A Dual-Stream-Modulated Learning Framework for Illuminating and Super-Resolving Ultra-Dark Images

Jiaxin Gao, Ziyu Yue, Yaohua Liu, Sihan Xie, Xin Fan, *Senior Member, IEEE*, and Risheng Liu, *Member, IEEE*

***Abstract*— Enhancement of image resolution for scenes captured under extremely dim conditions represents a practical yet challenging problem that has received little attention. In such low-light scenarios, the limited lighting and minimal signal clarity tend to intensify issues such as diminished detail visibility and altered color accuracy, which are often more severe during the image enhancement process than in scenarios with adequate lighting. Consequently, standard methods for enhancing low-light images or improving their resolution, whether implemented independently or through a combined approach, generally face challenges in effectively restoring luminance, preserving color integrity, and detailing intricate features. To conquer these issues, this article introduces an innovative dual-stream (DS) modulated learning framework designed to tackle the real-world coupled degradation issues in super-resolution (SR) under low-light conditions. Leveraging natural image color characteristics, we introduce a self-regularized luminance constraint to specifically target uneven illumination. We develop illumination-semantic dual modulator (ISDM), a refinement middleware embedded in the decoding stage to bridge illumination and semantic features concurrently, aimed at safeguarding the integrity of lighting and color details at the feature level. Our approach replaces simple upsampling methods with the resolution-sensitive merging upsampler (RSMU) module, which integrates diverse sampling techniques to effectively reduce artifacts and halo effects. Comprehensive experiments on three benchmarks showcase the applicability and generalizability of our approach to diverse and challenging ultra-poorly lit settings, outperforming state-of-the-art methods with a notable improvement. The code and benchmark are publicly available at https://github.com/moriyaya/UltraIS.**

***Index Terms*— Dual-stream (DS) modulation, image super-resolution (SR), low-light, low-level vision.**

I. Introduction

BOTH low-light image enhancement (LLIE) and super-resolution (SR) are foundational topics in image processing [1], [2], [3], [4], [5]. Nevertheless, research toward low-light image SR (LLISR) has been relatively neglected, despite its practical importance for applications such as security monitoring, automated driving, and medical image analysis [6], [7], [8]. Indeed, enhancing visibility and resolution under poorly lit settings is both a practical and challenging task. Although professional and sophisticated optical lenses can capture high-quality photos, their expensive cost makes them difficult to be widely adopted [9], [10], [11], [12], [13], [14]. In addition, the internet abounds with numerous images that are both underexposed and of low resolution, necessitating enhancement interventions to elevate their perceptual clarity. This research endeavors to refine the quality of images captured in near darkness by enhancing both their luminance and scale, striving to transform them into high-resolution images with balanced lighting.

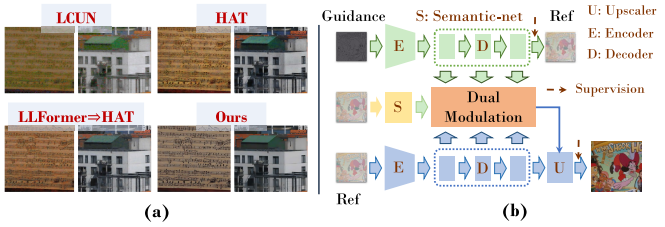


Fig. 1. (a) We illustrate three technical schemes (i.e., low-light SR, standard-light SR, and LLIE⇒SR) for addressing the LLISR problem. A recent LLIE method (i.e., LLFormer [15]), a low-light SR method (i.e., LCUN [16]), and a standard-light SR method (HAT [17]) are presented as examples for visualization. (b) Our dual-modulated framework can simultaneously integrate semantic features and illumination features, thereby generating rich textures.

Compared to sufficiently lit conditions, LLISR tasks intrinsically encounter coupled degradation at the data acquisition level due to severe low light and low resolution. The central challenges are focused on three key dimensions.

1) *Uneven Exposure:* In environments with minimal light, the unbalanced distribution of lighting often leads to pronounced shadows and bright areas in images. Super-resolving tasks can amplify excessively dark or bright local areas, leading to color distortion phenomena.
2) *Loss of Detail:* Directly increasing the resolution of dimly lit images, particularly when done extensively, can intensify existing noise and obscure delicate textures. Ensuring the visual authenticity and naturalness of the output images becomes challenging under these conditions.

Manuscript received 29 September 2023; revised 20 January 2024 and 22 April 2024; accepted 9 May 2024. *(Corresponding author: Jiaxin Gao.)*

Jiaxin Gao, Yaohua Liu, Sihan Xie, Xin Fan, and Risheng Liu are with the School of Software Technology, Dalian University of Technology, Dalian 116024, China (e-mail: jiaxinn.gao@outlook.com; liuyaohua_918@163.com; XSH2018@mail.dlut.edu.cn; xin.fan@dlut.edu.cn; rsliu@dlut.edu.cn).

Ziyu Yue is with the School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China (e-mail: 11901015@mail.dlut.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2024.3409056

3) *Distorted Artifacts:* Simply increasing the luminance of dark images often unavoidably amplifies hidden edge ghosting and artifacts in dark areas, especially evident in the finer details of the image, such as thin lines or texture edges.

As shown in Fig. 1(a), we illustrate three technical schemes (i.e., low-light SR, standard-light SR, and LLIE⇒SR) for addressing the LLISR problem, yet empirical evidence points to their inherent limitations. Indeed, it has been confirmed that current techniques intended for LLIE and standard-light SR are not effective when directly applied to LLISR problems. Furthermore, simply applying these task-specific frameworks in a cascaded manner (e.g., LLIE⇒SR) still fails to yield satisfactory results. Recently, several methods have been introduced specifically targeting this challenge, and however, they lack delivering the expected enhancement results [16], [18], [19], [20]. Notably, two of these methods [18], [20] employ synthetic datasets generated by manually modifying brightness settings, such as through gamma adjustment, for purposes of enhancing brightness and SR. However, these techniques often lack successful application in real-world nighttime conditions, leading to noticeable hue anomalies and visual artifacts. In addition, other methods [16], [19] fail to generate fine high-frequency details, making it challenging to maintain the natural appearance and visual realism of the enhanced images. In conclusion, existing methods demonstrate distinct limitations in simultaneously addressing the aforementioned three core challenges. As illustrated, the standard-light SR technique (i.e., HAT [17]) struggles to produce delicate high-frequency details, making it difficult to preserve the natural look and visual authenticity of the enhanced images. The LLIE method (i.e., LLFormer [15]) cascaded with HAT exhibits noticeable color shifts and unclear textures. In comparison, our proposed method shows more authentic and true-to-life colors and structural features.[1] From the above analysis, we summarize the underlying causes of the limitations in current methods as follows: 1) the tendency to build complex networks without adequately considering the physical principles unique to dark conditions, such as those informed by the Retinex theory for illumination assessment; 2) a lack of attention to finely modulating the low-resolution features, which can lead to color bias and blurriness; and 3) an overreliance on single-layer upsampling ways for enlarging features and enhancing resolution, which often results in numerous artifacts and halos, thereby compromising the clarity of details.

To address these issues, this article proposes a specialized dual-stream (DS) modulated learning framework for simultaneously Illuminating and Super-resolving images captured in Ultra-dark scenes, dubbed UltraIS. As illustrated in Fig. 1(b), UltraIS utilizes a DS learning framework, which represents a pioneering effort to comprehensively analyze the essence of the LLISR task. During the initial stage, we propose incorporating illumination loss to impose constraints on low-light scenes to learn uniform illumination features and initially brightened reflection maps, whereas, in the subsequent stage, we introduce a semantic knowledge base and build illumination-semantic dual modulator (ISDM) as refinement middleware. ISDM runs through the entire network layer of decoding for cross-branch interaction, facilitating the far-flung retention of illuminance details and semantic color details, to enable the SR network to focus more on faithful texture details. Meanwhile, experiments confirm its ability to effectively preserve the illumination details and texture details of the image even when the resolution is increased. Furthermore, we introduce a resolution-sensitive merging upsampler (RSMU) module, a strategic departure from singular upsampling, successfully mitigating artifacts and halos. We evaluate the applicability and generalizability of our framework in diverse extremely ultralow-light settings and demonstrate its strengths through comprehensive analysis. The main contributions can be summarized as follows.

1) We present a specialized DS-modulated learning framework that comprehensively dissects the LLISR task by utilizing ultra-dark scene constraints as principled priors. This framework is designed to tackle the intricate task of simultaneously brightening and enlarging images captured under extremely dark conditions.
2) We develop the ISDM, positioned within the decoding stage to concurrently bridge illumination features and semantic features. This middleware modulates reflection features from top to bottom, effectively avoiding color discrepancies and emphasizing faithful texture details.
3) Incorporating multiple substrate sampling techniques in place of a singular upsampling layer, we have developed the RSMU, which proficiently diminishes artifacts and halos, specifically aimed at improving high-fidelity images.
4) Comprehensive testing on three benchmarks highlights the broad generalizability of our approach in severely low-light settings, surpassing existing state-of-the-art techniques in six crucial performance indicators.

II. Related Work

A. Low-Light Image Enhancement

LLIE aims to make images hidden in the dark visible. Early efforts often focused on employing handcrafted priors and empirical observations to tackle LLIE challenges, i.e., Retinex model [21], [22], [23]. These conventional, manually crafted techniques frequently struggle to retain image exposure and colors, often leading to issues such as lost details or artifacts. Recent years have witnessed substantial developments in the design of convolutional neural network (CNN) models to meet these critical challenges [24], [25], [26], [27], [28], [29], [30]. For example, Guo et al. [24] redefined the task as estimating image-specific curves for luminance correction and included various prior-associated constraints (i.e., illumination, lightness, and color) to maintain tonal authenticity. Nevertheless, when applied to extremely dark images, color bias issues persisted. Ma et al. [25] pursued iterative correction of luminance and noise reduction to generate textures, yet the enhancements often lacked clarity in detail. Recently, methods employing transformers for LLIE

[1]For more detailed experimental analysis, please refer to the Section IV-C.

have gained popularity [15], [31], [32], [33]. For example, Wang et al. [15] adopted an architecture leveraging transformer technology with axis-specific multihead self-attention and cross-layer attention fusion block to improve dark image quality. Indeed, while these methods excel in LLIE, they are not effectively transferable to LLISR tasks. In addition, simply combining these methods with standard-light SR techniques does not adequately address several critical issues, including exposure problems, color inaccuracies, and artifact generation. In contrast, we develop a refined middleware (i.e., ISDM) that acts as a bridge between illumination features and semantic features. This middleware operates in a top–down manner to modulate reflected features, thereby avoiding exacerbation of color distortion and enhancing the reproduction of fine textural details.

B. Standard-Light Image SR

Standard-light SR task involves generating high-resolution images from low-resolution inputs under standard lighting conditions. In recent decades, a large number of methods based on CNNs have emerged to continuously refresh the performance [34], [35], [36], [37], [38], [39]. Haris et al. [36] leveraged iterative upsampling and downsampling layers to address the mutual dependencies of low- and high-resolution images. By employing attention mechanisms, several derivative methods [40], [41], [42] have achieved enhanced improvements in the realm of reconstruction fidelity. For instance, Zhao et al. [40] presented a streamlined network design for image SR featuring 3-D attention maps and self-adjusting pixel attention mechanisms. It should be highlighted that Zamir et al. [41] proposed the image restoration framework incorporating a nonlocal attention mechanism and multiscale feature aggregation, exhibiting exceptional efficacy in standard-light SR tasks. Recently, various transformer-inspired approaches [17], [43], [44], [45] have been developed for SR enhancement. Among these, notable approaches include SwinIR [43], Restormer [44], the recently proposed SRFormer [45], and HAT [17]. However, these methods tailored for standard-light conditions fail to address the challenges posed by low-light settings, often resulting in undesirable outcomes such as artifact spreading and texture blurring. Contrarily, we integrate multiple sampling modalities as foundations to construct a resolution-sensitive upsampler module. This upsampler approach is expertly designed to reduce artifacts and halos, specifically engineered to enhance the refinement of high-fidelity images.

C. Low-Light Image SR

Recent advances in LLISR research [16], [18], [19], [20], [46] have not attained ideal results, with the practical outcomes falling far short of expectations. For instance, Guo et al. [18] developed a generative adversarial network for enhancing low-resolution facial details in low-light settings. However, since it only performed simple brightness correction on a synthetic dataset, the super-resolved images suffer from severe color distortion and artifacts. Cheng et al. [16] constructed a light-guided and cross-fusion U-Net for SR of uneven-light images. However, due to the sole use of the pixel shuffle operation for improvement in resolution, significant color distortion and unclear structural features have emerged. Rasheed and Shi [20] utilized backprojection iterative learning to brighten dark images and perform SR tasks, but due to the focus on synthetic datasets only, there is poor generalizability to real-world scenes, resulting in overexposure and artifact phenomena. Aakerberg et al. [19] proposed a multiscale hierarchical encoder–decoder network based on transformers for the joint task of LLIE and SR. However, due to the lack of consideration for the physical principles of low-light scenes, issues, such as color bias and blurriness, have arisen. Therefore, this article capitalizes on expert domain knowledge in extremely low-light environments to develop a custom DS learning architecture. Within the low-frequency domain, we impose dark scene constraints (specifically, brightness constraints) to guide the decoupled learning of uniform illumination features and initial brightness reflection maps. This approach prevents the direct SR of dark images, which would otherwise amplify undesirable dark-light features.

III. Methodology

Given a low-resolution image captured in low-light environments (i.e., $\mathbf{x}^{LL} \in \mathbb{R}^{H\times W\times 3}$), our UltraIS aims to establish an end-to-end learning framework for recovering a standard-light SR image (i.e., $\mathbf{y}^{NS} \in \mathbb{R}^{H\times W\times 3}$) that closely resembles its reference counterpart (i.e., $\mathbf{y}^{NH} \in \mathbb{R}^{H\times W\times 3}$) (which is standard-light, high-resolution) in terms of both sharpness and fidelity. In what follows, we provide a detailed description of the proposed UltraIS (as illustrated in Fig. 2), encompassing two core components: ISDM and RSMU.

A. Retinex-Inspired DS Framework

Drawing inspiration from the classical Retinex theory [21], [22], [23], we initially construct a foundational illumination learning network to capture the underlying physical principles of low-light scenes. As illustrated in Fig. 2(a), we introduce the neighborhood difference operator [6], $\mathbf{u}^{IG} = F_{diff}(\mathbf{x}^{LL}) \in \mathbb{R}^{H\times W\times 1}$, to compute the initial illumination guidance $\mathbf{u}^{IG}$ as a spatially varying illumination operator to obtain edge-aware structural lighting, i.e., $F_{diff}(\mathbf{x}^{LL}) = \mathcal{A}(\mathbf{x}_g^{LL}) + \mathcal{B}(\mathbf{x}_g^{LL})$, where $\mathbf{x}_g^{LL}$ represents a grayscale image. Furthermore, we define $\mathcal{A}$ and $\mathcal{B}$, respectively, denoting maximum value operator and neighborhood difference operator, i.e.,

$$\mathcal{A}(\cdot) := \mathtt{Max}_{\vec{h}}(\mathtt{Max}_{\vec{v}}(\cdot)),\ \ \mathcal{B}(\cdot) := |\mathtt{Sub}_{\vec{h}}(\cdot)| + |\mathtt{Sub}_{\vec{v}}(\cdot)| \quad (1)$$

where $\mathtt{Max}_{\vec{v}}$ and $\mathtt{Max}_{\vec{h}}$, respectively, represent taking the maximum values in the horizontal or vertical direction, respectively, while $\mathtt{Sub}_{\vec{h}}$ and $\mathtt{Sub}_{\vec{v}}$ correspond to performing subtraction operations on pixels in the horizontal or vertical direction, respectively. Subsequently, $\mathbf{u}^{IG}$ is fed into a U-Net-style network to generate the illumination map $\mathbf{u}^{NL}$, with such learning process being supervised by the introduced scene constraint for ultra-dark environments.[2] Furthermore, we obtain the

[2]Please refer to constraint losses as outlined in (6) and (7).

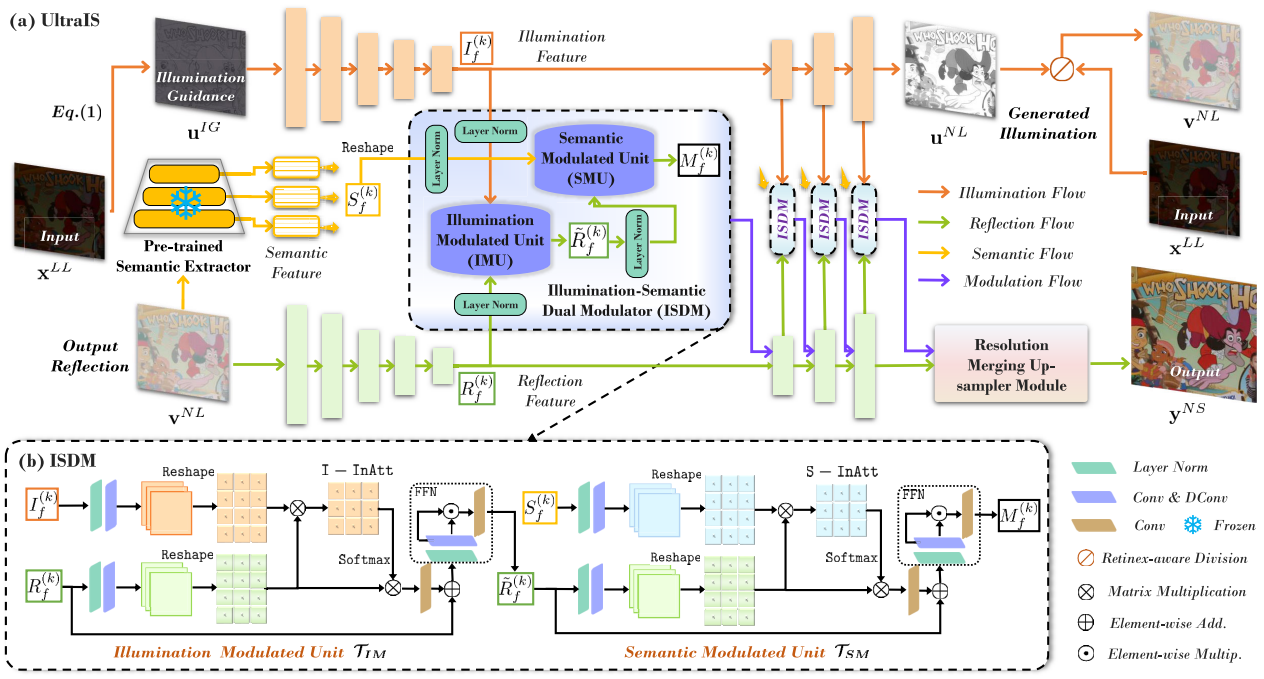


Fig. 2. (a) Overview of the proposed UltraIS. Given low-light low-resolution image $\mathbf{x}^{LL}$, we construct the DS framework to generate the initial reflection image $\mathbf{v}^{NL}$ with illumination guidance $\mathbf{u}^{IG}$. Subsequently, $\mathbf{v}^{NL}$ is fed into a subsequent refinement SR network to restore a standard-light, SR image $\mathbf{y}^{NS}$. (b) Refinement middleware, ISDM, that operates in a top–down manner to modulate reflection features, which effectively avoids exacerbating hue anomalies and emphasizes faithful texture details.

low-resolution reflection map $\mathbf{v}^{NL}$ by employing the Retinex-based element-wise division, formulated as $\mathbf{v}^{NL} = \mathbf{x}^{LL} \oslash \mathbf{u}^{NL}$ and $\mathbf{u}^{NL} = Unet(\mathbf{u}^{IG})$.

In the subsequent stage, $\mathbf{v}^{NL}$ is transmitted to another Unet-style network with similar architecture to complete the refined SR process. We design a refined intermediary, namely, ISDM, depicted as the blue shaded box in Fig. 2(b), positioned within the decoding stage to facilitate progressive interbranch interactions. Regarding semantic feature acquisition, we utilize HRNet [47] that has been pretrained (with fixed parameters for the training process) to extract multiscale semantic features $S_f^{(k)}$, where k denotes the layer index. Experiments confirm that ISDM can effectively avoid exacerbating color shifts and distortions during image SR. Finally, we devised an RSMU module to obtain the enhanced image $\mathbf{y}^{NS}$. It is worth noting that the Unet is constructed based on context units [CUs, as refer to Fig. 3(a)] as the base module, which transforms the scale using max-pooling and bilinear interpolation operations in the encoding and decoding phases, respectively. In the following, we expand on the architectural details of two key modules designed, including ISDM and RSMU.

B. Illumination-Semantic Dual Modulator

Following the well-established fact that is the introduction of pretrained semantic segmentation networks on large-scale datasets as semantic prior knowledge can improve the network's representational capabilities, we utilize HRNet [47] to extract multiscale semantic features $S_f^{(k)}$ from $\mathbf{x}^{LL}$, where $k \in \{1, 2, \ldots, 5\}$ denotes the layer index. Admittedly, the ISDM module is designed in order to better preserve image illuminance details and semantic details in subsequent SR processes.

As shown in Fig. 2(a), the ISDM consists of the illumination modulated unit (IMU, $\mathcal{T}_{IM}$) and semantic modulated unit (SMU, $\mathcal{T}_{SM}$) with a similar architectural design. The illumination features and reflection features extracted from the decoders of the top and down branches are denoted as $I_f^{(k)}$ and $R_f^{(k)}$, respectively. The ISDM facilitates a top–down information flow of illumination and semantic features and calculates a similarity matrix as modulation response to guide the refinement of reflection features. Fig. 2(b) depicts the sequential propagation pathway, where semantic and illumination features successively guide and refine the modulation of reflectance features, which can be formalized as follows:

$$\left\{R_f^{(k)}, I_f^{(k)}\right\} \xrightarrow{\mathcal{T}_{IM}} \tilde{R}_f^{(k)}; \quad \left\{\tilde{R}_f^{(k)}, S_f^{(k)}\right\} \xrightarrow{\mathcal{T}_{SM}} \tilde{R}_f^{(k)}. \tag{2}$$

As depicted in Fig. 2(b), $I_f^{(k)}$ and $R_f^{(k)}$ are individually transformed through layer normalization, 1×1 convolution, and 3×3 depth-wise convolution, further reshaping to obtain illumination query ($\tilde{Q}_i \in \mathbb{R}^{\tilde{H}\tilde{W}\times\tilde{C}}$), reflection key ($\tilde{K}_r \in \mathbb{R}^{\tilde{C}\times\tilde{H}\tilde{W}}$), and reflection value ($\tilde{V}_r \in \mathbb{R}^{\tilde{H}\tilde{W}\times\tilde{C}}$) projections. Subsequently, we calculate the Illumination-Induced Attention (I-InAtt) map, $\mathbf{Z}_{\mathtt{I-InAtt}} \in \mathbb{R}^{\tilde{C}\times\tilde{C}}$. This map undergoes normalization via the softmax function $\mathtt{Softmax}$. The reflectance feature $R_f^{(k)}$ is then dynamically updated through the transposed Illumination-Induced Response (I-InR, $W_{\mathtt{I-InR}}$)

$W_{\mathtt{I-InR}}$

$$= \mathrm{Conv}\left[\Phi_R\left(\tilde{V}_r \otimes \mathrm{Softmax}\left(\frac{\mathbf{Z}_{\mathrm{I-InAtt}}}{(\tilde{Q}_i \otimes \tilde{K}_r)/\alpha}\right)\right)\right] + R_f^{(k)} \tag{3}$$

where α represents a learnable scaling factor that adjusts the magnitude of the product of $\tilde{K}_r$ and $\tilde{Q}_i$ prior to the application of the softmax function. $\Phi_R(\cdot)$ is the reshape transform. $\otimes$ denotes the element-wise multiplication operation. In the final step, we introduce a feedforward network ψ_{FFN}, following methodologies in [44] and [48], for enhanced content reconstruction, expressed as: $\bar{R}_f^{(k)} = \psi_{FFN}(W_{\mathrm{I-InR}})$. It is designed as the product of elements from two parallel paths of linear transformation layers, one of which is activated by Gaussian error linear unit (GELU) nonlinearity [49]. This architecture encodes the position of adjacent pixels in the information space through depth-wise convolution, focusing on enriching features with contextual information and learning useful local image structures for effective restoration.

Similarly, $\bar{R}_f^{(k)}$ and $S_f^{(k)}$ are fed into $\mathcal{T}_{SM}$ to generate the Semantic-Induced Attention (S-InAtt) map, $\mathbf{Z}_{\mathrm{S-InAtt}} \in \mathbb{R}^{\tilde{C}\times\tilde{C}}$. First, they are also individually transformed through layer normalization, 1×1 convolution, and 3×3 depth-wise convolution, further reshaping to obtain semantic query ($\tilde{Q}_s \in \mathbb{R}^{\tilde{H}\tilde{W}\times\tilde{C}}$), reflection key ($\tilde{K}_r \in \mathbb{R}^{\tilde{C}\times\tilde{H}\tilde{W}}$), and reflection value ($\tilde{V}_r \in \mathbb{R}^{\tilde{H}\tilde{W}\times\tilde{C}}$) projections. The reflectance feature $\bar{R}_f^{(k)}$ is then dynamically updated through the transposed Semantic-Induced Response (S-InR, $W_{\mathrm{S-InR}}$)

$$W_{\mathrm{S-InR}} = \mathrm{Conv}\left[\Phi_R\left(\tilde{V}_r \otimes \mathrm{Softmax}\left(\frac{\mathbf{Z}_{\mathrm{S-InAtt}}}{(\tilde{Q}_s \otimes \tilde{K}_r)/\beta}\right)\right)\right] + \bar{R}_f^{(k)} \tag{4}$$

where β represents a learnable scaling factor that adjusts the magnitude of the product of $\tilde{K}_r$ and $\tilde{Q}_s$ prior to the application of the softmax function. Then, $W_{\mathrm{S-InR}}$ undergoes a transformation executed by ψ_{FFN}, resulting in the doubly modulated reflection feature $\tilde{R}_f^{(k)} = \psi_{FFN}(W_{\mathrm{S-InR}})$.

C. Resolution-Sensitive Merging Upsampler

Traditional SR methods often depend on basic single-layer interpolation upsampling techniques (such as bilinear or bicubic interpolation) to enhance image resolution [50]. Lacking the capacity to fully harness the synergistic potential of scale-sensitive features, these methods struggle to produce intricate details of high quality and often result in the presence of unsightly artifacts and halos. Motivated by this, we construct the RSMU module that amalgamates distinct sampling modalities as foundational constituents to progressively learn the mapping from low-resolution space to high-resolution space.

As illustrated in Fig. 3(c), RSMU comprises three stages, each employing different sampling modalities as substrates, including pixel shuffle upsampling, bilinear interpolation, and bicubic upsampling. Initially, it employs pixel shuffle simultaneously producing three separate scale features $U_S^{i\times}$, where

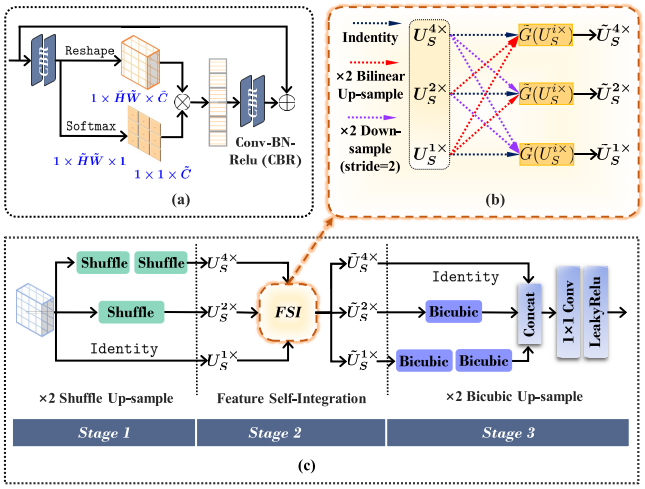


Fig. 3. Illustrations of (a) CU element, (b) FSI for clustering of features and basic, and (c) RSMU with multiple resolution-sensitive upsampler substrates.

$i \in \{1, 2, 4\}$ indexes the scale layer. Next, these three groups of distinct features are converged within the feature self-integration (FSI) module for both feature selection and fusion. As depicted in Fig. 3(b), FSI aggregates and transforms the input features $U_S^{i\times}$, yielding corresponding outputs denoted as $\tilde{U}_S^{j\times}$, where j obtains values from $\{1, 2, 4\}$ to represent different final resolution scales. Through the introduced selective attention mechanism [51], denoted as $\mathcal{M}_{skff}$, each $\tilde{U}_S^{j\times}$ adaptively chooses critical features, which can be defined as follows:

$$\tilde{U}_S^{j\times} = \lfloor \tilde{G}(U_S^{i\times}) \rfloor_{i=1,2,4}, \ \tilde{G} = \mathcal{M}_{skff} \circ \begin{cases} 1 & \forall\ i = j \\ \frac{j}{i} \uparrow & \forall\ j > i \\ \frac{i}{j} \downarrow & \forall\ j < i. \end{cases} \tag{5}$$

In the precise mathematical expression of $\tilde{G}$: when $i = j$, an identity operation (with unchanged size) is used (blue dashed line); when $j < i$, a transposed convolution downsampling with a step size of 2 is employed (yellow dashed line); and when $j > i$, bilinear interpolation is used (red dashed line).

D. Loss Function

1) Self-Regularized Luminance Loss: Inspired by the color statistical regularities of natural image distributions, we propose the self-regularized luminance loss $\mathcal{L}_{SL}$, to encourage the naturalness of colors for $\mathbf{v}^{NL}$, which can be formulated as

$$\mathcal{L}_{SL}(\mathbf{v}^{NL}) = e^{|\bar{\mathbf{v}}_c^{NL} - \boldsymbol{\mu}_c - \boldsymbol{\sigma}_c|} - 1, \ \ c \in \{R, G, B\} \tag{6}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$ denote the mean and standard deviation of the natural image distribution of Imagenet, respectively [52], with $\boldsymbol{\mu}_c = [0.485, 0.456, 0.406]$ and $\boldsymbol{\sigma}_c = [0.229, 0.224, 0.225]$. $\bar{\mathbf{v}}^{NL}$ denotes the channel-wise mean.

2) Illumination Smooth Loss: Drawing insights from [6], we utilize the $\mathrm{Smooth}_{\mathcal{L}_1}$ loss to ensure the structural consistency between the generated illuminance image $\mathbf{u}^{NL}$ and the gray image $\mathbf{x}_G^{LL}$ of input, presented as

$$\mathcal{L}_{IS}(\mathbf{u}^{NL}, \mathbf{x}_G^{LL}) = \mathrm{Smooth}_{\mathcal{L}_1}(\mathbf{u}^{NL} - \mathbf{x}_G^{LL}) \tag{7}$$

where $\texttt{Smooth}_{\mathcal{L}_1}(x) = 0.5x^2$ when $|x| \leq 1$; otherwise, $\texttt{Smooth}_{\mathcal{L}_1}(x) = |x| - 0.5$.

3) Reconstruction Loss: To maintain the content consistency between the generated image $\mathbf{y}^{NS}$ and the reference image (standard-light high-resolution image) $\mathbf{y}^{NH}$, we introduce the reconstruction loss, i.e.,

$$\mathcal{L}_R\left(\mathbf{y}^{NS}, \mathbf{y}^{NH}\right) = ||\mathbf{y}^{NS} - \mathbf{y}^{NH}||_1 \tag{8}$$

4) Perceptual Loss: To maintain perceptual consistency, we introduce a perceptual loss [53] to calculate the disparity between $\mathbf{y}^{NS}$ and $\mathbf{y}^{NH}$

$$\mathcal{L}_P\left(\mathbf{y}^{NS}, \mathbf{y}^{NH}\right) = \frac{1}{c_j h_j w_j} ||\phi_j\left(\mathbf{y}^{NS}\right) - \phi_j\left(\mathbf{y}^{NH}\right)||_1 \tag{9}$$

where ϕ is the VGG-19 model and $c_j h_j w_j$ indicates the feature map size at the jth layer. We use $\{\texttt{conv1}, \ldots, \texttt{conv5}\}$ feature layers with weights $\{0.1, 0.1, 1, 1, 1\}$.

5) Overall Loss: We train our network by minimizing the following overall loss:

$$\mathcal{L}_{total} = \lambda_1 * \mathcal{L}_{SL} + \lambda_2 * \mathcal{L}_{IS} + \lambda_3 * \mathcal{L}_R + \lambda_4 * \mathcal{L}_P \tag{10}$$

where the weights $\{\lambda_i\}_{i=1}^4$ are set to 1.0, 1.0, 1.0, and 1.2.

E. Discussion

1) Naive Cascade Learning Versus DS Learning: The proposed UltraIS model is not a mere cascade that simply combines two distinct tasks, following a "First LLIE then SR" approach of initially brightening images and subsequently performing SR. Instead, the DS architecture is embodied on both the "macro–micro" levels.

1) *Macro-Principle Level:* Rooted in Retinex theory, we apply distinct stream constraints to the two branches of the entire network. The first stream, operating in the low-frequency domain, adheres to dark scene constraints (i.e., luminance constraint), driving it to learn uniform illumination features and initially brightened reflection maps (this step also aims to avoid direct SR of dark images, which would amplify dark features). The second stream, in the high-frequency domain, is guided by reconstruction constraints, ensuring the post-SR image reconstructs high-quality content and texture details.
2) *Microfeature Level:* Our innovative ISDM works in a "top–down" manner, receiving dual modulation from both "illumination stream features" and "semantic stream features." Positioned within the decoding stage, the ISDM modulates reflection features in a way that effectively prevents the exacerbation of hue anomalies and enhances the accuracy of texture details. This DS guidance enables dynamic and precise learning and adaptation of reflection features, further enhancing the model's capability to handle low-illumination scenarios.

IV. Experiments

A. Dataset and Experimental Setup

1) Dataset and Evaluation Metrics: We evaluate the performance on three widely used datasets: RELLISUR,[3] DarkFace[4] [54], and Dark-Zurich[5] [55]. We utilize the RELLISUR dataset for training our method since RELLISUR offers a collection of 1045 image pairs at three distinct resolution scales ($\times 1$, $\times 2$, and $\times 4$) and five varying low-light levels (ranging from -2.5 to -5.0 EV), encompassing both low-resolution low-light images and high-resolution standard-light images. To broadly assess the real-world performance of our proposed method under various authentic low-light open scenes and different exposure conditions, the remaining two datasets (i.e., DarkFace and Dark-Zurich) are employed as extensive test sets. Specifically, the DarkFace dataset, comprising 1000 images with dimensions of 1024×720, was utilized to further gauge the generalization capability of our method in actual night settings scenarios. In addition, we selected a subset of 151 nighttime autonomous driving scenes from the Dark-Zurich open dataset, with dimensions downscaled to 480×270, to serve as a test set.

We evaluate the performance using three widely used full-reference metrics, including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) [56], as well as three non-reference metrics, including blind/referenceless image spatial quality evaluator (BRISQUE) [57], MetaIQA [58], and natural image quality evaluator (NIQE) [59].

2) Implementation Details: All experiments were performed on a PC equipped with an NVIDIA GeForce GTX 2080Ti GPU. Our model was trained using the Adam optimizer for a total of 150 000 iterations. The initial learning rate is set to 2×10^{-4} and the weight decay parameter is 1×10^{-4} with $\beta = [0.9, 0.999]$. The progressive training strategy with mini-batch sizes sets to $[8, 5, 4, 2, 1, 1]$.

3) Comparisons With Cutting-Edge Methods: To ensure a comprehensive comparison, we compare our method against three distinct schemes: 1) *standard-light SR methods*, i.e., RDN [60], SRFBN [38], PAN [40], SwinIR [43], MIRNet [41], Restormer [44], SRFormer [45], and HAT [17]; 2) *low-light SR methods*, i.e., LCUN [16] and CollaBA [46]; and 3) *cascaded LLIE⇒SR methods*, i.e., ZeroDCE⇒HAT, SCI⇒HAT, and LLFormer⇒HAT. Note that we meticulously selected three representative LLIE methods, including ZeroDCE [10], SCI [25], and LLFormer [15]. The enhanced results of LCUN [16] on the RELLISUR dataset are provided by the authors. For all other methods, we retrain their released code on the training set of RELLISUR dataset. Over half of these methods have been proposed within the past two years.

Among numerous SR methods, we have chosen the latest standard-light SR method, HAT, as the subsequent magnification model, cascaded behind three LLIE methods. This setup allows us to observe the performance of the concatenated methods for the LLISR problem. Notably, considering that ZeroDCE and SCI are unsupervised methods, we solely trained them on the $\times 1$ low-light RELLISUR dataset for the preliminary brightening task. In addition, as HAT serves as a posterior module, used for SR on the brightened images, we trained the

[3]https://vap.aau.dk/rellisur/

[4]https://flyywh.github.io/CVPRW2019LowLight/

[5]https://www.trace.ethz.ch/publications/2019/GCMA_UIoU/

TABLE I
EXPLORING THE PERFORMANCE COMPARISON BETWEEN ONE-STREAM (OS) AND DS ARCHITECTURES

Configuration	PSNR↑	SSIM↑	LPIPS↓
OS	$20.213_{\downarrow 2.051}$	$0.676_{\downarrow 0.067}$	$0.316_{\uparrow 0.062}$
DS (w/ $\mathbf{x}^{LL}$)	$21.821_{\downarrow 0.443}$	$0.713_{\downarrow 0.030}$	$0.285_{\uparrow 0.031}$
DS (w/ $\mathbf{v}^{LL}$)	**22.264**	**0.743**	**0.254**

model on the $\times 2$ and $\times 4$ standard-light RELLISUR datasets, denoting it as $\text{HAT}_{\ddagger}$ to distinguish it from the original HAT.

B. Structure Evaluation and Generalization Analysis

1) Why Design the DS Learning Framework: We conducted an ablation study to compare the performance of single-stream versus DS architectures. As depicted in Table I, removing the DS branch and retaining only a single second-stage learning process results in a substantial 9% reduction in the most severe PSNR score degradation. In addition, substituting the input of the second stage with $\mathbf{x}^{LL}$ leads to a decrease of 0.443 dB in PSNR. These observations collectively underscore the significance of applying prior constraints to low-light scenes.

2) Generalization Analysis Across Diverse Darkness Levels: Fig. 4 showcases the adaptability to various low-light levels. For the same input image, five different low exposure levels were created by adjusting exposure time to produce corresponding dark images (e.g., -2.0, -2.5, -3.0, -3.5, and -4.0 EV), as depicted in the first row of the figure. Rows 2–5 display the results of the three latest methods (i.e., SRFormer, HAT, and Restormer) and our proposed approach for the $\times 2$ LLISR task. For better visualization, we present input and output at equal sizes. Upon observation, it can be seen that SRFormer and HAT exhibit extremely pronounced artifacts along the borders, which intensify as the image becomes darker. Restormer, ranking second in quantitative performance, demonstrates overwhitened (overexposed) visual effects at lower darkness levels (i.e., -2.0, -2.5, and -3.0 EV), exhibiting a certain degree of color deviation when compared to the reference image. In comparison, our method maintains consistent restoration effects across different darkness levels, yielding natural and detailed textures.

C. Benchmark Evaluation

1) Experiments on the RELLISUR Dataset: Table II presents the quantitative results among various standard-light SR methods and cascaded methods on the *RELLISUR* dataset for $\times 2$ and $\times 4$ LLISR tasks. Our method outperforms existing state-of-the-art (SOTA) approaches, securing the highest scores in four evaluation metrics, and ranking second in the remaining two indicators. Among the competing methods, Restormer and LLFormer⇒HAT emerge as the second and third overall, respectively, based on their aggregate performance across all metrics. Compared to the standard-light SR method Restormer, our approach achieved a significant improvement (i.e., ↑5% in PSNR, ↑2% in SSIM, and ↑34% in LPIPS). Compared to the second-ranking low-light SR method, CollaBA, our approach achieves the top performance in eight out of 12 metrics for both $\times 2$ and $\times 4$ tasks. Specifically, in terms of PSNR scores, our method achieves an improvement of 0.64 and 0.61 dB over the CollaBA method for $\times 2$ and $\times 4$ tasks, respectively. This demonstrates the superiority of our designed method for LLISR tasks.

Qualitative results on realistic *RELLISUR* dataset for $\times 2$ task are displayed in Figs. 5 and 6. Figs. 7 and 8 display the simultaneous illumination correction and magnification results for the $\times 4$ task on the realistic RELLISUR dataset. The input image size is 624×624, and the output image sizes for the two models are 1248×1248 for $\times 2$ task and 2496×2496 for $\times 4$ task. As illustrated in Figs. 5 and 7, the first four methods [see (a)–(d)] generate significant noise and artifacts after performing SR on low-light images, severely limiting the clarity of the images; the middle four methods [see (e)–(h)], although with fewer artifacts, fail to generate refined high-frequency details, such as the edges of windows and the textures of bricks. Upon zooming in on regions, our method compared to other methods produces authentic images with vivid lightness and an exceptional ability to restore high-frequency structural features. As illustrated in Fig. 8, the cascaded methods [see (b)–(d)] induce overexposure, struggling to maintain or enhance the color fidelity of the images; method [see (e)] also produces color bias, resulting in a whitened color appearance, while methods [see (f)–(g)] exhibit gray artifacts and blurring in areas such as window edges and brick textures. The probability density curves in Figs. 6 and 8 provide an intuitive visual representation of the differences in enhancement effects between various methods and the reference image at the pixel level. A more consistent curve, as seen in (h), corroborates the superiority of our method.

2) Experiments on the Real DarkFace Dataset: To better evaluate the performance of simultaneous brightening and magnification tasks in real nighttime scenes, we calculated two unpaired metrics on the DarkFace dataset to assess the quantitative scores of nine different methods, as shown in Fig. 9. It can be observed in the two histograms that our results, represented in red, consistently achieve the best scoring ranks, particularly with MetaIQA, where our score improved by 19.13% over the second-best score. This demonstrates the effectiveness of our method in removing various types of degradation such as noise, blurring, and exposure issues. Furthermore, Fig. 10 depicts a violin plot to showcase the statistical distribution trends of scores across the entire DarkFace dataset. Note that in the violin plot, the scatter points and vertical lines represent the minimum value, first quartile (25%), median (50%), third quartile (75%), and maximum value of the data. Upon observation, it can be seen that for the first metric, our method ranks first in the first quartile, median, and third quartile; similarly, for the second metric, it also ranks first in the first quartile, median, and third quartile, with the maximum value significantly exceeding that of other methods.

Fig. 11 displays the visualization results on the DarkFace dataset. Through observation, it is clear that other methods exhibit notable color deviations (leaning toward red) and fail to generate relatively fine texture details, as seen in the

TABLE II

QUANTITATIVE COMPARISON ON *RELLISUR* DATASET FOR @×**2** AND @×**4** TASKS. NOTICE THAT ALL METHODS ARE RETRAINED ON *RELLISUR*. ARROW SYMBOLS ↑ / ↓ FOR A METRIC INDICATE WHETHER A HIGHER OR LOWER VALUE IS MORE DESIRABLE. THE TOP TWO PERFORMANCES ARE HIGHLIGHTED IN BOLD WITH RED AND GREEN

Method \ Metric	*RELLISUR* @ ×2						*RELLISUR* @ ×4					
	❶PSNR↑	❷SSIM↑	❸LPIPS↓	❹BRISQUE↓	❺MetaIQA↑	❻NIQE↓	❶PSNR↑	❷SSIM↑	❸LPIPS↓	❹BRISQUE	❺MetaIQA	❻NIQE↓
RDN	18.795	0.701	0.455	60.074	0.310	7.864	18.213	0.703	0.584	56.573	0.300	9.098
SRFBN	18.421	0.662	0.510	57.389	0.278	**7.202**	17.679	0.665	0.640	58.878	0.271	**7.766**
PAN	18.783	0.693	0.450	59.232	0.295	7.370	18.106	0.700	0.559	62.353	0.303	8.737
SwinIR	18.386	0.640	0.577	63.179	0.276	7.231	17.534	0.663	0.688	58.362	0.277	9.489
MIRNet	21.053	0.720	0.436	62.136	0.310	7.881	19.783	0.704	0.599	57.262	0.294	8.803
Restormer	21.215	0.727	0.385	56.210	0.329	7.636	20.298	0.720	0.492	62.071	0.333	9.065
††LCUN	18.911	0.684	0.531	56.104	0.210	8.543	18.463	0.657	0.644	55.676	0.212	9.992
SRFormer	19.556	0.704	0.469	53.418	0.296	7.550	18.723	0.705	0.613	61.658	0.299	9.664
HAT	20.213	0.719	0.454	61.668	0.310	8.332	19.751	0.715	0.561	65.003	0.303	10.028
CollaBA	**21.621**	**0.787**	**0.247**	51.510	**0.434**	8.083	**20.423**	**0.734**	**0.371**	**54.894**	**0.423**	8.554
†ZeroDCE⇒‡HAT	12.927	0.354	0.698	54.873	0.228	7.841	12.524	0.321	0.739	59.721	0.275	8.132
†SCI⇒‡HAT	14.963	0.439	0.591	**51.509**	0.248	7.625	14.776	0.452	0.697	61.690	0.244	8.235
‡LLFormer⇒†HAT	21.218	0.720	0.455	65.497	0.290	8.574	20.135	0.718	0.575	68.499	0.295	9.231
Ours	**22.264**	**0.743**	**0.254**	**49.348**	**0.438**	**7.216**	**21.036**	**0.726**	**0.371**	**53.043**	**0.434**	**7.537**

† signifies training using the ×*1* low-light RELLISUR dataset for LLIE. ‡ indicates training using the ×*2* or ×*4* *RELLISUR* dataset for standard-light SR.
†† The enhanced results on RELLISUR test datasets are provided by the authors.

Fig. 4. Generalization analysis across diverse darkness levels on the RELLISUR dataset for ×2 LLISR task, ranging from −2.0 to −4.0 EV.

blurred texture of the tree trunk within the dashed rectangle. In contrast, our method generates more vibrant nighttime scenes with realistic contrasted architectural elements. Furthermore, in terms of probability density distribution in the figure, our method shows a greater overlap in probability distribution across the RGB channels compared to other methods, with the trends of the three curves being more consistent. This is largely attributed to our model's prior constraints on nighttime scenes and the intricate design of our fine-grained architecture.

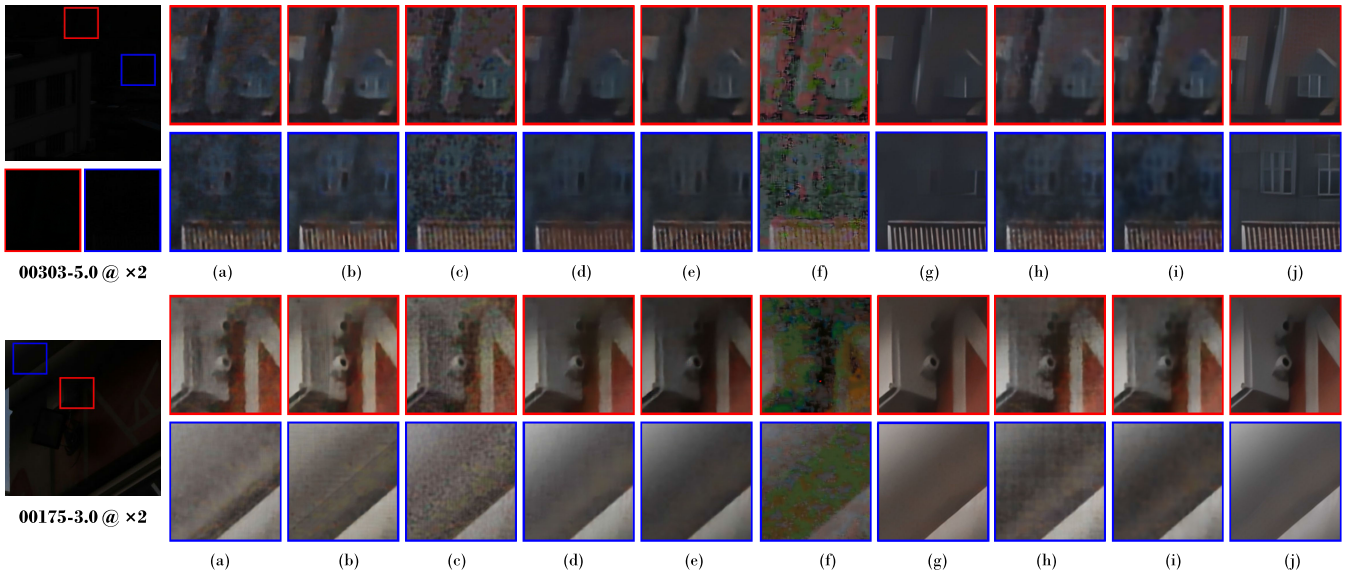


Fig. 5. Qualitative comparison results with two examples (i.e., 00303-5.0 and 00175-3.0) on the RELLISUR dataset for ×2 LLISR task. The resolution of the image is enhanced from 624 × 624 to 1248 × 1248. (a) SRFBN. (b) PAN. (c) SwinIR. (d) MIRnet. (e) Restormer. (f) LCUN. (g) CollaBA. (h) SRFormer. (i) HAT. (j) Ours.

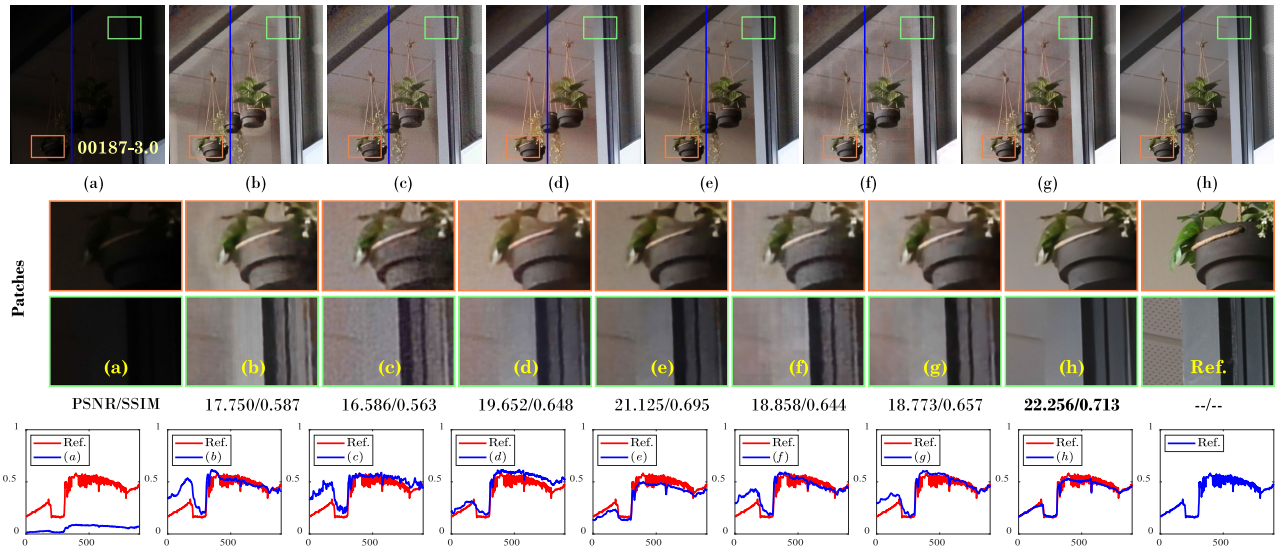


Fig. 6. Visual comparisons on the RELLISUR dataset of our method with SOTA methods for ×2 LLISR task. For the input image 00187-3.0, the enlarged orange and green boxes provide a clearer and more distinct visualization of the comparative results of the seven different advanced methods. Please note that we have annotated two quantitative metrics (i.e., PSNR/SSIM) below the images for comparison. Below signal maps provide the differences of pixel value between created images and the reference image across the blue line segment. (a) Input. (b) PAN. (c) SwinIR. (d) MIRNet. (e) Restormer. (f) SRFormer. (g) HAT. (h) Ours.

3) Experiments on the Real Dark-Zurich Dataset: We assess the quantitative comparison results of related methods on the actual night settings open-road scenario of the Dark-Zurich dataset, as shown in Fig. 12. Our method achieves the best results across three metrics, which indicates that our approach can effectively handle a variety of degradations, such as noise, blurring, and underexposure. Fig. 13 presents a qualitative visual effect comparison on the Dark-Zurich dataset. Upon observation, compared to other methods (such as Restormer and SRFormer), which either exhibit obvious overexposure or widespread underexposure and overexposure issues (as seen with LLFormer⇒HAT), it is clear that our method produces more realistically illuminated and evenly lit street scenes, with more vivid textural details (such as vegetation and pedestrians).

D. Assessment of Computational Resource Efficiency

To examine model efficiency, we report the *parameters* (MB)↓, *FLOPs* (G)↓, *inference* (S)↓, and *frames per second* (FPS)↑ of compared SOTA methods in Table III. The measurements are conducted on a single 2080Titan GPU using images of size 128 × 128. Excluding the parameters of the additionally loaded semantic network, the network of our method has a relatively small footprint, with a total parameter count of less than 3.8 MB. It is clear that our approach not only

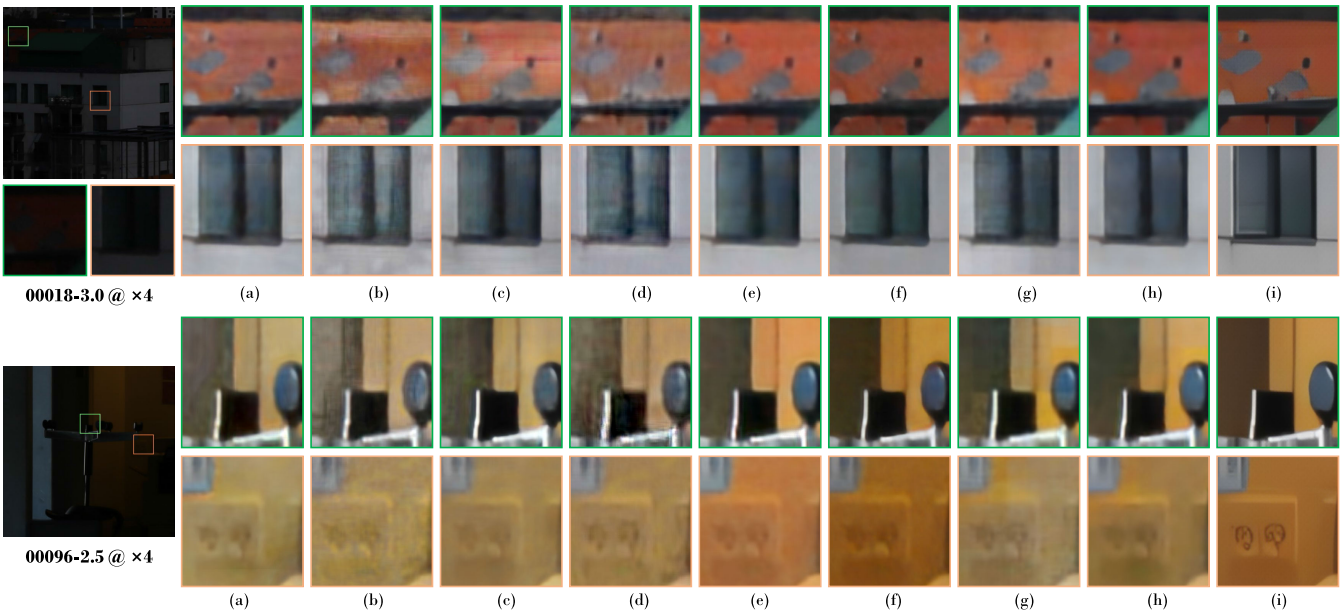


Fig. 7. Qualitative comparisons with two examples (i.e., 00018-3.0 and 00096-2.5) on the RELLISUR dataset for ×4 LLISR task. The image expands from 624 × 624 in the input to 2496 × 2496 in the output. (a) RDN. (b) SRFBN. (c) PAN. (d) SwinIR. (e) MIRnet. (f) Restormer. (g) SRFormer. (h) HAT. (i) Ours.

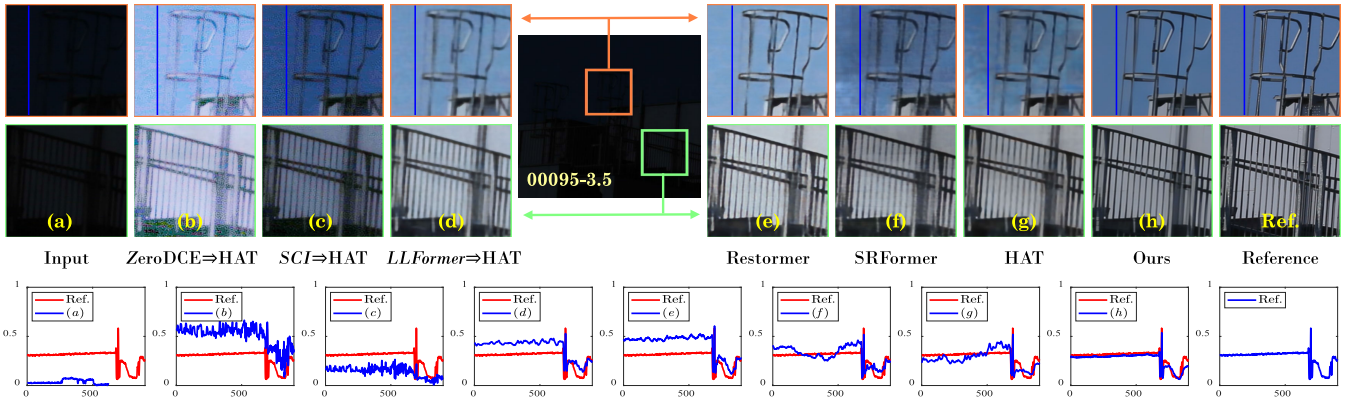


Fig. 8. Visual comparisons on RELLISUR dataset for ×4 LLISR task. The image expands from 624 × 624 in the input to 2496 × 2496 in the output. Below provides the differences of pixel intensity. (a) Input. (b) ZeroDCE⇒HAT. (c) SCI⇒HAT. (d) LLFormer⇒HAT. (e) Restormer. (f) SRFormer. (g) HAT. (h) Ours.

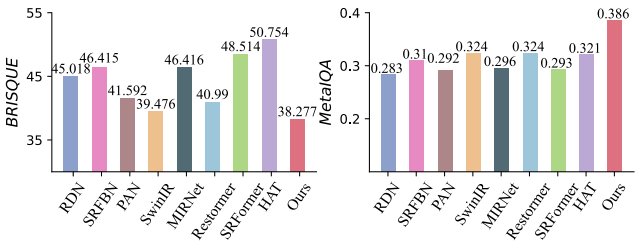


Fig. 9. Quantitative evaluation of actual night settings images for LLISR, conducted on DarkFace samples, illustrating the performance scores across two distinct unpaired metrics, i.e., BRISQUE↓ and MetaIQA↑.

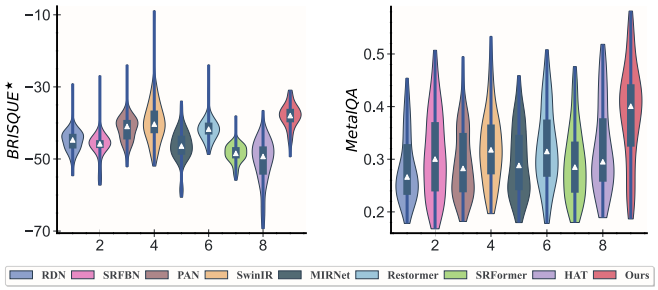


Fig. 10. Quantitative evaluation using a violin plot on the DarkFace dataset, showcasing the statistical distribution trends of two unpaired indicators, i.e., BRISQUE↓ and MetaIQA↑. Note that the median is particularly emphasized with a white triangle marker. The asterisk symbol (⋆) denotes negation.

achieves the lowest FLOPs but also the quickest inference time when benchmarked against the four most recent methodologies. In terms of real-time performance, our method ranks just behind Restormer with regard to two key metrics: inference time and FPS. It demonstrates that our model can process nearly 24 frames per image within 1 s. This performance is sufficient for devices with moderate requirements for frame smoothness, ensuring basic visual quality. Consequently, our method strikes an optimal balance between high performance and computational efficiency.

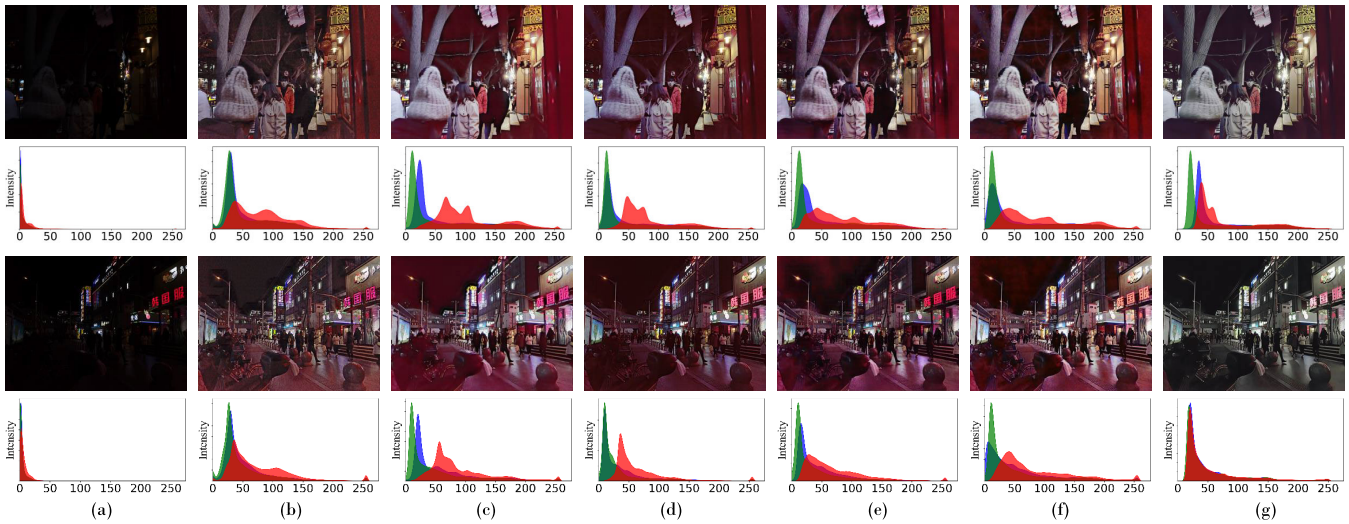


Fig. 11. Visual comparison on DarkFace dataset for ×2 LLISR task. Below is the probability density histogram about RGB. A more consistent probability density distribution across the RGB channels indicates more authentic colors. (a) Input. (b) SwinIR. (c) MIRNet. (d) Restormer. (e) SRFormer. (f) HAT. (g) Ours.

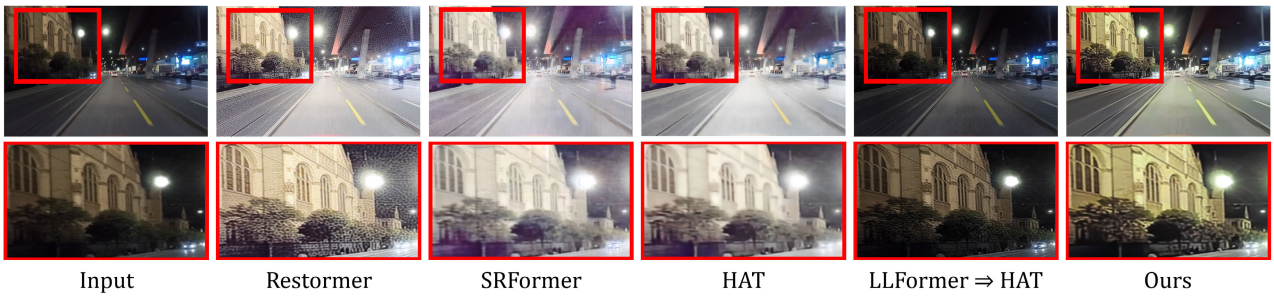


Fig. 12. Qualitative comparisons on the Dark-Zurich dataset for ×2 LLISR task. The image size scales from an input of 480 × 270 to an output of 960 × 540.

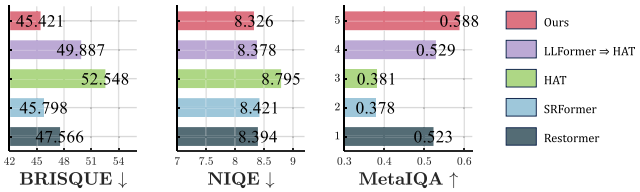


Fig. 13. Quantitative evaluation of actual night settings images for LLISR, conducted on Dark-Zurich samples across three unpaired metrics.

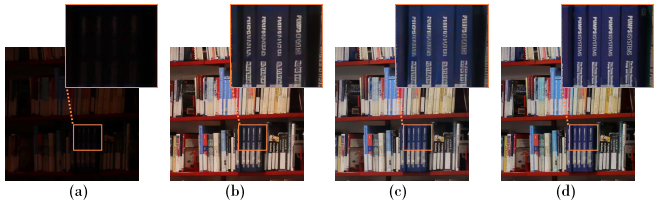


Fig. 14. Effect of the proposed ISDM module. (a) Input. (b) w/o ISDM. (c) Full model. (d) Reference.

TABLE III

COMPUTATIONAL EFFICIENCY OF SOTA METHODS, CONDUCTING ON IMAGES OF SIZE 128 × 128. ∗ IMPLIES THAT WE DID NOT ACCOUNT FOR THE PARAMETERS OF THE PRETRAINED SEGMENTATION NETWORK. THE INDICATORS RANKED FIRST AND SECOND ARE DENOTED WITH BOLD AND UNDERLINE, RESPECTIVELY

Method	❶Parameters (MB)	❷FLOPs (G)	❸Inference (S)	❹FPS
SwinIR	11.683	57.154	0.118	8.46
Restormer	26.126	<u>35.375</u>	**0.033**	**26.87**
SRFormer	10.162	81.797	0.218	3.01
HAT	<u>9.473</u>	58.990	0.184	5.31
Ours*	**3.665**	**31.397**	<u>0.041</u>	<u>24.19</u>

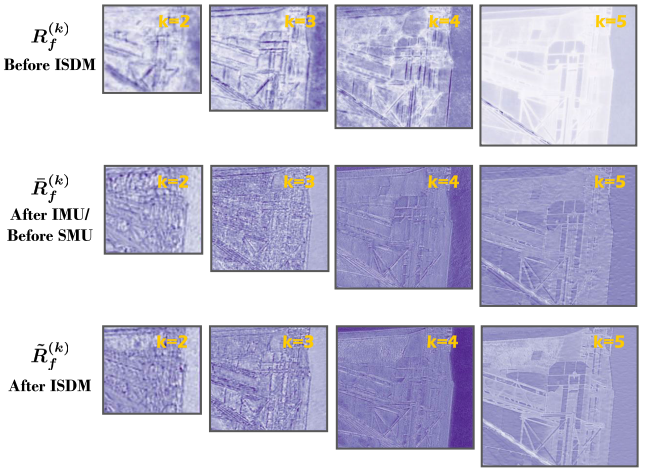


Fig. 15. Comparison of intermediate layer feature visualization for ISDM.

V. ABLATION STUDY

A. Impact of ISDM

Table IV in [*Config. (b)–Config. (d)*] demonstrates the effectiveness of the ISDM module, including the efficacy of its intermediate components IMU and SMU. Compared to the baseline model [*Config. (a)*], removing the ISDM

TABLE IV

ABLATION OF THE ISDM AND RSMU. THE SUBSCRIPT FOR MODELS *(b)–(f)* INDICATES THE PERFORMANCE GAP COMPARED TO MODEL *(a)*. SBMNET† DENOTES TRAINING W/O THE LOSS TERM $\mathcal{L}_{SL}$

Configuration	PSNR↑	SSIM↑	LPIPS↓
(a) SBMNet†	**21.954**	**0.735**	**0.268**
(b) w/o ISDM	21.401$_{\downarrow 0.553}$	0.726$_{\downarrow 0.009}$	0.285$_{\uparrow 0.017}$
(c) ISDM - SMU	21.635$_{\downarrow 0.319}$	0.728$_{\downarrow 0.007}$	0.279$_{\uparrow 0.011}$
(d) ISDM - IMU	21.732$_{\downarrow 0.222}$	0.730$_{\downarrow 0.005}$	0.281$_{\uparrow 0.013}$
(e) Bilinear	21.550$_{\downarrow 0.404}$	0.725$_{\downarrow 0.010}$	0.277$_{\uparrow 0.009}$
(f) RSMU - FSI	21.653$_{\downarrow 0.301}$	0.731$_{\downarrow 0.004}$	0.275$_{\uparrow 0.007}$

TABLE V

ABLATION OF THE DIFFERENT LOSSES ($\mathcal{L}_{SL}$, $\mathcal{L}_{IS}$, AND $\mathcal{L}_P$) ON THE RELLISUR DATASET, FOCUSING ON A $\times 2$ SCALING TASK

No.	Loss				Metric		
	$\mathcal{L}_{SL}$	$\mathcal{L}_{IS}$	$\mathcal{L}_R$	$\mathcal{L}_P$	PSNR↑	SSIM↑	LPIPS↓
0			✓	✓	20.213$_{\downarrow 2.051}$	0.676$_{\downarrow 0.067}$	0.316$_{\uparrow 0.062}$
1		✓	✓		21.614$_{\downarrow 0.650}$	0.723$_{\downarrow 0.020}$	0.462$_{\uparrow 0.208}$
2		✓	✓	✓	21.954$_{\downarrow 0.310}$	0.735$_{\downarrow 0.008}$	0.268$_{\uparrow 0.014}$
3	✓	✓	✓		22.149$_{\downarrow 0.115}$	0.734$_{\downarrow 0.009}$	0.417$_{\uparrow 0.163}$
4	✓	✓	✓	✓	**22.264**	**0.743**	**0.254**

structure results in a significant performance drop (e.g., 0.553-dB decrease in terms of PSNR). Similarly, removing its components (i.e., w/o SMU [*Config. (c)*] or [*Config. (d)*]) also leads to a decrease in performance. Fig. 14 presents the qualitative ablation results of the proposed ISDM module. Upon observing the zoomed-in region, it is clear that removing the ISDM module results in significant color deviation issues in the image compared to the reference image. This confirms the effectiveness of the ISDM in suppressing color deviation.

Fig. 15 also presents the comparison results of feature visualization for ISDM. From the first to the third row, we have visualized the reflection features at three stages through ISMU: before ISMU, after IMU (before SMU), and after ISMU, i.e., $R_f^{(k)}$, $\bar{R}_f^{(k)}$, and $\tilde{R}_f^{(k)}$. Upon comparison, it can be observed that features not processed by IMU and SMU are more generic and dispersed, whereas those processed through IMU and SMU achieve more abstract and enriched visual forms. This indicates that the incorporation of IMU and SMU into the model leads to a greater focus on capturing essential information and higher level semantics.

B. Effects of RSMU

Table IV in [*Config. (e)* and *Config. (f)*] presents the effectiveness of RSMU. In comparison to a simple upsampling layer (e.g., Bilinear), RSMU facilitates a 0.404-dB PSNR increase [*Config. (e)*]. The FSI module [*Config. (f)*] also contribute to the improvement in the final results. Fig. 16 illustrates the qualitative ablation results of the proposed RSMU module. Upon observing the zoomed-in region, it is clear that removing the RSMU module leads to the emergence of artifacts in the image, as indicated by the dashed rectangle. This experiment confirms the effectiveness of the RSMU in suppressing artifacts during the SR process.

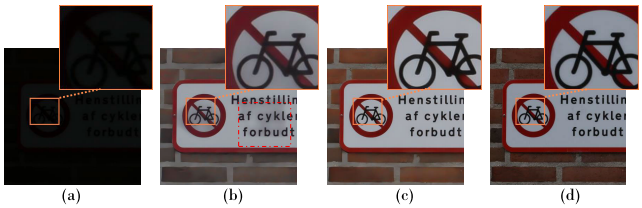


Fig. 16. Effect of the proposed RSMU module. (a) Input. (b) Bilinear, w/o RSMU. (c) Full model. (d) Reference.

Fig. 17. Demonstration of limitations in simultaneous brightening and zooming on four actual night setting images, focusing on a $\times 2$ scaling task. It is clear that for extremely low-light actual night setting images, *particularly in areas with extensive dark skies,* our method struggles to generate fine texture details, as indicated by the green arrowheads.

C. Analysis of Different Losses

Table V illustrates the ablation results for various loss components. As illustrated, when comparing the model (i.e., No.1) solely with $\mathcal{L}_{IS}$ and $\mathcal{L}_R$ to our full model (i.e., No.4), which incorporates all loss components, there is a marked improvement of 0.65 dB in PSNR and a 0.208 enhancement in LPIPS. Comparing No.2 with No.4, we observe that the illuminance constraint imposed on our model (i.e., $\mathcal{L}_{SL}$) leads to an increase of 0.31 dB and 0.014 in PSNR and LPIPS scores, respectively. In addition, comparing No.3 with No.4, the perceptual loss $\mathcal{L}_P$ has the most significant impact on the perceptual LPIPS scores, with a 0.163 enhancement in LPIPS.

D. Discussion of Limitations and Future Prospects

Despite the significant performance gains achieved, the UltraIS model still faces some limitations. Fig. 17 illustrates the limitations of our method when applied to real nighttime images, especially in areas with large backgrounds, such as night skies, where it struggles to generate fine texture details. In future work, our aim is to concentrate on developing robust models to further tackle extremely adverse conditions with multidegradation in low-exposure scenarios. In addition, the increased computation in the DS Unet architecture, combined with the dual-modulation attention mechanism, results in higher GPU memory requirements and increased computational resources. This highlights the necessity of developing a more lightweight structure to efficiently tackle LLISR problems.

VI. Conclusion

This study addresses the relatively unexplored challenge of SR in low-illumination scenarios. Leveraging a novel dual-modulated learning framework, we introduced specialized components to enhance feature-level preservation of illumination and color details while reducing artifacts through an RSMU. Our approach's remarkable applicability and generalizability across diverse ultralow-light conditions were validated through comprehensive experiments. Our work holds significance beyond the specific problem of ultra-dark SR. This broader perspective reinvigorates the research landscape, encouraging exploration of joint multiple image processing tasks (i.e., low-light derain and low-light defog) in diverse adverse conditions.

References

[1] X. Liu, Q. Xie, Q. Zhao, H. Wang, and D. Meng, "Low-light image enhancement by retinex-based algorithm unrolling and adjustment," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2023.

[2] T. Wu, W. Wu, Y. Yang, F.-L. Fan, and T. Zeng, "Retinex image enhancement based on sequential decomposition with a plug-and-play framework," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2023.

[3] Z. Yan et al., "Learning complementary correlations for depth super-resolution with incomplete data in real world," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5616–5626, May 2024.

[4] M. Zhang, Q. Wu, J. Guo, Y. Li, and X. Gao, "Heat transfer-inspired network for image super-resolution reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1810–1820, Feb. 2024.

[5] G. Wu, J. Jiang, and X. Liu, "A practical contrastive learning framework for single-image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12, 2023.

[6] L. Ma, R. Liu, J. Zhang, X. Fan, and Z. Luo, "Learning deep context-sensitive decomposition for low-light image enhancement," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5666–5680, Oct. 2022.

[7] X. Hu, Z. Zhang, C. Shan, Z. Wang, L. Wang, and T. Tan, "Meta-USR: A unified super-resolution network for multiple degradation parameters," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4151–4165, Sep. 2021.

[8] K. Zhang, D. Tao, X. Gao, X. Li, and J. Li, "Coarse-to-fine learning for single-image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1109–1122, May 2017.

[9] C. Li et al., "Low-light image and video enhancement using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9396–9416, Dec. 2022.

[10] 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.

[11] R. Liu, L. Ma, T. Ma, X. Fan, and Z. Luo, "Learning with nested scene modeling and cooperative architecture search for low-light vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5953–5969, May 2022.

[12] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, and X. Fan, "CoCoNet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1748–1775, May 2024.

[13] J. Liu, G. Wu, J. Luan, Z. Jiang, R. Liu, and X. Fan, "HoLoCo: Holistic and local contrastive learning network for multi-exposure image fusion," *Inf. Fusion*, vol. 95, pp. 237–249, Jul. 2023.

[14] J. Liu, J. Shang, R. Liu, and X. Fan, "Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5026–5040, Aug. 2022.

[15] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 3, 2023, pp. 2654–2662.

[16] D. Cheng, L. Chen, C. Lv, L. Guo, and Q. Kou, "Light-guided and cross-fusion U-Net for anti-illumination image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8436–8449, Dec. 2022.

[17] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. Comput. Vis. Pattern Recognit.*, 2023, pp. 22367–22377.

[18] K. Guo et al., "Deep illumination-enhanced face super-resolution network for low-light images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3, pp. 1–19, Aug. 2022.

[19] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, "RELIEF: Joint low-light image enhancement and super-resolution with transformers," in *Proc. Scandin. Conf. Image Anal.*, 2023, pp. 157–173.

[20] M. T. Rasheed and D. Shi, "LSR: Lightening super-resolution deep network for low-light image enhancement," *Neurocomputing*, vol. 505, pp. 263–275, Sep. 2022.

[21] E. Land and J. McCann, "Lightness and Retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, p. 1–11, 1971.

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

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

[24] 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.

[25] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5637–5646.

[26] 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.

[27] D. Jin, L. Ma, R. Liu, and X. Fan, "Bridging the gap between low-light scenes: Bilevel learning for fast adaptation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2401–2409.

[28] L. Ma et al., "Bilevel fast scene adaptation for low-light image enhancement," *Int. J. Comput. Vis.*, vol. 2023, pp. 1–19, Oct. 2023.

[29] R. Liu, L. Ma, Y. Zhang, X. Fan, and Z. Luo, "Underexposed image correction via hybrid priors navigated deep propagation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3425–3436, Aug. 2022.

[30] R. Liu, J. Gao, X. Liu, and X. Fan, "Learning with constraint learning: New perspective, solution strategy and various applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 5026–5043, Jul. 2024.

[31] N. Jiang, J. Lin, T. Zhang, H. Zheng, and T. Zhao, "Low-light image enhancement via stage-transformer-guided network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3701–3712, Aug. 2023.

[32] S. Zhang, N. Meng, and E. Y. Lam, "LRT: An efficient low-light restoration transformer for dark light field images," *IEEE Trans. Image Process.*, vol. 32, pp. 4314–4326, 2023.

[33] X. Shang, G. Li, Z. Jiang, S. Zhang, N. Ding, and J. Liu, "Holistic dynamic frequency transformer for image fusion and exposure correction," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102073.

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

[35] D. Zhang, J. Shao, Z. Liang, L. Gao, and H. T. Shen, "Large factor image super-resolution with cascaded convolutional neural networks," *IEEE Trans. Multimedia*, vol. 23, pp. 2172–2184, 2021.

[36] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[37] J. Xin, J. Li, X. Jiang, N. Wang, H. Huang, and X. Gao, "Wavelet-based dual recursive network for image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 707–720, Feb. 2020.

[38] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 3867–3876.

[39] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Comput. Vis. (ECCV) Workshops*, vol. 11133, Munich, Germany, 2018, pp. 63–79.

[40] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 56–72.

[41] S. W. Zamir et al., "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12370, 2020, pp. 492–511.

[42] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 191–207.
[43] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[44] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.
[45] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, "SRFormer: Permuted self-attention for single image super-resolution," 2023, *arXiv:2303.09735*.
[46] J. Gao, Y. Liu, Z. Yue, X. Fan, and R. Liu, "Collaborative brightening and amplification of low-light imagery via bi-level adversarial learning," *Pattern Recognit.*, vol. 154, Oct. 2024, Art. no. 110558.
[47] Z. Wang, J. Chen, and S. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Mar. 2020.
[48] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.
[49] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[50] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.
[51] S. W. Zamir et al., "Learning enriched features for fast image restoration and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1934–1948, Feb. 2023.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[53] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[54] W. Yang et al., "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020.
[55] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3139–3153, Jun. 2022.
[56] S. S. Agaian, B. Silver, and K. A. Panetta, "Transform coefficient histogram-based image enhancement algorithms using contrast entropy," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 741–758, Mar. 2007.
[57] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[58] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14131–14140.
[59] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Jul. 2012.
[60] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

Jiaxin Gao received the B.S. degree in applied mathematics from Dalian University of Technology, Dalian, China, in 2018, where she is currently pursuing the Ph.D. degree in software engineering.

She is currently with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology. Her research interests include computer vision, machine learning, and optimization.

Ziyu Yue received the B.S. degree in information and computing science from Dalian University of Technology, Dalian, China, in 2017, where he is currently pursuing the Ph.D. degree in computer and graphic imaging.

He is currently with Liaoning Key Laboratory of Computational Mathematics and Data Intelligence, Dalian University of Technology. His research interests include super-resolution, low-light enhancement, and nerf.

Yaohua Liu received the M.S. degree in software engineering from Dalian University of Technology, Dalian, China, in 2021, where he is currently pursuing the Ph.D. degree in software engineering.

His research interests include computer vision, bilevel optimization, adversarial attack and defense, and deep learning.

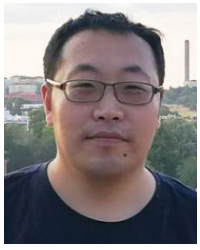
Sihan Xie received the B.E. degree in software engineering from Dalian University of Technology, Dalian, China, in 2022, where she is currently pursuing the master's degree.

Her research interests include super-resolution, underwater enhancement, and action recognition.

Xin Fan (Senior Member, IEEE) received the B.E. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1998 and 2004, respectively.

He was with Oklahoma State University, Stillwater, OK, USA, and the University of Texas Southwestern Medical Center, Dallas, TX, USA, from 2006 to 2009, as a Post-Doctoral Research Fellow. He joined Dalian University of Technology, Dalian, China, in 2009, where he is currently a Full Professor. His current research interests include image processing and machine vision.

Dr. Fan received the 2015 IEEE International Conference on Multimedia and Expo (ICME), Best Student Award as the corresponding author. Two papers were selected as the Finalist of the Best Paper Award at ICME 2017.



Risheng Liu (Member, IEEE) received the B.Sc. and Ph.D. degrees from Dalian University of Technology, Dalian, China, in 2007 and 2012, respectively.

From 2010 to 2012, he was doing research as a joint Ph.D. student at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. From 2016 to 2018, he was doing research as a Hong Kong Scholar at The Hong Kong Polytechnic University, Hong Kong. He is currently a Full Professor with the School of Software Technology, Dalian University of Technology. He was awarded the "Outstanding Youth Science Foundation" of the National Natural Science Foundation of China. His research interests include optimization, computer vision, and multimedia.