

Evaluating Low-Light Image Enhancement Across Multiple Intensity Levels

Maria Pilligua^{1,2} David Serrano-Lozano^{1,2} Pai Peng³
 Ramon Baldrich^{1,2} Michael S. Brown⁴ Javier Vazquez-Corral^{1,2}

¹Computer Vision Center ²Universitat Autònoma de Barcelona

³University of Wisconsin-Madison ⁴York University

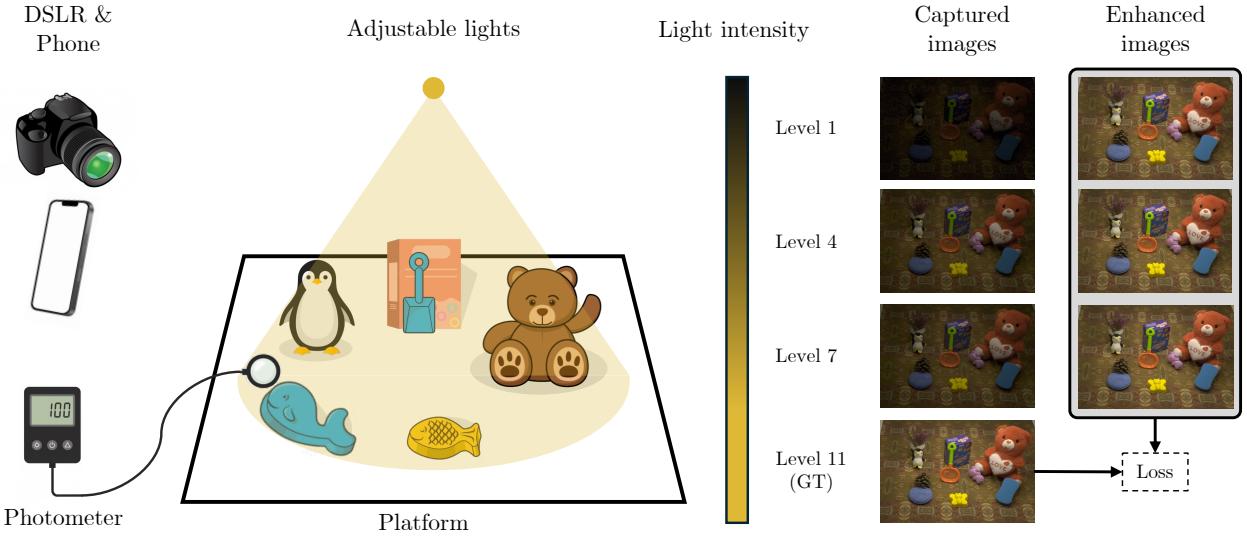


Figure 1. Illustration of our capture setup. A set of controllable lights illuminates the scene. For each scene, we capture 11 images by varying the light intensity from minimum to maximum at 10% intervals (levels). Camera parameters (aperture, exposure time, ISO) remain fixed, and images are captured in unprocessed RAW format. The image captured at maximum intensity serves as the ground truth, while all other images serve as low-light inputs for training, validation, and testing. Scenes are captured with both a DSLR and a smartphone, and scene illuminance is measured with a photometer.

Abstract

Imaging in low-light environments is challenging due to reduced scene radiance, which leads to elevated sensor noise and reduced color saturation. Most learning-based low-light enhancement methods rely on paired training data captured under a single low-light condition and a well-lit reference. The lack of radiance diversity limits our understanding of how enhancement techniques perform across varying illumination intensities. We introduce the Multi-Illumination Low-Light (MILL) dataset, containing images captured at diverse light intensities under controlled conditions with fixed camera settings and precise illuminance measurements. MILL enables comprehensive evaluation

of enhancement algorithms across variable lighting conditions. We benchmark several state-of-the-art methods and reveal significant performance variations across intensity levels. Leveraging the unique multi-illumination structure of our dataset, we propose improvements that enhance robustness across diverse illumination scenarios. Our modifications achieve up to 10 dB PSNR improvement for DSLR and 2 dB for the smartphone on Full HD images.

1. Introduction

Images taken in low-light environments are corrupted by sensor noise and diminished color saturation. Simple digital exposure adjustments, such as scaling the image's digital values, result in poor image quality due to high levels

of sensor noise. Consequently, deep learning techniques have been developed to directly enhance low-light images, efficiently reducing noise and improving color and texture (e.g., [2, 23, 33, 42, 44]). The success of these approaches is heavily dependent on how the training data is collected.

Existing low-light image enhancement (LLIE) datasets obtain paired data either by varying camera settings or through post-processing, but nearly all capture images under a single low-light condition. This fails to reflect real-world scenarios where low-light images span a wide range of brightness levels, limiting the robustness of LLIE methods when deployed in practice.

To address this limitation, we present the Multi-Illumination Low-Light (MILL) dataset. Unlike existing datasets, MILL captures the same scene under 11 systematically varied light intensities, ranging from minimum to maximum brightness with equispaced intervals, while maintaining fixed camera parameters (see Fig. 1). Each capture is accompanied by precise illuminance measurements (lux) from a calibrated sensor and the input parameters of the programmable lights. We use the maximum-intensity image as ground truth and the remaining 10 images as low-light inputs. All images are captured in RAW format, ensuring no camera-processed artifacts.

Using the MILL dataset, we analyze how current state-of-the-art methods perform under varying lighting conditions and find that model performance varies significantly across different intensity ranges. Based on our findings, we propose an improvement over the best-performing method, Retinexformer [2]. We propose to disentangle scene and illumination information in the network’s latent features by leveraging the multi-level nature of our dataset. We demonstrate that our simple modification improves PSNR by 10 dB for the DSLR and 2 dB for the smartphone camera on Full HD images.

Our contributions can be summarized as follows:

- We introduce **MILL**, a new low-light image enhancement dataset in which each scene is captured at 11 distinct illumination levels. Every image is paired with a photometer-measured lux value and the corresponding input setting of the programmable lights. Using fixed camera parameters on both a DSLR and a smartphone, we collected a total of 1100 images.
- We benchmark several state-of-the-art enhancement methods on our dataset to evaluate their robustness across a broad range of illumination levels. Our analysis reveals that certain methods display unexpected performance fluctuations at different intensity ranges.
- We further propose two loss terms that exploit the auxiliary illumination information (i.e., intensity level) provided by our dataset. Integrating these terms leads to substantial improvements over prior state-of-the-art models.

2. Related Work

2.1. Low-Light Datasets

Early LLIE datasets, such as VV [26] and LIME [8], contained only unpaired low-light images (15 and 10 samples, respectively) without corresponding well-exposed references. For this reason, Wei et al. [35] introduced the LLow Light paired dataset (LOL) to allow end-to-end training of LLIE models. The LOL dataset proved valuable to the research community, enabling end-to-end training of methods. LoLv1 contains images captured under different camera settings to capture the same scene under low-light and well-lit conditions. The LOLv1 dataset consists of 500 images, of which 485 are for training and 15 are for testing. An extension of the LoLv1 dataset, LoLv2, was later introduced [38]. In LoLv2, the authors introduced two variants: one following the LoLv1 methodology and the other generating the low-light image synthetically from the well-lit counterpart. LoLv2 contains 689 training scenes and 100 test scenes. A major issue with these datasets is that they contain images from the same scene in both the training and test sets, potentially affecting generalization.

Several datasets were subsequently introduced to address the limitations of early LLIE benchmarks, primarily in terms of scale and diversity. The DPED [11] dataset provided images captured across multiple smartphone cameras, enabling cross-device evaluation. Deep-UPE [29] emphasized extremely low-light scenarios with more challenging exposure conditions. The LSRW [9] dataset expanded camera diversity by including both DSLR and smartphone captures, recognizing the distinct image formation characteristics of different sensor types. More recently, LLIV-Phone [15] introduced temporal information by capturing video sequences under low-light conditions, allowing methods to exploit inter-frame correlations. This video-based approach was further explored in the DID [5] and SDSD [30] datasets, which provided paired low-light and normal-light video sequences for dynamic scene enhancement.

A significant methodological shift came with the SID [3] dataset, which captured image pairs in RAW format rather than processed RGB, enabling methods to leverage the complete sensor information before in-camera processing. To scale dataset creation, VE-LOL [17] adopted a synthetic approach, darkening well-exposed images and adding synthetic noise patterns to simulate sensor characteristics at high ISO settings. Recently, the BVI-LowLight dataset [19] was introduced, containing 40,000 images of objects captured at different ISO settings. Additionally, to obtain more training data, some LLIE methods use exposure-correction datasets such as PHOS [27] and SICE [1]. Despite these advances in scale, sensor diversity, temporal modeling, and data modality, existing datasets either capture each scene under a single low-light condition or rely on modifying the

Table 1. Performance degradation of LLIE methods across varying brightness levels. Models trained on the original LoLv1 dataset show diminished performance when tested on blended images (20% and 50% ground truth mixing), despite reduced information loss. Lower ΔE_{76} and higher $PSNR_L$, the better.

	Retinexformer [2]		CIDNet [36]	
	ΔE_{76}	$PSNR_L$	ΔE_{76}	$PSNR_L$
Original	8.810	28.819	10.587	26.381
20%	11.450	21.910	16.981	17.721
50%	16.165	17.804	24.811	14.115



Figure 2. Impact of brightness variation on LLIE model performance. Blending input images with ground truth at 20% and 50% ratios degrades Retinexformer performance.

camera parameters.

In this work, we introduce the Multi-Illumination Low-Light (MILL) dataset to address the limitations of existing low-light datasets. MILL is the first dataset to capture multiple low-light images of the same scene at varying illumination levels under fixed camera settings (i.e., constant ISO and shutter speed) by systematically controlling light intensity in a controlled environment. Each low-light image is paired with a corresponding ground truth captured under normal lighting. Additionally, we provide RAW files and accompanying metadata, including illumination intensity values and LUX measurements for each capture.

2.2. Low-Light Image Enhancement Methods

Early LLIE methods built upon the seminal Retinex algorithm [13], which decomposed images into reflectance and illumination components, inspiring variants including Multi-Scale Retinex [22], SRIE [6], and Milano-Retinex [25]. Methods such as LIME [8] and NPE [31] demonstrated strong performance by leveraging natural image statistics without training data. However, these traditional approaches have been largely superseded by deep learning methods.

Early end-to-end deep learning methods include SID [3] for RAW images and RetinexNet [35] for RGB inputs. Subsequent approaches introduced various architectural innovations: GLADNet [34] combined global illumination and

local detail modules; KinD [44] and KinD++ [45] adopted Retinex-inspired decomposition strategies; and Yang et al. [37] incorporated adversarial learning. Recent methods leverage transformer architectures (LLFormer [32, 42], Retinexformer [2]) and diffusion models (Diff-Retinex [39], PyDiff [46]). Alternative formulations include specialized color spaces [36] and pixel-wise mean estimation losses [16]. Some methods jointly address enhancement and degradation removal, such as DarkIR [4] and LED-Net [47]. To avoid paired training data requirements, unsupervised approaches have been proposed, including EnlightenGAN [12], Zero-DCE [7], SCI [18], and lightweight RUAS [23].

However, all existing methods have been evaluated exclusively on fixed single low-light inputs without considering behavior across varying illumination levels. We benchmark state-of-the-art LLIE methods across different brightness conditions and propose two simple modifications to Retinexformer [2] that leverage our multi-illumination dataset to improve robustness across intensity levels.

3. Multi-Illumination Low-Light (MILL) Dataset

Existing LLIE datasets present two critical limitations: (1) they either contain a single severely underexposed image per scene (e.g. [35, 38]) or (2) simulate brightness variations via camera parameter adjustments or post-processing (e.g. [1, 19, 27]). This constraint limits real-world applicability, where low-light conditions span a continuous range of intensities.

To quantify this limitation, we simulated varying brightness levels on the LoLv1 dataset [33] by blending input images with their ground truth counterparts at different ratios. This blending reduces degradation severity by simulating intermediate brightness levels, theoretically making enhancement easier. As shown in Table 1, both Retinexformer [2] and CIDNet [36] performed worse on the blended versions (with 0.2 and 0.5 ground truth mixing ratios) than on the original dataset, as measured by ΔE_{76} and $PSNR_L$. This result occurs because models trained on fixed brightness levels fail to generalize across different intensities. Figure 2 illustrates this problem. Oversaturation in the output increases proportionally with input brightness, a clear evidence that intermediate brightness levels are absent from training data. This lack of intensity diversity severely limits the practical applicability of LLIE methods, as real-world deployment requires robustness across varying brightness conditions. To study this problem and address this limitation, we introduce a novel LLIE dataset featuring multiple brightness levels per scene with fixed camera parameters, enabling more robust training, benchmarking, and evaluation of LLIE methods across multiple intensity levels.

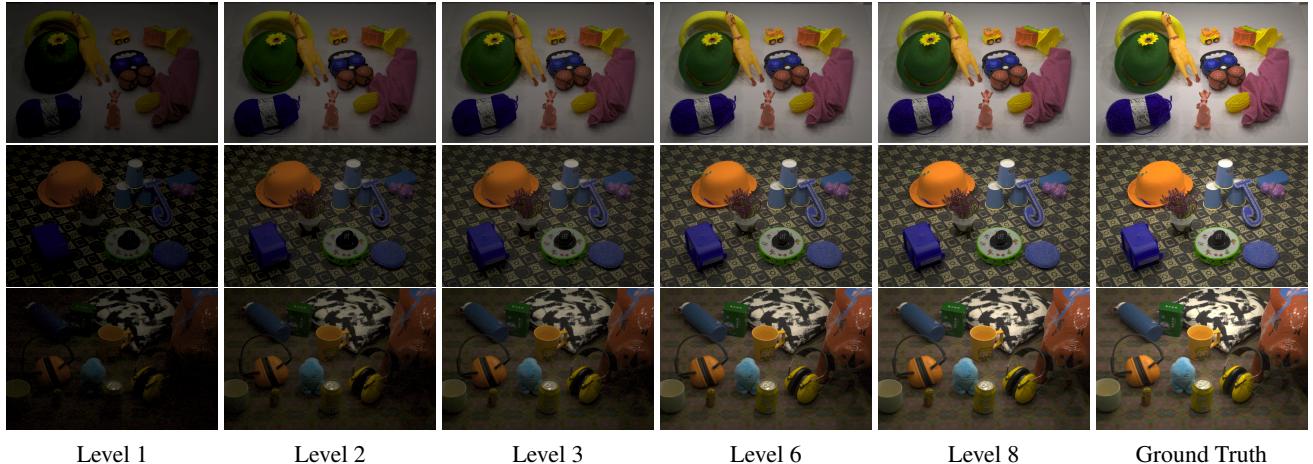


Figure 3. Example scenes from our dataset at different levels for both the DSLR camera (first and second rows) and the smartphone camera (third row). We show the three lowest values to illustrate how intensity varies across successive levels. We also display higher levels to show they remain noticeably underexposed compared to the ground truth.

Our primary objective is to capture a high-quality, well-calibrated dataset for evaluating and training LLIE methods across different intensity levels. To ensure consistency and eliminate uncontrollable variables, we captured all images in a controlled indoor environment without windows or external light sources. We used a dedicated room with a platform for placing different floor backgrounds and objects, equipped with programmable lighting to precisely control brightness levels. Images were captured using two devices: a Nikon D5200 DSLR camera and a Samsung Galaxy S7 smartphone. The smartphone provides a contrasting capture profile compared to the DSLR, enabling evaluation across different sensor characteristics. Figure 1 provides a schematic overview.

During image capture, all camera parameters remained fixed while scene light intensity was adjusted to achieve the desired brightness level. To capture well-lit ground-truth images, we set programmable light sources to maximum power without oversaturating the scene. We fixed the ISO to 100 and placed a Macbeth color chart at the platform center. For the DSLR, we systematically tested all aperture-shutter speed combinations, while for the smartphone (fixed aperture), we tested all available shutter speeds. We analyzed the RGB values of the white patch in the color chart and selected the image with values closest to 95% of the maximum intensity in the camera-RAW format before saturation. This process determined optimal settings of f/9 and 1/5 seconds for the DSLR, and f/1.7 (default) and 1/10 seconds for the smartphone.

We captured the lowest-intensity images (Level 1) with lights at minimum power. To obtain intermediate brightness levels, we computed 10 evenly spaced intervals based on lux meter readings between Level 1 and the ground truth,

creating Levels 2-10, where lower numbers correspond to lower illumination intensity.

We assembled 6 different backgrounds and 98 different objects, with no overlap between train/validation and test sets. The dataset comprises 4 backgrounds in training/validation scenes and 2 in test scenes, with 46 unique objects for training, 24 for validation, and 28 for testing. We captured 50 scenes using both the DSLR and the smartphone across all 11 intensity levels, totaling 1,100 images. The dataset is split into 30 training, 12 validation, and 8 test scenes. Figure 3 shows three representative scenes displaying some of the intensity levels to illustrate the illumination intervals. We display the three lowest values to show how the intensity levels change in consecutive levels. Note that the highest levels remain noticeably underexposed compared to the ground truth, demonstrating the continuous range of realistic low-light conditions.

All images were captured in RAW format (NEF for DSLR, DNG for smartphone) and processed using Camera RAW. DSLR images have a native resolution of 6036×4020 pixels, while smartphone images are 1560×1040 pixels. Following prior LLIE datasets, we created a small version (MILL-s) by bilinearly resizing all images to 600×400 pixels to enable evaluation of methods with computational or memory constraints. Additionally, we divided each DSLR image into 9 non-overlapping patches of 2012×1340 pixels, expanding the dataset to 5,500 Full-HD resolution images. Smartphone images remained at their original resolution due to their comparable full-HD size. We refer to this higher-resolution variant as MILL-f.

Table 2. Performance of different LLIE methods across different intensity levels on our DSLR split of the MILL-s dataset. We report the mean ΔE_{76} and PSNR on the luminance channel ($PSNR_L$). Best, second best, and third best results are highlighted.

	Params (M)	Venue	Level 1		Level 3		Level 5		Level 7		Level 9	
			ΔE_{76}	$PSNR_L$								
Unprocessed	-	-	30.34	13.46	18.62	17.68	11.90	21.48	7.60	25.56	3.62	36.64
RUAS [23]	0.003	CVPR'21	25.46	16.63	45.33	9.86	57.47	6.93	62.99	5.86	67.14	5.18
LLFormer [32]	24.52	AAAI'23	16.37	20.88	13.73	21.79	13.34	22.06	13.17	22.25	12.90	22.55
KinD [44]	1.20	ACMM'19	23.88	16.71	17.62	21.79	14.87	21.34	14.49	21.63	15.00	21.36
FourLLIE [28]	0.12	ACMM'23	24.51	17.29	22.79	17.66	26.64	14.97	28.79	14.07	30.79	13.35
SCI [18]	0.0003	CVPR'22	24.05	16.02	17.99	21.42	25.89	15.69	31.66	13.24	38.38	11.23
MirNet [40]	5.86	CVPR'20	14.03	26.46	11.11	25.34	11.39	24.81	11.65	24.49	11.72	24.96
Retinexformer [2]	1.61	ICCV'23	14.15	25.09	10.45	26.39	10.35	26.55	10.41	26.48	10.46	27.41
DarkIR [4]	3.31	CVPR'25	14.39	24.65	11.29	25.23	11.58	24.74	10.41	23.91	12.15	24.63
HVI-CIDNet [36]	1.88	CVPR'25	14.78	24.08	13.71	22.49	14.58	21.44	15.22	20.83	15.85	20.63
PromptNorm [24]	44.80	CVPRW'25	13.47	25.89	10.51	26.06	10.59	25.94	10.82	25.66	10.89	26.28
GT-Mean [16]	1.88	ICCV'25	14.59	24.32	12.48	23.76	13.23	22.80	13.80	22.19	13.57	22.88
Ours	1.61	-	13.90	25.53	9.11	31.47	9.09	31.52	8.94	32.31	9.17	32.48

4. Method using New Loss Terms

We introduce two auxiliary loss terms that leverage the multi-level nature of our dataset to improve existing LLIE methods. Our goal is to explicitly disentangle the latent features into illumination-related and scene-related components. To this end, we introduce two complementary constraints: (1) an intensity prediction loss that uses the first latent channel to predict the input illumination level, and (2) a scene consistency loss that encourages the remaining channels to encode illumination-invariant scene content across different brightness conditions. The following subsections describe each loss term in detail before presenting the full objective.

Most current LLIE architectures follow a UNet-like structure, comprising an encoder and a decoder. We aim to disentangle the latent features extracted by the architecture’s bottleneck. We adopt Retinexformer [2] as our baseline architecture due to its strong performance in our benchmark evaluation (see Section 5.1).

4.1. Intensity Prediction Loss

We propose a straightforward approach to encode the scene intensity level using the latent features. Specifically, we constrain the first latent feature channel to predict the normalized intensity value of the scene, $i_{in} \in [0, 1]$.

To accomplish this, we introduce a loss component, \mathcal{L}_{ip} , that minimizes the L_1 distance between the predicted intensity at each spatial location and the known scene illumination intensity. Let $\mathbf{Z}_I \in \mathbb{R}^{H \times W}$ denote the first channel of the latent features and $I_{in} \in \mathbb{R}^{H \times W}$ denote the spatially-replicated version of i_{in} matching the spatial dimensions of \mathbf{Z}_I . The intensity prediction loss is defined as:

$$\mathcal{L}_i = \|\mathbf{Z}_I - I_{in}\|_1. \quad (1)$$

4.2. Scene Content Loss

While the intensity prediction loss constrains the first channel to encode illumination information, we enforce the remaining channels to focus on scene content independent of lighting conditions. We achieve this through a triplet loss that encourages images of the same scene captured under different illumination levels to have similar latent representations (excluding the intensity channel), while pushing apart representations of different scenes captured at the same intensity level.

The scene content loss \mathcal{L}_s is defined as:

$$\mathcal{L}_s = \max (\|\mathbf{Z}_q - \mathbf{Z}_p\|^2 + m - \|\mathbf{Z}_q - \mathbf{Z}_n\|^2, 0), \quad (2)$$

where $\mathbf{Z}_q, \mathbf{Z}_p, \mathbf{Z}_n \in \mathbb{R}^{H \times W \times (C-1)}$ correspond to the latent features (excluding the intensity channel) of three images: the query input image, a positive sample from the same scene with different illumination, and a negative sample from a different scene with the same brightness level as the query. The margin m defines the minimum desired distance between positive and negative pairs; we set $m = 1$ in all experiments.

4.3. Combined Objective Function

In addition to the proposed loss terms, we employ a reconstruction loss, \mathcal{L}_{re} , defined as the L_1 distance between the network output and the ground truth image. The complete objective function combines all three components:

$$\mathcal{L} = \mathcal{L}_{re} + \mathcal{L}_i + \mathcal{L}_s. \quad (3)$$

Table 3. Quantitative comparisons on our MILL-s for the DSLR and the smartphone splits. Results are averaged over all the images. Best, second best, and third best results are highlighted.

	$\text{PSNR}_L \uparrow$	$\text{PSNR}_C \uparrow$	$\text{SSIM} \uparrow$	$\text{LPIPS} \downarrow$	$\Delta E_{76} \downarrow$	$\text{MS-SWD} \downarrow$	$\text{NIQE} \downarrow$	$\text{Brisque} \downarrow$
DSLR	Unprocessed	23.237	21.718	0.740	0.151	13.211	1.615	5.419
	RUAS [23]	8.310	6.497	0.355	0.499	53.752	6.781	6.202
	LLFormer [42]	22.029	19.731	0.850	0.155	13.622	1.490	3.876
	KinD [44]	20.409	18.157	0.779	0.219	16.782	1.871	4.383
	FourLLIE [28]	15.486	13.534	0.687	0.241	26.759	3.514	5.398
	SCI [18]	15.329	13.006	0.609	0.270	28.669	3.407	5.791
	MirNet [40]	25.141	21.821	0.882	0.139	11.763	1.403	3.953
	Retinexformer [2]	26.557	22.782	0.888	0.140	10.944	1.254	3.880
	DarkIR [4]	24.700	21.502	0.876	0.142	12.177	1.382	3.788
	HVI-CIDNet [36]	21.755	19.173	0.844	0.155	14.873	1.699	3.727
Smartphone	PromptNorm [24]	26.061	22.497	0.893	0.144	11.059	1.308	3.779
	GT-Mean [16]	23.083	20.257	0.861	0.147	13.520	1.539	3.744
	Ours	31.209	26.197	0.896	0.135	9.681	1.036	3.754
	Unprocessed	19,155	17,573	0.511	0.215	17,832	2.718	5,088
	RUAS [23]	6,920	5,082	0.267	0.722	62,707	6,246	10,556
	LLFormer [42]	23,015	20,062	0.580	0.203	12,510	1.518	4,194
	KinD [44]	19,462	17,452	0.536	0.259	17,252	1.980	3,546
	FourLLIE [28]	21,377	18,510	0.539	0.236	16,976	3.279	4,831
	SCI [18]	15,960	13,208	0.437	0.330	29,477	3.314	5,475
	MirNet [40]	20,760	18,100	0.614	0.219	15,231	1.720	3,537
	Retinexformer [2]	23,230	20,485	0.629	0.195	12,162	1.682	3,162
	DarkIR [4]	22,540	19,898	0.622	0.192	13,047	1.679	3,070
	HVI-CIDNet [36]	20,517	18,170	0.598	0.196	15,823	1.782	3,080
	PromptNorm [24]	22,198	19,464	0.627	0.206	13,238	1.692	3,477
	GT-Mean [16]	21,503	19,018	0,610	0.192	14,438	1.716	3,037
	Ours	23,870	21,166	0.629	0.195	11,671	1.619	3.235

5. Experiments

5.1. Benchmark on MILL

We benchmark mainstream LLIE methods by retraining them on our MILL-s dataset using their officially released code. Our evaluation includes unsupervised methods (RUAS [23], SCI [18]), a Retinex-based approach (KinD [44]), transformer-based methods built on Restormer [41] (LLFormer [42], Retinexformer [2], PromptNorm [24]), a frequency-domain method (FourLLIE [28]), image restoration approaches (MIRNet [40], DarkIR [4]), and specialized LLIE methods (HVI-CIDNet [36], GT-Mean [16]). We also evaluate our proposed modifications to Retinexformer.

Table 2 presents ΔE_{76} and PSNR_L results on the DSLR split of MILL-s across different brightness levels. Several interesting patterns emerge. First, certain methods, such as RUAS and FourLLIE, exhibit degraded performance as brightness increases, suggesting specialization for extremely low-light conditions at the expense of failing at correcting images at higher intensity levels. Conversely, recent methods demonstrate greater consistency across intensity levels. The first row shows input image quality: while all methods successfully enhance severely underex-

Table 4. Quantitative comparisons on MILL-f.

		PSNR_L	PSNR_C	SSIM	ΔE_{76}
DSLR	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
Smartphone	Baseline [2]	22.53	20.62	0.668	10.85
	S-Retinexformer	21.02	19.27	0.645	12.95
	I-Retinexformer	23.53	21.55	0.672	9.63
	Ours	24.45	22.37	0.682	8.53

posed images (lower levels), most fail to improve moderately underexposed images (higher levels), indicating that robustness across varying intensities remains an open challenge. Notably, our modifications to Retinexformer yield improvements across all intensity levels (except Level 1, where PromptNorm achieves the best performance with 40 times more parameters).

Table 3 reports performance averaged across all intensity levels for both the DSLR and smartphone splits of MILL-s, using three full-reference metrics: PSNR_L , PSNR_C (on RGB), and SSIM ; one perceptual full-reference metric: LPIPS [43]; two color similarity metrics: ΔE_{76} and MS-

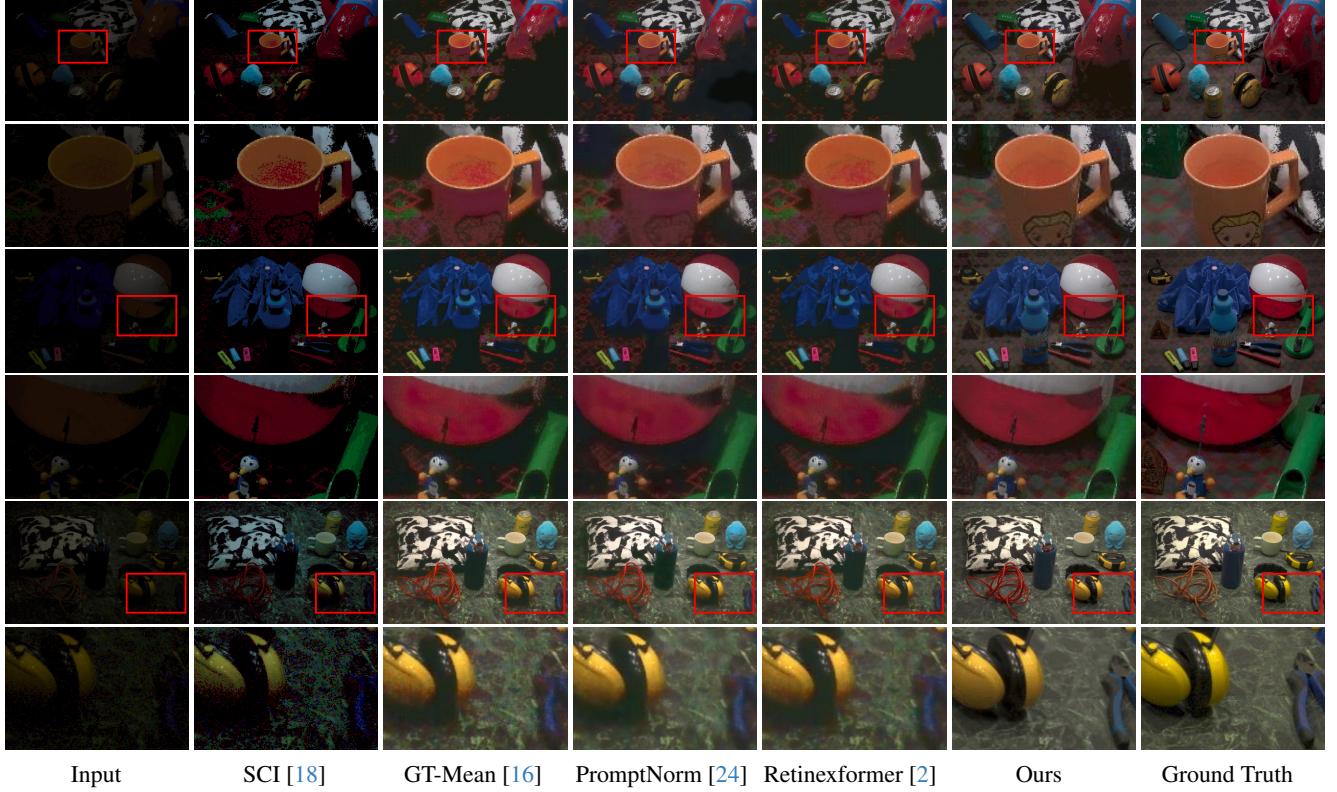


Figure 4. Visual comparison on MILL-s. From left to right: input, SCI [18], GT-Mean [16], PromptNorm [24], Retinexformer [2], Ours, and ground truth. We show three examples with zoomed-in regions below each. First two images: DSLR; last image: smartphone.

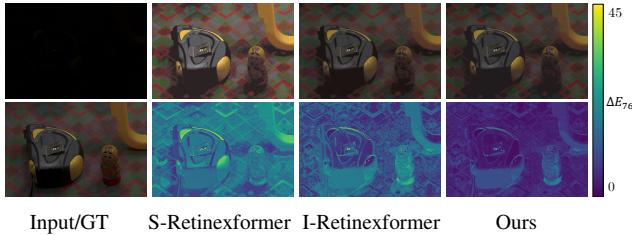


Figure 5. Ablation Study for the different components of our loss term. On the second row, we show the ΔE_{76} error maps.

SWD [10]; and two non-reference metrics: NIQE [21], and Brisque [20]. Surprisingly, several methods fail to improve upon the input images on average. As demonstrated in Table 2, these methods enhance extremely low-light images but degrade moderately underexposed images, resulting in net negative impact. This observation highlights the difficulty of achieving robust LLIE across variable intensity levels. Retinexformer achieves the best performance, followed closely by PromptNorm and LLFormer. This indicates that Restormer-based architectures perform well for this task. We therefore select Retinexformer as our baseline due to its superior performance and parameter efficiency. Our proposed method further improves upon Retinexformer by 4.6dB in $PSNR_L$, 3.8dB in $PSNR_C$, and 1.2 in ΔE_{76} on the DSLR split and 0.5dB in $PSNR_L$, $PSNR_C$, and 0.5 ΔE_{76} units on the smartphone split. The smartphone split

proves more challenging across all methods due to inferior sensor quality. However, our proposed auxiliary loss terms achieve superior performance in most metrics.

5.2. Qualitative Results

Figure 4 presents a qualitative comparison of our method against state-of-the-art approaches. We show three representative examples with zoomed-in regions to highlight enhancement differences. The first two examples are from the DSLR split, while the third is from the smartphone split. Across all cases, our method produces substantially better enhancements. In the first example, SCI, GT-Mean, and Retinexformer present noise and color artifacts in the background next to the orange mug and within the mug interior. PromptNorm reduces noise but fails to recover the mug’s texture details. In contrast, our method produces outputs closer to ground truth with sharper object boundaries and effective noise reduction. In the second example, all competing methods generate washed-out colors with unnatural saturation and color artifacts, particularly visible in the shadow cast by the inflatable ball and the green object. In the final example, competing methods produce noisy outputs, while our method’s enhanced image has higher quality with better-preserved background texture.

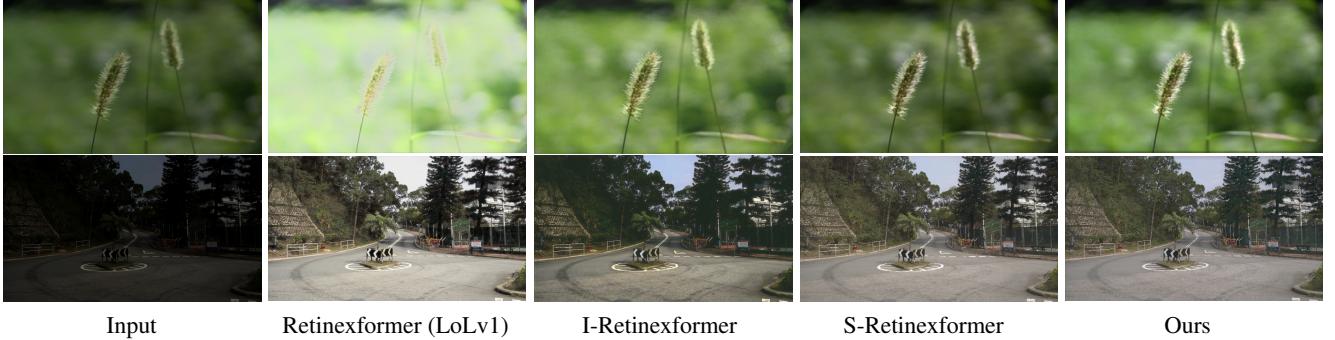


Figure 6. Outdoor examples from the DICM [14] (first row) and SICE [1] (second row) of Retinexformer trained on LoLv1, our baseline with the two proposed additional loss terms independently, and our final approach.

5.3. FullHD Experiments and Ablation

While the previous analysis was conducted on MILL-s due to computational constraints of older methods, we now evaluate our modifications against the best-performing baseline, Retinexformer, including an ablation study on the Full-HD MILL-f dataset. This enables assessment of our improvements without image downsampling and provides more detailed analysis of our proposed loss components. We compare Retinexformer with variants incorporating our intensity prediction loss (I-Retinexformer) and scene content loss (S-Retinexformer) independently alongside the reconstruction loss, as well as our complete method combining both losses. Table 4 reports PSNR_L , PSNR_C , SSIM, and ΔE_{76} metrics.

Our proposed modifications outperform the baseline model across all metrics. On the DSLR split, we observe improvements of approximately 10 dB in PSNR_L and PSNR_C , 0.03 in SSIM, and 5 in ΔE_{76} . The smartphone split exhibits smaller but consistent gains due to sensor limitations; nevertheless, our method maintains a clear performance advantage over the baseline. Notably, the intensity prediction loss yields larger improvements than the scene content loss when applied independently. However, combining both losses delivers the strongest performance, as effective feature disentanglement requires their joint optimization.

Figure 5 shows one example of the MILL-f dataset comparing individual and combined loss terms, with corresponding ΔE_{76} error maps displayed below each output. The combined use of both loss terms achieves better performance. The error maps reveal complementary behavior: the scene content loss alone produces spatially uniform ΔE_{76} values across the image, while the intensity prediction loss concentrates errors in specific regions. Combining both loss terms reduces ΔE_{76} values both globally and locally, yielding the best overall results. This demonstrates that proper feature disentanglement is only achieved through the joint application of both loss terms.

5.4. Outdoor Images

We evaluate our method on underexposed outdoor images from the DICM [14] and SICE [1] datasets. Figure 6 presents results comparing Retinexformer trained on LoLv1 with models trained on our dataset using each of our loss terms individually and our complete approach.

In the first example, Retinexformer overexposes the scene due to the moderately low-light input, an expected limitation since LoLv1 lacks images captured at varying intensity levels. In contrast, the intensity prediction loss produces accurate exposure, while the scene content loss enhances fine details in the plant. The combination of both losses yields optimal results, balancing exposure and detail enhancement. In the second example, while Retinexformer enhances overall brightness, it oversaturates the sky region due to its higher input intensity. This highlights the fundamental limitation of LoLv1 and similar datasets that contain limited diversity and only a single fixed low-light intensity level. Our multi-level dataset mitigates this issue by learning robust LLIE across different intensity levels.

6. Conclusion

We introduced the MILL dataset, which captures images under systematically varied intensity levels with all camera parameters fixed. For each scene, MILL contains 11 images spanning the complete illumination range. The highest-intensity image serves as ground truth, while the remaining images serve as low-light inputs. Leveraging the multi-level structure of our dataset, we analyzed how current LLIE methods perform under different input intensities, revealing that performance varies significantly across methods and that robust LLIE across varying intensities remains challenging. We also propose two new loss terms that disentangle latent features into illumination intensity and scene content components, yielding substantial gains across all MILL splits. We believe the MILL dataset and our proposed enhancements will advance future research in LLIE.

Acknowledgements

This work was supported by Grants PID2021-128178OB-I00 and PID2024-162555OB-I00 funded by MCIN/AEI/10.13039/ 501100011033 and by ERDF "A way of making Europe", and by the Generalitat de Catalunya CERCA Program. DSL also acknowledges the FPI grant from Spanish Ministry of Science and Innovation (PRE2022-101525). JVC also acknowledges the 2025 Leonardo Grant for Scientific Research and Cultural Creation from the BBVA Foundation. The BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors. This research was also supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs (CRC) program.

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. [2](#), [3](#), [8](#)
- [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. [2](#), [3](#), [5](#), [6](#), [7](#)
- [3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. [2](#), [3](#)
- [4] Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *CVPR*, 2025. [3](#), [5](#), [6](#)
- [5] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. Dancing in the dark: A benchmark towards general low-light video enhancement. In *ICCV*, 2023. [2](#)
- [6] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, 2016. [3](#)
- [7] 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. [3](#)
- [8] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016. [2](#), [3](#)
- [9] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90:103712, 2023. [2](#)
- [10] Jiaqi He, Zhihua Wang, Leon Wang, Tsein-I Liu, Yuming Fang, Qilin Sun, and Kede Ma. Multiscale sliced wasserstein distances as perceptual color difference measures. In *ECCV*, 2024. [7](#)
- [11] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017. [2](#)
- [12] 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 TIP*, 30:2340–2349, 2021. [3](#)
- [13] Edwin H Land. The retinex. *American Scientist*, 52(2):247–264, 1964. [3](#)
- [14] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation. In *ICIP*, 2012. [8](#)
- [15] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE TPAMI*, 44(12):9396–9416, 2021. [2](#)
- [16] 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. [3](#), [5](#), [6](#), [7](#)
- [17] Jiaying Liu, Xu Dejia, Wenhan Yang, Minhao Fan, and Haofeng Huang. Benchmarking low-light image enhancement and beyond. *IJCV*, 129:1153–1184, 2021. [2](#)
- [18] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. [3](#), [5](#), [6](#), [7](#)
- [19] Alexandra Malyugina, Nantheera Anantrasirichai, and David Bull. A topological loss function for image denoising on a new bvi-lowlight dataset. *Signal Processing*, 211:109081, 2023. [2](#), [3](#)
- [20] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. [7](#)
- [21] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. [7](#)
- [22] Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Multi-scale retinex for color image enhancement. In *ICIP*, 1996. [3](#)
- [23] 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. [2](#), [3](#), [5](#), [6](#)
- [24] 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. [5](#), [6](#), [7](#)
- [25] Gabriele Simone, Michela Lecca, Gabriele Gianini, and Alessandro Rizzi. Survey of methods and evaluation of retinex-inspired image enhancers. *J. Electron. Imaging*, 31(6):063055–063055, 2022. [3](#)
- [26] Vassilios Vonikakis. Tm-died: The most difficult image enhancement dataset, Accessed 10/2025. [2](#)
- [27] Vassilios Vonikakis, Dimitrios Chrysostomou, Rigas Kouskouridas, and Antonios Gasteratos. A biologically inspired scale-space for illumination invariant feature detection. *Meas. Sci. Technol.*, 24(7):074024, 2013. [2](#), [3](#)

- [28] Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *ACM MM*, 2023. 5, 6
- [29] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 2019. 2
- [30] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: High-quality video dataset with mechatronic alignment. In *ICCV*, 2021. 2
- [31] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE TIP*, 22(9):3538–3548, 2013. 3
- [32] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *AAAI*, 2023. 3, 5
- [33] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 2, 3
- [34] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. In *Face and Gestures*, 2018. 3
- [35] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 2, 3
- [36] Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yan-ning Zhang. Hvi: A new color space for low-light image enhancement. In *CVPR*, 2025. 3, 5, 6
- [37] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, 2020. 3
- [38] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE TIP*, 30:2072–2086, 2021. 2, 3
- [39] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *ICCV*, 2023. 3
- [40] 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. 5, 6
- [41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 6
- [42] 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. 2, 3, 6
- [43] 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
- [44] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, 2019. 2, 3, 5, 6
- [45] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *IJCV*, 129(4):1013–1037, 2021. 3
- [46] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *IJCAI*, 2023. 3
- [47] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, 2022. 3