LightenDiffusion: Unsupervised Low-Light Image Enhancement with Latent-Retinex Diffusion Models

[1] Sichuan University, Chengdu, China
jianghai@stu.scu.edu.cn, hansongchen@scu.edu.cn
[2] Southwest Jiaotong University, Chengdu, China
aoluo@swjtu.edu.cn
[3] University of Electronic Science and Technology of China, Chengdu, China
liushuaicheng@uestc.edu.cn
[4] Shanghai Jiao Tong University, Shanghai, China
xiaohongliu@sjtu.edu.cn
[5] Megvii Technology, Beijing, China

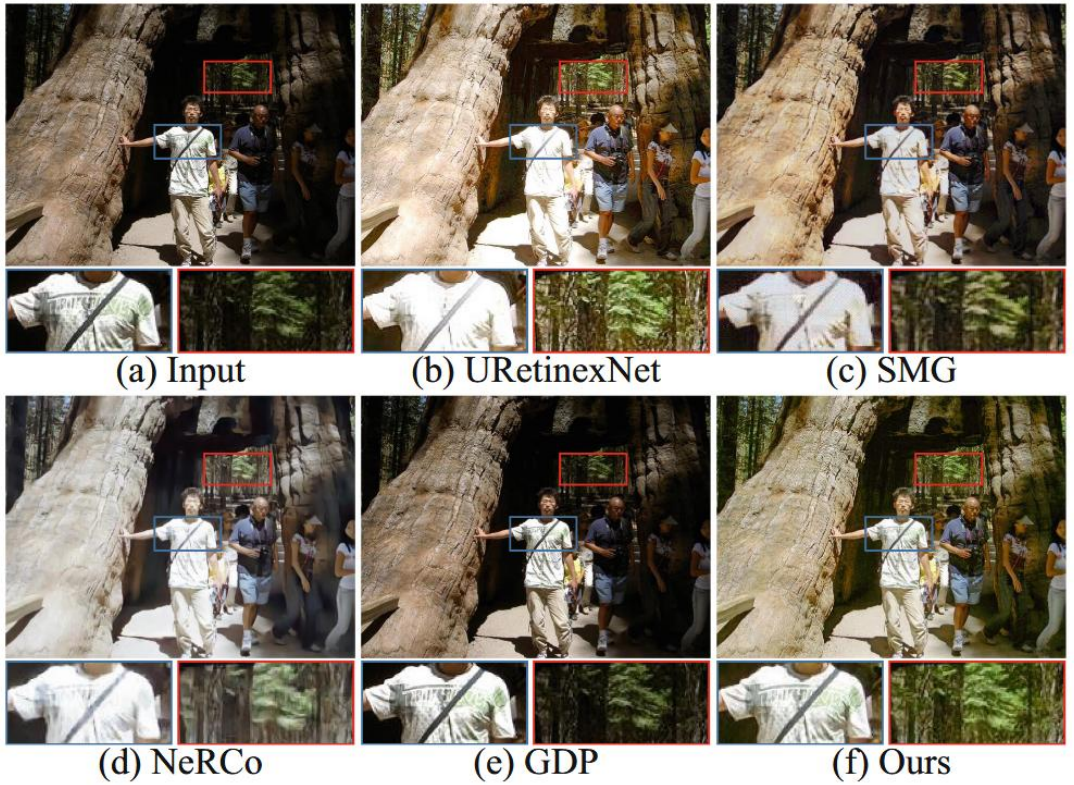
ECCV 2024

Introduction

- Traditional methods mainly adopt hand-crafted priors, such as histogram equalization (HE) and Retinex theory to improve contrast and restore details.

- learning-based methods can directly learn the mapping from low-light images to normal-light images through powerful network architectures and sophisticated learning strategies, which present more robustness than traditional methods.

- To leverage the label-free characteristic of unsupervised learning to improve the generalization of diffusion models, some methods employ zero-shot solutions that utilize well-established priors from pre-trained diffusion models for restoration without training from scratch.



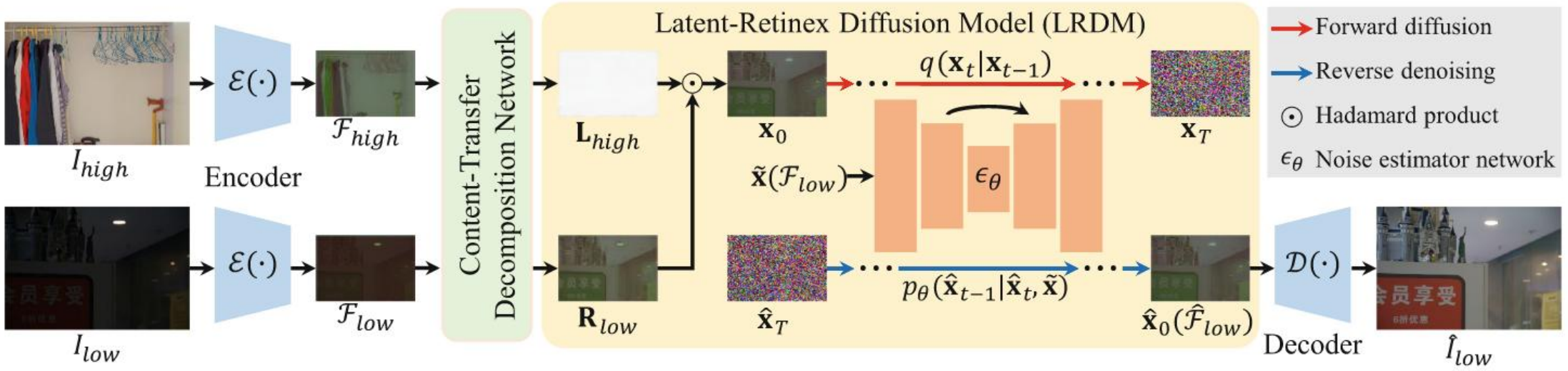
Introduction

Contribution:

- Propose a diffusion-based framework, termed LightenDiffusion, that leverages the advantages of Retinex theory and the generative ability of diffusion models for unsupervised low-light image enhancement, with a self-constrained consistency loss further proposed to improve visual quality.

- A content-transfer decomposition network that performs decomposition in the latent space, aiming to obtain content-rich reflectance maps and content-free illumination maps to promote unsupervised restoration.

Proposed Network



- Employ an encoder E(·) to convert the unpaired low-light image Ilow and normal-light image Ihigh into latent space
- The encoded features are sent to the proposed content-transfer decomposition network (CTDN)
- The reflectance map of the low-light image R_{low} and the illumination of the normal-light image Lhigh are taken as the input of the diffusion model
- Send it to a decoder D(·) to produce the final result $\hat{I}_{low}$.

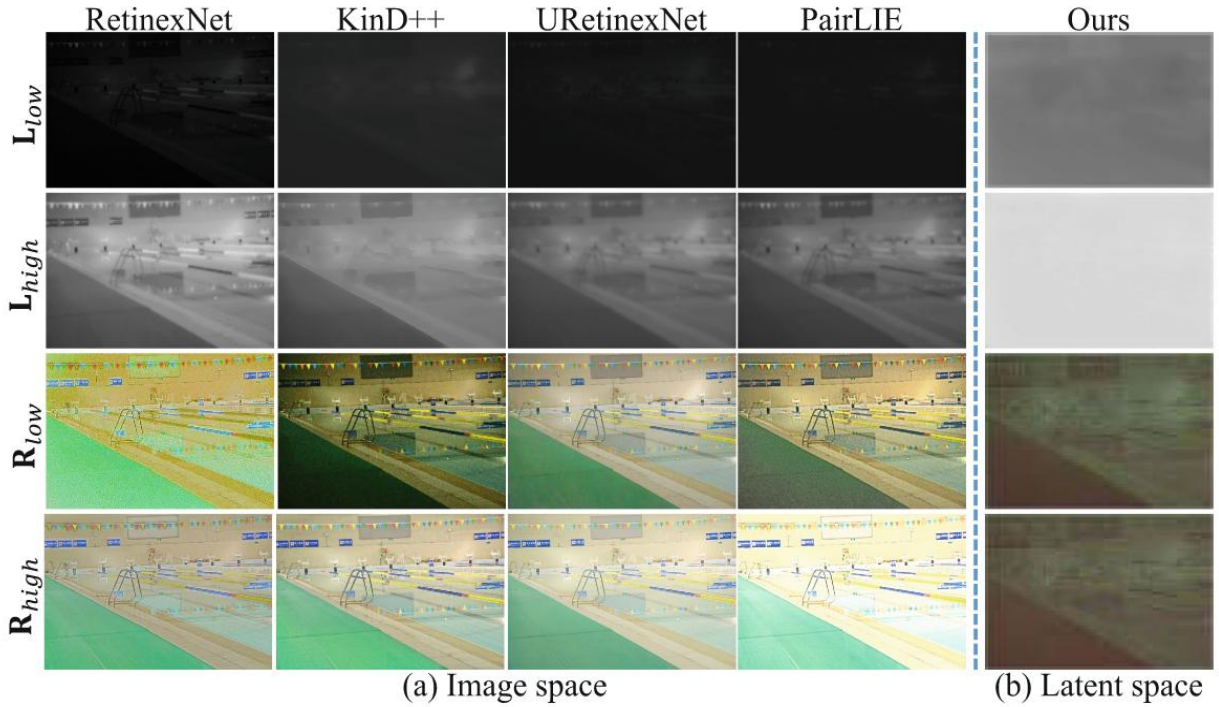
Proposed Network

Content-Transfer Decomposition Network

The Retinex theory： $I = \mathbf{R} \odot \mathbf{L}$

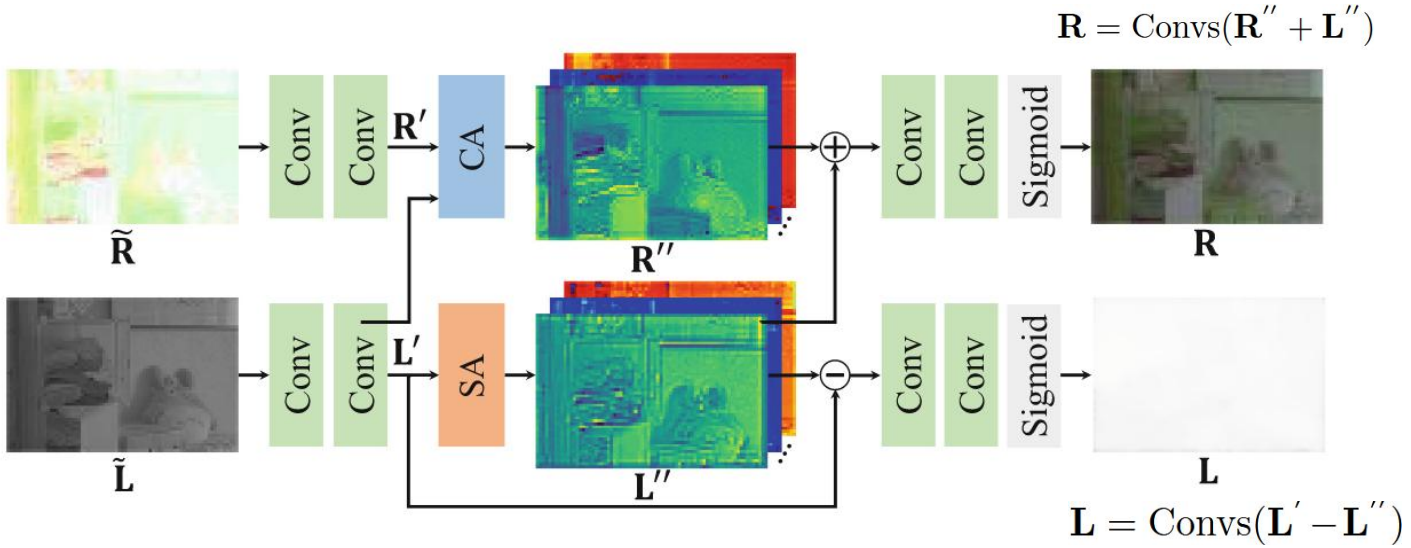
- Existing methods typically perform decomposition in the image space to obtain the above components, which results in the content information not being fully decomposed into the reflectance map and partially retained in the illumination map

- CTDN can generate content-rich reflectance maps that fully represent the intrinsic information of the image, and content-free illumination maps that only reveal the lighting conditions



Proposed Network

Content-Transfer Decomposition Network



estimate the initial reflectance and illumination maps

$$\tilde{\mathbf{L}}(x) = \max_{c \in [0,C)} \mathcal{F}^c(x), \tilde{\mathbf{R}}(x) = \mathcal{F}(x)/(\tilde{\mathbf{L}}(x) + \tau),$$

- cross-attention (CA) module to leverage the illumination map to reinforce the content information in the reflectance map
- self-attention module (SA) is adopted to further extract content information in the illumination map

Proposed Network

Network Training

a two-stage strategy for network training：

1. Using SICE dataset to optimize the encoder E(·), CTDN, and decoder D(·), while freezing the parameters of the diffusion model.

   encoder and decoder are optimized with the content loss

$$\mathcal{L}_{con} = \sum_{i=1}^{2} \|I_{low}^{i} - \mathcal{D}(\mathcal{E}(I_{low}^{i}))\|_2.$$

   The CTDN is optimized with the decomposition loss $\mathcal{L}_{dec}$

$$\mathcal{L}_{dec} = \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{ref} + \lambda_3 \mathcal{L}_{ill}.$$

$$\mathcal{L}_{rec} = \sum_{i=1}^{2}\sum_{j=1}^{2} \|\mathcal{F}_{low}^{j} - \mathbf{R}_{low}^{i} \odot \mathbf{L}_{low}^{j}\|_1.$$

$$\mathcal{L}_{ref} = \|\mathbf{R}_{low}^{1} - \mathbf{R}_{low}^{2}\|_1$$

$$\mathcal{L}_{ill} = \sum_{i=1}^{2} \|\nabla \mathbf{L}_{low}^{i} \cdot \exp(-\lambda_g \nabla \mathbf{R}_{low}^{i})\|_2$$

   λ2, λ3, and λg are empirically set to 0.1, 0.01, and 10

Proposed Network

Network Training

2. Collect ~180k unpaired low/normal-light image pairs to optimize the diffusion model while freezing the parameters of other modules.

the L_{scc} aims to enable the restored feature to share the same intrinsic information as the input low-light image

$$\mathcal{L}_{scc} = \|\tilde{\mathcal{F}}_{low} - \hat{\mathcal{F}}_{low}\|_1.$$

first perform the reverse denoising process in the training phase following to generate the restored feature and construct a pseudo label $\tilde{F}_{low}$ from decomposition results of the low-light image as a reference based on traditional Gamma correction approaches as

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_1 \mathcal{L}_{scc}.$$

Experiments

Settings

- Implement on four NVIDIA RTX 2080Ti
- batch size and patch size are set to 12 and 512 ×512
- the initial learning rate set to 1×10^{-4} in the first stage and decays by a factor of 0.8 while reinitializing it to a fixed value of 2×10^{-5} in the second stage.
- the U-Net architecture is adopted as the noise estimator network with the time step T and sampling step S set to 1000 and 20 for the forward diffusion and reverse denoising processes, respectively.

Datasets

- two paired datasets：LOL and LSRW
  ——using PSNR,SSIM and LPIPS
- three real-world unpaired benchmarks：DICM, NPE, and VV
  ——use two non-reference perceptual metrics NIQE and PI

Experiments

Quantitative Comparisons

Type	Method	LOL [58]			LSRW [16]			DICM [28]		NPE [53]		VV [51]	
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	PI ↓	NIQE ↓	PI ↓	NIQE↓	PI ↓
T	LIME [14]	17.546	0.531	0.290	17.342	0.520	0.416	4.476	4.216	4.170	3.789	3.713	3.335
	SDDLLE [17]	13.342	0.634	0.261	14.708	0.486	0.382	4.581	3.828	4.179	3.315	4.274	3.382
	CDEF [30]	16.335	0.585	0.351	16.758	0.465	0.314	4.142	4.242	3.862	2.910	5.051	3.272
	BrainRetinex [4]	11.063	0.475	0.327	12.506	0.390	0.374	4.350	3.555	3.707	3.044	4.031	3.114
SL	RetinexNet [58]	16.774	0.462	0.390	15.609	0.414	0.393	4.487	3.242	4.732	3.219	5.881	3.727
	KinD++ [74]	17.752	0.758	0.198	16.085	0.394	0.366	4.027	3.399	4.005	3.144	3.586	2.773
	LCDPNet [52]	14.506	0.575	0.312	15.689	0.474	0.344	4.110	3.250	4.106	3.127	5.039	3.347
	URetinexNet [59]	19.842	0.824	0.128	18.271	0.518	0.295	4.774	3.565	4.028	3.153	3.851	2.891
	SMG [64]	**23.814**	0.809	0.144	17.579	0.538	0.456	6.224	4.228	5.300	3.627	5.752	3.757
	PyDiff [76]	23.275	**0.859**	**0.108**	17.264	0.510	0.335	4.499	3.792	4.082	3.268	4.360	3.678
	GSAD [20]	22.021	0.848	0.137	17.414	0.507	**0.294**	4.496	3.593	4.489	3.361	5.252	3.657
SSL	DRBN [68]	16.677	0.730	0.252	16.734	0.507	0.376	4.369	3.800	3.921	3.267	3.671	3.117
	BL [39]	10.305	0.401	0.382	12.444	0.333	0.384	5.046	4.055	4.885	3.870	5.740	4.030
UL	Zero-DCE [12]	14.861	0.562	0.330	15.867	0.443	0.315	3.951	3.149	3.826	2.918	5.080	3.307
	EnlightenGAN [24]	17.606	0.653	0.319	17.106	0.463	0.322	3.832	3.256	3.775	2.953	3.689	2.749
	RUAS [32]	16.405	0.503	0.257	14.271	0.461	0.455	7.306	5.700	7.198	5.651	4.987	4.329
	SCI [40]	14.784	0.525	0.333	15.242	0.419	0.321	4.519	3.700	4.124	3.534	5.312	3.648
	GDP [8]	15.896	0.542	0.337	12.887	0.362	0.386	4.358	3.552	4.032	3.097	4.683	3.431
	PairLIE [10]	19.514	0.731	0.254	17.602	0.501	0.323	4.282	3.469	4.661	3.543	3.373	2.734
	NeRCo [67]	19.738	0.740	0.239	17.844	0.535	0.371	4.107	3.345	3.902	3.037	3.765	3.094
	Ours	20.453	0.803	0.192	**18.555**	**0.539**	0.311	**3.724**	**3.144**	**3.618**	**2.879**	**2.941**	**2.558**

'T', 'SL', 'SSL', and 'UL' indicate that the methods belong to traditional, supervised, semi-supervised, and unsupervised methods, respectively.

Experiments

Qualitative Comparisons

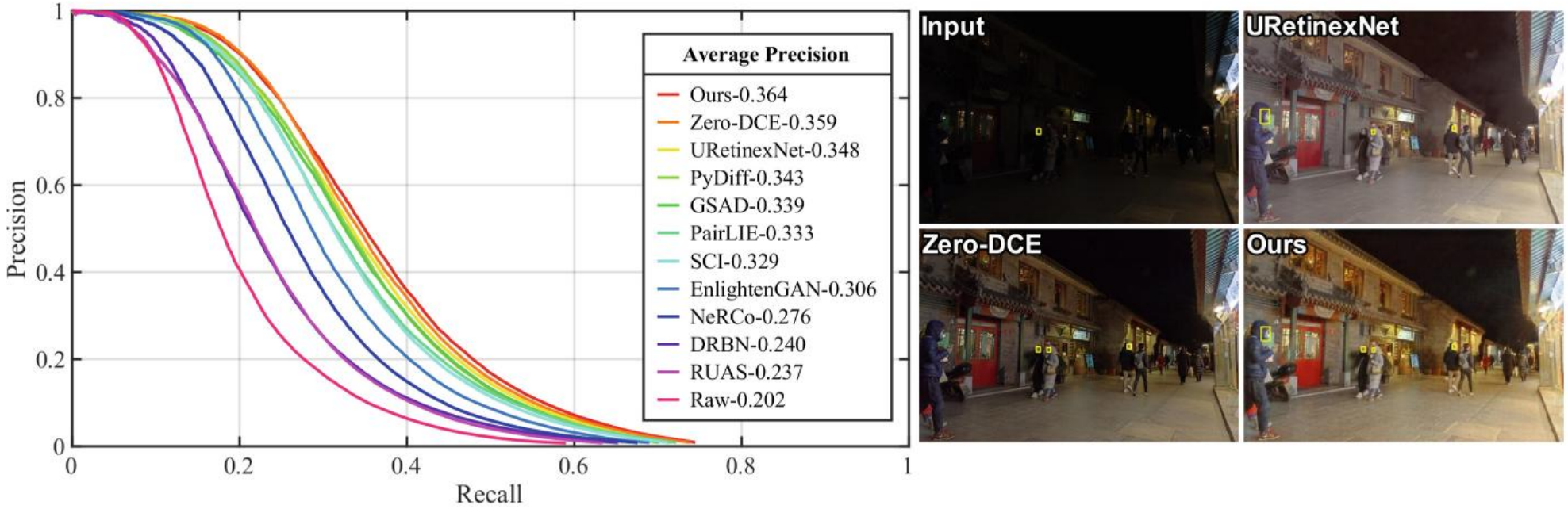
Qualitative comparison on the LOL and LSRW test sets.

Experiments

Qualitative comparison on the DICM, NPE, and VV datasets.

Experiments

Conduct experiments on the DARK FACE dataset

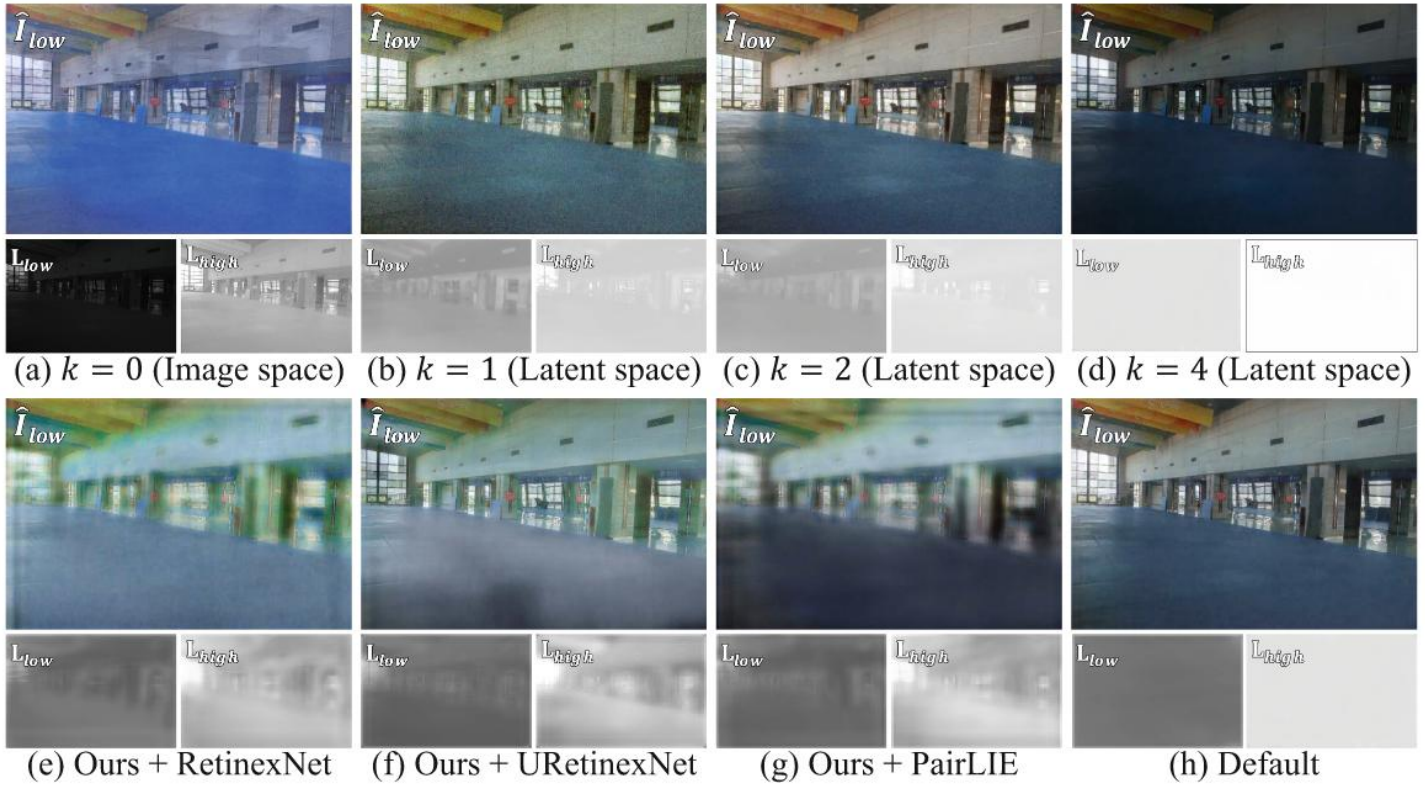


Comparison of low-light face detection results on the DARK FACE dataset

Ablation Study

	Method	LOL [58]			DICM [28]		Time (s) ↓
		PSNR ↑	SSIM ↑	LPIPS ↓	NIQE↓	PI↓	
1)	$k = 0$ (Image Space)	17.054	0.715	0.372	4.519	4.377	4.733
2)	$k = 1$ (Latent Space)	19.228	0.728	0.355	4.101	3.457	0.872
3)	$k = 2$ (Latent Space)	20.097	0.798	0.210	4.021	3.402	0.411
4)	$k = 4$ (Latent Space)	20.104	0.785	0.195	3.906	3.332	0.256
5)	RetinexNet [58]	16.616	0.563	0.579	5.859	6.056	0.296
6)	URetinexNet [59]	17.916	0.703	0.391	4.371	4.561	0.293
7)	PairLIE [10]	17.089	0.605	0.568	6.017	6.349	0.295
8)	w/o $\mathcal{L}_{scc}$ ($S = 20$)	19.184	0.785	0.213	4.045	3.408	0.314
9)	w/o $\mathcal{L}_{scc}$ ($S = 50$)	19.473	0.791	0.209	3.998	3.392	0.687
10)	w/o $\mathcal{L}_{scc}$ ($S = 100$)	20.255	0.801	0.209	3.831	3.228	1.208
11)	Default	20.453	0.803	0.192	3.724	3.144	0.314

Ablation Study



(a) $k = 0$ (Image space) (b) $k = 1$ (Latent space) (c) $k = 2$ (Latent space) (d) $k = 4$ (Latent space)

(e) Ours + RetinexNet (f) Ours + URetinexNet (g) Ours + PairLIE (h) Default

- It is difficult to achieve satisfactory decomposition in the image space
- Increasing k improves the overall performance and inference speed, while showing slight performance degradation at k = 4

- Previous decomposition networks are unable to obtain content-free illumination maps
- CTDN enables the generation of content-rich reflectance maps and content-free illumination maps

Ablation Study

(a) w/o $\mathcal{L}_{scc}$ $(S = 20)$ (b) w/o $\mathcal{L}_{scc}$ $(S = 50)$ (b) w/o $\mathcal{L}_{scc}$ $(S = 100)$ (d) w/ $\mathcal{L}_{scc}$ (default)

- Removing Lscc results in decreased overall performance
- Increase the sampling step S to 50 and 100 to evaluate the performance of the diffusion model trained with vanilla diffusion loss——the quality of generated results from diffusion models would improve with increasing S but slower inference speed

Thanks