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/

ICCV 2023

Background

Low-light images:

(a) Insufficient lighting

(b) back lighting

(c) Uneven lighting

Backlit image enhancement is more challenging as it requires **preserving well-lit areas** while **enhancing underexposed areas**.

Background

- **Conventional methods.**

  Difficult to process real-world backlit images.

- **Supervised methods.** Retinex-Net、KID、 URetinex-Net

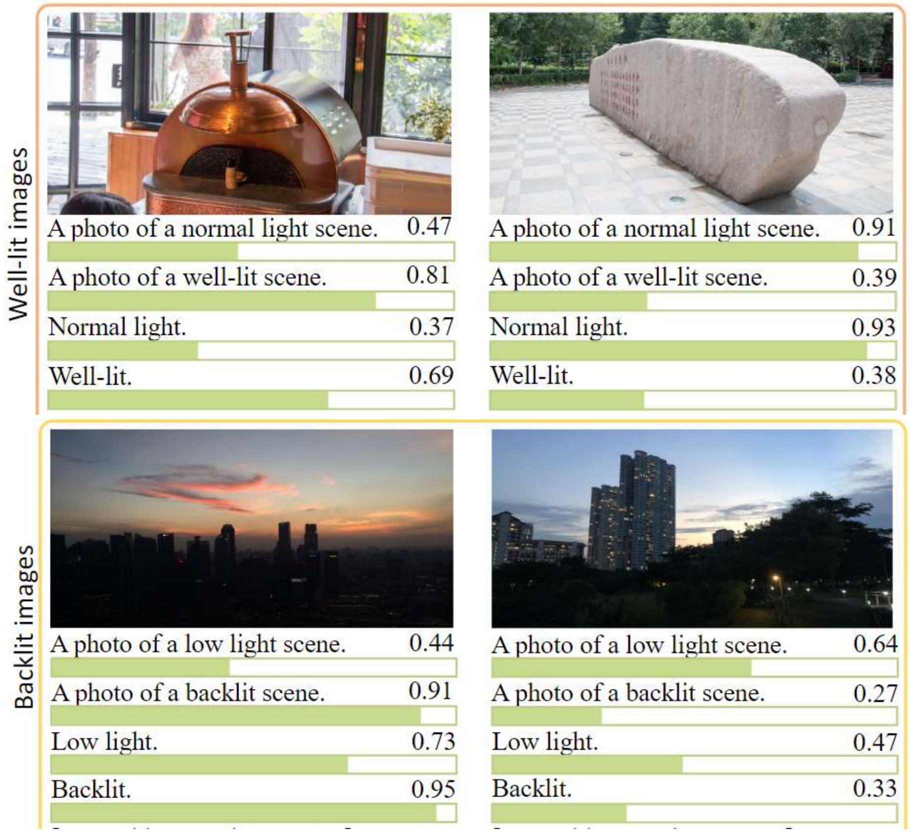
  Over-enhancement in well-lit areas or under-enhancement in low-light areas.

- **Unsupervised methods.** **Zero-DCE、EnlightenGAN**

  Rely on ideal assumptions such as mean brightness and gray world models or directly learn the distribution of normal light images through adversarial training.

Introduction

$$s = \frac{\boldsymbol{x} \odot \boldsymbol{t}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{t}||},$$

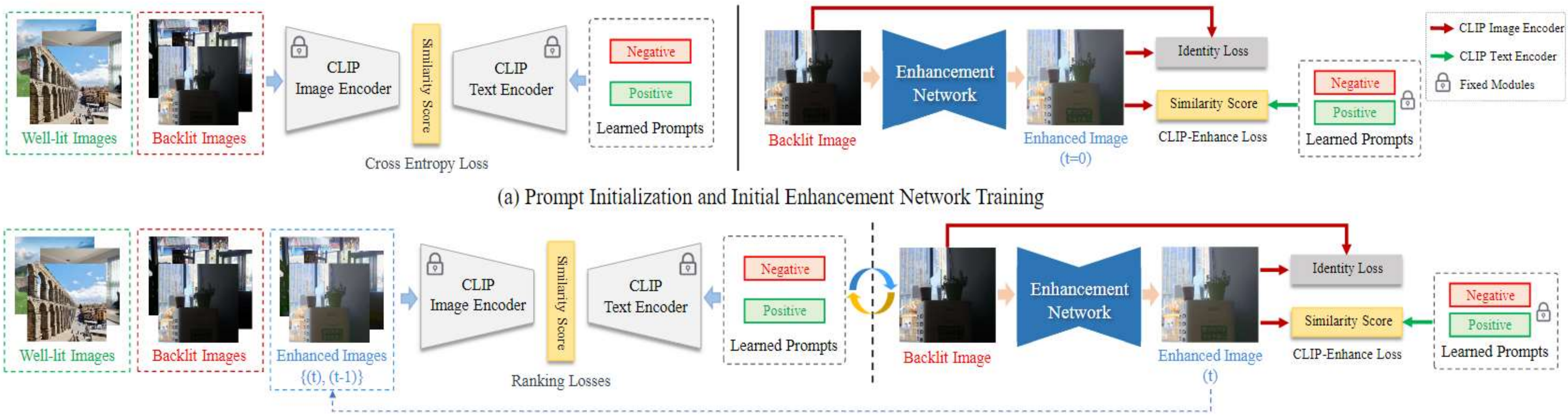


The optimal prompts could vary on a case-by-case basis due to the complex illuminations in the scene.

It is unlikely to achieve optimal performance with **fixed prompts**.

Replace fixed prompts with **learnable prompts** and continuous **prompt refinement** helps achieve image enhancement

Method

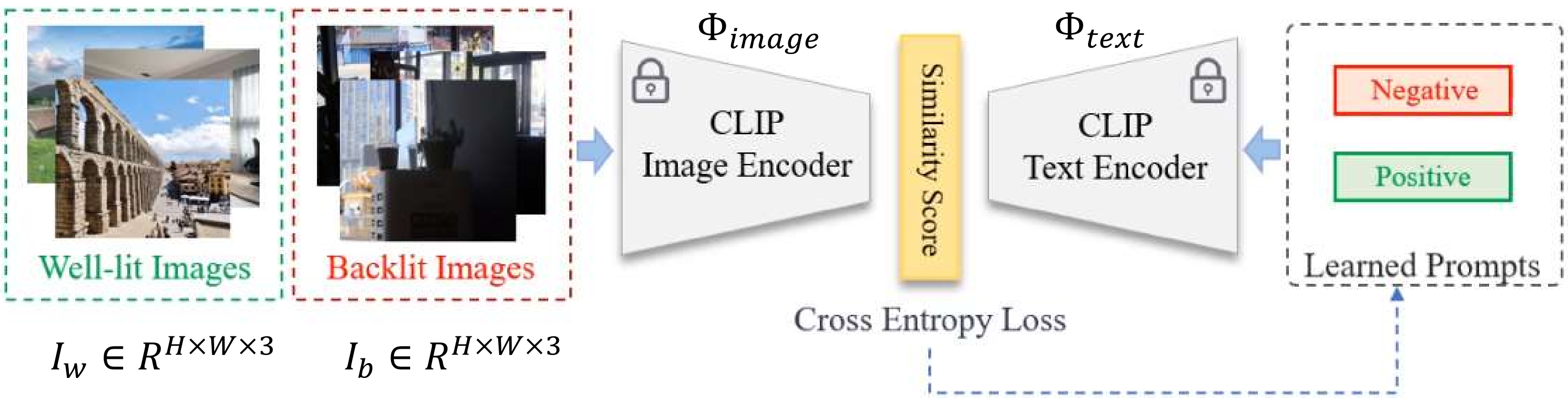


(a) Prompt Initialization and Initial Enhancement Network Training

(b) Prompt Refinement and Enhancement Model Fine-tuning

Method

The first stage 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**.



$T_n \in R^{N\times 512}$

$T_p \in R^{N\times 512}$

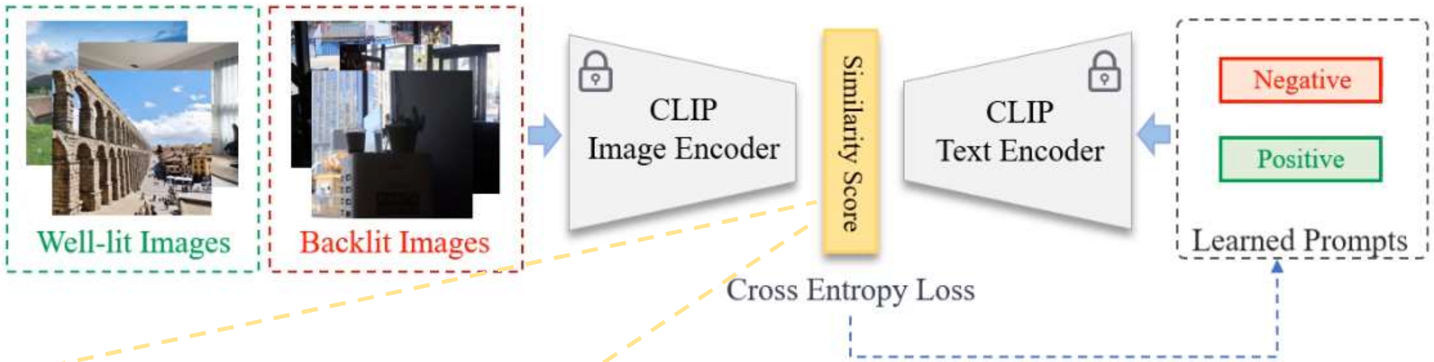
➢ Antonym prompt pairing

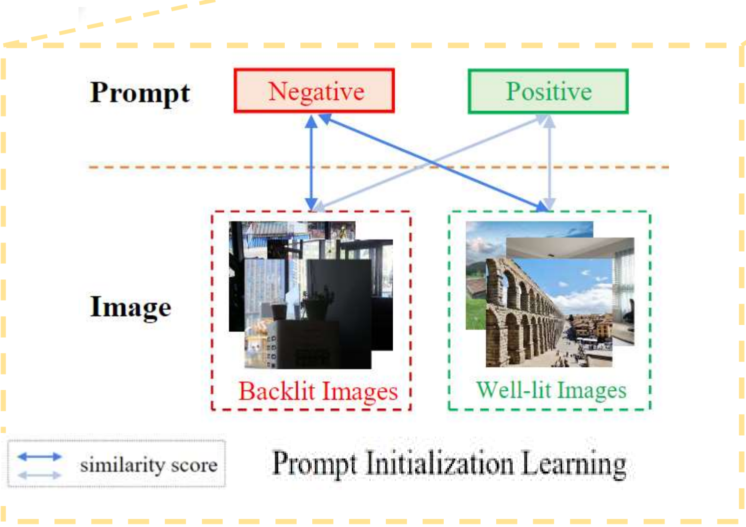
$$s = \frac{\boldsymbol{x} \odot \boldsymbol{t}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{t}||},$$

$$s_i = \frac{\boldsymbol{x} \odot \boldsymbol{t}_i}{||\boldsymbol{x}|| \cdot ||\boldsymbol{t}_i||}, \quad i \in \{1, 2\},$$

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}.$$

Method



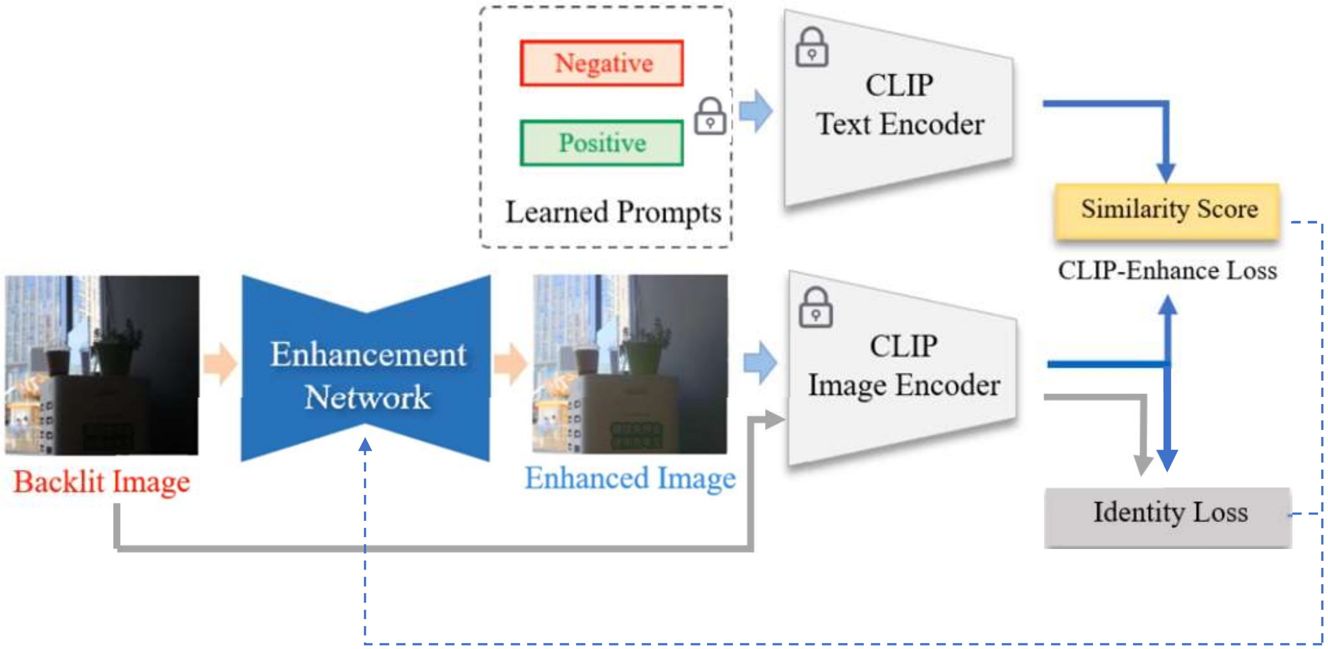


$$\mathcal{L}_{initial} = -(y * \log(\hat{y}) + (1 - y) * \log(1 - \hat{y})), \quad (1)$$

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

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

Method



The CLIP-Enhance loss L_{clip}

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

The identity loss $L_{identity}$

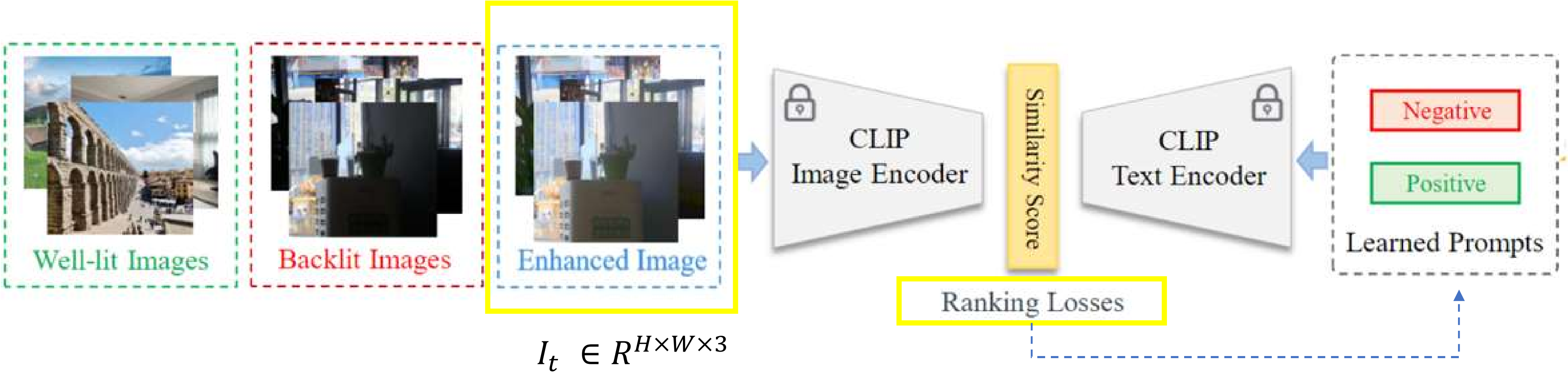
$$\mathcal{L}_{identity} = \sum_{l=0}^{4} \alpha_l \cdot ||\Phi^l_{image}(I_b) - \Phi^l_{image}(I_t)||_2,$$

The enhancement loss $L_{enhance}$

$$\mathcal{L}_{enhance} = \mathcal{L}_{clip} + w \cdot \mathcal{L}_{identity},$$

Method

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**.



$I_t \in R^{H\times W\times 3}$

The negative similarity score between the prompt pair and an image:

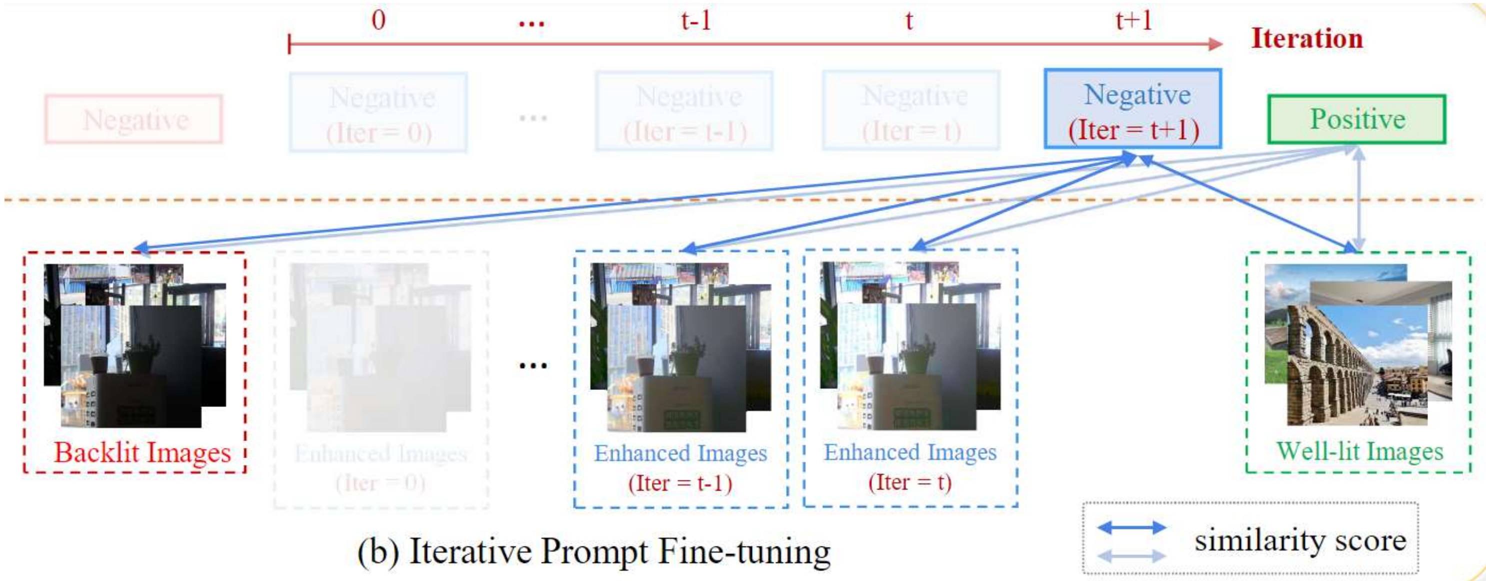
$$S(I) = \frac{e^{cos(\Phi_{image}(I),\Phi_{text}(T_n))}}{\sum_{i\in\{n,p\}} e^{cos(\Phi_{image}(I),\Phi_{text}(T_i))}},$$

Method

Margin ranking loss:

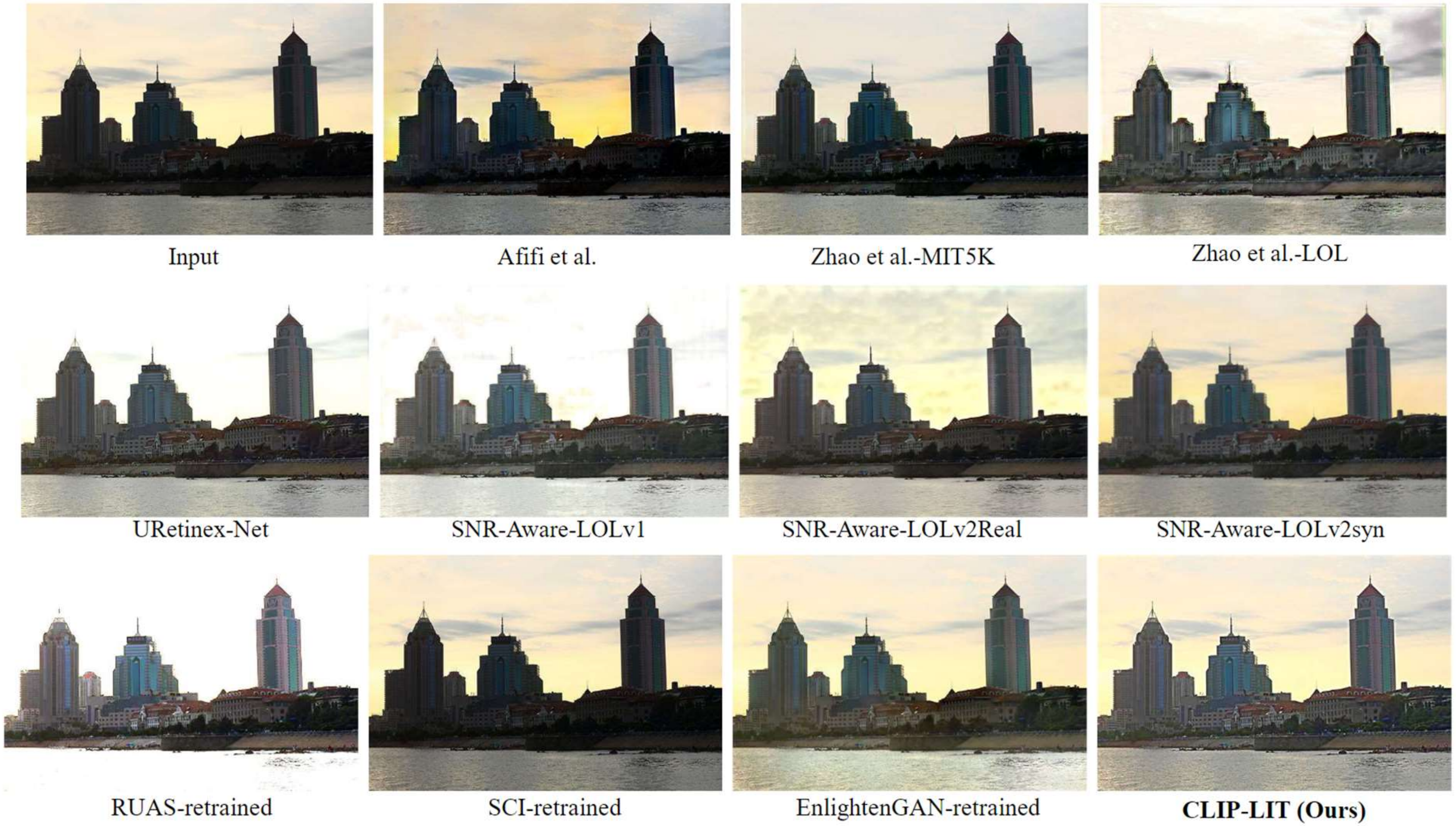
$$\begin{aligned}\mathcal{L}_{prompt1} &= \max(0, S(I_w) - S(I_b) + m_0) \\ &+ \max(0, S(I_t) - S(I_b) + m_0) \\ &+ \max(0, S(I_w) - S(I_t) + m_1),\end{aligned}$$

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



(b) Iterative Prompt Fine-tuning

Experiments



Experiments

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

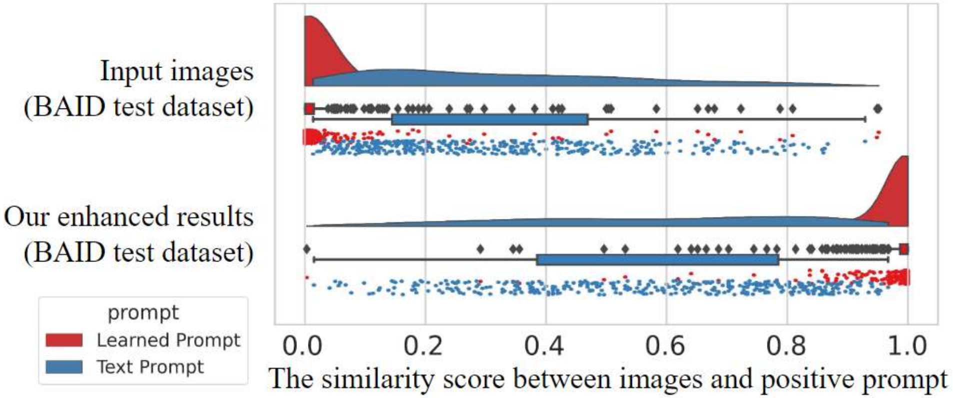
Type	Methods	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑
	Input	16.641	0.768	0.197	52.115
Supervised	Afifi et al. [1]	15.904	0.745	0.227	52.863
	Zhao et al.-MIT5K [34]	18.228	0.774	0.189	51.457
	Zhao et al.-LOL [34]	17.947	0.822	0.272	49.334
	URetinex-Net [27]	18.925	0.865	0.211	**54.402**
	SNR-Aware-LOLv1 [28]	15.472	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**

Ablation Studies

- Necessity of Learned Prompt



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**

- Effectiveness of Iterative Learning

Ablation Studies

- Impact of Training Data

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.

- Advantage of CLIP-Enhance Loss over the Adversarial Loss.

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**

Ablation Studies

➢ Impact of Different Prompt Initialization

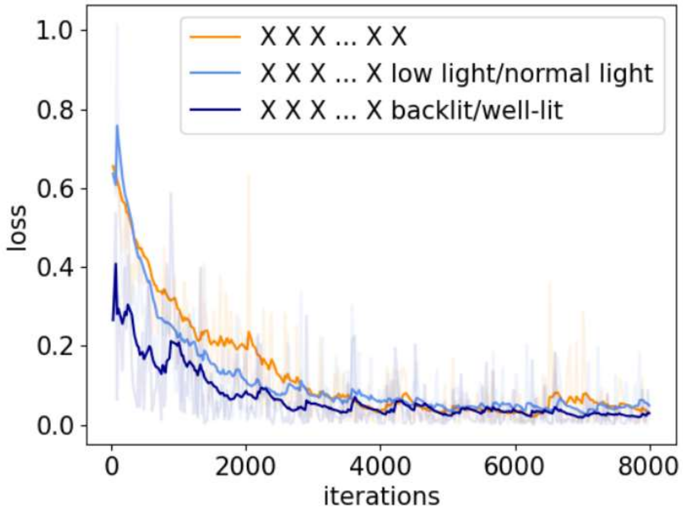


Figure 15: Prompt initialization learning investigation.

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

Thanks