Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhancement

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, Yulun Zhang
Tsinghua University, University of W¨urzburg, ETH Z¨urich

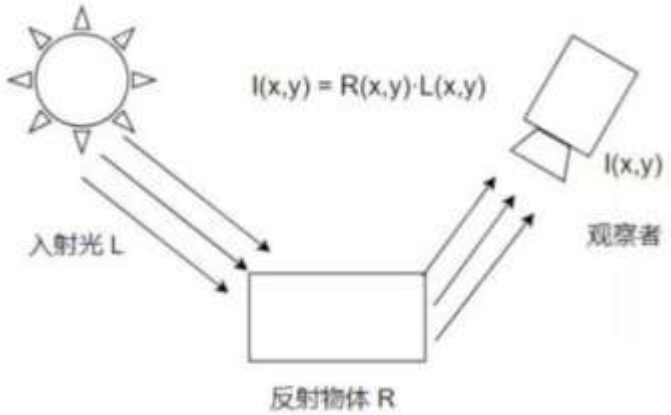
ICCV 2023

Introduction

Retinex theory： $I = R \odot L$

R: reflectance image ➡ contains the "real" content, details, and colors of the image

L: illumination map ➡ the intensity of light shining on an object



- the Retinex model does not consider the corruptions hidden in the dark or introduced by the light-up process.

➡

- We revise the original Retinex model by introducing perturbation terms to the reflectance and illumination for modeling the corruptions.

- Directly applying original vision Transformers for low-light image enhancement may encounter an issue. The computational complexity is quadratic to the input spatial size.

➡

- We design a new self-attention mechanism, IG-MSA, that utilizes the illumination information as a key clue to guide the modeling of long-range dependences.

Model the Corruption

Retinex theory： I = R ⊙ L R: reflectance image L: illumination map

Corruptions:
•the high-ISO and long-exposure imaging settings of dark scenes inevitably introduce noise and artifacts.
•the light-up process may amplify the noise and artifacts and also cause under-/over-exposure and color distortion

$$\mathbf{I} = (\mathbf{R} + \hat{\mathbf{R}}) \odot (\mathbf{L} + \hat{\mathbf{L}})$$
$$= \mathbf{R} \odot \mathbf{L} + \mathbf{R} \odot \hat{\mathbf{L}} + \hat{\mathbf{R}} \odot (\mathbf{L} + \hat{\mathbf{L}}),$$

$\hat{\mathbf{R}} \in \mathbb{R}^{H \times W \times 3}$ and $\hat{\mathbf{L}} \in \mathbb{R}^{H \times W}$ denote the perturbations

- Multiply the two sides of by a light-up map to light up： $\bar{\mathbf{L}} \odot \mathbf{L} = \mathbf{1}$

$$\mathbf{I} \odot \bar{\mathbf{L}} = \mathbf{R} + \mathbf{R} \odot (\hat{\mathbf{L}} \odot \bar{\mathbf{L}}) + (\hat{\mathbf{R}} \odot (\mathbf{L} + \hat{\mathbf{L}})) \odot \bar{\mathbf{L}}, \longrightarrow \mathbf{I}_{lu} = \mathbf{I} \odot \bar{\mathbf{L}} = \mathbf{R} + \mathbf{C}$$

$\hat{\mathbf{R}} \odot (\mathbf{L} + \hat{\mathbf{L}})$： noise and artifacts hidden in the dark scenes and are amplified
$\mathbf{R} \odot (\hat{\mathbf{L}} \odot \bar{\mathbf{L}})$： the under-/over-exposure and color distortion caused by the light-up process

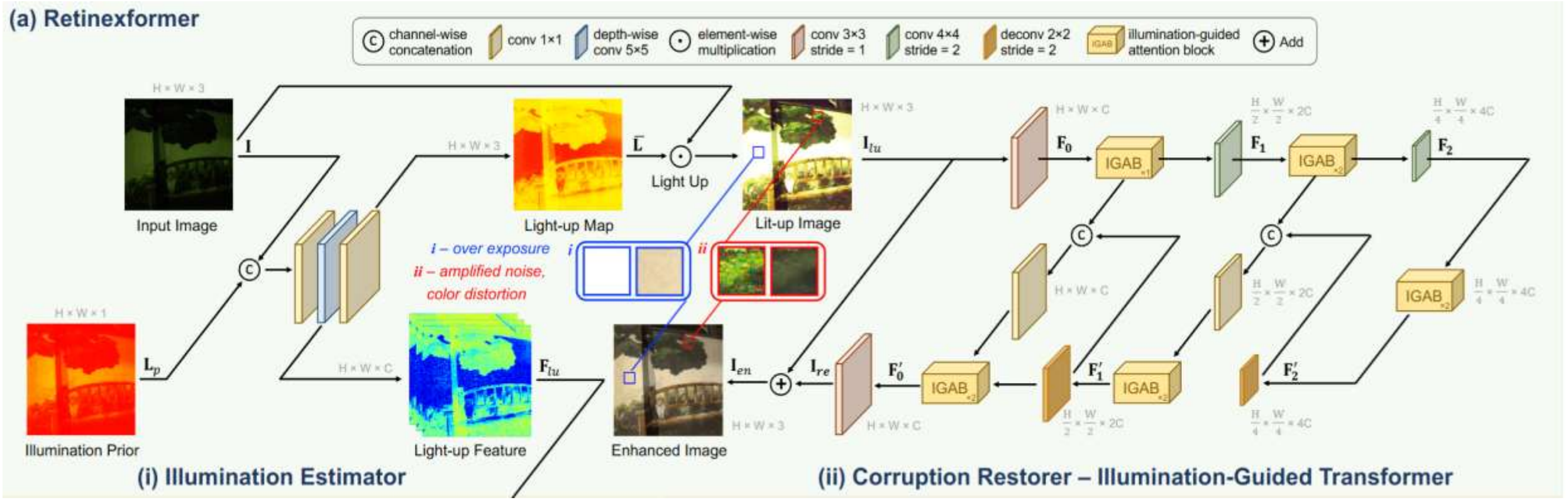
- One-stage Retinex-based Framework： $(\mathbf{I}_{lu}, \mathbf{F}_{lu}) = \mathcal{E}(\mathbf{I}, \mathbf{L}_p), \quad \mathbf{I}_{en} = \mathcal{R}(\mathbf{I}_{lu}, \mathbf{F}_{lu})$

$\mathcal{E}$ takes I and its illumination prior map $\mathbf{L}_p \in \mathbb{R}^{H \times W}$ as inputs

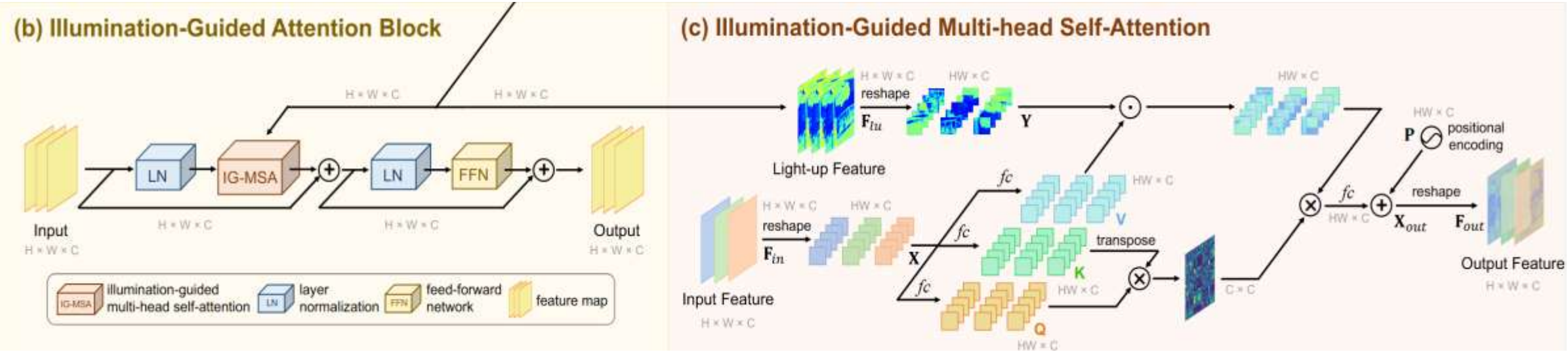
$\mathbf{L}_p = \text{mean}_c(\mathbf{I})$: the operation that calculates the mean values for each pixel along the channel dimension

$\mathbf{F}_{lu} \in \mathbb{R}^{H \times W \times C}$: model the interactions of regions with different lighting conditions

The Overview of The Retinexformer



IGAB&IG-MSA



$$\mathbf{F}_{in} \in \mathbb{R}^{H\times W\times C} \xrightarrow{\text{reshape}} \mathbf{X} \in \mathbb{R}^{HW\times C} \xrightarrow{\text{split}} \mathbf{X} = [\mathbf{X}_1,\ \mathbf{X}_2,\ \cdots,\ \mathbf{X}_k] \quad \mathbf{X}_i \in \mathbb{R}^{HW\times d_k}, d_k = \frac{C}{k}$$

For each head, three fully connected layers (fc) without bias are used to linearly project $\mathbf{X}_i$ into **query elements** $\mathbf{Q}_i \in \mathbb{R}^{HW\times d_k}$, **key elements** $\mathbf{K}_i \in \mathbb{R}^{HW\times d_k}$, and **value elements** $\mathbf{V}_i \in \mathbb{R}^{HW\times d_k}$

Regions with better lighting conditions can provide semantic contextual representations to help enhance the dark regions

⟹ use the light-up feature $\mathbf{F}_{lu}$ to direct the computation of self-attention

$$\mathbf{F}_{lu} \xrightarrow{\text{reshape}} \mathbf{Y} \in \mathbb{R}^{HW\times C} \xrightarrow{\text{split}} \mathbf{Y} = [\mathbf{Y}_1,\ \mathbf{Y}_2,\ \cdots,\ \mathbf{Y}_k] \quad \mathbf{Y}_i \in \mathbb{R}^{HW\times d_k}$$

IGAB&IG-MSA

Complexity

IG-MSA： $\text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i, \mathbf{Y}_i) = (\mathbf{Y}_i \odot \mathbf{V}_i)\text{softmax}(\frac{\mathbf{K}_i^{\mathrm{I}} \mathbf{Q}_i}{\alpha_i})$

the computational complexity mainly comes from $\mathbb{R}^{d_k \times HW} \times \mathbb{R}^{HW \times d_k}$ and $\mathbb{R}^{HW \times d_k} \times \mathbb{R}^{d_k \times d_k}$

$$\begin{aligned}\mathcal{O}(\text{IG-MSA}) &= k \cdot [d_k \cdot (d_k \cdot HW) + HW \cdot (d_k \cdot d_k)], \\ &= 2HWkd_k^2 = 2HWk(\frac{C}{k})^2 = \frac{2HWC^2}{k}.\end{aligned}$$

G-MSA(global MSA):

$$\mathcal{O}(\text{G-MSA}) = 2(HW)^2C$$

Experiments

Methods	Complexity		LOL-v1		LOL-v2-real		LOL-v2-syn		SID		SMID		SDSD-in		SDSD-out	
	FLOPS (G)	Params (M)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SID [9]	13.73	7.76	14.35	0.436	13.24	0.442	15.04	0.610	16.97	0.591	24.78	0.718	23.29	0.703	24.90	0.693
3DLUT [63]	0.075	0.59	14.35	0.445	17.59	0.721	18.04	0.800	20.11	0.592	23.86	0.678	21.66	0.655	21.89	0.649
DeepUPE [49]	21.10	1.02	14.38	0.446	13.27	0.452	15.08	0.623	17.01	0.604	23.91	0.690	21.70	0.662	21.94	0.698
RF [26]	46.23	21.54	15.23	0.452	14.05	0.458	15.97	0.632	16.44	0.596	23.11	0.681	20.97	0.655	21.21	0.689
DeepLPF [38]	5.86	1.77	15.28	0.473	14.10	0.480	16.02	0.587	18.07	0.600	24.36	0.688	22.21	0.664	22.76	0.658
IPT [11]	6887	115.31	16.27	0.504	19.80	0.813	18.30	0.811	20.53	0.561	27.03	0.783	26.11	0.831	27.55	0.850
UFormer [52]	12.00	5.29	16.36	0.771	18.82	0.771	19.66	0.871	18.54	0.577	27.20	0.792	23.17	0.859	23.85	0.748
RetinexNet [54]	587.47	0.84	16.77	0.560	15.47	0.567	17.13	0.798	16.48	0.578	22.83	0.684	20.84	0.617	20.96	0.629
Sparse [59]	53.26	2.33	17.20	0.640	20.06	0.816	22.05	0.905	18.68	0.606	25.48	0.766	23.25	0.863	25.28	0.804
EnGAN [22]	61.01	114.35	17.48	0.650	18.23	0.617	16.57	0.734	17.23	0.543	22.62	0.674	20.02	0.604	20.10	0.616
RUAS [30]	0.83	0.003	18.23	0.720	18.37	0.723	16.55	0.652	18.44	0.581	25.88	0.744	23.17	0.696	23.84	0.743
FIDE [56]	28.51	8.62	18.27	0.665	16.85	0.678	15.20	0.612	18.34	0.578	24.42	0.692	22.41	0.659	22.20	0.629
DRBN [58]	48.61	5.27	20.13	0.830	20.29	0.831	23.22	0.927	19.02	0.577	26.60	0.781	24.08	0.868	25.77	0.841
KinD [66]	34.99	8.02	20.86	0.790	14.74	0.641	13.29	0.578	18.02	0.583	22.18	0.634	21.95	0.672	21.97	0.654
Restormer [60]	144.25	26.13	22.43	0.823	19.94	0.827	21.41	0.830	22.27	0.649	26.97	0.758	25.67	0.827	24.79	0.802
MIRNet [61]	785	31.76	24.14	0.830	20.02	0.820	21.94	0.876	20.84	0.605	25.66	0.762	24.38	0.864	27.13	0.837
SNR-Net [57]	26.35	4.01	24.61	0.842	21.48	0.849	24.14	0.928	22.87	0.625	28.49	0.805	29.44	0.894	28.66	0.866
Retinexformer	15.57	1.61	25.16	0.845	22.80	0.840	25.67	0.930	24.44	0.680	29.15	0.815	29.77	0.896	29.84	0.877

Table 1. Quantitative comparisons on LOL (v1 [54] and v2 [59]), SID [9], SMID [10], and SDSD [48] (indoor and outdoor) datasets. The highest result is in red color while the second highest result is in blue color. Our Retinexformer significantly outperforms SOTA algorithms.

Methods	DeepUPE [49]	MIRNet [61]	SNR-Net [57]	Restormer [60]	**Ours**
PSNR (dB)	23.04	23.73	23.81	24.13	**24.94**
FLOPS (G)	21.10	785.0	26.35	144.3	**15.57**

Experiments

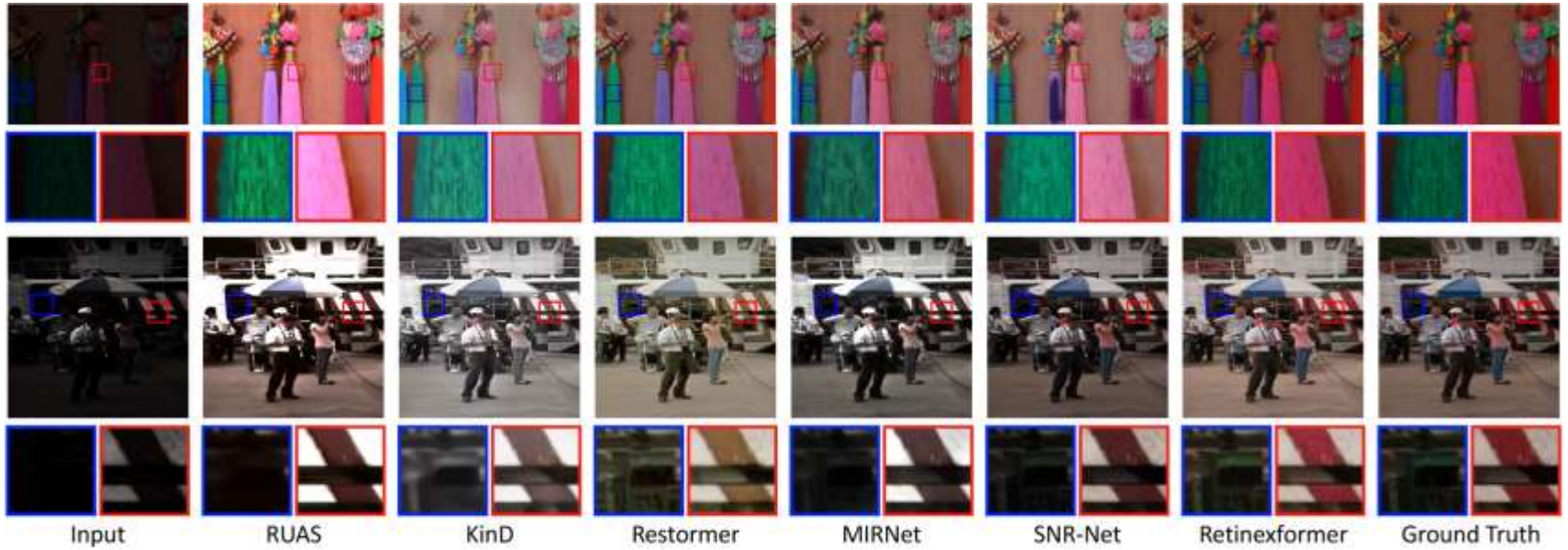


Figure 3. Results on LOL-v1 [54] (top) and LOL-v2 [59] (bottom). Our method effectively enhances the visibility and preserves the color.

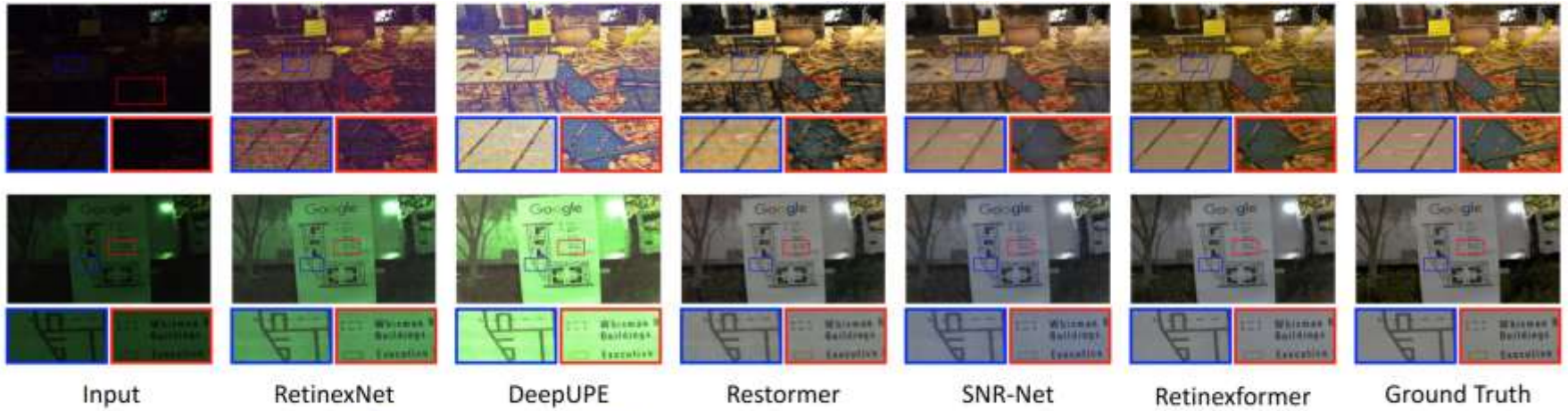


Figure 4. Visual results on SID [9] (top) and SMID [10] (bottom). Previous methods either collapse by noise, or distort color, or produce blurry and under-/over-exposed images. While our algorithm can effectively remove the noise and reconstruct well-exposed image details.

Experiments

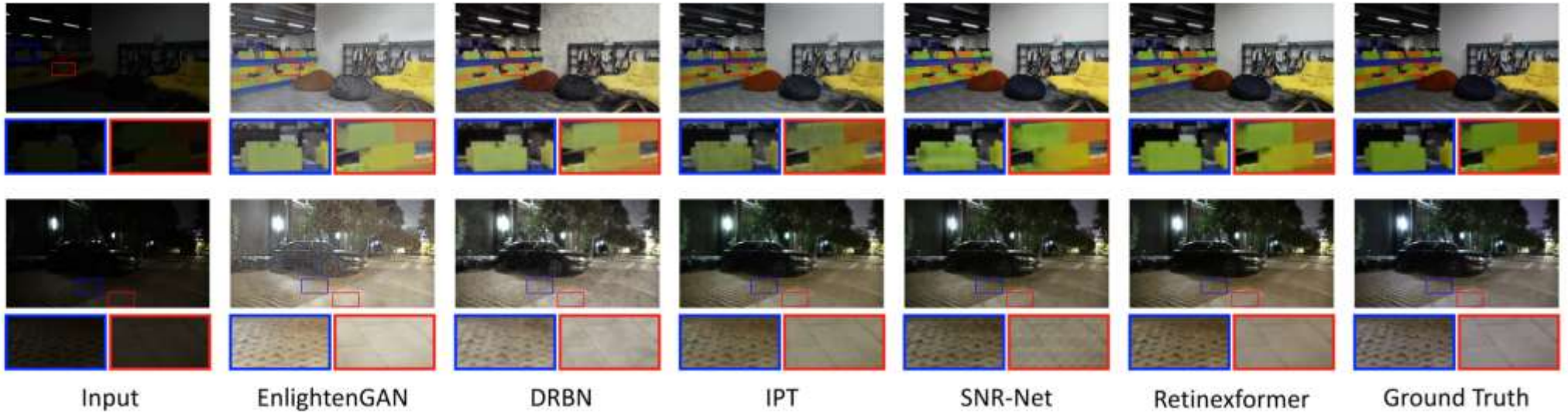


Figure 5. Visual resulst on SDSD [48]-indoor (top) and out-door (bottom). Other algorithms either generate over-exposed and noisy images, or introduce black spot corruptions and unnatural artifacts. While Retinexformer can restore well-exposed structural contents and textures.

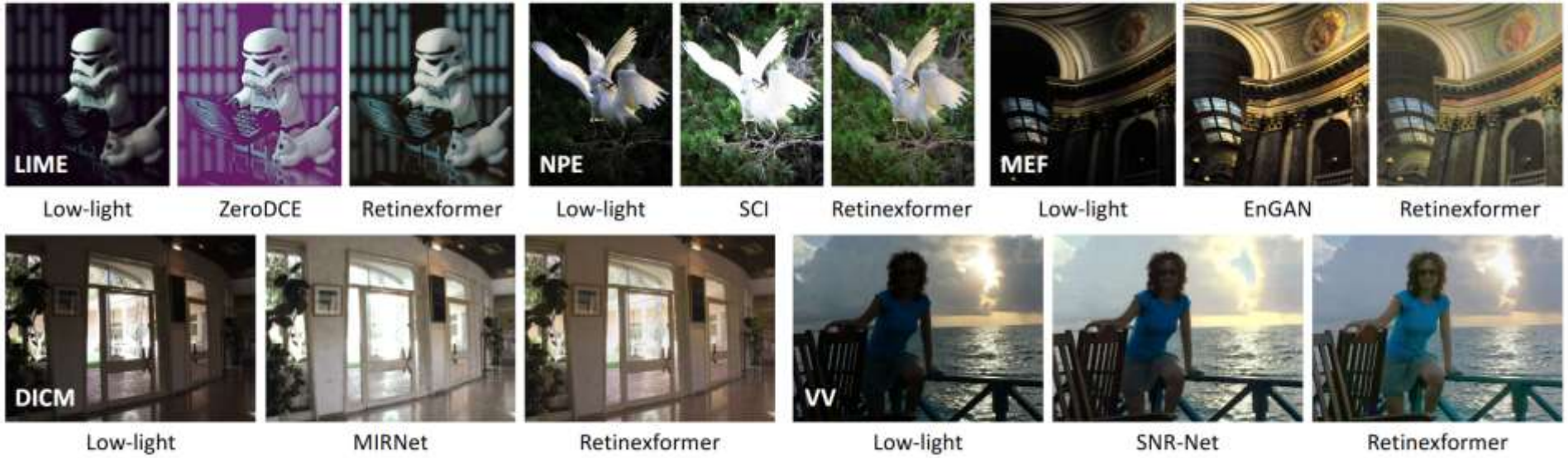


Figure 7. Visual results on the LIME [18], NPE [50], MEF [36], DICM [28], and VV [47] datasets. Our Retinexformer performs better.

Experiments

Low-light Object Detection

- Experiment Settings

  DataSet：ExDark
  5890 images training -1473 images testing

  YOLO-v3 is employed as the detector trained from scratch

Methods	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motor	People	Table	Mean
MIRNet [61]	71.8	63.8	62.9	81.4	71.1	58.8	58.9	61.3	63.1	52.0	68.8	45.5	63.6
RetinexNet [54]	73.8	62.8	64.8	84.9	80.8	53.4	57.2	68.3	61.5	51.3	65.9	43.1	64.0
RUAS [30]	72.0	62.2	65.2	72.9	78.1	57.3	62.4	61.8	60.2	61.5	69.4	46.8	64.2
Restormer [60]	76.2	65.1	64.2	84.0	76.3	59.2	53.0	58.7	66.1	62.9	68.6	45.0	64.9
KinD [66]	72.2	66.5	58.9	83.7	74.5	55.4	61.7	61.3	63.8	63.0	70.5	47.8	65.0
ZeroDCE [17]	75.8	66.5	65.6	84.9	77.2	56.3	53.8	59.0	63.5	64.0	68.3	46.3	65.1
SNR-Net [57]	75.3	64.4	63.6	85.3	77.5	59.1	54.1	59.6	66.3	65.2	69.1	44.6	65.3
SCI [37]	74.6	65.3	65.8	85.4	76.3	59.4	57.1	60.5	65.6	63.9	69.1	45.9	65.6
Retinexformer	76.3	66.7	65.9	84.7	77.6	61.2	53.5	60.7	67.5	63.4	69.5	46.0	66.1

(b) Low-light detection results on ExDark [32] enhanced by different algorithms.

Figure 6. Visual comparison of object detection in low-light (left) and enhanced (right) scenes by our method on the Exdark dataset.

Ablation Study

- Break-down Ablation

  Baseline-1 is derived by removing ORF and IG-MSA from Retinexformer

Baseline-1	ORF	IG-MSA	PSNR	SSIM	Params (M)	FLOPS (G)
✓			26.47	0.843	1.01	9.18
✓	✓		27.92	0.857	1.27	11.37
✓		✓	28.86	0.868	1.34	13.38
✓	✓	✓	**29.84**	**0.877**	1.61	15.57

- One-stage Retinex-based Framework

  Remove ORF from Retinexformer and set the input of $\mathcal{R}$ as $\mathbf{I}_{lu} = \mathbf{I}$

Method	$\mathbf{I}_{lu} = \mathbf{I}$	$\mathbf{I}_{lu} = \mathbf{I}./\mathbf{L}$	$\mathbf{I}_{lu} = \mathbf{I} \odot \bar{\mathbf{L}}$	$+\mathbf{F}_{lu}$
PSNR	28.86	28.97	29.26	**29.84**
SSIM	0.868	0.868	0.870	**0.877**
Params (M)	1.34	1.61	1.61	1.61
FLOPS (G)	13.38	14.01	14.01	15.57

- Self-Attention Scheme

  Baseline-2： removing IG-MSA from Retinexformer

  G-MSA： used by previous CNN-Transformer hybrid methods，downscaled into 1/4 size

  W-MSA： window-based MSA proposed by Swin Transformer

Method	Baseline-2	G-MSA	W-MSA	IG-MSA
PSNR	27.92	28.43	28.65	**29.84**
SSIM	0.857	0.841	0.845	**0.877**
Params (M)	1.27	1.61	1.61	1.61
FLOPS (G)	11.37	17.65	16.43	15.57

Thanks