# Iterative Prompt Learning for Unsupervised Backlit Image Enhancement

Zhexin Liang Chongyi Li Shangchen Zhou Ruicheng Feng Chen Change Loy
S-Lab, Nanyang Technological University
{zliang008, chongyi.li, s200094, ruicheng002, ccloy}@ntu.edu.sg
https://zhexinliang.github.io/CLIP_LIT_page/

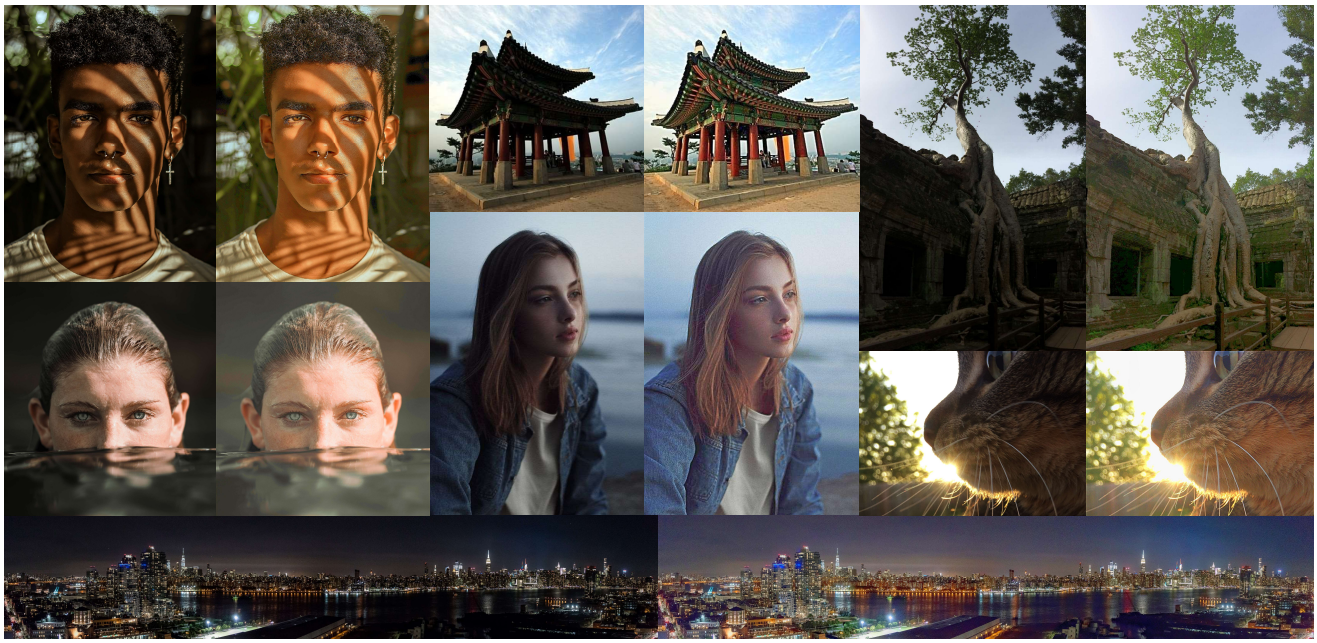


Figure 1: The proposed method trained using only hundreds of images yields favorable results on unseen backlit images captured in various scenarios, including the human face, natural landscape, animal, architecture, and night scene.

## Abstract

*We propose a novel unsupervised backlit image enhancement method, abbreviated as CLIP-LIT, by exploring the potential of Contrastive Language-Image Pre-Training (CLIP) for pixel-level image enhancement. We show that the open-world CLIP prior not only aids in distinguishing between backlit and well-lit images, but also in perceiving heterogeneous regions with different luminance, facilitating the optimization of the enhancement network. Unlike high-level and image manipulation tasks, directly applying CLIP to enhancement tasks is non-trivial, owing to the difficulty in finding accurate prompts. To solve this issue, we devise a prompt learning framework that first learns an initial prompt pair by constraining the text-image similarity between the prompt (negative/positive sample) and the corresponding image (backlit image/well-lit image) in the CLIP latent space. Then, we train the enhancement network based on the text-image similarity between the enhanced result and the initial prompt pair. To further improve the accuracy of the initial prompt pair, we iteratively fine-tune the prompt learning framework to reduce the distribution gaps between the backlit images, enhanced results, and well-lit images via rank learning, boosting the enhancement performance. Our method alternates between updating the prompt learning framework and enhancement network until visually pleasing results are achieved. Extensive experiments demonstrate that our method outperforms state-of-the-art methods in terms of visual quality and generalization ability, without requiring any paired data. Code for our method will be made available.*

## 1. Introduction

Backlit images are captured when the primary light source is behind some objects. The images often suffer from highly imbalanced illuminance distribution, which affects the visual quality or accuracy of subsequent perception algorithms.

Correcting backlit images manually is a laborious task given the intricate challenge of preserving the well-lit re-

gions while enhancing underexposed regions. One could apply an automatic light enhancement approach but will find that existing approaches could not cope well with backlit images [14]. For instance, many existing supervised light enhancement methods [26, 27, 33] cannot precisely perceive the bright and dark areas, and thus process these regions using the same pipeline, causing over-enhancement in well-lit areas or under-enhancement in low-light areas. Unsupervised light enhancement methods, on the other hand, either rely on ideal assumptions such as average luminance and a gray world model [9, 15] or directly learn the distribution of reference images via adversarial training [10]. The robustness and generalization capability of these methods are limited. As for conventional exposure correction methods [1, 29], they struggle in coping with real-world backlit images due to the diverse backlit scenes and luminance intensities. The problem cannot be well resolved by collecting backlit images that consist of ground truth images that are retouched by photographers [19], since these images can never match the true distribution of real backlit photos.

In this work, we propose an unsupervised method for backlit image enhancement. Different from previous unsupervised methods that learn curves or functions based on some physical hypothesis or learn the distribution of well-lit images via adversarial training that relies on task-specific data, we explore the rich visual-language prior encapsulated in a Contrastive Language-Image Pre-Training (CLIP) [21] model for pixel-level image enhancement. While CLIP can serve as an indicator to distinguish well-lit and backlit images to a certain extent, using it directly for training a backlit image enhancement network is still non-trivial. For example, for a well-lit image (Fig. 2 top left), replacing similar concepts "normal light" with "well-lit" brings a huge increase in CLIP score. In the opposite case (Fig. 2 top right), "normal light" becomes the correct prompt. This indicates the optimal prompts could vary on a case-by-case basis due to the complex illuminations in the scene. In addition, it is barely possible to find accurate 'word' prompts to describe the precise luminance conditions. Prompt engineering is labor-intensive and time-consuming to annotate each image in the dataset. Moreover, the CLIP embedding is often interfered by high-level semantic information in an image. Thus, it is unlikely to achieve optimal performance with fixed prompts or prompt engineering.

To overcome the problems, we present a new pipeline to tailor the CLIP model for our task. It consists of the following components: 1) *Prompt Initialization.* We first encode the backlit and well-lit images along with a learnable prompt pair (positive and negative samples) into the latent space using the pre-trained CLIP's image and text encoder. By narrowing the distance between the images and text in the latent space, we obtain an initial prompt pair that can effectively distinguish between backlit and well-lit images. 2) *CLIP-aware Enhancement Training.* With the initialized prompt, we train an enhancement network using the text-image similarity constraints in the CLIP embedding space. 3) *Prompt Refinement.* We introduce a prompt fine-tuning mechanism, in which we update the prompt by further distinguishing the distribution gaps among backlit images, enhanced results, and well-lit images via rank learning. We iteratively update the enhancement network and prompt learning framework until achieving visually pleasing results.

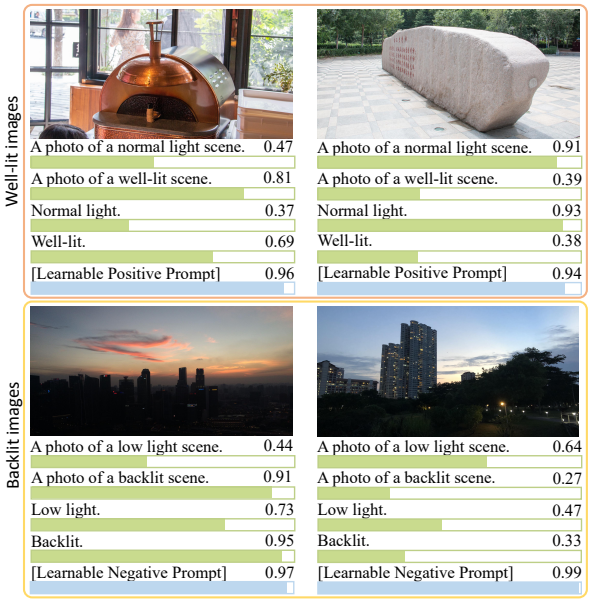



Figure 2: Motivation. CLIP scores of proper prompts demonstrate alignment with human annotations (*e.g.*, well-lit images), suggesting that CLIP can serve as an indicator to differentiate between well-lit and backlit images. However, the best wordings could differ on a case-by-case basis due to complex illumination. In contrast, the learnable positive/negative prompts are more robust and consistent with the labels.

Our method stands apart from existing backlit image enhancement techniques as we leverage the intrinsic perceptual capability of CLIP. Rather than solely utilizing CLIP as a loss objective [7, 35], we incorporate prompt refinement as an essential component of the optimization process to further enhance performance. Our approach surpasses state-of-the-art methods in both qualitative and quantitative metrics, without requiring any paired training data. We demonstrate the generalization capability and robustness of our method through the preview of our results shown in Fig. 1, and we compare our results with existing methods in Fig. 3.

## 2. Related Work

**Backlit Image Enhancement.** Several approaches have been proposed in the literature. Li and Wu [17] employ a region segmentation technique in combination with a learning-based restoration network to separately process the back-


Figure 3: Visual comparison between our method and the state-of-the-art light enhancement methods, including exposure correction method (Afifi et al. [1]), backlit enhancement method (ExCNet [31]), low-light image enhancement methods (SCI [20], Zero-DCE [9], SNR-aware [28], EnlightenGAN [10]). Our method effectively enhances the backlit image without introducing artifacts and over-/under-enhancement.

lit and front-lit areas of an image. Buades et al. [3] and Wang et al. [24] use fusion-based techniques to combine pre-processed images. Zhang et al. [31] learn a parametric "S-curve" using a small image-specific network, ExCNet, to correct ill-exposed images. More recently, Lv et al. [19] have created the first paired backlit dataset, named BAID, in which the ground truth images are edited by photographers so that the quality is still sub-optimal, shown in Fig. 8.

**Light Enhancement.** Backlit image enhancement is closely related to low-light image enhancement and exposure correction. Traditional methods for low-light image enhancement [8, 16] typically employ the Retinex model to restore normal-light images. With the availability of paired data [26] and simulated data [38], several supervised methods [27, 28] have been proposed, which design various networks for low-light image enhancement. Despite their success, supervised methods suffer from limited generalization capability. Consequently, unsupervised methods[9, 15, 18, 20] have garnered increasing attention. Since low-light image enhancement cannot effectively process both underexposed and overexposed regions, exposure correction methods [1, 5, 29] have also been proposed. For example, Afifi et al. [1] propose an exposure correction network based on Laplacian pyramid decomposition and reconstruction.

**CLIP and Prompting in Vision.** CLIP [21] has shown remarkable performance in zero-shot classification, thanks to the knowledge learned from large-scale image-text data. Its generalizability has been shown in high-level tasks[30, 12, 35]. A recent study[23] shows that the rich visual language prior encapsulated in CLIP can be used for assessing both the quality and abstract perception of images in a zero-shot manner. These studies inspire our work to exploit CLIP for backlit image enhancement. Prompt learning, as the core of vision-and-language models, is a recent emerging research direction. CoOp [37] introduces prompt learning into the adaptation of vision-language models for downstream vision tasks. CoCoOp [36] further improves the generalizability by allowing a prompt to be conditioned on each input instance rather than fixed once learned. Existing prompt learning methods focus solely on obtaining better prompts for high-level vision tasks. In contrast, our approach uses prompt learning to extract more accurate low-level image representations, such as color, exposure, and saturation, while ignoring high-level semantic information in CLIP.

## 3. Methodology

**Overview.** Our proposed approach consists of two stages, as illustrated in Fig. 4. In the first stage, we learn an initial prompt pair (negative/positive prompts referring to backlit/well-lit images) by constraining the text-image similarity between the prompt and the corresponding image in the CLIP embedding space. With the initial prompt pair, we use a frozen CLIP model to compute the text-image similarity between the prompts and the enhanced results to train the initial enhancement network. In the second stage, we refine the learnable prompts by utilizing backlit images, enhanced results, and well-lit images through rank learning. The refined prompts can be used to fine-tune the enhancement network for further performance improvement. We alternate the prompt refinement and fine-tuning of the enhancement network until we achieve visually pleasing results. It should be noted that the CLIP model remains fixed throughout the learning process, and our method does not introduce any additional computational burden apart from prompt initialization and refinement. We provide further details on the key components of our approach below.

### 3.1. Initial Prompts and Enhancement Training

The first stage of our approach involves the initialization of negative and positive (learnable) prompts to roughly characterize backlit and well-lit images, as well as the training of the initial enhancement network.

**Prompt Initialization.** The process of prompt initialization is depicted in Fig. 5(a). Given a backlit image $I_b \in \mathbb{R}^{H\times W\times 3}$ and a well-lit image $I_w \in \mathbb{R}^{H\times W\times 3}$ (as reference), we randomly initialize a positive prompt $T_p \in \mathbb{R}^{N\times 512}$ and a negative prompt $T_n \in \mathbb{R}^{N\times 512}$. $N$ represents the number of embedded tokens in each prompt. Then, we feed the backlit and well-lit images to the image encoder $\Phi_{image}$ of the pre-trained CLIP to obtain their latent code. Meanwhile, we also extract the latent code of the positive and negative prompts by feeding them to the text

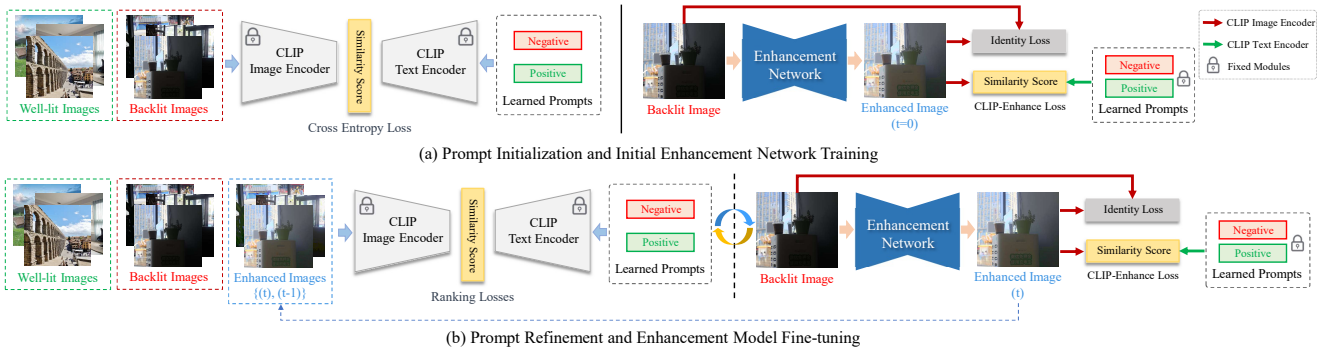



Figure 4: Our proposed method involves two main stages. (a) The first stage constitutes prompt initialization and the initial training of an enhancement network. (b) The second stage involves prompt refinement and enhancement model fine-tuning. The two components here are updated in an alternating manner. The prompt refinement in the second stage aims at learning accurate prompts that distinguish among backlit images, enhanced results, and well-lit images. By employing these learned prompts, the enhancement network produces enhanced results that are similar to well-lit images and distinct from backlit images in the CLIP embedding space, ultimately leading to visually pleasing results.

encoder $\Phi_{text}$. Based on the text-image similarity in the CLIP latent space, we use the binary cross entropy loss of classifying the backlit and well-lit images to learn the initial prompt pair:

$$\mathcal{L}_{initial} = -(y * \log(\hat{y}) + (1-y) * \log(1-\hat{y})), \quad (1)$$

$$\hat{y} = \frac{e^{cos(\Phi_{image}(I), \Phi_{text}(T_p))}}{\sum_{i \in \{n,p\}} e^{cos(\Phi_{image}(I), \Phi_{text}(T_i))}}, \quad (2)$$

where $I \in \{I_b, I_w\}$ and $y$ is the label of the current image, 0 is for negative sample $I_b$ and 1 is for positive sample $I_w$.

**Training the Initial Enhancement Network.** Given the initial prompts obtained from the first stage, we can train an enhancement network with a CLIP-aware loss. As a baseline model, we use a simple Unet [22] to enhance the backlit images, though more complex networks can also be employed. Inspired by the Retinex model [13], which is widely used for light enhancement, the enhancement network estimates the illumination map $I_i \in \mathbb{R}^{H \times W \times 1}$ and then produces the final result via $I_t = I_b / I_i$. To train the enhancement network, we employ CLIP-Enhance loss $\mathcal{L}_{clip}$ and identity loss $\mathcal{L}_{identity}$.

The CLIP-Enhance loss measures the similarity between the enhanced result and the prompts in the CLIP space:

$$\mathcal{L}_{clip} = \frac{e^{cos(\Phi_{image}(I_t), \Phi_{text}(T_n))}}{\sum_{i \in \{n,p\}} e^{cos(\Phi_{image}(I_t), \Phi_{text}(T_i))}}. \quad (3)$$

The identity loss encourages the enhanced result to be similar to the backlit image in terms of content and structure:

$$\mathcal{L}_{identity} = \sum_{l=0}^{4} \alpha_l \cdot ||\Phi^l_{image}(I_b) - \Phi^l_{image}(I_t)||_2, \quad (4)$$

where $\alpha_l$ is the weight of the $l^{th}$ layer of the image encoder in the ResNet101 CLIP model. The final loss for training the enhancement network is the combination of the two losses:

$$\mathcal{L}_{enhance} = \mathcal{L}_{clip} + w \cdot \mathcal{L}_{identity}, \quad (5)$$

where $w$ is the weight to balance the magnitude of different loss terms. We divide the training schedule into two parts. First, we use the identity loss to implement self-reconstruction as it encourages the enhanced result to be similar to the backlit image in the pixel space. Then, we use both the identity loss and the CLIP-Enhance loss to train the network. For the identity loss, we set $\alpha_{l=0,1,...,4}$ in Eq. (4) to 1.0 during the self-reconstruction stage. During training of the backlit enhancement network, we set $\alpha_{l=0,1,2,3} = 1.0$ and $\alpha_{l=4} = 0.5$. This is because we found that the features of the last layer are more related to the color of the images, which is what we want to adjust.

### 3.2. Prompt Refinement and Enhancement Tuning

In the second stage, we iteratively perform prompt refinement and enhancement network tuning. The prompt refinement and the tuning of the enhancement network are conducted in an alternating manner. The goal is to improve the accuracy of learned prompts for distinguishing backlit images, enhanced results, and well-lit images, as well as perceiving heterogeneous regions with different luminance.

**Prompt Refinement.** We observed that in some cases, using only the initial prompts obtained from the backlit and well-lit images is insufficient for enhancing the color and illuminance. This is because the initial prompts may fail to capture the fine-grained differences among the backlit images, enhanced results, and well-lit images. To address this, we propose a further refinement of the learnable positive and negative prompts. Given the result $I_t \in \mathbb{R}^{H \times W \times 3}$ enhanced by the current enhancement network, we use a margin ranking loss to update the prompts. The process of prompt refinement is illustrated in Fig. 5(b).

Formally, we define the negative similarity score between

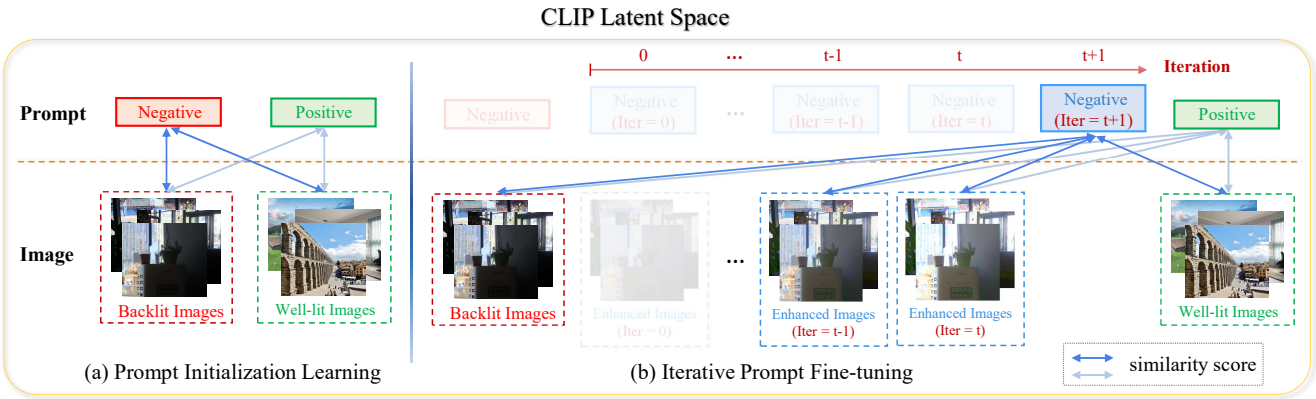



Figure 5: Illustration of the prompt learning framework. **1.** Prompt Initialization. A cross-entropy loss constrains the learned prompts, maximizing the distance between the representation of negative and positive samples in the CLIP latent space. **2.** Adding the enhanced results from the current round $I_t$ into the ranking process (i.e., ranking loss) to make the enhanced results $I_t$ closer to the representation of the well-lit images $I_w$ in the CLIP latent space and far from the representation of the input image $I_b$. **3.** Adding the images inferred from the previous round $I_{t-1}$ to constrain the result of updated enhancement network. $I_t$ being closer to the representation of positive samples than the previous round $I_{t-1}$, and far from the representation of negative ones $I_b$ in CLIP latent space.

the prompt pair and an image as:

$$S(I) = \frac{e^{cos(\Phi_{image}(I), \Phi_{text}(T_n))}}{\sum_{i \in \{n,p\}} e^{cos(\Phi_{image}(I), \Phi_{text}(T_i))}}, \tag{6}$$

Then, the margin ranking loss can be expressed as:

$$\begin{aligned} \mathcal{L}_{prompt1} = & \max(0, S(I_w) - S(I_b) + m_0) \\ & + \max(0, S(I_t) - S(I_b) + m_0) \\ & + \max(0, S(I_w) - S(I_t) + m_1), \end{aligned} \tag{7}$$

where $m_0 \in [0, 1]$ represents the margin between the score of well-lit/enhanced results and the backlit images in the CLIP embedding space. We set $m_0$ to 0.9 to extend the distance between backlit images and well-lit images as much as possible. Meanwhile, $m_1$ represents the margin between the score of the enhanced results and the well-lit images in the CLIP embedding space. We set $m_1$ to 0.2 to ensure that the enhanced results are similar to well-lit images. These hyper-parameters are chosen empirically based on the performance of the algorithm on the validation set.

To ensure that the iterative learning can improve the performance in each iterative round, we preserve the previous enhanced results $I_{t-1}$ obtained by the previous enhancement network in the ranking process. We add the two groups of enhanced results, $I_{t-1}$ and $I_t$, into the constraints, enabling the newly learned prompts to focus more on the light and color distribution of images, rather than high-level content in the image (see Fig. 10). The loss function is modified as:

$$\begin{aligned} \mathcal{L}_{prompt2} = & \max(0, S(I_w) - S(I_b) + m_0) \\ & + \max(0, S(I_{t-1}) - S(I_b) + m_0) \\ & + \max(0, S(I_w) - S(I_t) + m_1) \\ & + \max(0, S(I_t) - S(I_{t-1}) + m_2), \end{aligned} \tag{8}$$

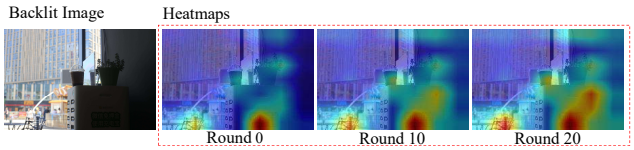


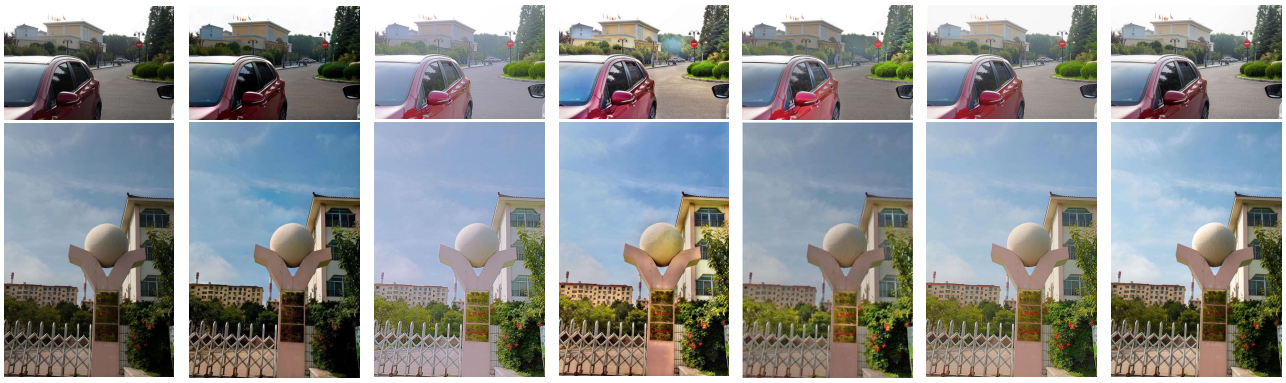

Figure 6: Attention map changes with iterative learning.


Figure 7: Enhanced results of different iteration rounds.

where $m_2$ represents the margin between the newly enhanced results and previously enhanced results. We set $m_2 = m_1$ as the margins $m_1$ and $m_2$ have the same target, keeping the two image groups similar.

**Tuning the Enhancement Network.** The tuning of the enhancement network follows the same process in Sec. 3.1 except we use the refined prompts to compute for the CLIP-Enhance loss $\mathcal{L}_{clip}$ and generate the enhanced training data from the updated network to further refine the prompt.

**Discussion.** To show the effectiveness of iterative learning, following Chefer *et al.* [6], we visualize the attention maps in the CLIP model for the interaction between the learned negative prompt and an input image at different alternate rounds. The heatmap, as shown in Fig. 6, represents the relevance between each pixel in the image and the learned prompt. The heatmap shows that during iterations, the learned negative prompt becomes increasingly relevant to the regions with unpleasant lighting and color. We also show the enhanced results with different iterative rounds in Fig. 7. At the intermediate round, the color in some enhanced regions of the outputs is over-saturated. After enough iterations, the over-saturation is corrected while the dark regions are closer to the well-lit state compared with the previous outputs. The



Input | Afifi et al. [1] | Zero-DCE [9] | EnlightenGAN [10] | ExCNet [31] | CLIP-LIT (Ours) | Ground Truth

Figure 8: Visual comparison on the backlit images sampled from the Backlit300 test dataset.

observation here suggests the capability of our approach in perceiving heterogeneous regions with different luminance. We will provide the quantitative comparison in Sec. 4.3.

## 4. Experiments

**Dataset.** For training, we randomly select 380 backlit images from BAID [19] training dataset as input images and select 384 well-lit images from DIV2K [2] dataset as reference images. We test our methods on the BAID test dataset, which includes 368 backlit images taken in diverse light scenarios and scenes. To examine the generalization ability, we collected a new evaluation dataset, named Backlit300, which consists of 305 backlit images from Internet, Pexels, and Flickr. The data will be made available.

**Training.** We implement our method with PyTorch on a single NVIDIA GTX 3090Ti GPU. We use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The number $N$ of embedded tokens in each learnable prompt is set to 16. We set the total training iterations to $50K$, within which, the number of self-reconstruction iterations is set to $1K$, the number of prompt pair initialization learning iterations is set to $10K$. We set the learning rate for the prompt initialization/refinement and enhancement network training to $5 \cdot 10^{-6}$ and $2 \cdot 10^{-5}$. The batch size for prompt initialization/refinement and enhancement network training is set to 8 and 16. During training, we resize the input images to $512 \times 512$ and use flip, zoom, and rotate as augmentations.

**Inference.** The sizes of some input images from the BAID and Backlit300 test datasets are large, and some methods are unable to handle such high-resolution images directly. To ensure a fair comparison, we resize all test images to have a long side of 2048 pixels if their size is larger than $2048 \times 2048$.

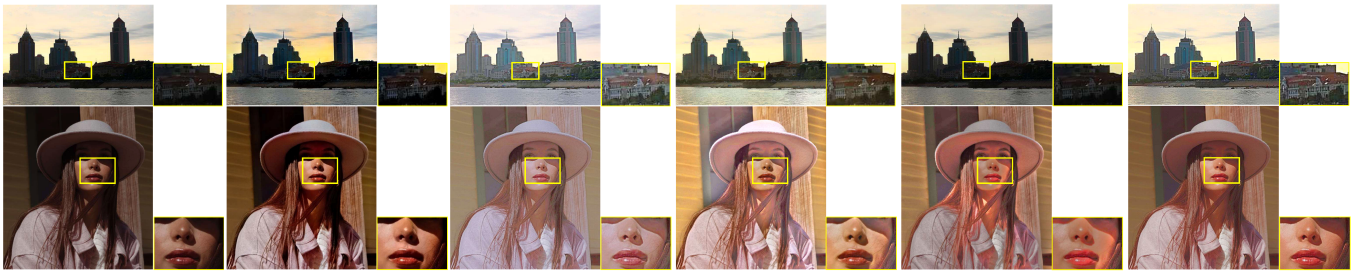
**Compared Methods.** As there are very few publicly available deep learning-based methods for backlit image enhancement, we compare our approach with representative methods that solve related tasks, including low-light image enhancement methods such as Zero-DCE [9], Zero-DCE++ [15], SCI [20], URetinex-Net [27], SNR-Aware [28], Zhao et al. citeINN, and EnlightenGAN [10]; exposure correction methods such as Afifi et al. [1]; and backlit enhancement methods such as ExCNet [31]. Some methods provide different models trained on different datasets. We compare our method with all released models of different methods to ensure a fair comparison. To further validation, we also provide retrained supervised methods' results in supplementary material. For unsupervised methods, we retrained them on the same training data as our method to ensure that they are evaluated under the same conditions.

### 4.1. Results

**Visual Comparison.** We present visual comparisons of some typical samples from the BAID test dataset in Fig. 8. Due to space limitations, we only show the results of the best-performing methods. The complete comparisons of all methods can be found in the supplementary material. Our method consistently produces visually pleasing results with improved color and luminance without over- or under-exposure. Moreover, our method excels in handling challenging backlit regions, restoring clear texture details and satisfactory luminance without introducing any artifacts, while other methods may either fail to address such regions or produce unsatisfactory results with visible artifacts.

We also evaluate our method on the Backlit300 test dataset, and present the comparison results in Fig. 9. We can see that compared to EnlightenGAN [10] and ExCNet [31], our method produces results without visible distortion artifacts. Our method is also more effective in enhancing dark regions, unlike Afifi et al. [1] and EXCNet [31]. Moreover, our results exhibit better color contrast and input-output consistency in well-lit regions. We emphasize that our method achieves these results without the need for paired data, which is not available in many real-world scenarios.

**Quantitative Comparison.** We use three full-reference image quality evaluation (IQA) metrics, i.e., PSNR, SSIM [25], and LPIPS [32] (Alex version) and one non-reference IQA metric MUSIQ [11] to evaluate the quantitative results. As current non-reference IQA metrics only evaluate the overall image quality, they may not accurately measure the results of



Input | Afifi et al. [1] | Zero-DCE [9] | EnlightenGAN [10] | ExCNet [31] | CLIP-LIT (Ours)

Figure 9: Visual comparison on the backlit images sampled from the Backlit300 test dataset.

Table 1: Quantitative comparison on the BAID test dataset. The best and second performance are marked in red and blue.

| Type | Methods | PSNR↑ | SSIM↑ | LPIPS↓ | MUSIQ↑ |
|---|---|---|---|---|---|
| | Input | 16.641 | 0.768 | 0.197 | 52.115 |
| Supervised | Afifi et al. [1] | 15.904 | 0.745 | 0.227 | 52.863 |
| | Zhao et al.-MIT5K [34] | 18.228 | 0.774 | 0.189 | 51.457 |
| | Zhao et al.-LOL [34] | 17.947 | 0.822 | 0.272 | 49.334 |
| | URetinex-Net [27] | 18.925 | 0.865 | 0.211 | **54.402** |
| | SNR-Aware-LOLv1 [28] | 15.472 | 0.747 | 0.408 | 26.425 |
| | SNR-Aware-LOLv2real [28] | 17.307 | 0.754 | 0.398 | 26.438 |
| | SNR-Aware-LOLv2synthetic [28] | 17.364 | 0.752 | 0.403 | 23.960 |
| Unsupervised | Zero-DCE [9] | **19.740** | 0.871 | 0.183 | 51.804 |
| | Zero-DCE++ [15] | 19.658 | **0.883** | 0.182 | 48.573 |
| | RUAS-LOL [18] | 9.920 | 0.656 | 0.523 | 37.207 |
| | RUAS-MIT5K [18] | 13.312 | 0.758 | 0.347 | 45.008 |
| | RUAS-DarkFace [18] | 9.696 | 0.642 | 0.517 | 39.655 |
| | SCI-easy [20] | 17.819 | 0.840 | 0.210 | 51.984 |
| | SCI-medium [20] | 12.766 | 0.762 | 0.347 | 44.176 |
| | SCI-diffucult [20] | 16.993 | 0.837 | 0.232 | 52.369 |
| | EnlightenGAN [10] | 17.550 | 0.864 | 0.196 | 48.417 |
| | ExCNet [31] | 19.437 | 0.865 | **0.168** | 52.576 |
| Unsupervised (retrained) | Zero-DCE [9] | 18.553 | 0.863 | 0.194 | 49.436 |
| | Zero-DCE++ [15] | 16.018 | 0.832 | 0.240 | 47.253 |
| | RUAS [18] | 12.922 | 0.743 | 0.362 | 45.056 |
| | SCI [20] | 16.639 | 0.768 | 0.197 | 52.265 |
| | EnlightenGAN [10] | 17.957 | 0.849 | 0.182 | 53.871 |
| | CLIP-LIT (Ours) | **21.579** | **0.883** | **0.159** | **55.682** |

Table 2: Quantitative comparison on the Backlit300 test dataset.

| Methods | MUSIQ↑ |
|---|---|
| Input | 51.900 |
| Afifi et al. [1] | 51.930 |
| Zhao et al.-MIT5K [34] | 50.354 |
| Zhao et al.-LOL [34] | 48.334 |
| URetinex-Net [27] | 51.551 |
| SNR-Aware-LOLv1 [28] | 29.915 |
| SNR-Aware-LOLv2real [28] | 30.903 |
| SNR-Aware-LOLv2synthetic [28] | 29.149 |
| Zero-DCE [9] | 51.250 |
| Zero-DCE++ [15] | 48.216 |
| RUAS-LOL [18] | 40.329 |
| RUAS-MIT5K [18] | 44.523 |
| RUAS-DarkFace [18] | 48.216 |
| SCI-easy [20] | 50.642 |
| SCI-medium [20] | 48.216 |
| SCI-diffucult [20] | 49.428 |
| EnlightenGAN [10] | 48.308 |
| ExCNet [31] | 50.278 |
| Zero-DCE [9] | 48.491 |
| Zero-DCE++ [15] | 46.000 |
| RUAS [18] | 45.251 |
| SCI [20] | **51.960** |
| EnlightenGAN [10] | 48.261 |
| CLIP-LIT (Ours) | **52.921** |

backlit image enhancement. Hence, we primarily rely on the state-of-the-art MUSIQ metric to evaluate the performance.

The quantitative comparison on the BAID test dataset is presented in Tab. 1. Our method outperforms all state-of-the-art methods in terms of the full-reference IQA metrics, indicating that the results generated by our method preserve the content and structure of the original images well and are close to the reference images retouched by photographers. Our method also performs the best in the non-reference MUSIQ metric when compared to other methods, demonstrating the good image quality of our results. We also report the quantitative comparison on the Backlit300 test dataset in Tab. 2, where our method continues to achieve the best performance, further indicating the effectiveness of our method.

### 4.2. User Study

We conducted a user study to more comprehensively evaluate the visual quality of enhanced results obtained by different methods. In addition to our results, we chose the results obtained from the top-3 PSNR methods: Zero-DCE [9], ExCNet [31], and URetinex [27], as well as EnlightenGAN [10] since it is a related work to our method. We randomly selected 20 images from the Backlit300 test partition as the evaluation set. For each image, we provided the input backlit image, the corresponding images enhanced by our method and a baseline. A total of 40 participants were invited to select their preferred image.

The statistics of the user study are summarized in Fig. 11. The vote distribution shows that our results are the most favored by participants, with obvious advantages over the other methods. For each image, over 60% of the participants voted for our result, indicating that our method generates more visually pleasing results when compared to other methods.

### 4.3. Ablation Studies

**Effectiveness of Iterative Learning.** In addition to the observation provided in Sec. 3.2, to further validate the effectiveness of iterative learning, we provide the quantitative comparison in Tab. 3. As presented, fine-tuning the prompts using the loss functions Eq. (7) and Eq. (8) improve the


Figure 10: Comparison of similarity scores: learned positive prompt vs. the same images with gradually improved luminance and color conditions, indicating the learned prompts' sensitivity to light and color distribution rather than high-level content.

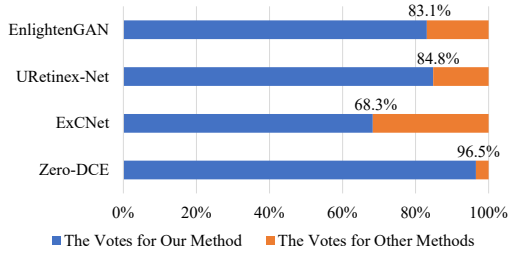



Figure 11: User Study. Voting statistics of different methods versus our method.

Table 3: Quantitative comparisons of the iterative learning on the BAID test set.

| Method | PSNR↑ | SSIM↑ |
|---|---|---|
| fixed prompts (backlit/well-lit) | 14.748 | 0.823 |
| w/o ranking losses (w/o Eqs. (7) and (8)) | 20.884 | 0.865 |
| w/o $t-1$ outputs (w/o Eq. (8)) | 20.146 | 0.866 |
| Ours | **21.579** | **0.883** |

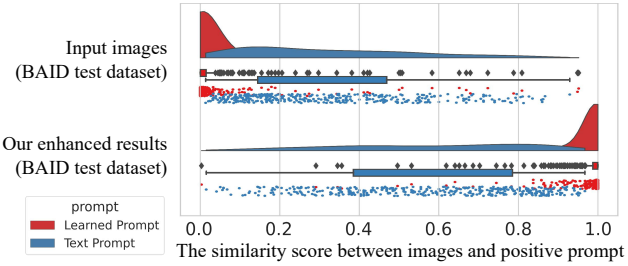



Figure 12: The distribution of similarity scores between the learned positive prompt and images across the BAID test dataset. The upper part is the kernel density estimation curve, and the lower part is the point and box plot. Our learned prompts have more precise presentation of the images' luminance quality than text prompts, such as backlit/well-lit.

enhancement performance.

**Necessity of Prompt Refinement.** Compared to the selected words or sentences, our learned prompts can better distinguish between backlit and well-lit images (see Fig. 12). Results in Tab. 3 also indicate that the enhancement model trained under the constraint of our refined prompts performs better than a fixed prompts.

**Impact of Training Data.** To investigate the impact of the reference data (the well-lit images) on our method, we conducted an experiment where we retrained our method on another dataset containing 1000 images selected from DIV2K[2] and MTI5K[4], which has more diverse well-lit images. The results, as shown in Fig. 13 and Tab. 4, indicate

Table 4: Comparison of training data impact. The quantitative comparisons are conducted on BAID test dataset.

| Reference images | PSNR↑ | SSIM↑ | LPIPS↓ | MUSIQ↑ |
|---|---|---|---|---|
| MIT5K [4]+DIV2K [2] | 21.413 | **0.881** | 0.162 | **56.494** |
| DIV2K [2] | **21.579** | 0.883 | **0.159** | 55.682 |


Figure 13: Visual comparisons of our method trained using different reference images.

Table 5: Comparison between CLIP-Enhance loss and adversarial loss on the BAID test dataset.

| loss | PSNR↑ | SSIM↑ | LPIPS↓ | MUSIQ↑ |
|---|---|---|---|---|
| Adversarial loss | 17.407 | 0.785 | 0.194 | 52.416 |
| CLIP-Enhance loss | **21.579** | **0.883** | **0.159** | **55.682** |

that the two sets of results obtained by our method using different training data are similar, and the quantitative scores only have slight differences. Such results demonstrate that the number and variety of well-lit images used as training data have little impact on the performance of our method.

**Advantage of CLIP-Enhance Loss over the Adversarial Loss.** To show the advantage of our CLIP-Enhance loss over the adversarial loss, we trained our enhancement network on the same unpaired training data using adversarial loss. We used the same discriminator as EnlightenGAN [10]. The results in Tab. 5 indicate that our CLIP-Enhance loss achieves better enhancement performance than adversarial loss. This may be due to the fact that the CLIP prior is more sensitive to color and luminance distribution, enabling it to differentiate between images with varied lighting conditions (see Fig. 10) and perceive unbalanced luminance regions (see Fig. 6). Visual comparison is provided in supplementary material.

## 5. Conclusion

We have introduced a novel approach for training a deep network to enhance backlit images using only a few hundred unpaired data. Our method exploits the rich priors embedded in a CLIP model and leverages an iterative prompt learning strategy to generate more precise prompts that can better characterize both backlit and well-lit images. Notably, our study is the first attempt to use CLIP for low-level restoration tasks, and we anticipate that this methodology will find additional applications in the future.

## References


[1] Mahmoud Afifi, Konstantinos G. Derpanis, Bjorn Ommer, and Michael S. Brown. Learning multi-scale photo exposure correction. In *CVPR*, 2021. 2, 3, 6, 7, 13

[2] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, July 2017. 6, 8

[3] Antoni Buades, Jose-Luis Lisani, Ana Belen Petro, and Catalina Sbert. Backlit images enhancement using global tone mappings and image fusion. *IET Image Processing*, 14(2):211–219, 2020. 3

[4] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Fredo Durand. Learning phtographic global tonal adjustment with a dataset of input/output image pairs. In *CVPR*, 2011. 8

[5] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure image. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018. 3

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021. 5

[7] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021. 2

[8] Xueyang Fu, Delu Zeng, Yue Huang, Xiaoping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, 2016. 3

[9] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, 2020. 2, 3, 6, 7, 13

[10] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 2, 3, 6, 7, 8, 13

[11] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, 2021. 6

[12] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*. 3

[13] Edwin H. Land and John J. McCann. Lightness and retinex theory. *Josa*, 1971. 4

[14] Chongyi Li, Chunle Guo, Linhao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[15] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 6, 7, 13

[16] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018. 3

[17] Zhenhao Li and Xiaolin Wu. Learning-based restoration of backlit images. *IEEE Transactions on Image Processing*, 2018. 2

[18] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 3, 7, 13

[19] Xiaoqian Lv, Shengping Zhang, Qinglin Liu, Haozhe Xie, Bineng Zhong, and Huiyu Zhou. Backlitnet: A dataset and network for backlit image enhancement. *Computer Vision and Image Understanding*, 2022. 2, 3, 6

[20] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 3, 6, 7, 13

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4, 11

[23] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 3

[24] Qiuhong Wang, Xueyang Fu, Xiao-Ping Zhang, and Xinghao Ding. A fusion-based method for single backlit image enhancement. In *ICIP*, 2016. 3

[25] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[26] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 2, 3

[27] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *CVPR*, 2022. 2, 3, 6, 7, 13

[28] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-aware low-light image enhancement. In *CVPR*, 2022. 3, 6, 7, 13

[29] Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *ECCV*, 2012. 2, 3

[30] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary DETR with conditional matching. In *ECCV*, 2022. 3

[31] Lin Zhang, Lijun Zhang, Xiao Liu, Ying Shen, Shaoming Zhang, and Shengjie Zhao. Zero-shot restoration of backlit images using deep internal learning. In *ACMMM*, pages 1623–1631, 2019. 3, 6, 7, 13

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[33] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACMMM*, 2019. 2

[34] Lin Zhao, Shaoping Lu, Tao Chen, Zhenglu Yang, and Ariel Shamir. Deep symmetric network for underexposed image enhancement with recurrent attentional learning. In *ICCV*, 2021. 7, 13

[35] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 2, 3

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022. 3

[38] Shangchen Zhou, Chongyi Li, and Chen Change Loy. LEDNet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, 2022. 3

# Iterative Prompt Learning for Unsupervised Backlit Image Enhancement

# – Supplementary Material –

In this supplementary material, we present more ablation studies (Section A), extra training details (Section B), further discussions (Section C), and more comparisons (Section D and Section E). In addition, a video demo is provided to showcase the effectiveness of our method in a separate file.

## A. More Ablation Studies

### A.1. Superiority in Generalization Ability

To show the superiority of our unsupervised method over the supervised method in terms of generalization ability, we present the visual comparison with a supervised method Unet$_{pair}$ (Unet [22] is trained using the paired training data of the BAID dataset). The visual comparisons are conducted on the Backlit300 test dataset.

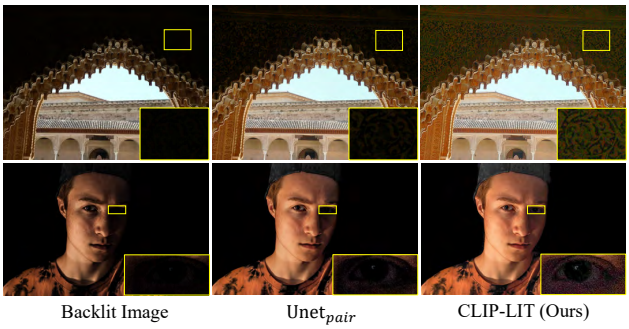



Figure 14: Visual comparisons between supervised method (Unet$_{pair}$) and our unsupervised method.

As shown in Fig. 14, our method produces visually pleasing results with clear details and sufficient luminance, while the results of Unet$_{pair}$ suffer from the under-exposure issue. Overall, our method yields a more realistic color distribution and a brighter image, indicating that an unsupervised backlit enhancement method is still desired even if having manually touched paired data.

### A.2. Impact of Labels on Prompt Initialization

While Fig. 15 shows using pure random initialization will take more iterations to converge, both Tab. 6 and Fig. 15 show whether to initialize with words does not have much impact on the final performance and model training.

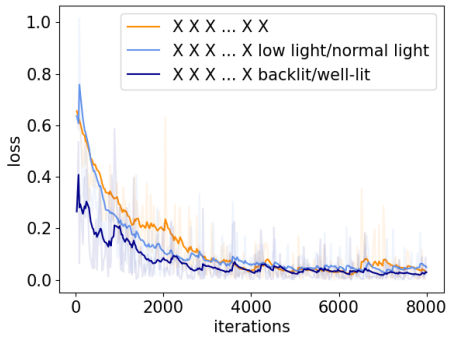



Figure 15: Prompt initialization learning investigation.

| Settings | PSNR↑ | SSIM↑ | LPIPS↓ | MUSIQ↑ |
|---|---|---|---|---|
| Pure random initialization | 21.237 | 0.884 | 0.158 | 55.959 |
| Random initialization+backlit/well-lit | 21.527 | 0.882 | 0.159 | 55.946 |
| Random initialization+low light/normal light | 21.579 | 0.883 | 0.159 | 55.682 |

Table 6: Quantitative comparison of different prompt initialization learning on BAID test dataset.


Figure 16: Visual comparisons between the model trained using adversarial loss and our method.

### A.3. Visual Comparisons with Adversarial Loss

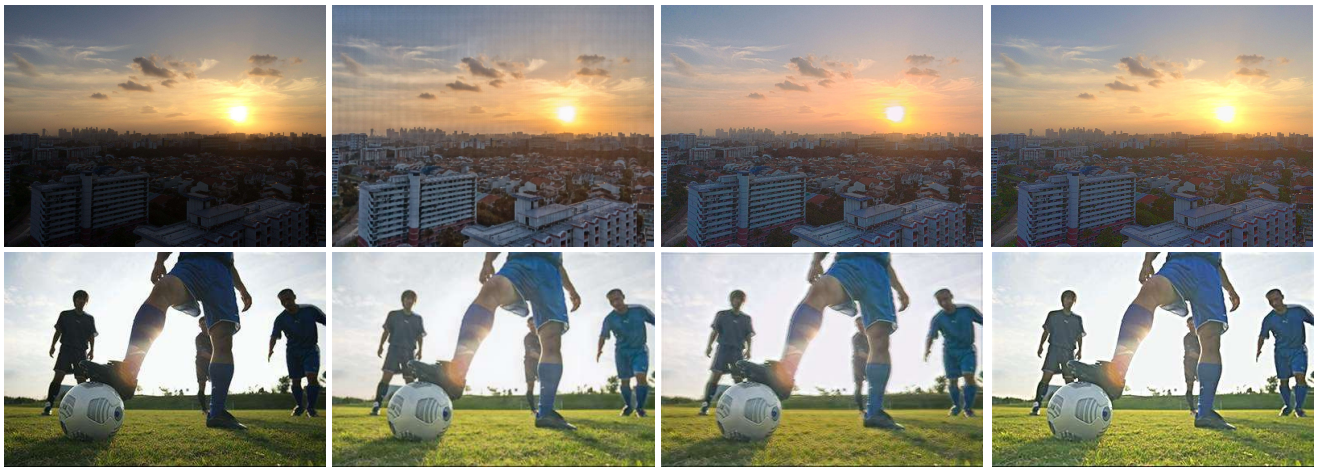
As shown in Fig. 16, compared to the results generated by the model trained with adversarial loss (keeping other settings fixed), our results are more consistent with the input and brighter in backlit areas.

## B. Extra Training Details

Apart from a fixed number of iterations in the process of prompt learning and enhancement training, we use two thresholds ($Thr_A$ and $Thr_B$) to control the alteration. Specifically, if the prompt learning's loss is lower than the

Input SNR (CVPR'22) Zhao et al. (ICCV'21) Ours

Figure 17: Visual comparisons between our method and the retrained supervised methods on the Backlit300 test dataset.



Input SNR (CVPR'22) Zhao et al. (ICCV'21) Ours

Figure 18: Visual comparisons between our method and the retrained supervised methods on the Backlit300 dataset.

threshold $Thr_A$, the learned prompts are frozen and the enhancement training will be triggered. If the enhancement model's training loss is lower than the threshold $Th_B$, the enhancement model is frozen and then infers the latest output images, and the prompt learning will be triggered. Otherwise, the alteration will be triggered when the iterations of the current process reach the pre-defined number (i.e. 800 for each stage). The threshold of the prompt learning loss ($Thr_A$) is set to 60, and the threshold of the enhancement training loss ($Thr_B$) is set to 90.

## C. Further Discussions

Some compared supervised methods provide various models that were pre-trained on different datasets. For a fair comparison, our paper shows results from existing supervised methods without retraining, as our method does not require paired data. Some retrained unsupervised methods perform worse than their original models. This is primarily due to 1) some methods perform global enhancement, which may be disturbed by the uneven brightness in backlit images. 2) some methods require the training data of diverse brightness levels, while the size of our training data (380 unpaired sets) is limited. *These results prove our method can perceive the pixel-level illuminance information and is free from the high requirement for sufficient training data.*

## D. Comparisons with Retrained Supervised Methods

We retrain the state-of-the-art supervised methods on the same 380 randomly selected paired data. In Tab. 7, our

method outperforms the retrained methods in most metrics. As shown in Fig. 17 and Fig. 18, the visual results of supervised methods are relatively blurred and under-exposed. Our visual quality is superior.

| Methods | BAID test dataset | | | | Backlit300 dataset |
|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | MUSIQ↑ | MUSIQ ↑ |
| Zhao et al. (ICCV'21) | **22.123** | 0.876 | 0.183 | 46.467 | 46.369 |
| SNR (CVPR'22) | 21.740 | 0.800 | 0.359 | 25.930 | 31.522 |
| Ours | 21.579 | **0.883** | **0.159** | **55.682** | **52.921** |

Table 7: Quantitative comparison with the retrained supervised method.

## E. More Visual Comparisons

Here we compare our method with the other methods mentioned in our paper (i.e. all the supervised methods mentioned in our paper (Afifi et al. [1], Zhao et al. [34], URetinex-Net [27], SNR-aware [28]) with all of their released pretrained models, and all the unsupervised methods mentioned in our paper ( ExCNet [31], SCI [20], Zero-DCE [9], Zero-DCE++ [15], EnlightenGAN [10], RUAS [18]) as well as the version retrained on the same unpaired dataset) on the two mentioned test datasets.

### E.1. Visual Comparisons on the BAID Test Dataset

Figs. 19, 20, 21, 22 and 23 show more visual comparisons between the results generated by our methods and the compared methods. Note that the BAID reference images are modified by photographers. The results show that our CLIP-LIT effectively enhances the backlit image without causing over/under-exposure and produces most natural appearance than the compared methods.

### E.2. Visual Comparisons on The Backlit300 Dataset

Figs. 24, 25, 26, and 27 show more visual comparisons between the results generated by our methods and the compared methods. The results show that our method stores the color and the content of details in the backlit area most clearly and realistically, and the enhanced details have best and natural color contrast while keeping the well-lit background remain unchanged. Our CLIP-LIT yields most visually favorable result in the night scene as well.

## F. Additional Illustration

Our model is lightweight and can handle a 4K image within 0.005 seconds using a single NVIDIA GeForce RTX 3090 GPU. We provide some 4K video samples enhanced by our CLIP-LIT in our video demo. Due to size limits, we resize the outputs to 2K to show in the demo.


Figure 19: **Complete comparisons with all methods and the reference image on the BAID test dataset.** Our CLIP-LIT effectively enhances the backlit image without causing over/under-exposure.


Figure 20: **Complete comparisons with all methods and the reference image on the BAID test dataset.** Our CLIP-LIT produces most natural appearance than the compared methods.


Figure 21: **Complete comparisons with all methods and the reference image on the BAID test dataset.** Our CLIP-LIT restores the human face most clearly and naturally.

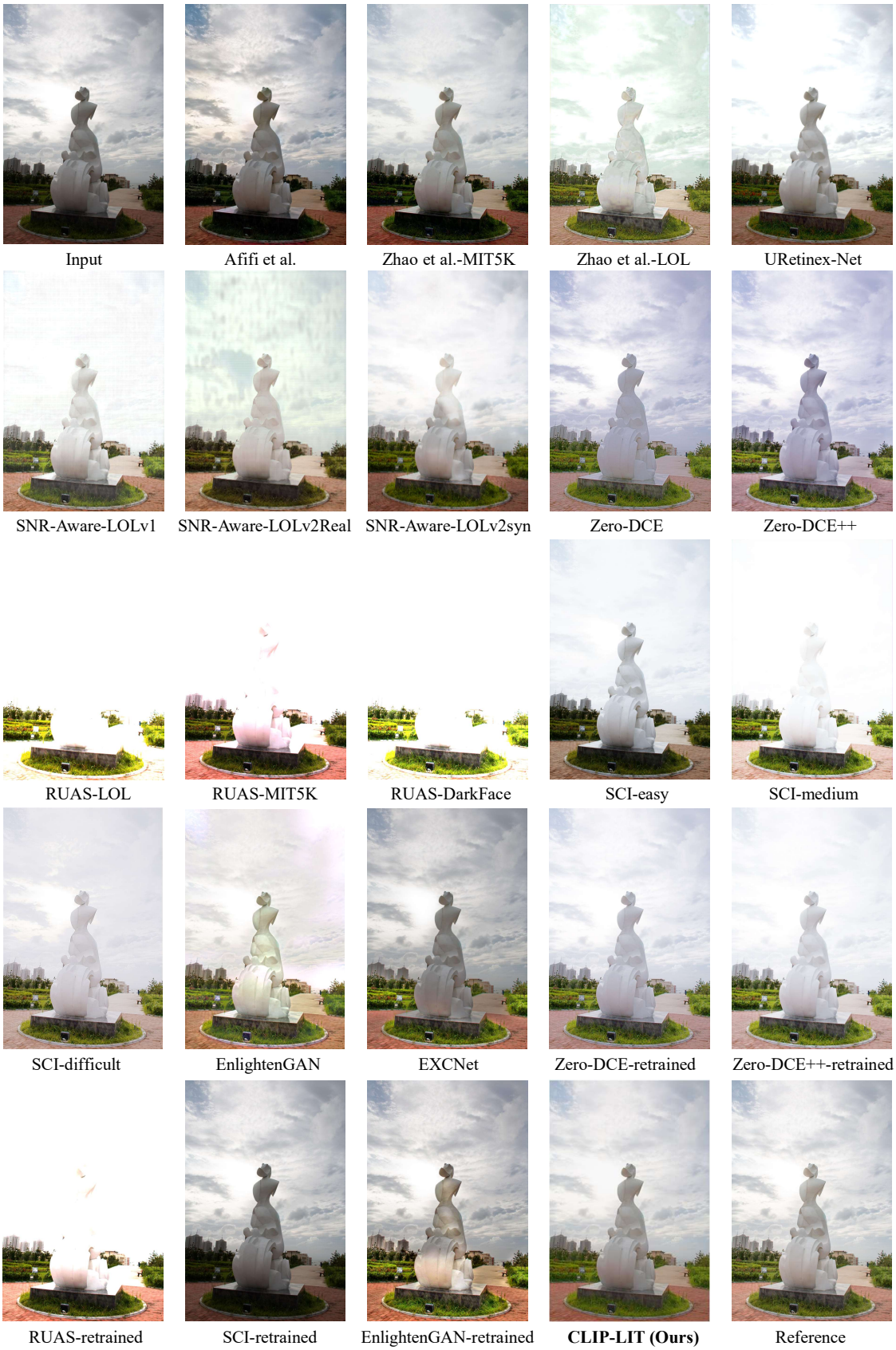



Figure 22: **Complete comparisons with all methods and the reference image on the BAID test dataset.** Our CLIP-LIT's result is visually closest to the reference image retouched by photographers.


Figure 23: **Complete comparisons with all methods and the reference image on the BAID test dataset.** Our method enlightens the dark area most naturally while preserving the color and content of the well-lit area well, that is, has the best input-output consistency in the well-lit regions.


Figure 24: **Complete comparisons with all methods on the Backlit300 test dataset.** Our CLIP-LIT restores the color and the content of the details in the backlit area most clearly and the enhanced details have best color contrast while keeping the well-lit background remain unchanged.

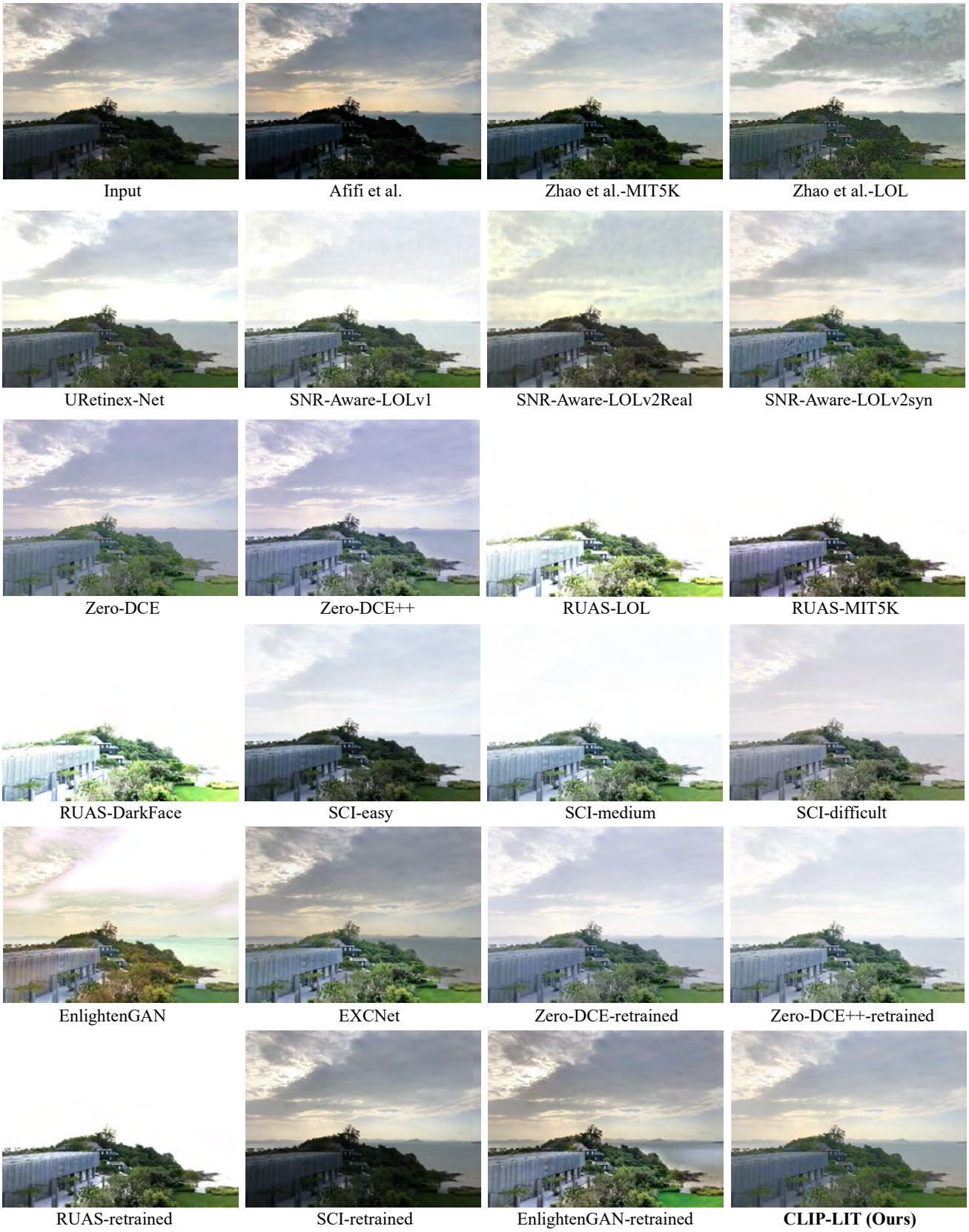



Figure 25: **Complete comparisons with all methods on the Backlit300 test dataset.** Our result is the most realistic and has input-output consistency in well-lit areas.

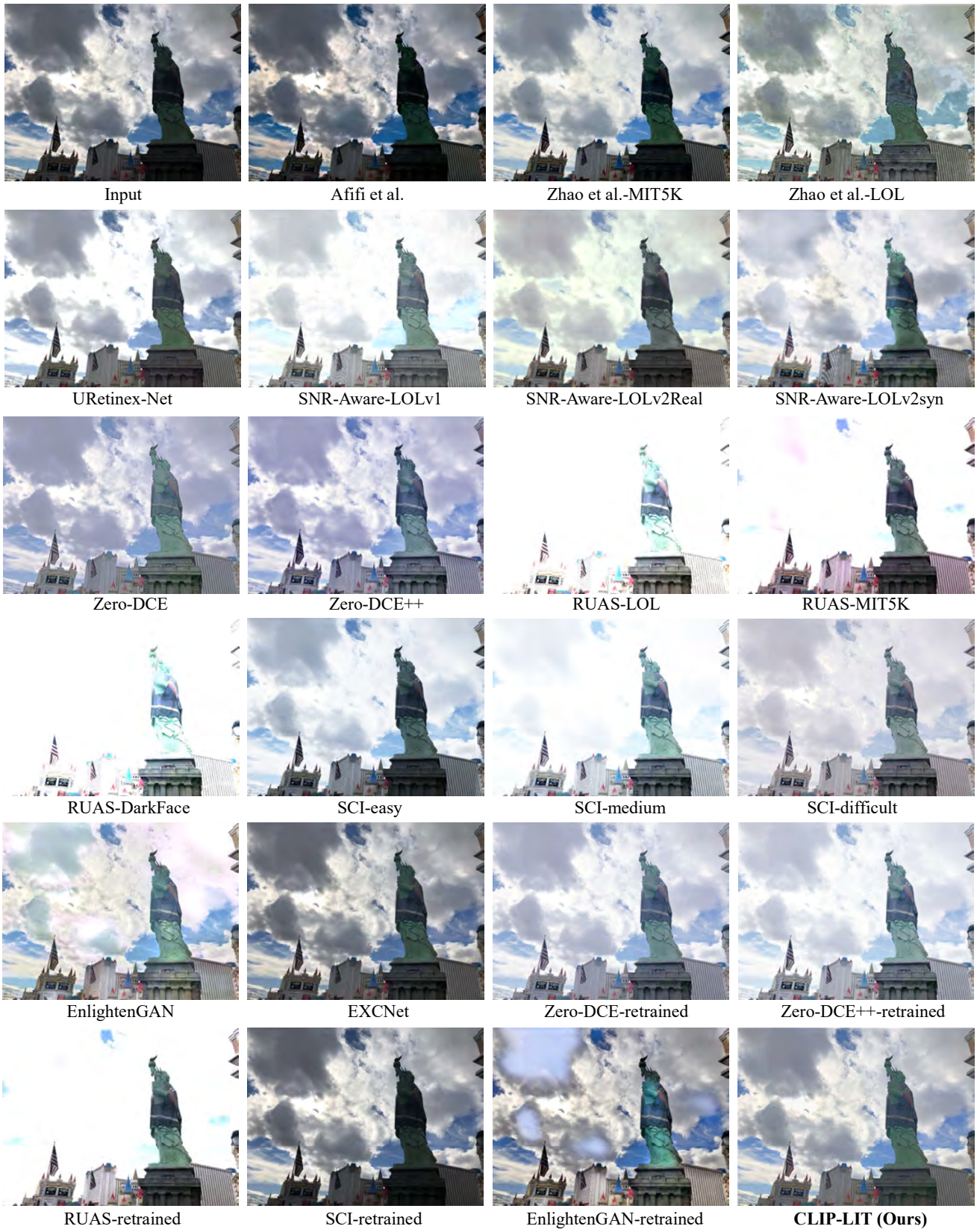



Figure 26: **Complete comparisons with all methods on the Backlit300 test dataset.** Our results do not contain artifacts and over-exposed regions. Moreover, Our result has the most natural color in the enlightened area and has the most pleasing contrast.

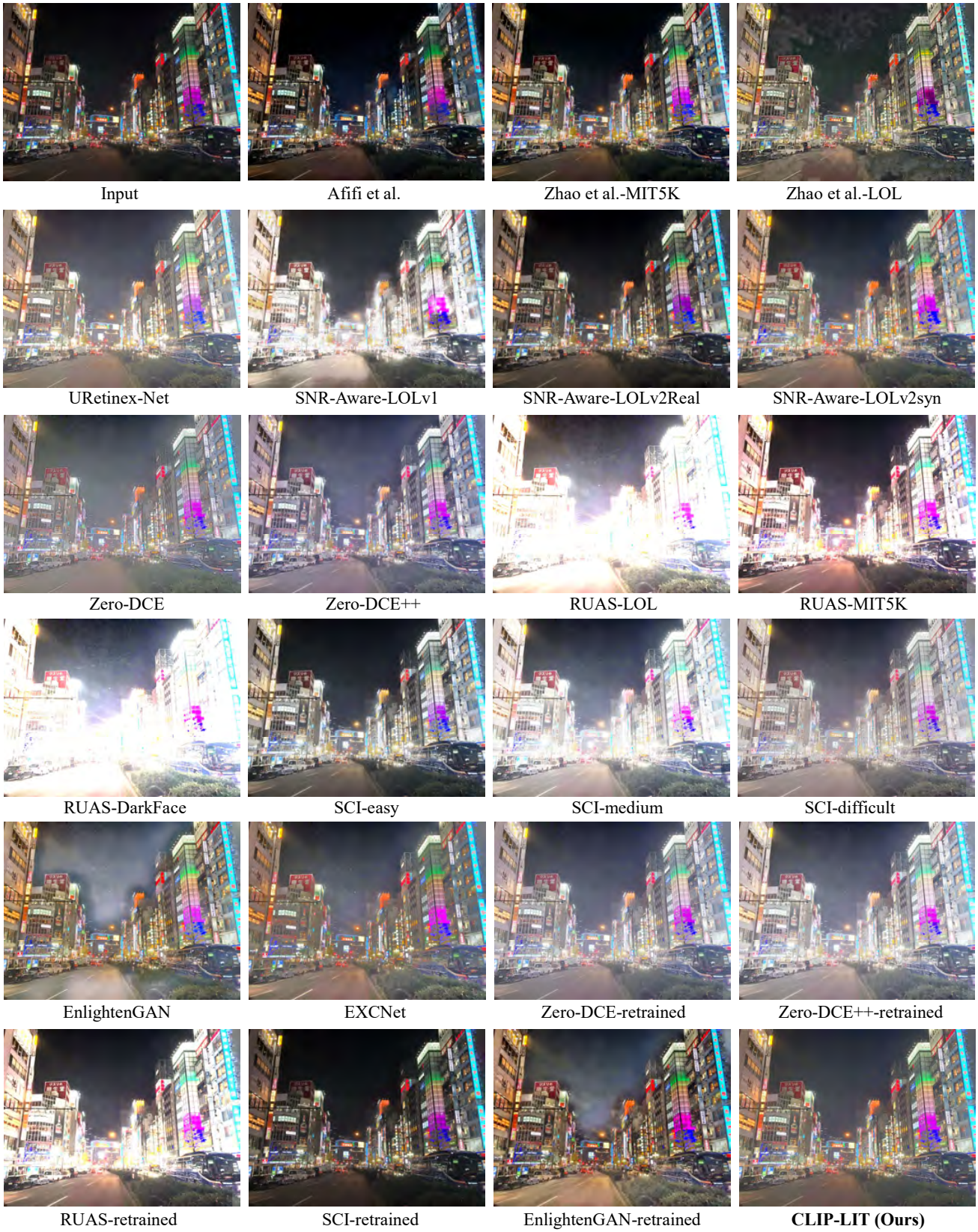



Figure 27: **Complete comparisons with all methods on the Backlit300 test dataset.** Our CLIP-LIT produces the most visually favorable result in the night scene as well, which enhances the backlit foreground well while not overly brightening the night sky.