face Recognition

Face recognition is a critical problem in the realm of computer vision as it has a wide range of real-world applications, such as access control, face unlocking, security surveillance, financial payment, etc. A typical face recognition pipeline involves the following four main steps.

1) *Face Detection:* This initial step aims at estimating the bounding box of the face in a given image. General object detection algorithms [25], [26], [27], [28] have shed light on the development of face detectors.
2) *Facial Landmark Detection:* The goal of this step is to identify key facial points, such as eyes, nose, and mouth [29]. These landmarks hold great importance in face alignment, contributing to recognition accuracy improvements.
3) *Facial Features Extraction:* This phase focuses on extracting essential facial features by taking advantage of effective network architectures.
4) *Facial Features Classification:* Facial features extracted in the previous step are employed for classification through various classifier algorithms.

Compared with general face recognition, limited research efforts have been dedicated to face recognition in low-light conditions. The prevailing method is to pre-process the low-light images by existing LLE approaches, and then feed the processed images into the subsequent recognition network. Although the overall quality of low-light images is improved, they do not necessarily guarantee an improvement in recognition performance. Recent research [30], [31], [32] also reveals that there is no straightforward cause-and-effect relation between the visual quality of enhanced images and the performance of high-level recognition tasks. In response to this challenge, we embrace a recognition task-driven paradigm that effectively bridges the gap between low-level enhancement and high-level recognition and achieves recognition-friendly enhancement.

III. Methodology

This section illustrates the details of our designed Low-FaceNet for LLE. Section III-A reveals our Retinex-based Enhance-Net network structure. The following Sections III-B–III-E demonstrate the four modules and loss terms adopted in the proposed framework at length.

Face recognition under adverse illumination conditions is still challenging. Intuitively, it may get better performance on the enhanced image. However, most of the previous LLE methods overlook how facial semantics and underexposed images can bring significant gain to both low-level enhancement and high-level recognition. Besides, LLE is typically applied as an independent pre-processing step, which might be suboptimal for the ultimate goal. As far as we know, few works pay attention to the improvement of recognition performance after enhancement. The problem of how low-level image processing could affect the high-level recognition task is still not thoroughly studied. In this work, we offer a new insight to investigate the logical relationship between LLE and face recognition by showing the mutual influence between them. We not only exploit facial semantics and underexposed images to generate both brighter and more realistic images but also propose a task-driven manner to solve both the low-level image enhancement and high-level recognition problems in a single unified framework.

As schematically illustrated in Fig. 3, Low-FaceNet is congregated by an image enhancement network and four key modules, i.e., a contrastive learning module, a feature extraction module, a semantic segmentation module, and a face recognition module. Since the joint framework includes multiple modules dedicated to image enhancement, semantic segmentation, and face recognition tasks, it inevitably introduces more uncertainties to the training process if they are initialized entirely randomly. To mitigate this challenge, the semantic segmentation and face recognition network are trained in advance and frozen during the learning process of the image enhancement network. This strategy ensures fast convergence and high performance of our method. Specifically, given a low-light input, the image enhancement network is first applied. Then the enhanced result is fed into the following four modules, where LLE is converted into four constraints for detail retention, color presentation, and exposure control. We provide details of the framework in what follows.

A. Retinex-Based Enhance-Net

Retinex theory models the color perception of human vision. It inspires us to integrate the strengths of model-based and learning-based methods. We formulate LLE as the

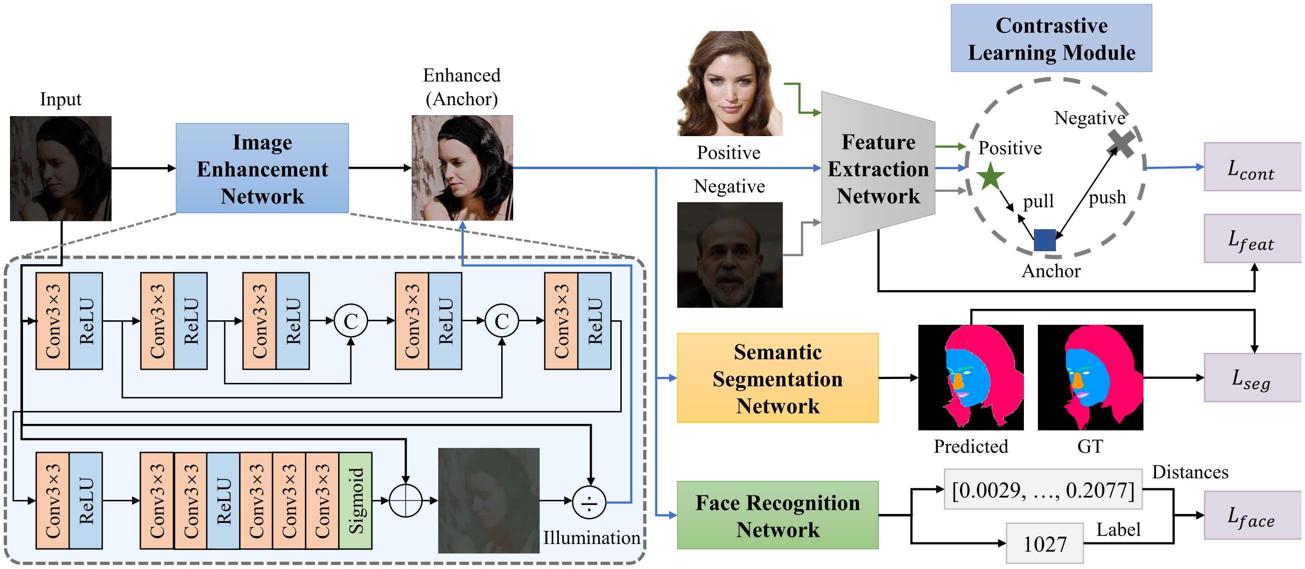


Fig. 3. Architecture of Low-FaceNet. Our network consists of an image enhancement network and four modules: a contrastive learning module, a feature extraction module, a semantic segmentation module, and a face recognition module, which perform contrastive brightness restoration and feature retention, semantic-guided smoothness, and face recognition, respectively.

combination of prior and learning, unfolding the update optimization steps into a neural network. Whereas, recovering two components from one single image is obviously an ill-posed problem. Instead of simultaneously estimating these two components in parallel, we alternatively resort to first estimating the illumination component I and then derive the reflectance component R by (1), where R is considered as the enhanced image.

The network architecture of Retinex-based Enhance-Net is illustrated in Fig. 3. We use simple convolution layers with concatenation operations to form multiscale features, and the intermediate connections compensate for the information loss during convolutions. Notably, to avoid the denominator being zero, the estimated illumination component I is adjusted by $I = \text{clamp}\{I, 0.0001, 1\}$, which is the clipping operator to drop the overflow value with the upper/lower bounds being $1/0.0001$. We learn the residual representation of I followed by a plus calculation instead of learning illumination directly, which not only guarantees exposure control and steadiness but also reduces the computational difficulty.

B. Contrastive Learning Module

Contrastive learning is enforced to learn a representation by encouraging the positive pairs closer while keeping the negative pairs further away. Therefore, our primary insight is that the features extracted from the enhanced results and the positive samples (i.e., normal-light images) should share some mutual properties, while the enhanced results and the negative samples (i.e., low-light images) should have a long distance between their embeddings. In the consideration of flexibility, the positive samples and negative samples can be randomly selected in different scenes, which means they are unpaired with each other. We enforce two contrastive constraints to learn LLE in the deep feature space. The detailed descriptions are provided below.

1) Feature CR: The goal is to learn a representation to pull together positive pairs in the latent feature space and push apart the representation between negative pairs. Inspired by [33], we employ contrastive regularization (CR) to cluster the latent feature space by

$$L_{\text{CR}_{\text{feature}}} = \sum_{i=1}^{n} w_i \cdot \frac{\left|F_i(I_e), F_i\left(I_p\right)\right|}{\left|F_i(I_e), F_i(I_n)\right| + \alpha} \tag{2}$$

where I_e, I_p, and I_n refer to the enhanced image, positive sample, and negative sample respectively. F_i denotes the ith layer of features extracted by the pretrained VGG-16 model [34]. n means the total number of feature layers. w_i is a weight coefficient. α is a small constant to prevent the denominator from being zero, which is set to 1×10^{-7}.

2) Brightness CR: Additionally, the brightness CR which has a similar form of Feature CR is proposed to brighten up the illumination and further constrain the optimization by

$$L_{\text{CR}_{\text{brightness}}} = \frac{\left|B_{I_e}, B_{I_p}\right|}{\left|B_{I_e}, B_{I_n}\right| + \alpha} \tag{3}$$

where B represents the brightness level of a given image. The rest of the variables have the same meanings as above.

The total contrastive learning loss is formulated as

$$L_{\text{cont}} = \lambda_f L_{\text{CR}_{\text{feature}}} + \lambda_b L_{\text{CR}_{\text{brightness}}} \tag{4}$$

where λ_f and λ_b represent two corresponding trade-off parameters. We set $\lambda_f = 0.62$, and $\lambda_l = 0.22$ in experiments.

C. Feature Extraction Module

To enhance the flexibility of our method, we do not use normal-light images as references to guide the training. In the absence of reference images, well-defined nonreference constraints are the key to stable enhancement. Consequently, we introduce three nonreference constraints, i.e., perceptual loss, color loss, and smooth loss. These constraints collectively ensure that the enhanced images appear to be more perceptually natural, maintain natural color tones, and exhibit a smooth overall structure with textural details.

1) Perceptual Loss: Perceptual Loss ensures the input and output images to be perceptually consistent by

$$L_{\text{perceptual}} = \frac{1}{C_l W_l H_l}\left(f^l(I_l) - f^l(I_e)\right)^2 \tag{5}$$

where $f^l(I_l)$ denotes the feature f of the input I_l in the layer l, and $f^l(I_e)$ is the feature of the enhanced image I_e in the layer l. $C_l W_l H_l$ refers to the size of feature map in the layer l. The features are extracted from a pre-trained VGG-16 model.

2) Color Loss: The color naturalness is one of the significant concerns of LLE. Inspired by [35], we compute the distance between the enhanced images and target images (i.e., positive samples) to measure the major color difference, which is invariance to small distortions. To be specific, we first define a fixed 2-D Gaussian blur operator, then compute the blur version of the enhanced image and random image picked from the positive samples, and finally compute L_2 loss between the two blur images, which can be denoted as

$$\begin{cases} G(k,l) = A \cdot \exp\left(-\frac{(k-\mu_x)^2}{2\sigma_x} - \frac{\left(l-\mu_y\right)^2}{2\sigma_y}\right) \\ I_{eb}(i,j) = \sum_{k,l} I_e(i+k, j+l) \cdot G(k,l) \\ I_{pb}(i,j) = \sum_{k,l} I_p(i+k, j+l) \cdot G(k,l) \\ L_{\text{color}} = ||I_{eb} - I_{pb}||_2^2 \end{cases} \tag{6}$$

where $G(k,l)$ refers to the 2-D Gaussian blur operator, $A = 0.053$, $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 3$. I_{eb}, and I_{pb} refer to the corresponding blurred image of the enhanced image I_e and a random positive sample I_p, respectively.

3) Smooth Loss: The smoothness property of illumination is vital for LLE, which has been revealed in [17] and [36]. Therefore, we present a smooth loss to encourage our network to output an illumination map with a smooth overall structure and textural details, which can be expressed as

$$\begin{cases} w_k = \exp\left(-\frac{\sum_{c\in\{R,G,B\}}\left(\nabla_k I_l^c\right)^2}{2\sigma^2}\right) \\ L_{\text{smooth}} = \sum_{k=1}^{K} w_k \cdot \left(\sum_{c\in\{R,G,B\}} \nabla_k I_i^c\right) \end{cases} \tag{7}$$

where c refers to the image channel in the RGB space. I_l^c refers to the cth channel of the input low-light image, and I_i^c refers to the cth channel of the illumination map learned by our network. k denotes the gradient operation in kth direction, and $K = 24$ denotes the total number of directions. $\sigma = 0.1$ is the standard deviation for the Gaussian kernels.

The total feature loss is formulated as

$$L_{\text{feat}} = \lambda_p L_{\text{perceptual}} + \lambda_c L_{\text{color}} + \lambda_s L_{\text{smooth}} \tag{8}$$

where λ_p, λ_c, λ_s represent several balancing hyperparameters. We set $\lambda_p = 1$, $\lambda_c = 0.05$, and $\lambda_s = 0.65$ We set $\lambda_f = 0.62$, and $\lambda_l = 0.22$ in experiments.

D. Semantic Segmentation Module

The significance of semantic information guidance is a broad consensus in LLE [19] and other low-level visual tasks. Most of the existing LLE methods focus on pixel-level loss functions and thus fail to adequately capture the geometric and semantic information, leading to unfavorable details and uneven exposures. To alleviate such issues, we integrate semantic information to promote the consistency of enhanced face images. We primarily consider that elements belonging to the same semantic category have an adjacent location and should exhibit consistent brightness. To this end, we introduce the semantic smooth loss to enhance the smoothness and consistency within each semantic part. By default, the well-known DeepLabv3+ [37] is employed as the semantic segmentation network, which is pretrained on LaPa-Face and frozen during the training process of our enhancement network. The semantic smooth loss can be denoted as

$$\begin{cases} B_s = \frac{1}{n}\sum_{i\in\theta_s}\left(B_{I_e}^i\right) \\ L_{\text{seg}} = \sum_{s=1}^{S}\sum_{i\in\theta_s}\left(B_{I_e}^i - B_s^i\right)^2 \end{cases} \tag{9}$$

where s represents the sth category. S denotes the total number of semantic categories. θ_s represents the collection of pixels belonging to the category s. n is the total number of pixels of θ_s. $B_{I_e}^i$ is the brightness level of the enhanced image at the ith pixel. B_s is defined as the average brightness level in the category s. That means all pixels of the same category are pushed closer to the average brightness level, leading to smoothness and consistency of each face part. As a result, the enhanced image will be more smooth and more consistent in the same semantic category, which is critical for avoiding local uneven exposures.

E. Face Recognition Module

The pipeline of face recognition for a given image is depicted in Fig. 4. The process begins with the construction of a face database and feature encoding. Subsequently, we calculate the cosine distances between the feature of the given image and the features of all faces within the database. The identity associated with the feature that yields the largest cosine distance is considered the outcome of the recognition process.

In the realm of face recognition in low-light conditions, existing LLE methods are often executed as an independent stage and are thus poorly related to the downstream tasks. Although the pre-processed low-light images exhibit enhanced visual quality, they are not guaranteed to bring gains in high-level computer vision tasks. We attribute this phenomenon to the divergent objectives of enhancement and recognition, which can cause potential conflicts between them. Images generated by enhancement models may contain some

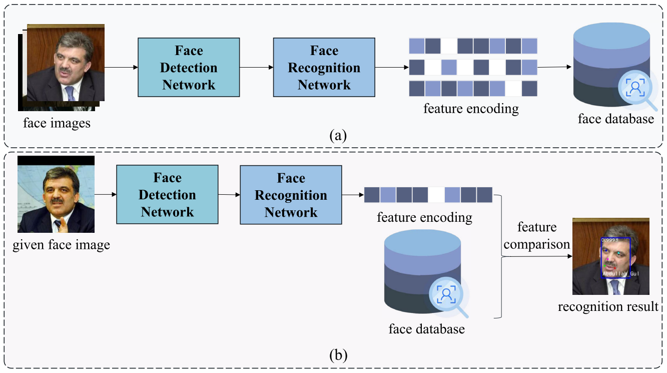


Fig. 4. Pipeline of face recognition. We first establish the face database and implement feature encoding. Given a face image, the face detection network is initially applied to locate the position of the face. It then proceeds to align and crop the facial area for further processing. Subsequently, the pre-processed image is fed into the face recognition network to extract feature encodings. Finally, feature comparisons are executed to predict the identity of the detected face. In our framework, we incorporate the recognition network and exploit a task-driven paradigm to make low-level LLE interact with high-level recognition and thus bring desirable accuracy improvements. (a) face database encoding and (b) face recognition flow.

noise that is invisible to human eyes and inadvertently lose critical details, thus hindering the performance of subsequent recognition models [32]. How to build a positive correlation between low-level enhancement and high-level recognition, especially in the absence of normal-light image labels, is essential yet under-studied in this field.

In response to this challenge, inspired by the insight of [31], we exploit a face recognition-driven strategy to build a positive connection between the LLE and the recognition task to achieve a win-win situation under a unified end-to-end framework. The task-driven paradigm is executed by the introduction of a high-level recognition task loss function into the training process of the low-level enhancement network. As illustrated in Fig. 3, given a low-light image, the image enhancement network is applied to obtain the enhanced features. In one direction, the enhanced features potentially promote face recognition performance improvements. In the opposite direction, face recognition performance plays a crucial role in guiding the learning process of the image enhancement network. In this manner, the bidirectional interaction between image enhancement and face recognition enforces the enhancement process to be more recognition-friendly. Our method is optimized not only for human-centric visibility but also for the high-level task models simultaneously. We formulate the face recognition loss as

$$L_{\text{face}} = -\sum_{f=1}^{F}\left(p_f * \log q_f\right) \tag{10}$$

where F is the total number of faces in the database. p_f and q_f represent the ground truth and the predicted identity, respectively. Notably, similar to the setting of the semantic segmentation module, the face recognition network is pretrained in advance and then frozen during the optimization of the enhancement network, out of consideration for reducing the training uncertainties and difficulties. Concretely, we employ RetinaFace [38] pretrained on the WIDER FACE dataset [39] as the detection network, and FaceNet [40] pretrained on the CASIA-WebFace dataset [41] as the recognition network.

Overall, the total loss function is formulated as

$$L_{\text{total}} = L_{\text{cont}} + L_{\text{feat}} + L_{\text{seg}} + L_{\text{face}}. \tag{11}$$

TABLE I
Descriptions of Datasets

Usage	Name	Total	Individual	Source
Train	LaPa-Face Contrast-Face	4000 360/360	2185 360	LaPa [42] CelebA-HQ [43]
Test	CASIA-Test LaPa-Test LFW CelebA-Test	500 1789 13233 1000	100 1313 5749 1000	CASIA-FaceV5 LaPa [42] LFW [44] CelebA-HQ [43]

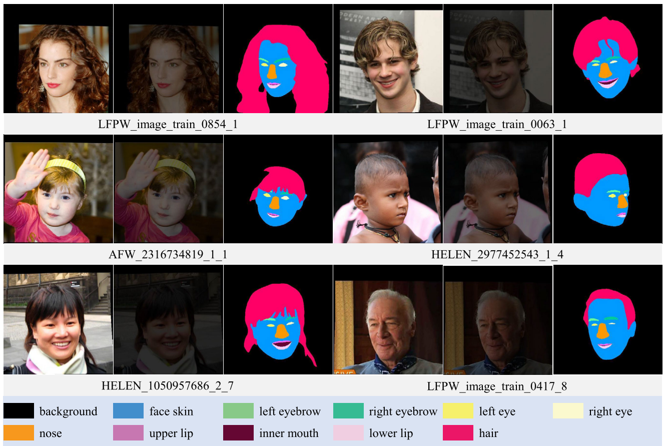


Fig. 5. Several examples of the proposed LaPa-Face dataset, which has a high degree of variability in race, age, pose, occlusion, expression, and appearance. The LaPa-Face dataset consists of 4000 normal-light/low-light images with 11-category semantic label maps and identity labels of 2185 individuals.

IV. Dataset

A. Observation and Consideration

To narrow the gap between the LLE task and the face recognition task, we exploit a task-driven paradigm to promote both tasks in an end-to-end manner. And the semantic information is leveraged to further boost the performance of our model. To satisfy the training requirements of Low-FaceNet and provide support for follow-up research, we note that the training dataset should have the following properties.

1) Low-light images for the LLE task.
2) Face images for face recognition applications.
3) Each individual should contain at least one face image. It is better if the same individual contains multiple face images with different expressions and postures.
4) Each image should have a corresponding semantic segmentation ground truth for integrating semantic information.

To the best of our knowledge, there are no publicly available benchmarks that satisfy all the above properties. We finally pick Landmark guided face Parsing, i.e., LaPa [42] as the source dataset to establish our new benchmark, termed LaPa-Face. We observe that there are several problems or limitations of LaPa.

1) It only contains normal-light images, lacking required low-light images.
2) There are numerous errors in individuals' identity labels. For instance, the same individual is labeled with different identities or different individuals are labeled with the same identity.
3) Multiple images of the same individual are mostly obtained by data enhancement, limiting the face recognition performance.

B. Synthetic Details

Based on the aforementioned observations, the first step is to manually pick and correct the identity labels, which is extremely time-consuming. But the correction results are not guaranteed to be completely accurate, due to the diverse expressions and postures. The second step is to synthesize low-light images based on normal-light images. We apply a linear transformation to obtain synthetic low-light images. The step-by-step description of the synthesis process is provided in Algorithm 1. In particular, we first discard images with a too-small size to facilitate training, then take a random number in the given interval as darken ratio to more closely approximate the real low-light environment with multiple brightness levels. Next, we apply pixel-wise multiplication by ratio to obtain the low-light image. After generating the whole set of low-light images, we calculate the average brightness of the set to verify if the average brightness meets the requirement by

$$l = 0.299r + 0.587g + 0.114b \tag{12}$$

where l refers to the average brightness of the image, r, g, and b denote three image channels, respectively. If the average brightness is not satisfied, the ratio interval needs to be fine-tuned and then re-synthesize the set.

LaPa-Face contains 4000 normal-light/low-light image pairs with large variations in race, age, pose, and facial expression. Each image has semantic annotations (11-category semantic label map) and identity labels (2185 individuals). Some examples from LaPa-Face are found in Fig. 5, and the total dataset has been released for public use and evaluation.

As for contrastive learning samples, 360 images are chosen from the CelebA-HQ dataset [43] as positive samples, and the negative samples are obtained by Algorithm 1, thus a total of 720 images are adopted as the contrast dataset. The other test datasets are synthesized in the same way with the fine-tuned ratio interval. More detailed descriptions of our datasets are tabulated in Table I.

Algorithm 1 Generation of the Set of Low-Light Images

```
Input: The set of normal-light images with total
       number $S$: $Set_{high} = (I_{high_1}, I_{high_2}, \ldots, I_{high_S})$,
       minimum image size $(height, width)$, darken
       ratio $(ratio_{min}, ratio_{max})$, required brightness
       $L$
Output: The corresponding set of low-light images
        with total number $K$:
        $Set_{low} = (I_{low_1}, I_{low_2}, \ldots, I_{low_K})$, $K \leq S$
1  $K \leftarrow 0$
2  for $I_{high} \in Set_{high}$ do
3      if $Size(I_{high}) < (height, width)$ then
4          delete $I_{high}$ from $Set_{high}$, continue    ▷ discard
             images with very small size
5      else
6          $ratio \leftarrow Random(ratio_{min}, ratio_{max})$
7          $I_{low} \leftarrow I_{high} \circ ratio$    ▷ darken process $I_{high}$ to
             generate $I_{low}$
8          $l \leftarrow 0.299r + 0.587g + 0.114b$    ▷ calculate the
             average brightness of $I_{low}$
9          save $I_{low}$ to $Set_{low}$
10         $K \leftarrow K + 1$
11     end
12 end
13 if $Average(l) \in (L - 0.01, L + 0.01)$ then
14     return $Set_{low}$ with total number $K$    ▷ the average
         brightness meets the requirement
15 else
16     clear $Set_{low}$, fine-tune $(ratio_{min}, ratio_{max})$, go to
         line 1                    ▷ regenerate $Set_{low}$
17 end
```

V. Experiment

In this section, we compare both the LLE and recognition performance of our Low-FaceNet with seven state-of-the-art approaches on several synthetic datasets and in wild scenarios. The results of the ablation study are presented to verify the effectiveness of each module in Low-FaceNet.

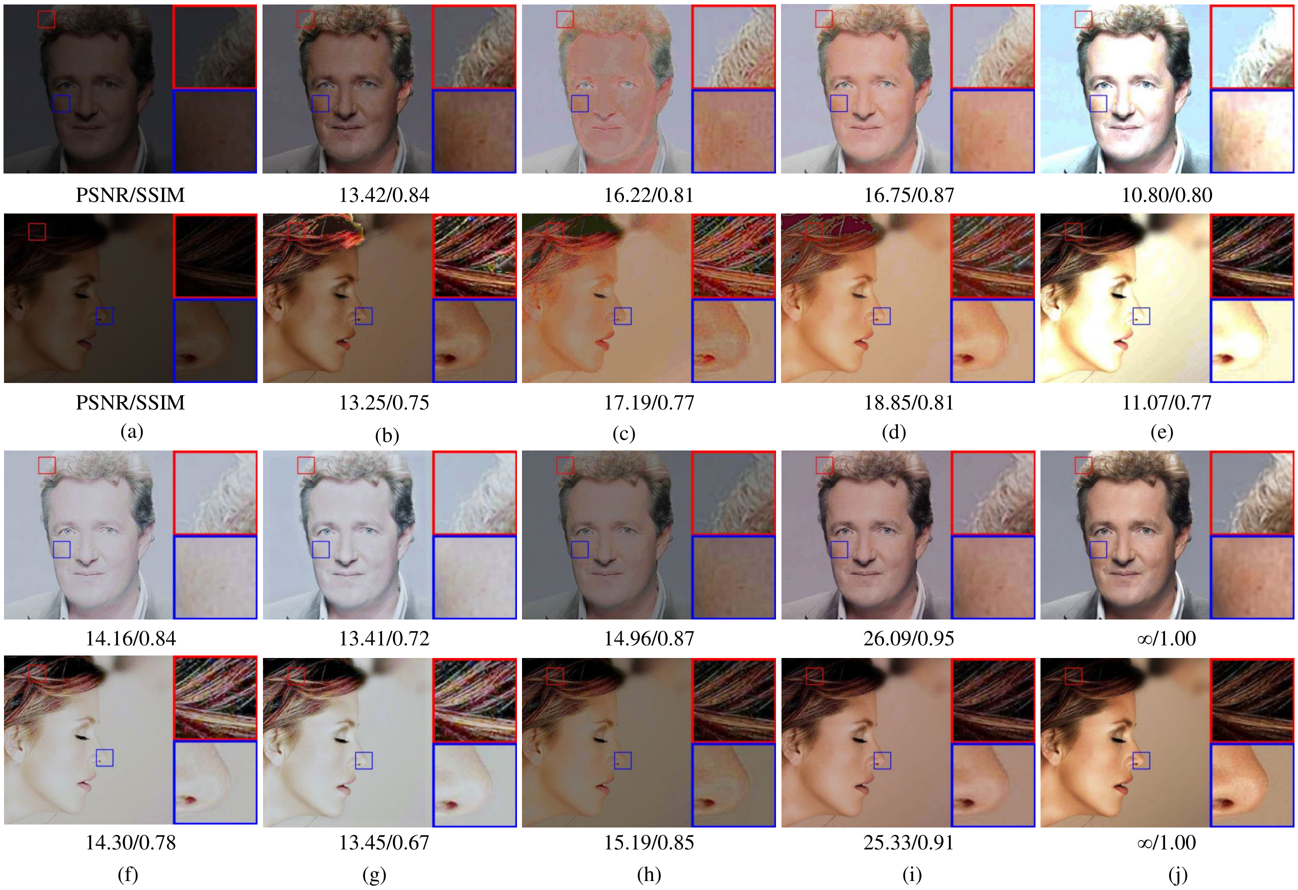
A. Implementation Details

1) Training Details: The experiments are implemented on Pytorch with an NVIDIA GeForce RTX 3060 GPU. The model is trained by Adam optimizer with a fixed learning rate of 1×10^{-4} for 100 epochs, and the batch size is set to 2. The training images are resized into 384 × 384 pixels. For the ablation study experiments, we reduce the number of epochs to 50 to save computational resources while still allowing the model to adequately fit the training data.

2) Evaluation Metrics: We assess the performance across both the image enhancement and face recognition tasks. Regarding image quality assessment, we employ a combination of traditional and deep learning-based metrics. For full-reference assessment, we employ two traditional metrics: peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [45] as well as two deep learning-based metrics: learned perceptual image patch similarity (LPIPS) [46] and Fréchet inception distance (FID) [47]. For no-reference assessment, we employ UNIQUE [48] to comprehensively evaluate image quality. Higher PSNR, SSIM, and UNIQUE, and lower LPIPS and FID indicate better quality. As for the face recognition task, we adopt the average accuracy as the evaluation metric.

B. Comparisons With State-of-the-Arts

To prove the effectiveness of our Low-FaceNet, we compare it with seven state-of-the-art LLE approaches, including



PSNR/SSIM 13.42/0.84 16.22/0.81 16.75/0.87 10.80/0.80

PSNR/SSIM 13.25/0.75 17.19/0.77 18.85/0.81 11.07/0.77

(a) (b) (c) (d) (e)

14.16/0.84 13.41/0.72 14.96/0.87 26.09/0.95 ∞/1.00

14.30/0.78 13.45/0.67 15.19/0.85 25.33/0.91 ∞/1.00

(f) (g) (h) (i) (j)

Fig. 6. Image enhancement results tested in CASIA-Test. (a) Input low-light image, and the enhancement results of (b) LIME [7], (c) RetinexNet [10], (d) SSIENet [16], (e) RUAS [17], (f) Zero-DCE [14], (g) Zero-DCE++ [18], (h) SCL-LLE [19], (i) our Low-FaceNet, and (j) ground-truth image. The numbers below every image are the PSNR and SSIM. Our Low-FaceNet can well brighten the illumination while preserving image details, but the other methods tend to cause color artifacts, structure distortions, and global or local overexposure.

LIME [7], RetinexNet [10], SSIENet [16], RUAS [17], Zero-DCE [14], Zero-DCE++ [18], and SCL-LLE [19]. Among them, RetinexNet, SSIENet, and RUAS are Retinex-based methods. All the methods are retrained on our LaPa-Face dataset with recommended settings for a fair comparison.

1) Comparisons on Synthetic Datasets: We visually compare the results of our method with the state-of-the-art methods on several synthetic datasets. As illustrated in Fig. 6, the images enhanced by LIME and SCL-LLE remain relatively dim, which poses a challenge in effectively capturing facial details. RetinexNet and SSIENet lead to noticeable distortions in image details and colors, resulting in unnatural effects that severely affect image quality. RUAS, Zero-DCE, and Zero-DCE++ potentially introduce overexposure in local or global regions of the enhanced images. Comparatively, our method can strikingly elevate the brightness to a natural level while minimizing color and detail distortions. The enhanced facial images maintain natural skin tones and recover minute details, which contributes to the performance of face recognition models. Overall, the proposed Low-FaceNet addresses the shortcomings of existing methods and excels in enhancing low-light images by optimally balancing color, detail, and brightness factors.

Besides, Table II reports the performance evaluation in the aspect of both the LLE and face recognition tasks on several synthetic datasets. Compared with the state-of-the-art approaches, our Low-FaceNet achieves the best performance in both image quality and face recognition accuracy. It is noteworthy that our method surpasses these approaches by large margins. The enhancements are particularly striking, with improvements of 0.64 db, 0.053, 0.08, and 5.01 in PSNR, SSIM, LPIPS, and FID over the second-best competitor, and remarkable gains of 1.29, 0.20, and 5.3 in recognition accuracy on three datasets, respectively.

We argue that the superior recognition performance is not only due to the gain brought about by changes in image quality. Albeit low-light images processed with low-level image enhancement approaches exhibit better visual quality, they may not confer the same benefits upon high-level computer vision tasks. Taking an example from the quantitative results presented in Table II, most methods bring gains in image quality compared to the results of unprocessed low-light data in the first row. However, some of them actually harm the recognition performance. For instance, RUAS yields 15.30 db (7.69 db gains) and 0.800 (0.413 gains) in PSNR and SSIM, while the accuracy drops to 89.34, significantly lower than 96.60 achieved with the low-light data without processing. In contrast, our Low-FaceNet delivers both

TABLE II

Quantitative Comparisons on Several Synthetic Datasets. **Bold** and Underlined Indicate the Best and Second Best, Respectively

Method	Publication	Image Enhancement Task					Face Recognition Task		
		CelebA-Test					Lapa-Test	CASIA-Test	LFW
		PSNR↑	SSIM↑	LPIPS↓	FID↓	UNIQUE↑		Accuracy ↑	
Low-light data	/	7.61	0.387	0.380	28.69	0.05	96.60	86.80	68.09
LIME [7]	TIP'17	12.10	0.747	0.222	33.03	0.34	96.48	97.60	70.33
RetinexNet [10]	BMVC'18	15.18	0.738	0.378	60.16	0.07	72.82	85.20	34.29
SSIENet [16]	arXiv'20	17.47	0.809	0.220	30.66	0.22	96.54	97.20	70.46
RUAS [17]	CVPR'20	15.30	0.800	0.254	22.10	-0.11	89.34	97.80	66.71
Zero-DCE [14]	CVPR'20	14.75	0.770	0.322	40.05	0.19	95.81	96.20	63.98
Zero-DCE++ [18]	TPAMI'21	14.43	0.665	0.308	35.73	0.20	96.21	96.40	66.98
SCL-LLE [19]	AAAI'22	12.70	0.802	0.235	28.63	0.23	96.43	96.40	69.11
Ours	/	**18.11**	**0.855**	**0.140**	**17.09**	**0.36**	**97.77**	**98.00**	**75.63**

high-quality enhancement results and superior recognition accuracy. It sufficiently proves the effectiveness of our method.

2) Comparisons on Real-World Scenes: LLE in real-world scenarios presents an extremely challenging task. The ability to control partial overexposure, correct overall color tones, and preserve intricate details is of great importance. To explore the generality of the proposed method in wild scenarios, we test our Low-FaceNet on some challenging real-world examples collected from the Internet. As depicted in Figs. 2 and 7, unfortunately, the performance of all the approaches has dropped. However, it is evident that our method consistently delivers more satisfactory visual results compared to the other competitors. It excels in enhancing dark regions while maintaining color tones and recovering the most desirable details. In contrast, LIME, Retinex, and SSIENet produce severe color deviations. SCL-LLE fail to enlighten the back-lit regions and achieve clear facial recovery. On the other hand, RUAS tends to amplify noise and produce over-exposed artifacts, particularly in the facial region. Zero-DCE tends to excessively smooth out intricate details. In comparison, our proposed Low-FaceNet yields natural exposure and structural detailing. This experiment verifies that Low-FaceNet trained on our proposed synthetic LaPa-Face can effectively cope with real-world low-light scenes with a remarkable generality ability.

3) Discussion on Data Synthesizing Method: To address concerns about the potential impact of our data synthesis method on the generality of our model, we have employed a more refined way to re-synthesize the CelebA-Test dataset, referred to as CelebA-Test-II. In this synthesis process, we take into account variations in gamma values and incorporate simulated noise in low-light images. Initially, we apply gamma correction to adjust the brightness of the images, using gamma values ranging from 2.15 to 2.25. Subsequently, we introduce random noise into the images, with a mean of 0 and a standard deviation in the range of 0.26–0.30. We conduct a fresh set of experiments utilizing the original pretrained model on LaPa-Face. The results are presented in Fig. 8 and Table III. The results provide clear evidence that even when the testing dataset is synthesized using this alternative method, our approach consistently outperforms other methods. Unfortunately, none of the methods are able to completely eliminate the noise. However, when compared to existing methods, the proposed Low-FaceNet exhibits a more natural and robust performance. Conversely, the other methods struggle to handle such scenarios and fall short of fully restoring the color and intricate details in the images. This further verifies the generality and effectiveness of our method.

TABLE III

Quantitative Results on Celeb-Test-II. **Bold** and Underlined Indicate the Best and the Second-Best Results, Respectively

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
LIME [7]	15.94	0.716	0.252	36.92
RetinexNet [10]	16.99	0.750	0.292	36.79
SSIENet [16]	15.44	0.759	0.260	29.97
RUAS [17]	11.34	0.660	0.398	40.96
Zero-DCE [14]	13.77	0.764	0.246	30.78
Zero-DCE++ [18]	13.85	0.617	0.225	26.83
SCL-LLE [19]	11.90	0.555	0.236	27.50
Ours	**17.07**	**0.789**	**0.173**	**20.38**

C. Ablation Study

We evaluate the effect of the contrastive negative samples and four modules (i.e., loss terms) on the performance of our proposed Low-FaceNet, as illustrated in Fig. 9 and Table IV.

1) Effect of Negative Samples: To verify the impact of negative samples, we conduct an experiment where we remove the negative samples from contrastive training data. It turns out that the model trained w/o S_{neg} presents notably poor performance, where the brightness level of the given image remains persistently low, with numerous concealed details.

2) Effect of Loss Terms: To investigate the contributions of the various loss terms, we conduct experiments where each loss term is alternately removed, allowing us to assess their impact on the quality of the enhanced images and recognition performance. As illustrated in Fig. 9, the numbers below each image indicate the average illumination.

1) The model trained w/o L_{cont} fails to elevate the brightness of the input image, which reveals the superiority of contrast learning-based brightness recovery.
2) The model trained w/o L_{feat} tends to produce several color deviations and lose some details, with an average illumination that is excessively high compared to the ground truth image. In contrast, the full model exhibits more desired colors and details, indicating the necessity

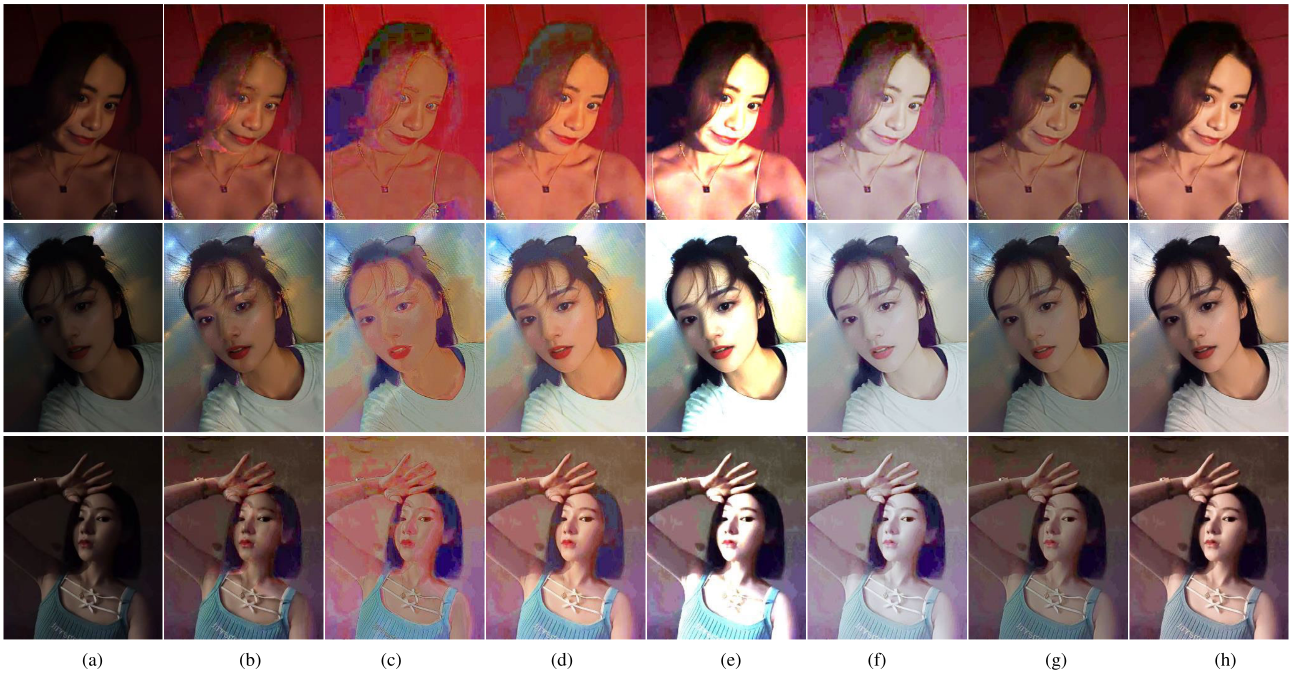


Fig. 7. Image enhancement results tested in real-world examples. (a) Input low-light image, and the enhancement results of (b) LIME [7], (c) RetinexNet [10], (d) SSIENet [16], (e) RUAS [17], (f) Zero-DCE [14], (g) SCL-LLE [19], and (h) our Low-FaceNet. Our method outperforms others, particularly in controlling exposure levels, representing natural color tones, and preserving intricate details.

TABLE IV

QUANTITATIVE RESULTS OF ABLATION STUDY. **BOLD** AND UNDERLINED INDICATE THE BEST AND THE SECOND-BEST RESULTS, RESPECTIVELY

Variant (50 epochs)	Image Enhancement Task					Face Recognition Task		
	CelebA-Test					Lapa-Test	CASIA-Test	LFW
	PSNR↑	SSIM↑	LPIPS↓	FID↓	UNIQUE↑	Accuracy ↑		
w/o S_{neg}	8.42	0.47	0.324	27.39	0.07	97.21	93.80	70.46
w/o L_{cont}	8.40	0.47	0.325	27.51	0.07	97.49	93.00	70.36
w/o L_{feat}	12.45	0.74	0.493	71.33	0.04	95.87	90.20	70.41
w/o L_{seg}	16.07	0.81	0.328	36.17	-0.05	97.21	80.00	70.23
w/o L_{face}	15.95	0.84	0.252	30.93	0.31	96.71	96.20	70.12
Ours	**18.26**	**0.86**	**0.134**	**16.93**	**0.35**	**97.60**	**97.60**	**75.55**

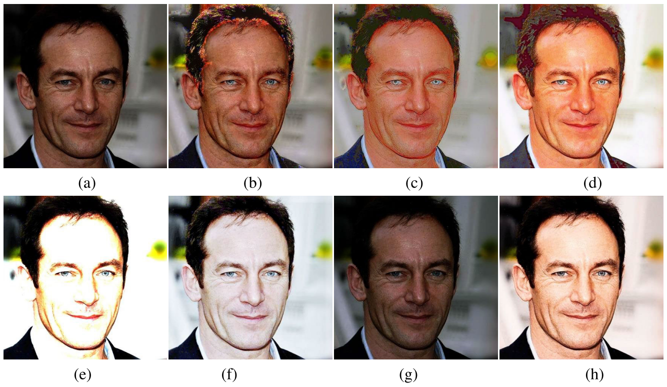


Fig. 8. Image enhancement results of a re-synthesized example from CelebA-Test-II. (a) Input low-light image, and the enhancement results of (b) LIME [7], (c) RetinexNet [10], (d) SSIENet [16], (e) RUAS [17], (f) Zero-DCE++ [18], (g) SCL-LLE [19], and (h) our Low-FaceNet. The proposed Low-FaceNet exhibits a more natural and robust performance.

of the feature loss in preserving colors and inherent features.

3) The model trained without L_{seg} has local overexposure areas (e.g., the forehead skin and the lip area), with a higher average illumination than the ground-truth images. In contrast, the full model exhibits a more natural exposure level, indicating the effectiveness of semantic information to ensure consistent and smooth illumination of the same semantic category.
4) The introduction of L_{face} primarily serves for recognition accuracy improvements. This term does not significantly affect the overall illumination of the results. The average illumination of images Fig. 9(f) w/o L_{face} is essentially the same as the average illumination of visual results produced by the full model. However, it brings significant gains in face recognition accuracy.

Clearly, our method with complete contrastive samples and all losses in joint training is the best-ranked approach that significantly outperforms the other options. The noticeable performance decline in the absence of negative samples or specific loss terms underscores the effectiveness of our comprehensive framework, which adeptly incorporates all four key modules and unpaired low-light images as negative samples.

D. Limitation Discussion

While Low-FaceNet demonstrates robust performance in many synthetic and real-world scenarios, there are observed

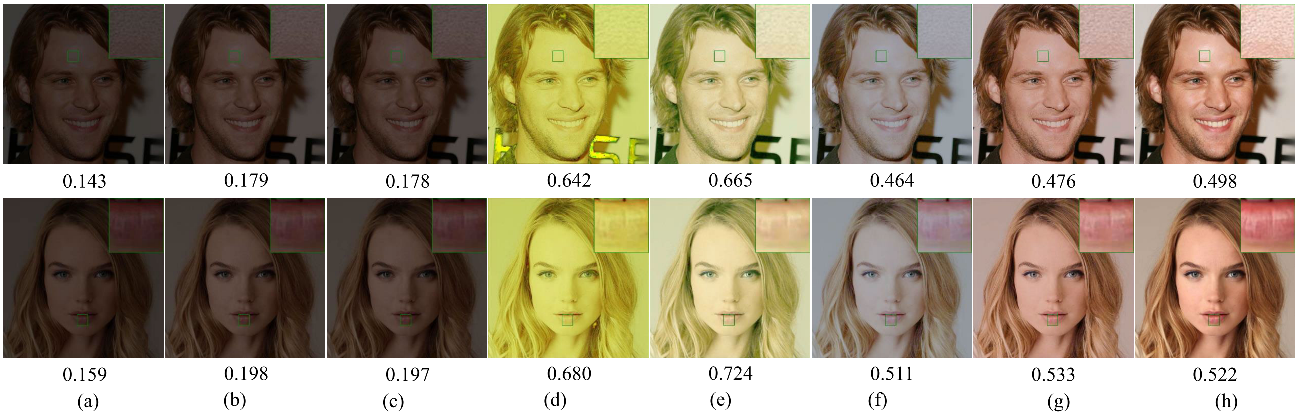


Fig. 9. Visual results of ablation study. (a) Input low-light image, and the results of (b) W/o S_{neg}, (c) W/o L_{cont}, (d) W/o L_{feat}, (e) W/o L_{seg}, (f) W/o L_{face}, (g) our Low-FaceNet, and (h) ground truth image. The numbers below every image indicate the average illumination of each image. Our Low-FaceNet trained with complete contrastive samples and modules can produce more realistic enhancement results with less color distortion and fewer artifacts. All components contribute to the overall superior performance.

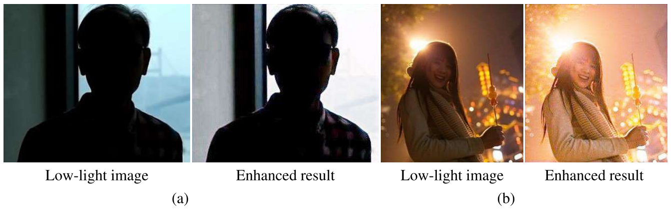


Fig. 10. Failure cases. Our Low-FaceNet can hardly handle images captured in (a) extremely dark conditions and (b) strong light effects, which rarely occurs in real-world face recognition systems.

failure cases that merit discussion. Under extremely dark conditions [see Fig. 10(a)], our method fails to perform illumination brightening and detail restoration. This can be attributed to the exceedingly severe underexposure in terms of intensity and coverage, there is little information available for the network to generate the missing details in the neighborhood. Even humans have difficulty recognizing the individual in the image. Conversely, when dealing with low-light images with strong light effects [see Fig. 10(b)], Low-FaceNet hardly struggles with these nighttime lighting effects and even mistakenly amplifies them. This is due to the lack of images with similar strong light effects in our training dataset. However, these cases rarely occur in real-world face recognition systems. Besides, we have to acknowledge that the quality of the proposed benchmark has certain limitations. Real-world low-light images often involve changes in contrast, gamma values, and loud noise in addition to brightness changes.

In our future work, we are dedicated to establishing a more refined benchmark that better simulates real-world scenarios. We believe this benchmark will further enhance the versatility and adaptability of our Low-FaceNet and also inspire more impactful research in the vision community. Furthermore, we are keen to explore how our framework can be extended to other tasks related to low-level vision.

VI. Conclusion

In this article, we propose a novel face recognition-driven low-light image enhancement network, called Low-FaceNet. To serve the training of Low-FaceNet and facilitate broad comparisons in the research community, we build a new benchmark named LaPa-Face, containing normal-light/low-light images with semantic and identity labels. Low-FaceNet possesses an image enhancement network and congregates four key modules, i.e., a contrastive learning module, a feature extraction module, a semantic segmentation module, and a face recognition module. The former three modules guarantee Low-FaceNet has the capacity of contrastive brightness enhancement, feature preservation, and semantic-guided smoothness, while the last one promotes the accuracy improvement of face recognition in low-light conditions. We illustrate that low-level and high-level tasks (i.e., LLE and face recognition) can promote each other and realize mutual benefits. Low-FaceNet reveals that underexposed images and semantic information that are easily overlooked can be beneficial in obtaining visual-pleasing results, even without the normal-light image labels. Extensive experiments show the superiority of Low-FaceNet for obtaining more natural and rich details and colors. The application of face recognition further reveals our potential in settling the downstream face recognition task to gain better performance.

References

[1] K. Xu et al., "HFMNet: Hierarchical feature mining network for low-light image enhancement," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[2] H. Cui, J. Li, Z. Hua, and L. Fan, "Progressive dual-branch network for low-light image enhancement," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–18, 2022.

[3] 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. (CVPR)*, Jun. 2019, pp. 4690–4699.

[4] Y. Li and J. Li, "An end-to-end defect detection method for mobile phone light guide plate via multitask learning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[5] Q. Liu, Q. Guo, W. Wang, Y. Zhang, and Q. Kang, "An automatic detection algorithm of metro passenger boarding and alighting based on deep learning and optical flow," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[6] H. Ibrahim and N. S. P. Kong, "Brightness preserving dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 4, pp. 1752–1758, Nov. 2007.

[7] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2016.

[8] N. Singh and A. K. Bhandari, "Principal component analysis-based low-light image enhancement using reflection model," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[9] 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, Jan. 2017.

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

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

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

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

[14] C. Guo et al., "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1780–1789.

[15] 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. (CVPR)*, Jun. 2020, pp. 3063–3072.

[16] Y. Zhang, X. Di, B. Zhang, and C. Wang, "Self-supervised image enhancement network: Training with low light images only," 2020, *arXiv:2002.11300*.

[17] 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. (CVPR)*, Jun. 2021, pp. 10561–10570.

[18] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022.

[19] D. Liang et al., "Semantically contrastive learning for low-light image enhancement," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, 2022, pp. 1555–1563.

[20] E. H. Land, "The Retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, 1977.

[21] M. Fan, W. Wang, W. Yang, and J. Liu, "Integrating semantic segmentation and retinex model for low-light image enhancement," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2317–2325.

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

[23] F. Lv, B. Liu, and F. Lu, "Fast enhancement for non-uniform illumination images using light-weight CNNs," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1450–1458.

[24] M. Zhu, P. Pan, W. Chen, and Y. Yang, "EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13106–13113.

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

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[27] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[29] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3067–3074, Dec. 2018.

[30] Y. Li, Y. Liu, Q. Yan, and K. Zhang, "Deep dehazing network with latent ensembling architecture and adversarial learning," *IEEE Trans. Image Process.*, vol. 30, pp. 1354–1368, 2021.

[31] Y. Lee, J. Jeon, Y. Ko, B. Jeon, and M. Jeon, "Task-driven deep image enhancement network for autonomous driving in bad weather," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13746–13753.

[32] C. Li et al., "Detection-friendly dehazing: Object detection in real-world hazy scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8284–8295, Jan. 2023.

[33] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10551–10560.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[35] A. Ignatov, N. Kobyshev, R. Timofte, and K. Vanhoey, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3277–3285.

[36] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1013–1037, Apr. 2021.

[37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[38] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5203–5212.

[39] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 5525–5533.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[41] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[42] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang, and T. Mei, "A new dataset and boundary-attention semantic segmentation for face parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11637–11644.

[43] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[44] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life'Images: Detection, Alignment, Recognit.*, 2008.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

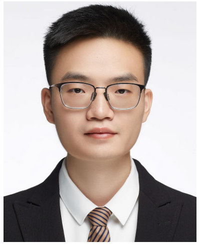
[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 586–595.

[47] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6626–6637.

[48] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, 2021.

Yihua Fan received the B.S. degree from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2022, where she is currently pursuing the Ph.D. degree with the School of Computer Science and Technology.

Her research interests include deep learning, image processing, and computer vision.



Yongzhen Wang received the Ph.D. degree in computer science and technology from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2023.

He is currently a Lecturer with Anhui University of Technology, Ma'anshan, China. He has published more than 20 research papers, including IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP) and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS). His research interests include deep learning and computer vision.

Dong Liang (Member, IEEE) received the Ph.D. degree from the Graduate School of IST, Hokkaido University, Sapporo, Japan, in 2015.

He is currently an Associate Professor with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include robust pattern recognition and large-scale video streaming processing.

Yiping Chen (Senior Member, IEEE) received the Ph.D. degree in information and communications engineering from the National University of Defense Technology, Changsha, China, in 2011.

She is currently an Associate Professor with Sun Yat-Sen University, Guangzhou, China. Her current research interests include remote sensing image processing, mobile laser scanning data analysis, 3-D point cloud computer vision, and autonomous driving.

Dr. Chen is currently an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Haoran Xie (Senior Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2013.

He is currently an Associate Professor with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong. His research interests include artificial intelligence, big data, and educational technology.

Prof. Xie is the Editor-in-Chief of Computers and Education: Artificial Intelligence.

Fu Lee Wang (Senior Member, IEEE) received the Ph.D. degree in systems engineering and engineering management from Chinese University of Hong Kong, Hong Kong, in 2003.

He is currently the Dean of the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong. He has over 250 publications in international journals and conferences and led more than 20 competitive grants with a total greater than HK$20 million. His current research interests include educational technology, information retrieval, computer graphics, and bioinformatics.

Jonathan Li (Fellow, IEEE) is currently a Professor with the Department of Geography and Environmental Management and cross-appointed with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada, and a Fellow of the Engineering Institute of Canada, Ottawa ON, Canada. His main research interests include image and point cloud analytics, mobile mapping, and AI-powered information extraction from LiDAR point clouds and earth observation images.

Dr. Li is currently serving as the Editor-in-Chief for *International Journal of Applied Earth Observation and Geoinformation*.

Mingqiang Wei (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from Chinese University of Hong Kong (CUHK), Hong Kong, in 2014.

He is currently a Professor with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His research interests include 3-D vision and computer graphics.

Dr. Wei is currently an Associate Editor of *ACM Transactions on Multimedia Computing, Communications, and Applications* (ACM TOMM), *The Visual Computer*, and a Guest Editor of IEEE TRANSACTIONS ON MULTIMEDIA.