Signal Processing: Image Communication 130 (2025) 117222



Contents lists available at ScienceDirect

Signal Processing: Image Communication

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





Text in the dark: Extremely low-light text image enhancement

Che-Tsung Lin [a,1], Chun Chet Ng [b,1], Zhi Qin Tan [b], Wan Jun Nah [b], Xinyu Wang [d], Jie Long Kew [b], Pohao Hsu [c], Shang Hong Lai [c], Chee Seng Chan [b,*], Christopher Zach [a]

[a] Chalmers University of Technology, Gothenburg, Sweden
[b] Universiti Malaya, Kuala Lumpur, Malaysia
[c] National Tsing Hua University, Hsinchu, Taiwan
[d] The University of Adelaide, Adelaide, Australia

ARTICLE INFO

Keywords:
Extremely low-light image enhancement
Edge attention
Text aware augmentation
Scene text detection
Scene text recognition

ABSTRACT

Extremely low-light text images pose significant challenges for scene text detection. Existing methods enhance these images using low-light image enhancement techniques before text detection. However, they fail to address the importance of low-level features, which are essential for optimal performance in downstream scene text tasks. Further research is also limited by the scarcity of extremely low-light text datasets. To address these limitations, we propose a novel, text-aware extremely low-light image enhancement framework. Our approach first integrates a Text-Aware Copy-Paste (Text-CP) augmentation method as a preprocessing step, followed by a dual-encoder–decoder architecture enhanced with Edge-Aware attention modules. We also introduce text detection and edge reconstruction losses to train the model to generate images with higher text visibility. Additionally, we propose a Supervised Deep Curve Estimation (Supervised-DCE) model for synthesizing extremely low-light images, allowing training on publicly available scene text datasets such as IC15. To further advance this domain, we annotated texts in the extremely low-light See In the Dark (SID) and ordinary LOw-Light (LOL) datasets. The proposed framework is rigorously tested against various traditional and deep learning-based methods on the newly labeled SID-Sony-Text, SID-Fuji-Text, LOL-Text, and synthetic extremely low-light IC15 datasets. Our extensive experiments demonstrate notable improvements in both image enhancement and scene text tasks, showcasing the model's efficacy in text detection under extremely low-light conditions. Code and datasets will be released publicly at https://github.com/chunchet-ng/Text-in-the-Dark.

1. Introduction

Scene text understanding involves extracting text information from images through text detection and recognition, which is a fundamental task in computer vision. However, performance drops sharply when images are captured under low-light conditions. The main difficulty in detecting text in low-light images is that low-level features, such as edges and character strokes, are no longer prominent or hardly visible. On the other hand, enhancing images captured in extremely low-light conditions poses a greater challenge than ordinary low-light images due to the higher noise levels and greater information loss. For instance, we show the difference in darkness level in Fig. 1, where it is evident that the See In the Dark (SID) datasets [1] are darker and, in theory, more difficult to enhance than the LOw-Light (LOL) dataset [2]. Quantitatively, we calculated the PSNR and SSIM values for two subsets of SID, SID-Sony and SID-Fuji, and LOL by comparing each image against pure black images in Table 1. Based on each dataset's average perceptual lightness (L* in the CIELAB color space), images in SID are at least 15 times darker than those in LOL. Hence, low-light image enhancement is a necessary pre-processing step for scene text extraction under such conditions.

Over the years, many general or low-light image enhancement models have been proposed to improve the interpretability and extraction of information in images by providing better input for subsequent image content analysis. Early methods [3–5] typically attempted to restore the statistical properties of low-light images to those of long-exposure images from a mathematical perspective. On the other hand, deep learning-based methods [1,2,6,7] aim to learn the mapping between low-light images and their corresponding long-exposure versions via regression. To the best of our knowledge, most existing low-light image enhancement works have not explicitly addressed the restored image quality in terms of downstream scene text tasks.

* Corresponding author.
E-mail address: cs.chan@um.edu.my (C.S. Chan).
[1] These authors contributed equally to this work and are listed in alphabetical order by first name.

https://doi.org/10.1016/j.image.2024.117222
Received 6 February 2024; Received in revised form 6 September 2024; Accepted 16 October 2024
Available online 28 October 2024
0923-5965/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

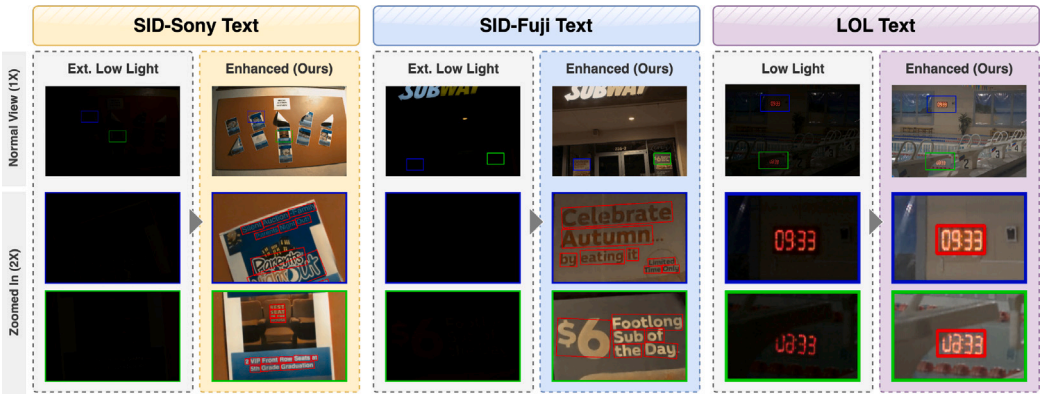


Fig. 1. Comparison of low-light and enhanced text images across different datasets. From left to right: SID-Sony-Text, SID-Fuji-Text, and LOL-Text. Each column shows extremely low-light images (left) and their enhanced versions (right). The top row displays full images, while the bottom row shows 2x zoomed-in regions highlighted by blue and green bounding boxes. Extremely low-light images in the SID datasets are significantly darker than those in the LOL dataset, and our model enhances the images to the extent that text is clearly visible with sharp edges.

Table 1
The difference between the extremely low-light dataset, SID, and the ordinary low-light dataset, LOL, is shown in terms of PSNR and SSIM values, computed by comparing short-exposure images against pure black images. Avg. L* is the average perceptual lightness in the CIELAB color space, calculated based on short-exposure images. Scores are averaged across training and test sets. Higher PSNR and SSIM values, along with lower Avg. L*, indicate darker images that are more challenging for image enhancement and scene text extraction.

Dataset	PSNR ↑	SSIM ↑	Avg. L* ↓
SID-Sony [1]	44.350	0.907	0.009
SID-Fuji [1]	41.987	0.820	0.004
LOL [2]	23.892	0.195	0.142
Pure black	∞	1.000	0.000

Recent advancements in visual attention mechanisms have demonstrated their effectiveness in identifying and boosting salient features in images. Channel-only attention [8–10], spatial attention [11,12] or the subsequent channel-spatial attention [13,14] modules were proposed to emphasize the most informative areas. However, these methods cannot preserve texture details, especially fine-grained edge information that is intuitively needed to enhance extremely low-light images with complex textures. To overcome this limitation, we introduce Edge-Aware Attention (Edge-Att). This novel attention module simultaneously performs channel and spatial attention-based feature learning on high-level image and edge features. Our model also considers text information in the image through a text-aware loss function. This way, our model can effectively enhance low-light images while preserving fine-grained edge information, texture details, and legibility of text.

The scarcity of extremely low-light text datasets presents a hurdle for further research. To address this, we annotated all text instances in both the training and testing sets of the SID and LOL datasets, creating three new low-light text datasets: SID-Sony-Text, SID-Fuji-Text, and LOL-Text. We then proposed a novel Supervised Deep Curve Estimation (Supervised-DCE) model to synthesize extremely low-light scene text images based on the commonly used ICDAR15 (IC15) scene text dataset. It allows researchers to easily translate naive scene text datasets into extremely low-light text datasets. In addition to the previously published conference version of this work [15], we have made four significant extensions. Firstly, we propose a novel dual encoder–decoder framework that can achieve superior performance on low-light scene text tasks (Section 3.1). Secondly, we introduce a new image synthesis method capable of generating more realistic extremely low-light text images (Section 4.1). Thirdly, we have further annotated texts in the Fuji and LOL datasets, thereby forming the largest low-light scene text datasets to date (Section 5). Fourthly, comprehensive experiments and analyses are carried out to study the latest methods along with our proposed methods on all synthetic and real low-light text datasets.

An overview of our extremely low-light image enhancement and text detection results compared to existing methods is showcased in Fig. 2. We demonstrate that our proposed method outperforms existing methods in terms of PSNR and SSIM, while also achieving the highest H-Mean metric on the SID-Fuji-Text dataset. The main contributions of our work are as follows:

- We present a novel scene text-aware extremely low-light image enhancement framework with dual encoders and decoders to enhance extremely low-light images, especially scene text regions within them. Our proposed method is equipped with Edge-Aware Attention modules and trained with new Text-Aware Copy-Paste (Text-CP) augmentation. Our model can restore images in challenging lighting conditions without losing low-level features.
- We developed a Supervised-DCE model to synthesize extremely low-light images. This allows us to use existing publicly available scene text datasets such as IC15 to train our model alongside genuine ones for scene text research under such extreme lighting conditions.
- We labeled the texts in the SID-Sony, SID-Fuji, and LOL datasets and named them SID-Sony-Text, SID-Fuji-Text, and LOL-Text, respectively. This provides a new perspective for objectively assessing enhanced extremely low-light images through scene text tasks.

2. Related works

The overall pipeline of the extremely low-light text detection framework is presented in Fig. 3. The process is divided into two steps: first, image enhancement is performed, followed by text extraction tasks. The architectural difference between existing low-light image enhancement methods and our proposed method is also illustrated in Fig. 3.

Low-light Image Enhancement. Retinex theory assumes that an image can be decomposed into illumination and reflectance. Most Retinex-based methods enhance results by removing the illumination part [3], while others such as LIME [4] keep a portion of the illumination to preserve naturalness. BIMEF [5] further designs a dual-exposure fusion framework for accurate contrast and lightness enhancement. RetinexNet [2] combines deep learning and Retinex theory, adjusting illumination for enhancement after image decomposition. The recent successes of generative adversarial networks (GANs) [16] have attracted attention from low-light image enhancement because GANs have proven successful in image translation. Pix2pix [17] and CycleGAN [18] have shown good image-translation results in paired and unpaired image settings, respectively. To overcome the complexity of CycleGAN, EnlightenGAN [6] proposed an unsupervised one-path GAN

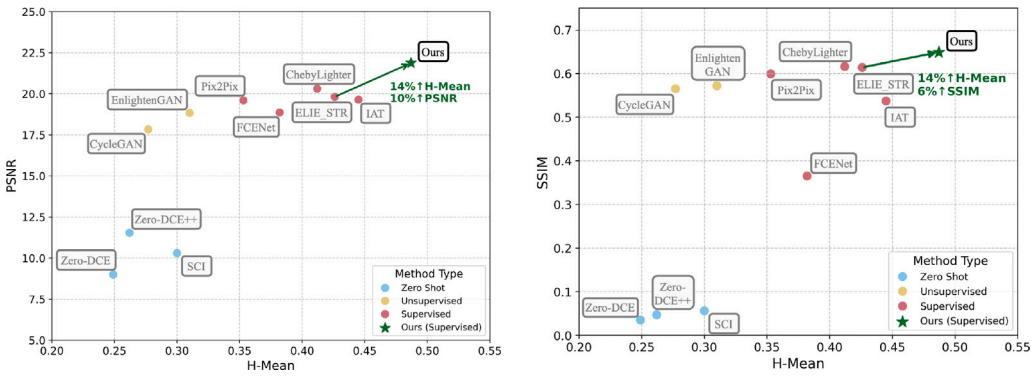


Fig. 2. Tradeoff between PSNR (left) and SSIM (right) against the H-Mean of the CRAFT model for our proposed method and other low-light image enhancement methods on the SID-Fuji-Text dataset.

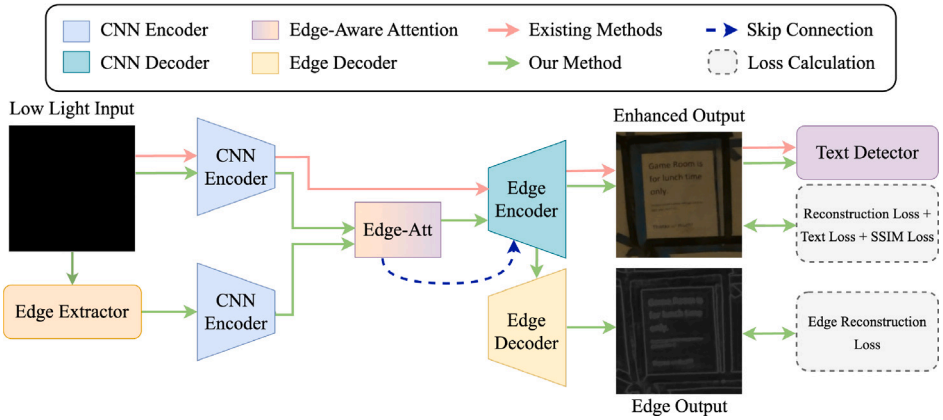


Fig. 3. Architectural Comparison of Low-Light Image Enhancement Frameworks: The *naive* extremely low-light image enhancement framework is guided by red arrows. Our proposed text-aware framework, indicated by green arrows, consists of a dual-encoder–decoder network with Edge-Aware Attention modules. The outputs are used to train the model to focus on text regions using novel text detection and edge reconstruction losses.

structure. Besides general image translation, [1] proposed learning-based low-light image enhancement on raw sensor data to replace much of the traditional image processing pipeline, which tends to perform poorly on such data. EEMEFN [19] also attempted to enhance images using multi-exposure raw data that is not always available.

Zero-Reference Deep Curve Estimation (Zero-DCE) [7] designed a light-weight CNN to estimate pixel-wise high-order curves for dynamic range adjustment of a given image without needing paired images. [20] designed a novel Self-Calibrated Illumination (SCI) learning with an unsupervised training loss to constrain the output at each stage under the effects of a self-calibrated module. ChebyLighter [21] learns to estimate an optimal pixel-wise adjustment curve under the paired setting. Recently, the Transformer [22] architecture has become the de-facto standard for Natural Language Processing (NLP) tasks. ViT [23] applied the attention mechanism in the vision task by splitting the image into tokens before sending it into Transformer. Illumination Adaptive Transformer (IAT) [24] uses attention queries to represent and adjust ISP-related parameters. Most existing models enhance images in the spatial domain. Fourier-based Exposure Correction Network (FECNet) [25] presents a new perspective for exposure correction with spatial-frequency interaction and has shown that their model can be extended to low-light image enhancement.

Scene Text Extraction. Deep neural networks have been widely used for scene text detection. CRAFT [26] predicts two heatmaps: the character region score map and the affinity score map. The region score map localizes individual characters in the image, while the affinity score map groups each character into a single word instance. Another notable scene text detection method is Pixel Aggregation Network (PAN) [27] which is trained to predict text regions, kernels, and similarity vectors. Both text segmentation models have proven to work well on commonly used scene text datasets such as IC15 [28] and TotalText [29]. Inspired by them, we introduced a text detection loss in our proposed model to focus on scene text regions during extremely low-light image enhancement. Furthermore, state-of-the-art text recognition methods such as ASTER [30] and TRBA [31] are known to perform well on images captured in complex scenarios. ASTER [30] employs a flexible rectification module to straighten the word images before passing them to a sequence-to-sequence model with the bi-directional decoder. The experimental results of ASTER showed that the rectification module could achieve superior performance on multiple scene text recognition datasets, including the likes of IC15 and many more. Besides, TRBA [31] provided interesting insights by breaking down the scene text recognition framework into four main stages: spatial transformation, character feature extraction, followed by sequence modeling, and the prediction of character sequences. Given these methods' robustness on difficult texts, they are well-suited to recognize texts from enhanced low-light images.

3. Extremely low-light text image enhancement

The overall architecture of our network is illustrated in Fig. 4, which depicts the dual-encoder–decoder design of our model. The first encoder processes the extremely low-light input image, focusing on extracting general features such as color, texture, and intensity. The second encoder is specifically dedicated to extracting edge features, emphasizing the preservation of text edges and other critical structural elements. Our proposed Edge-Aware Attention (Edge-Att) module is pivotal for enhancing text visibility by modulating attention across different image regions based on edge features.

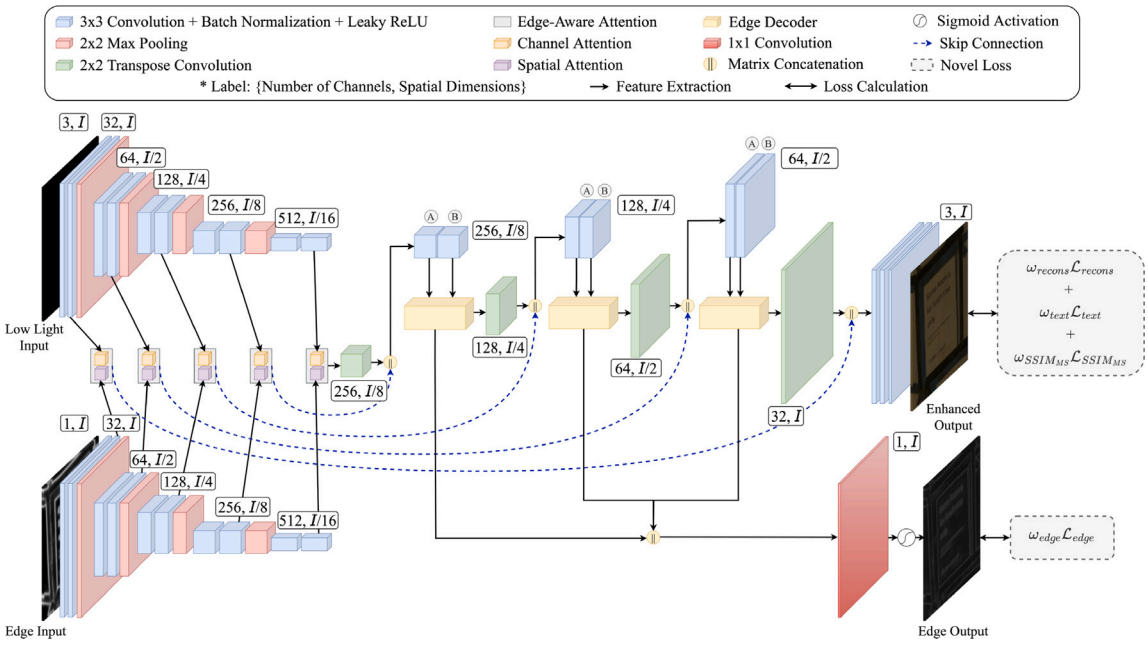


Fig. 4. Illustration of the architecture of our proposed framework, designed to enhance extremely low-light images while incorporating scene text awareness. I stands for the spatial image size.

3.1. Problem formulations

Let $x \in R^{W\times H\times 3}$ be a short-exposure image of width W and height H. An ideal image enhancement expects that a neural network $LE(x;\theta)$ parameterized by θ can restore this image to its corresponding long-exposure image, $y \in R^{W\times H\times 3}$, i.e., $LE(x;\theta) \simeq y$. However, previous works normally pursued the lowest per-pixel intensity difference, which should not be the goal for image enhancement because we usually expect that some high-level computer vision tasks can work reasonably well on those enhanced images. For example, in terms of text detection, the goal of the neural network can be the lowest detection bounding boxes discrepancy, i.e., $B(LE(x;\theta)) \simeq B(y)$.

Our novel image enhancement model consists of a U-Net accommodating extremely low-light images and edge maps using two independent encoders. During model training, instead of channel attention, the encoded edges guide the spatial attention sub-module in the proposed Edge-Att to attend to edge pixels related to text representations. Besides the image enhancement losses, our model incorporates text detection and edge reconstruction losses into the training process. This integration effectively guides the model's attention towards text-related features and regions, facilitating improved image textual content analysis. As a pre-processing step, we introduced a novel augmentation technique called Text-CP to increase the presence of non-overlapping and unique text instances in training images, thereby promoting comprehensive learning of text representations.

3.2. Network design

Our model was inspired by U-Net [1] with some refinements. Firstly, the network expects heterogeneous inputs, i.e., extremely low-light images, x, and the corresponding RCF [32] edge maps, e. Secondly, input-edge pairs are handled by two separate encoders with edge-aware attention modules between them. The attended features are then bridged with the decoder through skip connections. Finally, our multi-tasking network predicts the enhanced image, x', and the corresponding reconstructed edge, e'. The overall architecture of our network can be seen in Fig. 4 and modeled as:

$$x', e' = LE(x, e; \theta). \tag{1}$$

3.3. Objectives

Our proposed model is trained to optimize four loss functions. The first two, Smooth L1 loss and multi-scale SSIM loss focus on enhancing the overall image quality. The third, text detection loss, targets the enhancement of scene text regions specifically. The fourth, edge reconstruction loss, focuses on crucial low-level edge features.

Firstly, we employ smooth L1 loss as the reconstruction loss to better enforce low-frequency correctness [17] between x' and y as:

$$\mathcal{L}_{recons} = \begin{cases} 0.5 \cdot (x'-y)^2/\delta, & \text{if } |x'-y| < \delta \\ |x'-y| - 0.5 \cdot \delta, & \text{otherwise} \end{cases} \tag{2}$$

where we empirically found that $\delta = 1$ achieved good result. The authors of Pix2Pix [17] showed that by utilizing L1 loss, the model can achieve better results as the generated images are less blurry and proved that L1 loss can better enforce the learning of low-frequency details, which is also essential for OCR tasks. On the other hand, the L1 norm is less sensitive to outliers than the L2 norm, thus resulting in a more robust model towards extreme pixel intensities.

Secondly, the multi-scale SSIM metric was proposed in [33] for reference-based image quality assessment, focusing on image structure consistency. An M-scale SSIM between the enhanced image x' and ground truth image y is:

$$SSIM_{MS}(x', y) = [l_M(x', y)]^{\tau} \cdot \prod_{j=1}^{M} [c_j(x', y)]^{\phi} [s_j(x', y)]^{\psi}, \tag{3}$$

where l_M is the luminance at M-scale; c_j and s_j represent the contrast and the structure similarity measures at the jth scale; τ, ϕ, and ψ are parameters to adjust the importance of the three components. Inspired by [33], we adopted the M-scale SSIM loss function in our work to enforce the image structure of x' to be close to that of y:

$$\mathcal{L}_{SSIM_{MS}} = 1 - SSIM_{MS}(x', y). \tag{4}$$

Thirdly, a well-enhanced extremely low-light image implies that we could obtain similar text detection results on both the enhanced and ground truth images. As such, we propose to employ CRAFT [26] to localize texts in images through its region score heatmap. To implicitly

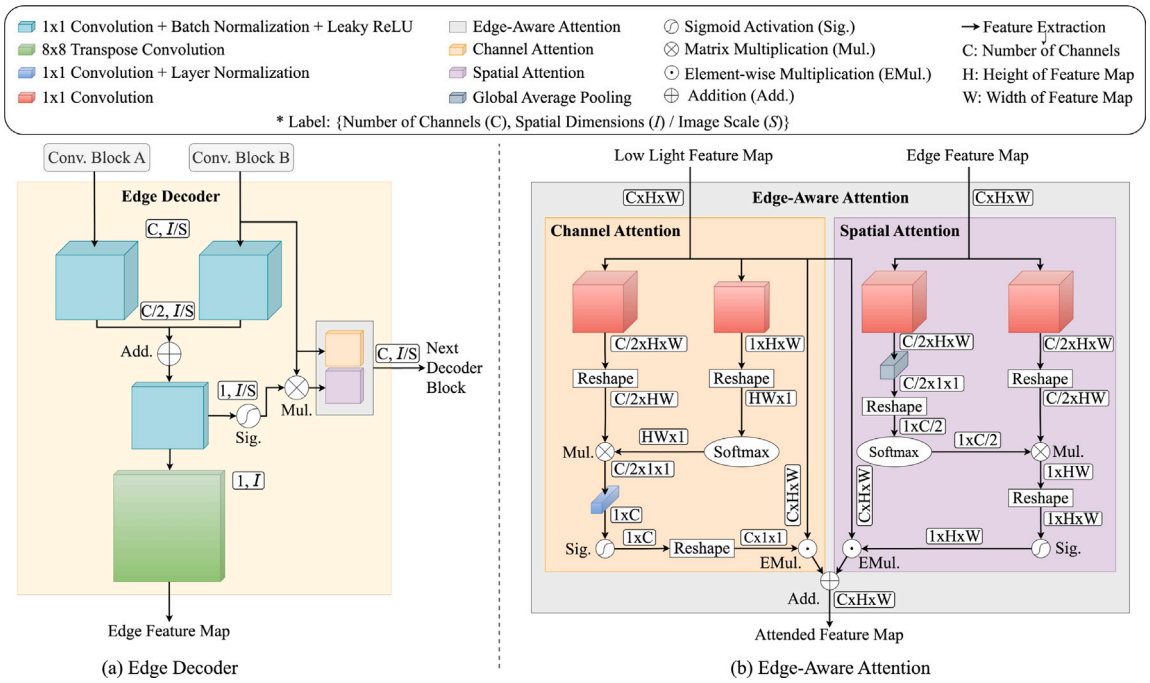


Fig. 5. A close-up of the Edge-Aware Attention (Edge-Att) module is provided as follows: (a) A visual representation of our edge decoder, where A and B represent the outputs from the corresponding convolution blocks (Conv. Block) in Fig. 4, and S denotes the scaling of the image. (b) An illustration of the proposed Edge-Aware Attention module. *I* stands for the spatial image size and S is the image scaling factor.

enforce our model to focus on scene text regions, we define the text detection loss, $\mathcal{L}_{text}$ as:

$$\mathcal{L}_{text} = \|R(x') - R(y)\|_1, \tag{5}$$

where $R(x')$ and $R(y)$ denote the region score heatmaps of the enhanced and ground truth images, respectively.

Fourthly, the edge reconstruction decoder in our model is designed to extract edges better, which are essential for text pixels. Fig. 5(a) shows an overview of the edge decoder. The loss at pixel i of detected edge, e_i, with respect to the ground truth edge, g_i is defined as:

$$l(e_i) = \begin{cases} \alpha \cdot log(1 - P(e_i)), & \text{if } g_i = 0 \\ \beta \cdot log P(e_i), & \text{if } g_i = 1 \end{cases} \tag{6}$$

where

$$\alpha = \lambda \cdot \frac{|Y^+|}{|Y^+| + |Y^-|}, \quad \beta = \frac{|Y^-|}{|Y^+| + |Y^-|}, \tag{7}$$

Y^+ and Y^- denote the positive and negative sample sets, respectively. λ is set to 1.1 to balance both types of samples. The ground truth edge is generated using a Canny edge detector [34], and P(e_i) is the sigmoid function. Then, the overall edge reconstruction loss can be formulated as:

$$\mathcal{L}_{edge} = \sum_{i=1}^{|I|} \sum_{j=1}^{J} l(e_i^j) + l(e_i'), \tag{8}$$

where $l(e_i^j)$ is the predicted edge at pixel i and level j. $J = 3$ is the number of side edge outputs in our model. e_i' is the final predicted edge map from the concatenation of side outputs. $|I|$ is the number of pixels in a cropped image during training.

Finally, the total joint loss function, $\mathcal{L}_{total_en}$ of our proposed model is:

$$\mathcal{L}_{total_en} = \omega_{recons}\mathcal{L}_{recons} + \omega_{text}\mathcal{L}_{text} + \omega_{SSIM_{MS}}\mathcal{L}_{SSIM_{MS}} + \omega_{edge}\mathcal{L}_{edge}, \tag{9}$$

where ω_{recons}, ω_{text}, $\omega_{SSIM_{MS}}$, and ω_{edge} are the weights to address the importance of each loss term during training.

3.4. Edge-aware attention

Polarized Self-Attention (PSA) [35] is one of the first works to propose an attention mechanism catered to high-quality pixel-wise regression tasks. However, we found that the original PSA module that only considers a single source of feature map for both channel and spatial attention is ineffective for extremely low-light image enhancement. Under low light conditions, the details of content such as the edges of the texts are barely discernible which is less effective in guiding the network to attend to spatial details. Therefore, we designed our Edge-Aware Attention (Edge-Att) module to take in feature maps from two encoders and process them differently, i.e., the feature maps of extremely low-light images from the image encoder are attended channel-wise, whereas the spatial attention submodule attends to feature maps from the edge encoder. By doing so, we can ensure that Edge-Att can attend to rich images and edge features simultaneously. The proposed attention module is illustrated in Fig. 5(b), and further mathematical details can be found in the supplementary material.

3.5. Text-aware copy-paste augmentation

This work aims to enhance extremely low-light images to improve text detection and recognition. However, the dataset's limited number of text instances could hinder the model's ability. Although Copy-Paste Augmentation [36] can increase the number of text instances, overlapping texts introduced by random placement might confuse CRAFT in text detection loss since CRAFT is not trained to detect such texts. In the commonly used scene text datasets such as ICDAR15 [28], overlapping texts are marked as "do not care" regions which are excluded from models' training and evaluation. Thus, to adhere to ICDAR's standard and to address overlapping text issues, we propose a novel approach called Text-Aware Copy-Paste Augmentation (Text-CP). Text-CP considers each text box's location and size by leveraging uniform and Gaussian distributions derived from the dataset. For a training image t of width w_t and height h_t to be augmented, we initialize a set of labeled text boxes in the training set as C, which is:

$$C = \left\{(u_1, v_1, w_1, h_1), \ldots, (u_{|C|}, v_{|C|}, w_{|C|}, h_{|C|})\right\}, \tag{10}$$

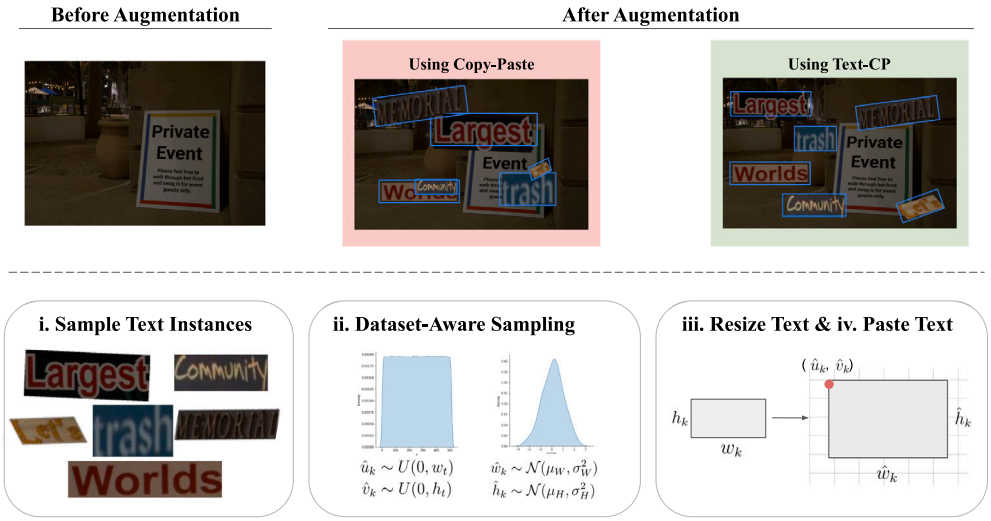


Fig. 6. Illustration of the Text-Aware Copy-Paste (Text-CP) data augmentation. Compared with the original Copy-Paste, our method generates images with non-overlapping text instances that allow the detection of texts outside their usual context.

where each tuple represents the top left position of a text located at u_k and v_k with width, w_k, and height, h_k with k representing the index of the current text's box in the set. We then sample a target number of text instances, n_{target}, from the set of C to form C_t, defined as the set of text boxes to be pasted on that training image, t. The next step is to crop and paste the sampled texts without overlapping. For each $c_k \in C_t$, we adopt two uniform distributions to model the position of the texts, $\hat{u}_k \sim U(0, w_t)$ and $\hat{v}_k \sim U(0, h_t)$. As for w_k and h_k, they are sampled from Gaussian distributions as $\hat{w}_k \sim \mathcal{N}(\mu_W, \sigma_W^2)$ and $\hat{h}_k \sim \mathcal{N}(\mu_H, \sigma_H^2)$ where μ and σ^2 are the estimated means and variances of width W and height H from all the labeled texts in the training set. We illustrate the overall data augmentation process of Text-CP and its augmented results in Fig. 6. The pseudocode of Text-CP is detailed in the supplementary material.

4. Extremely low-light image synthesis

4.1. Problem formulations

To the best of our knowledge, the research community has not extensively explored extremely low-light image synthesis, mainly due to the limited availability of datasets designed explicitly for extremely low-light scene text. While extremely low-light dataset, SID, and low-light dataset, LOL, exist, they are not primarily collected with scene text in mind. This scarcity of dedicated datasets for extremely low-light scene text poses challenges for evaluating the performance of existing image enhancement methods in terms of image quality and scene text metrics. In order to address this issue, we define the extremely low-light image synthesis problem as follows:

$$\hat{x} = LS(y; \theta_s), \tag{11}$$

where given a long-exposure image y, a low-light image synthesis neural network, $LS(y; \theta_s)$ parameterized by θ_s, will synthesize a set of images $\hat{x}$, such that $B(LS(y; \theta_s)) \simeq B(x)$. We want the synthesized extremely low-light images to be as realistic as possible to genuine low-light images, x.

Therefore, we introduce a Supervised-DCE model focusing on synthesizing a set of realistic extremely low-light images, enabling existing image enhancement techniques to leverage publicly available scene text datasets. Consequently, existing low-light image enhancement methods can benefit from training with synthetic data to the extent that they can perform better on the downstream scene text detection task, as detailed in Section 6.5.

4.2. Network design

Zero-DCE [7] was originally proposed to perform image enhancement through curve estimation. However, its network can only adjust brightness slightly since the per-pixel trainable curve parameter, α, in the quadratic curve limits the pixel variation. The advantage of performing intensity adjustment in terms of the quadratic curve is that the pixel range can be better constrained. In this work, we propose a Supervised-DCE model that learns to provide reconstructable extremely low-light images with paired short- and long-exposure images. The framework of our image synthesis network, Supervised-DCE, can be seen in Fig. 7. Our goal is to push most values closer to zero in the context of synthesizing extremely low-light images. Accordingly, we propose a reformulation of the DCE model as follows:

$$\hat{x} = -(H(y) + U(y))y^2 + (1 + H(y))y, \tag{12}$$

where y is the input (i.e., long-exposure image); $\hat{x}$ is the synthesized low-light image; $H(y)$ and $U(y)$ are the output of Tanh and ReLU branches, respectively. By introducing the second $U(y)$ branch, we eliminate the need for iteratively applying the model to produce the desired output, and drastic intensity adjustment can be done with only a single iteration.

In the original Zero-DCE model, image enhancement is learned by setting the exposure value to 0.6 in the exposure control loss. However, manually setting an exposure value to synthesize extremely low-light images is too heuristic and inefficient. In our proposed synthesis framework, the overall learning is done by training SID long-exposure images to be similar to their short-exposure counterparts. In contrast to the recently proposed LOL Synthetic [37], which is not dark enough and was not specifically designed for text-aware extremely low-light image enhancement. Our approach aims to replicate the natural loss of text legibility that occurs in such environments. Most importantly, these images are synthesized to simulate the text degradation observed in genuine extremely low-light conditions, making text reconstruction challenging. Then, the trained model can be used to transform scene text datasets in the public domain to boost the performance of extremely low-light image enhancement in terms of text detection.

4.3. Objectives

During extremely low-light image synthesis, we expect the output to maintain spatial consistency while reducing the overall proximity loss:

$$\mathcal{L}_{prox} = \|\hat{x} - x\|_1 + \mathcal{L}_{entropy}(\hat{x}, x) + \mathcal{L}_{smoothness}(\hat{x}, x), \tag{13}$$

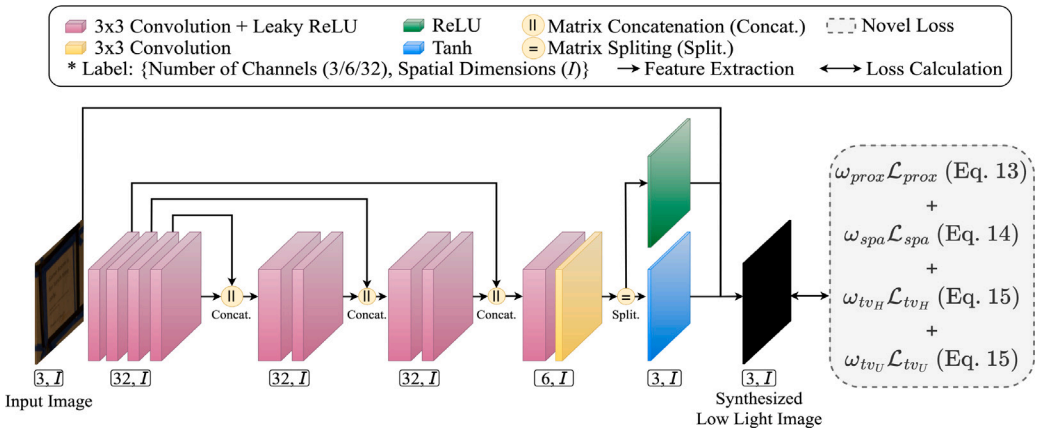


Fig. 7. Illustration of the proposed Supervised-DCE model for extremely low-light image synthesis. I stands for the spatial image size. Please refer to the respective equations for further details.

where $\hat{x}$ is the synthesized extremely low-light image given the long-exposure image y, and x, is the genuine low-light image, i.e., ground truth for $\hat{x}$. Entropy loss, $\mathcal{L}_{entropy}$, and smoothness loss, $\mathcal{L}_{smoothness}$ [38], are also used to encourage the differences to be both sparse and local. With the introduction of $\mathcal{L}_{prox}$, we removed the color constancy loss of the original Zero-DCE model since color constancy can be enforced through the supervised loss.

The spatial consistency loss, $\mathcal{L}_{spa}$ encourages spatial coherence of the synthesized image by preserving the difference of neighboring regions between the input image and its synthesized low-light version:

$$\mathcal{L}_{spa} = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \sum_{j \in \omega(i)} (|\hat{X}_i - \hat{X}_j| - \alpha_s \log_{10}(9|Y_i - Y_j| + 1))^2, \tag{14}$$

where $\mathcal{M}$ is the number of local regions, and $\omega(i)$ is four neighboring regions (top, down, left, right) centered at region i. $\hat{X}$ and Y are the averaged intensity values of local regions of the synthesized images and the long-exposure images, respectively. We introduced logarithm operation and α_s parameter to reduce the large spatial difference of Y where α_s is set to 0.05. We set the local region size to 4×4, following the original setting of Zero-DCE.

Besides spatial consistency, we also expect the monotonicity relation between neighboring pixels to be preserved. To achieve this, we reused the illumination smoothness loss:

$$\mathcal{L}_{tv_Z} = \sum_{\forall c \in \xi} (|\nabla_x Z^c| + |\nabla_y Z^c|)^2, \xi = \{R, G, B\}, \tag{15}$$

where ∇_x and ∇_y are gradient operations on the x-axis and y-axis, respectively. Illumination smoothness loss, $\mathcal{L}_{tv_Z}$, is applied on both $H(y)$ and $U(y)$, i.e., the curve parameter maps of the two branches, respectively, by substituting Z with H and U, resulting in $\mathcal{L}_{tv_H}$ and $\mathcal{L}_{tv_U}$.

In summary, the overall learning objective, $\mathcal{L}_{total_syn}$ to train our extremely low-light image synthesis network is defined as:

$$\mathcal{L}_{total_syn} = \omega_{prox}\mathcal{L}_{prox} + \omega_{spa}\mathcal{L}_{spa} + \omega_{tv_H}\mathcal{L}_{tv_H} + \omega_{tv_U}\mathcal{L}_{tv_U}. \tag{16}$$

5. New low-light text datasets

In this work, we annotated all text instances in the extremely low-light dataset, SID [1], and the ordinary low-light dataset, LOL [2]. SID has two subsets: SID-Sony, captured by Sony α7S II, and SID-Fuji, captured by Fujifilm X-T2. For this work, we included 878/810 short-exposure images and 211/176 long-exposure images at a resolution of 4240×2832/6000×4000 from SID-Sony and SID-Fuji, respectively. The short-exposure time is 1/30, 1/25, and 1/10, while the corresponding reference (long-exposure) images were captured with 100 to 300 times longer exposure, i.e., 10 to 30 s. In our experiments, we converted short- and long-exposure SID images to RGB format. The LOL dataset provides low/normal-light image pairs taken from real scenes by controlling exposure time and ISO. There are 485 and 15 images at a resolution of 600×400 in the training and test sets, respectively. We closely annotated text instances in the SID and LOL datasets following the common IC15 standard. We show some samples in Fig. 8. The newly annotated datasets are named SID-Sony-Text, SID-Fuji-Text, and LOL-Text to differentiate them from their low-light counterparts.

IC15 dataset was introduced in the ICDAR 2015 Robust Reading Competition for incidental scene text detection and recognition. It contains 1500 scene text images at a resolution of 1280×720. In this study, IC15 is primarily used to synthesize extremely low-light scene text images. Detailed statistics of the text annotations for SID-Sony-Text, SID-Fuji-Text, LOL-Text, and IC15 are shown in Table 2, where we included the statistics for long-exposure images only for the sake of brevity. In this table, we also report relevant statistics of the mean and standard deviation of labeled texts' width and height to be used by the proposed Text-Aware Copy-Paste augmentation. The text annotations for SID-Sony-Text, SID-Fuji-Text, and LOL-Text datasets will be released at https://github.com/chunchet-ng/Text-in-the-Dark.

Moreover, we synthesized extremely low-light images based on IC15 by using U-Net and our proposed Supervised-DCE model, respectively. To study the difference between these two variations of image synthesis methods, we generated a total of four sets of images by using the aforementioned two models trained on SID-Sony and SID-Fuji, individually. Naming convention of such synthetic datasets follows the format of "{Syn-IC15}-{Sony/Fuji}-{v1/v2}". "{Sony/Fuji}" is an indication of which dataset the image synthesis model is trained on, while "{v1/v2}" differentiates the image synthesis models where v1 is U-Net and v2 is our proposed Supervised-DCE model. For instance, the synthetic images generated by a U-Net trained on SID-Sony and SID-Fuji, are named Syn-IC15-Sony-v1 and Syn-IC15-Fuji-v1. And, synthetic images generated by our proposed Supervised-DCE model are denoted as Syn-IC15-Sony-v2 and Syn-IC15-Fuji-v2.

6. Experimental results

6.1. Experiment setup

Datasets and Metrics. All low-light image enhancement methods are trained and tested on the datasets detailed in Section 5. They are then evaluated in terms of intensity metrics (PSNR, SSIM), perceptual similarity (LPIPS), and text detection (H-Mean). For the SID-Sony-Text, SID-Fuji-Text, and LOL-Text datasets, which are annotated with text bounding boxes only, we used well-known and commonly used scene

Table 2
Statistics reported based on long-exposure images for all datasets. GT Img. stands for ground truth image count, where Leg. and Illeg. stand for legible and illegible text count, respectively.

Dataset	Training set							Testing set		
	GT Img.	Leg.	Illeg.	μ_W	μ_H	σ_W	σ_H	GT Img.	Leg.	Illeg.
SID-Sony-Text	161	5937	2128	79.270	34.122	123.635	50.920	50	611	359
SID-Fuji-Text	135	6213	4534	128.579	57.787	183.199	68.466	41	1018	1083
LOL-Text	485	613	1423	23.017	14.011	21.105	17.542	15	28	45
IC15	1000	4468	7418	78.410	29.991	55.947	24.183	500	2077	3153



(a) SID-Sony-Text (b) SID-Fuji-Text (c) LOL-Text

Fig. 8. Green boxes represent legible texts, and blue boxes represent illegible texts.

text detectors (CRAFT [26] and PAN [27]) to analyze the enhanced images. For IC15, which provides both text detection and text recognition labels, we conducted a two-stage text spotting experiment using the aforementioned text detectors (CRAFT, PAN) and two robust text recognizers (TRBA [31] and ASTER [30]) on the synthesized IC15 images after enhancement. The metric for text spotting is case-insensitive word accuracy.

Implementation Details. We trained our image enhancement model for 4000 epochs using the Adam optimizer [39] with a batch size of 2. The initial learning rate is set to $1e^{-4}$ and decreased to $1e^{-5}$ after 2000 epochs. At each training iteration, we randomly cropped a 512×512 patch with at least one labeled text box inside and applied random flipping and image transpose as data augmentation strategies. The weightings of each loss term, i.e., ω_{recons}, ω_{text}, $\omega_{SSIM_{MS}}$, and ω_{edge}, were empirically set to 0.2125, 0.425, 0.15, and 0.2125 respectively, following the work of ELIE_STR [15]. For other image enhancement methods, we re-trained them on all datasets using the best set of hyperparameters specified in their respective code repositories or papers.

As for the Supervised-DCE model, we used a batch size of 8 and trained for 200 epochs using the Adam optimizer with default parameters and a fixed learning rate of $1e^{-4}$. It was trained on 256×256 image patches with loss weightings of ω_{prox}, ω_{spa}, ω_{tv_A} and ω_{tv_B}, set to 1, 20, 10, and 10 respectively.

6.2. Results on SID-sony-text and SID-fuji-text datasets

Our model's performance is demonstrated in Table 3, achieving the highest H-Mean scores on all datasets with CRAFT and PAN. Following [15], we illustrate the CRAFT text detection results on SID-Sony-Text in Fig. 9. Qualitative results of existing methods on SID-Fuji-Text are presented in the supplementary material. The effectiveness of our model in enhancing extremely low-light images to a level where text can be accurately detected is readily apparent. In Fig. 9, only the images enhanced by our proposed model yield accurate text detection results. On the other hand, existing methods generally produce noisier images, resulting in inferior text detection results. While GAN-enhanced images tend to be less noisy, the text regions are blurry, making text detection challenging. Moreover, our model achieves the highest PSNR and SSIM scores on both SID-Sony-Text and SID-Fuji-Text datasets, showing that our enhanced images are the closest to the image quality of ground truth images. In short, better text detection is achieved on our enhanced images through the improvement of overall image quality and preservation of fine details within text regions.

6.3. Results on LOL-text dataset

To demonstrate the effectiveness of our model in enhancing low-light images with varying levels of darkness, we conducted experiments on the widely used LOL dataset, which is relatively brighter than the SID dataset, as depicted in Table 1. Interestingly, we found that our enhanced images achieved the best detection results on LOL-Text among existing methods, as shown in Table 3. Surprisingly, despite the lower resolution (600×400) of the images in LOL, our method's enhanced images with sharper and crisper low-level details surpassed the ground truth images' H-Mean scores. Qualitative results on the LOL-Text dataset are illustrated in the supplementary material. Although certain methods yielded output images with acceptable image quality (i.e., bright images without color shift), their text detection results were inferior to ours. It is worth noting that the LPIPS metric, designed to assess perceptual image quality, has shown less improvement compared to other metrics in the LOL dataset. This can be attributed to several factors inherent to the nature of low-light image enhancement. For example, images in LOL are at least 15 times brighter than those in SID (in terms of L* in the CIELAB color space), leading to limited room for improvement. Additionally, since our method prioritizes text clarity and legibility, our model might favor sharper text features at the cost of compromising other image details. Furthermore, our superior results on the LOL-Text dataset emphasize our method's ability to generalize well on both ordinary and extremely low-light images, effectively enhancing a broader range of low-light images while making the text clearly visible.

6.4. Effectiveness of the proposed supervised-DCE model

The goal of image synthesis in our work is to translate images captured in well-illuminated scenarios to extremely low light. In this work, we choose the commonly used IC15 scene text dataset as our main synthesis target. The synthesized dataset then serves as additional data to train better scene text-aware image enhancement models, which are studied in Section 6.5.

Intuitively, realistic synthesized images should possess similar characteristics to genuine extremely low-light images. To verify the effectiveness of our synthesis model, we compared our proposed Supervised-DCE model (v2) with the U-Net proposed in SID [1] (v1). Specifically, we trained the synthesizing models on the training set and synthesized the images based on the corresponding test set. Then, we calculated the PSNR and SSIM of the synthesized images by comparing them with the

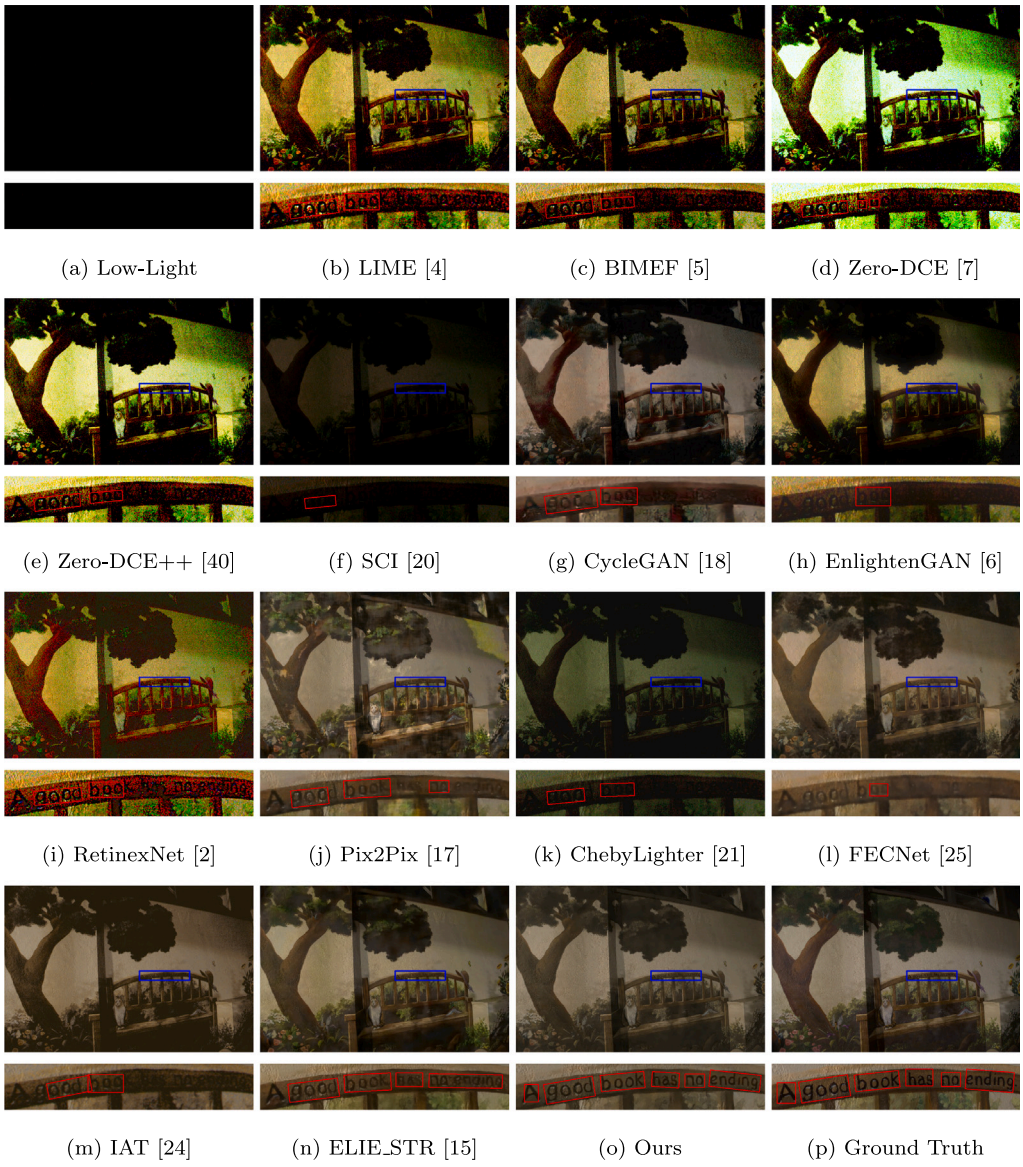


Fig. 9. Comparison with state-of-the-art methods on the SID-Sony-Text dataset is shown in the following manner: for each column, the first row displays enhanced images marked with blue boxes as regions of interest. The second row displays zoomed-in regions of enhanced images overlaid with red text detection boxes from CRAFT [26]. Column 9(a) displays the low-light image. Columns 9(b) to 9(o) show image enhancement results from all related methods. The last cell displays ground truth images.

genuine ones along with the average perceptual lightness in CIELAB color space. The comparison was made on two SID datasets, SID-Sony and SID-Fuji.

In Table 4, we show that v2's PSNR and SSIM are higher than v1's, indicating higher similarity between our synthesized and genuine images. Our new method (v2) also exhibits closer Avg. L* values and H-Mean scores to the genuine images than v1, indicating darker and more accurate deterioration of fine text details. In addition, qualitative results for the proposed Supervised-DCE model and results of synthetic IC15 datasets including Syn-IC15-Sony-v1, Syn-IC15-Sony-v2, Syn-IC15-Fuji-v1, and Syn-IC15-Fuji-v2 are presented in the supplementary material for comprehensive analyses.

6.5. Results on training with mixed datasets

We trained top-performing models from Section 6.2 using a mixture of genuine (SID) and synthetic low-light (IC15) datasets to test whether extremely low-light image enhancement can benefit from synthesized images. The trained models were evaluated on their respective genuine low-light datasets. Results in Table 5 showed a significant increase in H-Mean, and we found that both versions (v1 and v2) can fill the gap caused by the scarcity of genuine low-light images. This justifies the creation of a synthetic IC15 dataset for such a purpose. Furthermore, v2-images, i.e., extremely low-light images synthesized by our proposed Supervised-DCE, further pushed the limit of H-mean scores on genuine extremely low-light images, and our enhancement model benefited the most because it could learn more from text instances and reconstruct necessary details to represent texts. Despite our method's success, a noticeable gap exists between our results and the ground truth, emphasizing the need for further research and development to achieve even more accurate and reliable scene text extraction in low-light conditions.

6.6. Ablation study of proposed modules

To understand the effect of each component of our model, we conducted several ablation experiments by either adding or removing them one at a time. Results are presented in Table 6. The baseline was a plain U-Net without any proposed modules. We initiated the ablation

Table 3
Quantitative results of PSNR, SSIM, LPIPS, and text detection H-Mean for low-light image enhancement methods on SID-Sony-Text, SID-Fuji-Text, and LOL-Text datasets. Please note that TRAD, ZSL, UL, and SL stand for traditional methods, zero-shot learning, unsupervised learning, and supervised learning respectively. Scores in bold are the best of all.

	Type	Method	Image Quality			H-Mean	
			PSNR ↑	SSIM ↑	LPIPS ↓	CRAFT ↑	PAN ↑
SID-Sony-Text		Input	–	–	–	0.057	0.026
	TRAD	LIME [4]	13.870	0.135	0.873	0.127	0.057
		BIMEF [5]	12.870	0.110	0.808	0.136	0.079
	ZSL	Zero-DCE [7]	10.495	0.080	0.999	0.196	0.157
		Zero-DCE++ [40]	12.368	0.076	0.982	0.218	0.162
		SCI [20]	11.814	0.100	1.000	0.201	0.151
	UL	CycleGAN [18]	15.340	0.453	0.832	0.090	0.053
		EnlightenGAN [6]	14.590	0.426	0.793	0.146	0.075
	SL	RetinexNet [2]	15.490	0.368	0.785	0.115	0.040
		Pix2Pix [17]	21.070	0.662	0.837	0.266	0.190
		ChebyLighter [21]	15.418	0.381	0.787	0.260	0.184
		FECNet [25]	22.575	0.648	0.788	0.245	0.188
		IAT [24]	19.234	0.562	0.778	0.244	0.176
		ELIE_STR [15]	25.507	0.716	0.789	0.324	0.266
		Ours	**25.596**	**0.751**	**0.751**	**0.368**	**0.298**
		GT	–	–	–	0.842	0.661
SID-Fuji-Text		Input	–	–	–	0.048	0.005
	ZSL	Zero-DCE [7]	8.992	0.035	1.228	0.249	0.061
		Zero-DCE++ [40]	11.539	0.047	1.066	0.262	0.077
		SCI [20]	10.301	0.056	1.130	0.300	0.073
	UL	CycleGAN [18]	17.832	0.565	0.735	0.277	0.191
		EnlightenGAN [6]	18.834	0.572	0.822	0.310	0.277
	SL	Pix2Pix [17]	19.601	0.599	0.803	0.353	0.296
		ChebyLighter [21]	20.313	0.616	0.791	0.412	0.318
		FECNet [25]	18.863	0.365	0.829	0.382	0.185
		IAT [24]	19.647	0.537	0.844	0.445	0.277
		ELIE_STR [15]	19.816	0.614	0.801	0.426	0.333
		Ours	**21.880**	**0.649**	**0.788**	**0.487**	**0.356**
		GT	–	–	–	0.775	0.697
LOL-Text		Input	–	–	–	0.333	0.133
	ZSL	Zero-DCE [7]	14.928	0.587	0.328	0.421	0.229
		Zero-DCE++ [40]	15.829	0.537	0.408	0.389	0.242
		SCI [20]	14.835	0.549	0.335	0.421	0.171
	UL	CycleGAN [18]	19.826	0.734	0.288	0.250	0.133
		EnlightenGAN [6]	15.800	0.654	0.300	0.343	0.125
	SL	Pix2Pix [17]	20.581	0.771	0.247	0.353	0.129
		ChebyLighter [21]	19.820	0.769	0.199	0.353	0.176
		FECNet [25]	20.432	0.787	0.231	0.378	0.229
		IAT [24]	20.437	0.772	0.234	0.421	0.188
		ELIE_STR [15]	19.782	0.824	0.167	0.462	0.235
		Ours	**21.330**	**0.828**	**0.163**	**0.474**	**0.294**
		GT	–	–	–	0.439	0.205

Table 4
The difference between genuine extremely low-light dataset, SID, and synthetic extremely low-light images generated using U-Net (v1) and Supervised-DCE (v2). Please note that synthetic images' PSNR and SSIM values are based on a comparison against genuine low-light images in the test set instead of pure black images calculated in Table 1. Additionally, we can notice that v2-images are more realistic and darker, similar to genuine extremely low-light images due to their higher values of PSNR and SSIM, along with closer Avg. L*.

Dataset	PSNR	SSIM	Avg. L*	CRAFT	PAN
Syn-SID-Sony-v1	41.095	0.809	0.176	0.294	0.083
Syn-SID-Sony-v2	45.442	0.942	0.003	0.135	0.014
Genuine SID-Sony	–	–	0.008	0.057	0.026
Syn-SID-Fuji-v1	39.187	0.784	0.172	0.402	0.042
Syn-SID-Fuji-v2	41.881	0.863	0.002	0.093	0.002
Genuine SID-Fuji	–	–	0.004	0.048	0.005

study by adding Text-CP data augmentation, which improved CRAFT H-Mean from 0.283 to 0.304, indicating that involving more text instances during training is relevant to text-aware image enhancement for models to learn text representation. Moreover, scores increased steadily by gradually stacking the baseline with more modules. For instance, with the help of the dual encoder structure and Edge-Att module in our proposed framework, CRAFT H-Mean increased from 0.304 to 0.342. This shows that they can extract image features better and attend to edges that shape texts in enhanced images. The edge reconstruction loss calculated based on predictions from the edge decoder helped strengthen the learning of edge features and empowered encoders in our model. Interestingly, we found that removing one of the two most representative modules (i.e., dual encoder or Edge-Att module) led to significant differences in H-Mean because these two modules' designs allow them to extract and attend to significant image features independently. We concluded the experiment by showing that combining all proposed modules led to the highest scores, as each module played an integral role in our final network. Further analysis of Edge-Att

Table 5
Text detection H-Mean on genuine extremely low-light datasets when trained on a combination of genuine and synthetic datasets. Scores in bold are the best of all.

Type	Method	SID-Sony-Text + Syn-IC15-Sony-v1		SID-Sony-Text + Syn-IC15-Sony-v2		SID-Fuji-Text + Syn-IC15-Fuji-v1		SID-Fuji-Text + Syn-IC15-Fuji-v2	
		CRAFT ↑	PAN ↑	CRAFT ↑	PAN ↑	CRAFT ↑	PAN ↑	CRAFT ↑	PAN ↑
	Input	0.057	0.026	0.057	0.026	0.048	0.005	0.048	0.005
ZSL	Zero-DCE++ [40]	0.230	0.159	0.242	0.153	0.274	0.080	0.281	0.076
	SCI [20]	0.240	0.154	0.243	0.160	0.307	0.076	0.313	0.084
UL	CycleGAN [18]	0.180	0.071	0.219	0.143	0.297	0.284	0.310	0.277
	EnlightenGAN [6]	0.205	0.146	0.237	0.163	0.329	0.246	0.342	0.282
SL	ELIE_STR [15]	0.348	0.278	0.361	0.296	0.444	0.359	0.466	0.375
	Ours	**0.383**	**0.311**	**0.395**	**0.319**	**0.515**	**0.392**	**0.549**	**0.416**
	GT	0.842	0.661	0.842	0.661	0.775	0.697	0.775	0.697

Table 6
Ablation study of proposed modules in terms of PSNR, SSIM, LPIPS, and text detection H-Mean on the SID-Sony-Text dataset. Scores in bold are the best of all.

Proposed modules				Image quality			H-Mean	
Text-CP	Dual encoder	Edge-Att	Edge decoder	PSNR ↑	SSIM ↑	LPIPS ↓	CRAFT ↑	PAN ↑
–	–	–	–	21.847	0.698	0.783	0.283	0.205
✓	–	–	–	21.263	0.658	0.771	0.304	0.252
✓	✓	–	–	20.597	0.655	0.780	0.335	0.261
✓	✓	✓	–	21.440	0.669	0.776	0.342	0.256
✓	✓	–	✓	21.588	0.674	0.779	0.353	0.285
✓	–	✓	✓	23.074	0.712	0.783	0.350	0.281
–	✓	✓	✓	24.192	0.738	0.784	0.356	0.292
✓	✓	✓	✓	**25.596**	**0.751**	**0.751**	**0.368**	**0.298**

and Text-CP are included in the supplementary material to study their effectiveness as compared to the original versions.

7. Conclusion

This paper proposes a novel extremely low-light image enhancement model focusing on text regions. By leveraging the proposed Text-Aware Copy-Paste augmentation and a dual-encoder–decoder architecture equipped with Edge-Aware attention modules, our method substantially improves image quality in extremely low-light scenarios. The introduction of text detection and edge reconstruction losses enables our model to produce images with a focus on scene text regions that support downstream scene text tasks more effectively. Moreover, our proposed Supervised-DCE model facilitates better synthetic extremely low-light text image generation, enhancing the overall performance of our framework on public scene text datasets like IC15. Through comprehensive evaluations against state-of-the-art methods on datasets such as SID-Sony-Text, SID-Fuji-Text, LOL-Text, and synthetic IC15, our results consistently indicate the effectiveness of our model in enhancing extremely low-light images and extracting text. This work not only advances research in extremely low-light text image enhancement but also establishes a robust foundation for future research in the linkage of visual and textual image enhancement.

CRediT authorship contribution statement

Che-Tsung Lin: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Chun Chet Ng:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhi Qin Tan:** Visualization, Methodology, Formal analysis. **Wan Jun Nah:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis. **Xinyu Wang:** Writing – review & editing, Supervision. **Jie Long Kew:** Project administration. **Pohao Hsu:** Project administration. **Shang Hong Lai:** Resources. **Chee Seng Chan:** Writing – review & editing, Supervision. **Christopher Zach:** Supervision.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.image.2024.117222.

Data availability

Data will be made available on request.

References

[1] C. Chen, Q. Chen, J. Xu, V. Koltun, Learning to see in the dark, in: CVPR, 2018.
[2] C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement, 2018, arXiv preprint arXiv:1808.04560.
[3] D.J. Jobson, Z.-u. Rahman, G.A. Woodell, Properties and performance of a center/surround retinex, IEEE Trans. Image Process. 6 (3) (1997) 451–462.
[4] X. Guo, Y. Li, H. Ling, LIME: Low-light image enhancement via illumination map estimation, IEEE Trans. Image Process. 26 (2) (2016) 982–993.
[5] Z. Ying, G. Li, W. Gao, A bio-inspired multi-exposure fusion framework for low-light image enhancement, 2017, arXiv preprint arXiv:1711.00591.
[6] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, Z. Wang, EnlightenGAN: Deep light enhancement without paired supervision, 2019, arXiv arXiv:1906.06972.
[7] C.G. Guo, C. Li, J. Guo, C.C. Loy, J. Hou, S. Kwong, R. Cong, Zero-reference deep curve estimation for low-light image enhancement, in: CVPR, 2020.
[8] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: CVPR, 2018.
[9] J. Hu, L. Shen, S. Albanie, G. Sun, A. Vedaldi, Gather-excite: Exploiting feature context in convolutional neural networks, in: NIPS, 2018.
[10] C. Chen, B. Li, An interpretable channelwise attention mechanism based on asymmetric and skewed Gaussian distribution, Pattern Recognit. 139 (2023) 109467.
[11] L. Ju, J. Kittler, M.A. Rana, W. Yang, Z. Feng, Keep an eye on faces: Robust face detection with heatmap-assisted spatial attention and scale-aware layer attention, Pattern Recognit. 140 (2023) 109553.

[12] X. Hou, M. Liu, S. Zhang, P. Wei, B. Chen, CANet: Contextual information and spatial attention based network for detecting small defects in manufacturing industry, Pattern Recognit. 140 (2023) 109558.
[13] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: ECCV, 2018.
[14] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: CVPR, 2019.
[15] P.-H. Hsu, C.-T. Lin, C.C. Ng, J.L. Kew, M.Y. Tan, S.-H. Lai, C.S. Chan, C. Zach, Extremely low-light image enhancement with scene text restoration, in: ICPR, 2022.
[16] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014.
[17] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: CVPR, 2017.
[18] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networkss, in: ICCV, 2017.
[19] M. Zhu, P. Pan, W. Chen, Y. Yang, EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network, in: AAAI, 2020.
[20] L. Ma, T. Ma, R. Liu, X. Fan, Z. Luo, Toward fast, flexible, and robust low-light image enhancement, in: CVPR, 2022, pp. 5637–5646.
[21] J. Pan, D. Zhai, Y. Bai, J. Jiang, D. Zhao, X. Liu, ChebyLighter: Optimal curve estimation for low-light image enhancement, in: ACM MM, 2022.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, NIPS (2017).
[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.
[24] Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, Y. Qiao, T. Harada, You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction, in: BMVC, 2022.
[25] J. Huang, Y. Liu, F. Zhao, K. Yan, J. Zhang, Y. Huang, M. Zhou, Z. Xiong, Deep Fourier-based exposure correction network with spatial-frequency interaction, in: ECCV, 2022.
[26] Y. Baek, B. Lee, D. Han, S. Yun, H. Lee, Character region awareness for text detection, in: CVPR, 2019.
[27] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, C. Shen, Efficient and accurate arbitrary-shaped text detection with pixel aggregation network, in: ICCV, 2019.
[28] D. Karatzas, L.G.I. Bigorda, A. Nicolaou, S. Ghosh, A.D. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, E. Valveny, ICDAR 2015 competition on robust reading, in: ICDAR, 2015.
[29] C.-K. Ch'ng, C.S. Chan, C.-L. Liu, Total-text: toward orientation robustness in scene text detection, Int. J. Document Anal. Recognit. (IJDAR) 23 (1) (2020) 31–52.
[30] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, ASTER: An attentional scene text recognizer with flexible rectification, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2019) 2035–2048.
[31] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S.J. Oh, H. Lee, What is wrong with scene text recognition model comparisons? Dataset and model analysis, in: ICCV, 2019.
[32] Y. Liu, M.-M. Cheng, X. Hu, J. Bian, L. Zhang, X. Bai, J. Tang, Richer convolutional features for edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2019) 1939–1946.
[33] Z. Wang, E. Simoncelli, A. Bovik, Multiscale structural similarity for image quality assessment, in: ACSSC, 2003.
[34] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. (6) (1986) 679–698.
[35] H. Liu, F. Liu, X. Fan, D. Huang, Polarized self-attention: Towards high-quality pixel-wise regression, 2021, arXiv preprint arXiv:2107.00782.
[36] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E.D. Cubuk, Q.V. Le, B. Zoph, Simple copy-paste is a strong data augmentation method for instance segmentation, in: CVPR, 2021.
[37] W. Yang, W. Wang, H. Huang, S. Wang, J. Liu, Sparse gradient regularized deep retinex network for robust low-light image enhancement, IEEE Trans. Image Process. 30 (2021) 2072–2086.
[38] P. Samangouei, A. Saeedi, L. Nakagawa, N. Silberman, ExplainGAN: Model explanation via decision boundary crossing transformations, in: ECCV, 2018.
[39] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2015, arXiv preprint arXiv:1412.6980.
[40] C. Li, C. Guo, C.C. Loy, Learning to enhance low-light image via zero-reference deep curve estimation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (8) (2021) 4225–4238.