Evaluating Low-Light Image Enhancement Across Multiple Intensity Levels Supplementary Material

Maria Pilligua[1,2] David Serrano-Lozano[1,2] Pai Peng[3]
Ramon Baldrich[1,2] Michael S. Brown[4] Javier Vazquez-Corral[1,2]
[1]Computer Vision Center [2]Universitat Autònoma de Barcelona
[3]University of Wisconsin-Madison [4]York University
https://color.cvc.uab.cat/mill

We provide additional material to supplement our main submission, covering:

A. Extended Motivation.
B. Dataset Setup Details.
C. Additional Quantitative Results.
D. New Loss Terms in Other Baselines.
E. Additional Qualitative Results.

A. Extended Motivation

Existing Low-Light Image Enhancement (LLIE) datasets present a critical limitation: they either contain only a single severely underexposed image per scene, or they simulate brightness variations through camera parameter adjustments or post-processing operations. These constraints limit real-world applicability, where low-light conditions span a continuous range of intensities.

In the main submission, we demonstrate how Retinexformer [2] and HVI-CIDNet [10] exhibit degraded performance on the LoLv1 dataset when input image brightness is increased. Figure 1 provides additional examples with corresponding histograms to further demonstrate how this inherent limitation of existing LLIE datasets hinders the applicability of enhancement methods in real-world scenarios. For each example, the top row displays input images with varying brightness levels: the original image alongside versions blended with the ground truth at ratios of 0.2 and 0.5. The bottom row shows the corresponding HVI-CIDNet [10] outputs, with RGB histograms displayed in the bottom-left corner of each image.

In the first example, we observe that higher input intensity leads to increasingly saturated outputs. The output histograms reveal severe oversaturation, particularly evident in the large white regions of the image. The second example presents a more complex behavior: oversaturation is non-monotonic, with the 20% blend producing more saturation than the 50% blend. This non-linear response indicates that predicting when outputs will become oversaturated is challenging and depends on the specific image content and brightness level. Finally, the third example demonstrates how the model oversaturates the bright regions while attempting to enhance darker areas.

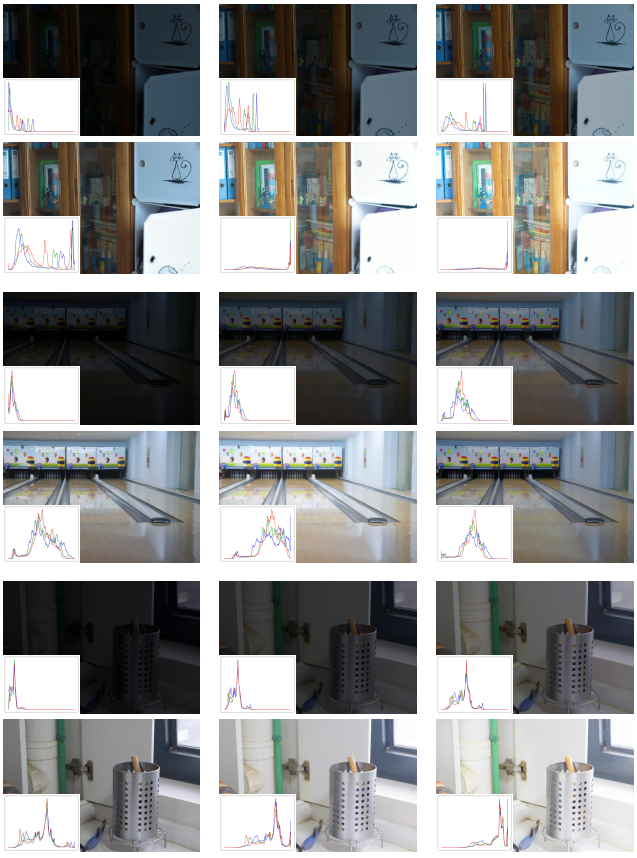

Figure 1. Impact of brightness variation on LLIE model performance. Blending input images with ground truth at 20% and 50% ratios degrades CIDNet [10] performance.

LLIE research has mainly focused on architectural improvements to enhance performance on existing bench-

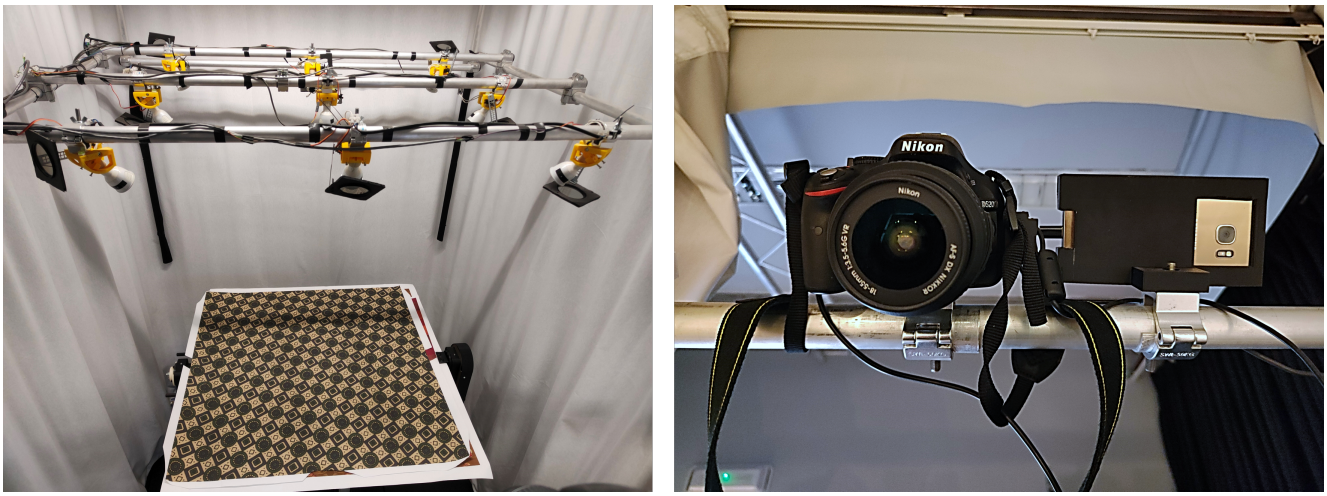

Figure 2. MILL dataset capture setup. Left: The capture platform with metallic overhead structure supporting programmable lighting arrays. Right: DSLR and smartphone cameras mounted on the structure and directed toward the platform.

(a) Training objects

(b) Validation objects

(c) Test objects

Figure 3. Objects used to construct the MILL dataset, separated according to their respective dataset splits.

(a) Training scenes

(b) Validation scenes

(c) Test scenes

Figure 4. Representative scenes from the MILL dataset, separated by splits.

marks. While these methods typically perform well on standard datasets, the datasets themselves have received considerably less attention. Through this analysis, we demonstrate that achieving robust LLIE methods requires equal focus on dataset design. This motivation drives our introduction of the Multi-Illuminant Low-Light (MILL) dataset. MILL enables systematic evaluation of current methods across different brightness intensities and provides a foundation for training more robust models, thereby advancing both LLIE research and the practical applicability of enhancement methods.

B. Dataset Setup Details

Acquiring images under a range of light intensities requires a controlled environment. We capture the MILL dataset in a room without windows or external light sources to eliminate uncontrolled illumination. As shown in Figure 2, our setup consists of a platform with a fixed metallic structure supporting nine programmable lights, a DSLR camera, and

a smartphone. The platform accepts interchangeable floor backgrounds and allows object placement under controlled conditions.

Figure 3 displays the objects used in MILL, organized by training, validation, and test splits. As mentioned in the main paper, we ensure that i) no objects appear in multiple splits, and ii) backgrounds are not shared between train/validation and test; thus maintaining strict separation to evaluate proper performance. Figure 4 shows representative images from each split, demonstrating the diversity achieved through varied object positions, orientations, and spatial arrangements.

C. Additional Quantitative Results

Tables 1 and 2 report the PSNR_L, SSIM, LPIPS, and ΔE_{76} metrics across all 10 illumination levels of the DSLR split in MILL-s.

D. New Loss Terms in Other Baselines

Table 3 presents additional results for the experiment in the camera split of the MILL-f dataset, including an ablation study. We report PSNR_L, PSNR_C, SSIM, and ΔE_{76} for a baseline LLIE model, with our scene content loss (S) and intensity prediction loss (I) added independently, as well as the combined loss (SI). We include DarkIR [3] and HVI-CIDNet [10] as representative state-of-the-art LLIE architectures. Each loss term independently improves performance; furthermore, combining both terms yields the best results, as effective disentanglement cannot be achieved using either term alone.

Retinexformer achieves the best performance on MILL-f and shows the greatest improvement when augmented with our loss terms (see red increments in Table 3). We hypothesize that this is because both DarkIR and HVI-CIDNet incorporate highly specialized components for LLIE, such as dedicated color space transformations and task-specific losses (edge losses, guiding losses, or losses in alternative color spaces). In contrast, Retinexformer serves as a powerful general-purpose baseline without such domain-specific design choices, making it more receptive to our proposed loss terms.

E. Additional Qualitative Results

Figure 7 shows a MILL-s scene captured at the first four illumination levels, with enhancement results from PromptNorm [8], Retinexformer [2], and our approach. For each image, we provide a zoomed-in region displayed below. Our method produces results closer to the ground truth, as evidenced by the accurate color reproduction in the red and yellow pen and the orange mug. Note that at Level 1 (the most challenging condition), all methods produce outputs that deviate considerably from the ground truth. However, as the illumination level increases, all methods improve, benefiting from the additional information preserved in the input images.

Figure 8 presents results on one MILL-f scene and one MILL-s scene, each evaluated at two illumination levels (Level 1 and Level 4). We compare Retinexformer [2], our loss terms applied independently, and our combined approach. Zoomed-in crops of visually salient regions are displayed below each result. Our combined approach achieves superior quality at Level 1 for both examples, while at Level 4 it produces sharper details and more accurate color reproduction compared to the ground truth.

Figure 6 presents qualitative comparisons between Retinexformer [2] and our approach on outdoor images from the SICE [1] and LIME [4] datasets. Both methods are trained on MILL-s. Figure 5 demonstrates the generalization capability of our method on in-the-wild outdoor images.

References

[1] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE TIP*, 27(4):2049–2062, 2018. 3, 6

[2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023. 1, 3, 4, 5, 6, 7, 8

[3] Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *CVPR*, 2025. 3, 4, 5

[4] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016. 3, 6

[5] Jingxi Liao, Shijie Hao, Richang Hong, and Meng Wang. Gt-mean loss: A simple yet effective solution for brightness mismatch in low-light image enhancement. In *ICCV*, 2025. 4, 5

[6] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 4, 5

[7] Liu Risheng, Ma Long, Zhang Jiaao, Fan Xin, and Luo Zhongxuan. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 4, 5

[8] David Serrano-Lozano, Francisco A Molina-Bakhos, Danna Xue, Yixiong Yang, Maria Pilligua, Ramon Baldrich, Maria Vanrell, and Javier Vazquez-Corral. Promptnorm: Image geometry guides ambient light normalization. In *CVPR Workshops*, 2025. 3, 4, 5, 7

[9] Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *ACM MM*, 2023. 4, 5

[10] Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *CVPR*, 2025. 1, 3, 4, 5

Table 1. Levels 1 to 9 of the DSLR MILL-s dataset.

DSLR	Level 1				Level 2				Level 3			
	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$
Unprocessed	13.457	0.132	0.483	30.337	15.884	0.443	0.258	23.108	17.681	0.623	0.157	18.621
RUAS [7]	16.635	0.316	0.457	25.462	13.068	0.438	0.394	36.526	9.861	0.426	0.399	45.332
LLFormer [12]	20.881	0.712	0.378	16.366	21.636	0.804	0.201	14.206	21.790	0.838	0.154	13.735
KinD [13]	16.706	0.571	0.439	23.884	17.900	0.647	0.324	21.865	19.756	0.738	0.247	17.615
FourLLIE [9]	17.287	0.434	0.458	24.510	19.726	0.671	0.306	21.114	17.658	0.741	0.240	22.791
SCI [6]	16.020	0.285	0.460	24.049	23.577	0.617	0.304	16.431	21.423	0.696	0.243	17.994
MirNet [11]	26.458	0.769	0.312	14.030	25.426	0.848	0.167	11.811	25.336	0.877	0.133	11.106
Retinexformer [2]	25.092	0.742	0.335	14.147	25.945	0.838	0.180	11.969	26.390	0.881	0.137	10.449
DarkIR [3]	24.651	0.736	0.336	14.392	25.522	0.833	0.182	12.367	25.233	0.873	0.139	11.293
CIDNet [10]	24.080	0.725	0.340	14.781	23.800	0.812	0.195	13.606	22.488	0.845	0.153	13.713
PromptNorm [8]	25.886	0.770	0.310	13.471	25.973	0.854	0.164	11.507	26.065	0.888	0.128	10.513
GT-Mean [5]	24.320	0.731	0.338	14.593	24.514	0.822	0.189	13.127	23.760	0.860	0.145	12.478
Ours	25.534	0.754	0.348	13.896	30.256	0.863	0.167	9.918	31.474	0.897	0.126	9.113

DSLR	Level 4				Level 5				Level 6			
	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$
Unprocessed	19.480	0.742	0.104	15.024	21.485	0.829	0.068	11.897	23.341	0.880	0.047	9.649
RUAS [7]	8.029	0.394	0.434	52.381	6.931	0.362	0.475	57.468	6.301	0.343	0.510	60.609
LLFormer [12]	21.918	0.854	0.134	13.516	22.061	0.868	0.123	13.343	22.158	0.874	0.118	13.246
KinD [13]	20.656	0.789	0.208	15.814	21.340	0.821	0.183	14.872	21.679	0.836	0.170	14.506
FourLLIE [9]	15.732	0.731	0.218	25.400	14.968	0.730	0.203	26.642	14.514	0.725	0.198	27.660
SCI [6]	17.983	0.698	0.228	21.918	15.691	0.681	0.228	25.890	14.314	0.661	0.233	28.892
MirNet [11]	24.974	0.885	0.121	11.295	24.814	0.895	0.115	11.394	24.605	0.899	0.113	11.551
Retinexformer [2]	26.548	0.894	0.122	10.463	26.548	0.907	0.113	10.353	26.476	0.912	0.108	10.389
DarkIR [3]	24.809	0.883	0.123	11.644	24.742	0.896	0.113	11.579	24.238	0.898	0.110	11.918
CIDNet [10]	21.670	0.851	0.137	14.508	21.442	0.863	0.126	14.585	21.049	0.864	0.122	14.965
PromptNorm [8]	26.090	0.899	0.116	10.568	25.944	0.910	0.109	10.591	25.777	0.913	0.106	10.716
GT-Mean [5]	22.910	0.866	0.130	13.252	22.801	0.879	0.119	13.226	22.571	0.882	0.115	13.383
Ours	31.360	0.895	0.114	9.091	31.517	0.908	0.107	9.088	32.094	0.917	0.101	9.017

DSLR	Level 7				Level 8				Level 9			
	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$
Unprocessed	25.557	0.915	0.030	7.595	28.553	0.942	0.022	6.132	32.063	0.952	0.020	5.318
RUAS [7]	5.859	0.329	0.540	62.992	5.507	0.318	0.577	65.126	5.478	0.313	0.583	65.994
LLFormer [12]	22.247	0.879	0.114	13.170	22.406	0.885	0.110	13.005	22.727	0.897	0.107	12.612
KinD [13]	21.627	0.842	0.161	14.493	21.680	0.851	0.155	14.658	21.531	0.856	0.150	14.879
FourLLIE [9]	14.069	0.716	0.195	28.788	13.802	0.711	0.193	29.569	13.503	0.710	0.194	30.625
SCI [6]	13.244	0.640	0.239	31.655	12.291	0.620	0.248	34.747	11.711	0.615	0.254	36.754
MirNet [11]	24.488	0.902	0.111	11.650	25.040	0.910	0.107	11.533	25.334	0.919	0.103	11.513
Retinexformer [2]	26.480	0.916	0.104	10.414	27.156	0.925	0.100	10.392	27.658	0.934	0.096	10.335
DarkIR [3]	23.910	0.900	0.106	12.213	24.533	0.909	0.102	12.113	24.793	0.919	0.100	12.088
CIDNet [10]	20.829	0.865	0.119	15.222	20.565	0.866	0.119	15.836	20.890	0.880	0.115	15.549
PromptNorm [8]	25.661	0.916	0.105	10.817	26.300	0.924	0.101	10.750	26.721	0.932	0.098	10.732
GT-Mean [5]	22.194	0.882	0.113	13.797	22.473	0.891	0.109	13.790	22.313	0.898	0.108	14.054
Ours	32.314	0.924	0.097	8.942	32.580	0.932	0.095	9.291	34.158	0.934	0.094	9.172

[11] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020. 4, 5

[12] Kaihao Zhang, Dongxu Li, Wenhan Luo, Wenqi Ren, Bjorn Stenger, Wei Liu, Hongdong Li, and Ming-Hsuan Yang. Benchmarking ultra-high-definition image super-resolution. In *ICCV*, 2021. 4, 5

[13] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM*

Table 2. Levels 10 of the DSLR MILL-s dataset.

DSLR	Level 10			
	$PSNR_L \uparrow$	SSIM ↑	LPIPS ↓	$\Delta E_{76} \downarrow$
Unprocessed	36.642	0.965	0.014	3.620
RUAS [7]	5.184	0.309	0.615	67.143
LLFormer [12]	22.552	0.892	0.108	12.898
KinD [13]	21.360	0.854	0.148	14.997
FourLLIE [9]	13.353	0.700	0.196	30.970
SCI [6]	11.231	0.594	0.262	38.379
MirNet [11]	24.960	0.916	0.105	11.716
Retinexformer [2]	27.412	0.932	0.098	10.457
DarkIR [3]	24.635	0.917	0.101	12.155
CIDNet [10]	20.631	0.874	0.116	15.847
PromptNorm [8]	26.281	0.930	0.100	10.889
GT-Mean [5]	22.877	0.902	0.105	13.571
Ours	32.476	0.937	0.094	9.170

Table 3. Quantitative comparison on the MILL-f dataset. We compare DarkIR, CIDNet and Retinexformer with our loss terms added independently (S and I) and our combined approach. We also report the gains with respect to the corresponding baseline.

DSLR	$PSNR_L \uparrow$	$PSNR_C \uparrow$	SSIM ↑	$\Delta E_{76} \downarrow$
DarkIR [3]	24.92	21.87	0.879	11.35
S-DarkIR	25.09	21.98	0.881	10.99
I-DarkIR	25.92	23.04	0.889	10.20
SI-DarkIR	26.83	23.80	0.896	9.03
	(+1.91)	(+1.93)	(+0.02)	(-2.32)
CIDNet [10]	22.58	20.49	0.857	13.76
S-CIDNet	23.16	21.36	0.864	12.71
I-CIDNet	25.87	23.61	0.880	11.04
SI-CIDNet	26.54	24.52	0.884	10.69
	(+3.96)	(+4.03)	(+0.03)	(-3.07)
Retinexformer [2]	27.47	25.41	0.895	8.27
S-Retinexformer	28.45	26.31	0.905	7.48
I-Retinexformer	36.36	33.09	0.924	4.25
Ours	37.55	34.05	0.929	3.67
	(+10.08)	(+8.64)	(+0.03)	(-4.60)

MM, 2019. 4, 5

Figure 5. Enhancement results of our approach on outdoor low-light images captured in the wild.

Input Retinexformer [2] Ours

(a) SICE [1]

Input Retinexformer [2] Ours

(b) LIME [4]

Figure 6. Outdoor images from SICE and LIME datasets. We display results from Retinexformer and our approach trained on MILL.

Figure 7. First four levels of a scene of MILL with results from PromptNorm [8], Retinexformer [2] and our approach.

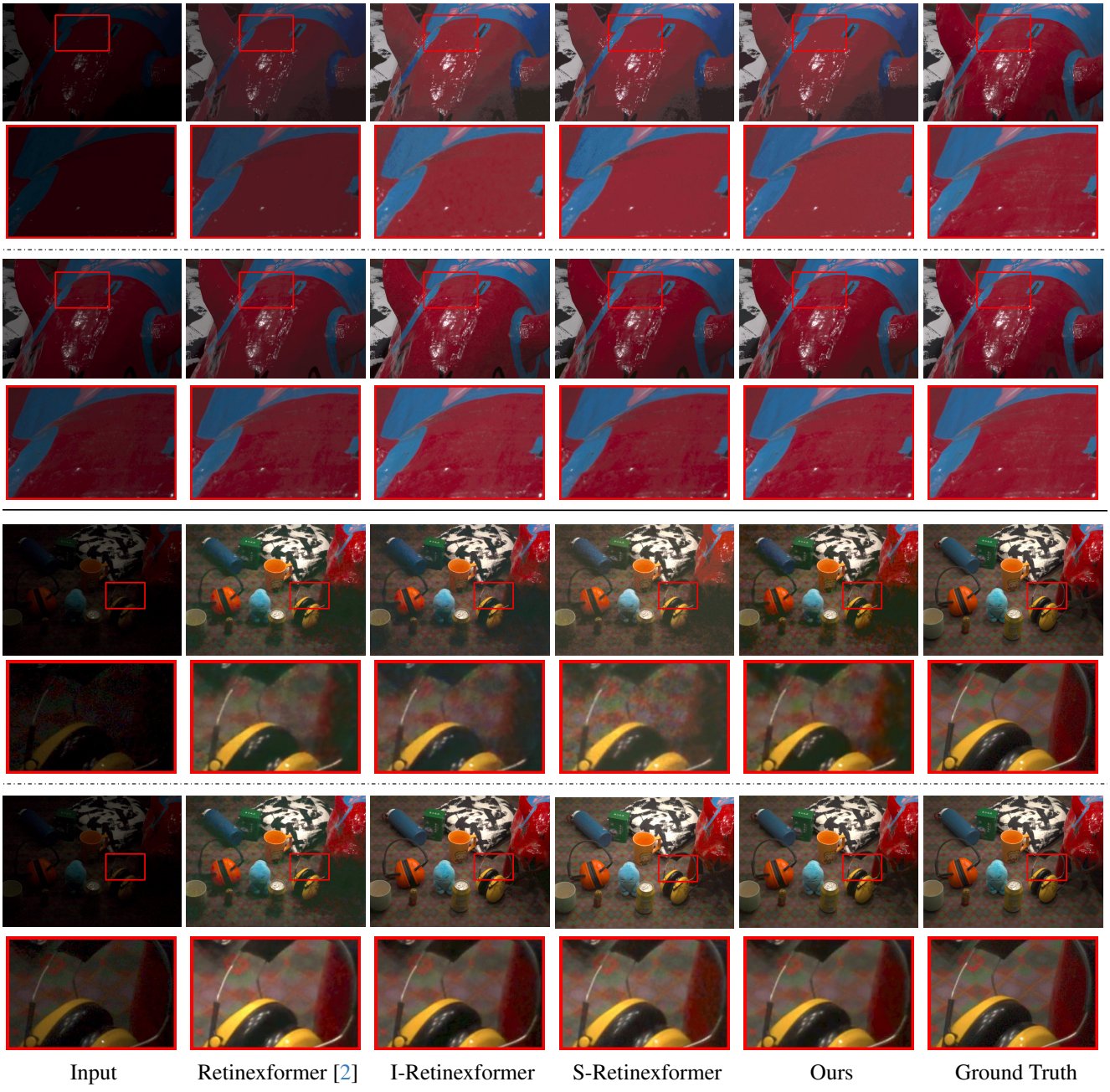

Figure 8. Qualitative comparison on MILL-f (top) and MILL-s (bottom) at Level 1 and Level 4. We compare Retinexformer [2] with our loss terms applied independently (S and I) and combined (Ours).