MESS: Multi-Exposure Sequence Synthesis for Generalizable Image Enhancement

Ruodai Cui[1,3] **Shuaizheng Liu**[1,2] **Rongyuan Wu**[1,2] **Lei Zhang**[1,2*]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]OPPO Research Institute [3]Qualcomm Technologies, Inc.
{ruodai.cui, shuaizheng.liu, rong-yuan.wu}@connect.polyu.hk
cslzhang@comp.polyu.edu.hk

Abstract

Despite the significant progress in image enhancement under adverse illuminations, such as low-light image enhancement, exposure correction, and backlit image enhancement, existing models show limited generalizability to different scenarios. A major bottleneck lies in the lack of large-scale and diverse training data, as these tasks typically rely on manually captured, precisely aligned image pairs that are expensive to scale up. Although synthetic data have been explored as an alternative, existing synthesis methods are limited in modeling complex adverse illumination conditions and camera imaging pipelines. To address this issue, we propose a novel framework to synthesize realistic training pairs for generalizable image enhancement in the wild. Specifically, we first derive multi-exposure sequences (MES) from high-bit RAW images by rendering them under different exposure settings through an emulated ISP pipeline. Using these RAW-derived MES as supervision, we train a diffusion-based RGB-to-MES generator to synthesize plausible exposure sequences from a single 8-bit RGB image, extending the synthesis pipeline from limited RAW collections to large-scale in-the-wild RGB data. For each source image, we randomly sample one synthesized exposure variant as the degraded input and use the original high-quality RGB image as the target. With the proposed **MES S**ynthesis (**MESS**) approach, we construct a dataset of 500K training pairs and train a lightweight network, which demonstrates significantly better generalization performance than existing models across various image enhancement tasks. Code and data are available at `https://github.com/PolyU-VCLab/MESS`.

1 Introduction

Images captured under adverse illumination conditions often suffer from underexposure, overexposure, and low contrast, resulting in poor visibility and degraded visual quality [64, 21, 44]. Many image enhancement methods have been developed for specific scenarios, including low-light image enhancement (LLIE) [10, 70, 20, 28, 45, 8], exposure correction (EC) [63, 1, 24, 66, 38], and backlit image enhancement (BIE) [42, 17]. Despite their progress, most existing methods are designed for a single enhancement task and trained on narrow data distributions, showing limited generalization performance. As shown in Fig. 1, a model trained for LLIE may fail when applied to exposure correction, producing inaccurate brightness adjustment and losing local details. Moreover, real-world images often contain mixed illumination degradations, e.g., underexposed and overexposed regions appearing simultaneously, which further increases the difficulty of image enhancement in the wild. Although recent methods attempt to improve generalization by using larger models or more complex designs [67, 68, 8, 65, 34], their performance under diverse real-world conditions remains limited.

*Corresponding author. This work is partly supported by the PolyU-OPPO Joint Innovative Research Center.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

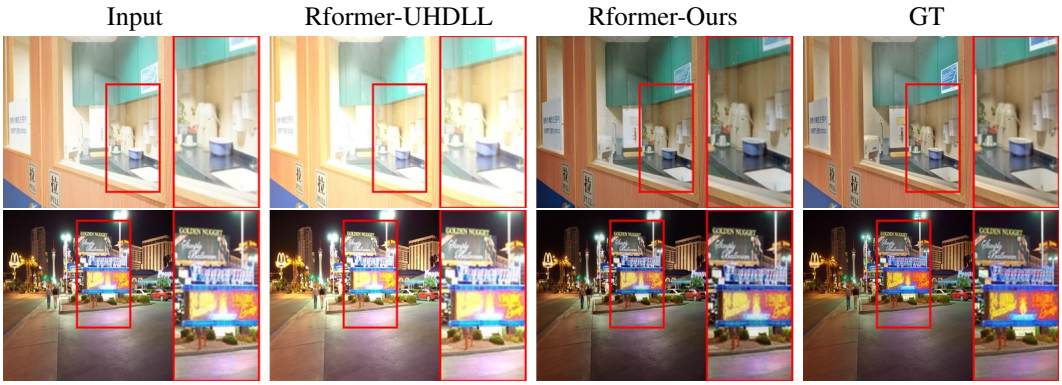

Figure 1: Visual comparisons of a recent state-of-the-art low-light image enhancement method, Rformer [8], trained on the UHDLL dataset [36] (2nd column) and on our large-scale synthetic dataset (3rd column). Zoomed-in regions are highlighted with red boxes for better comparison.

We argue that a key bottleneck for generalizable illumination enhancement is the lack of large-scale, diverse, and reliable paired training data. Existing datasets are typically built from manually captured and carefully aligned image pairs, where a degraded image is paired with a better-exposed reference [10, 70, 7, 44]. However, this process is costly and difficult to scale, resulting in limited coverage of real-world scenes and degradation types. As a consequence, models trained on such datasets often struggle to generalize beyond their target task or data domain.

A natural alternative is to synthesize paired training data by applying controlled degradations to high-quality images. However, realistic and scalable synthesis is a challenging task. As shown in the top-left of Fig. 2, one line of research performs exposure manipulation in the high-bit RAW domain using image signal processing (ISP) operators [1, 12, 11]. These methods produce realistic exposure variation, but their scalability is constrained by the limited availability of high-quality RAW data. Another line of work (top-right of Fig. 2) starts from 8-bit RGB images and reverse the ISP to mimic exposure changes in a pseudo-RAW space [73, 31]. While more scalable, such approximations often fail to capture the complexity of real camera pipelines and illumination variations. Once a high-bit RAW image has been rendered into 8-bit RGB space, its dynamic range is already clipped or compressed, making faithful inverse exposure manipulation fundamentally difficult.

In this work, we propose **M**ulti-**E**xposure **S**equences **S**ynthesis (**MESS**), a framework that transfers RAW-derived exposure variations to large-scale in-the-wild RGB images. Given a high-quality RGB image, MESS synthesizes a realistic multi-exposure sequence (MES), from which degraded-target pairs can be automatically constructed for training generalizable enhancement models. The key idea is to use RAW images to provide realistic supervision for exposure variation, while learning an RGB-to-MES generator that transfers such variation to ordinary RGB images at scale. In this way, MESS combines the physical realism of RAW-based exposure rendering with the scalability of large RGB image collections, as shown in the bottom of Fig. 2.

In particular, MESS is built upon three components. First, we exploit RAW images to construct two types of supervision: multi-style high-quality renderings of the same scene, and camera-consistent exposure renderings generated by varying only the exposure setting while keeping the underlying ISP fixed. Based on this design, we introduce a frequency-based enhancement module to reduce rendering-style discrepancies while preserving scene structure and local details. Second, we incorporate a diffusion prior to compensate for information loss in 8-bit RGB images and synthesize structurally plausible missing content. Third, we design an exposure-conditioned rendering module that injects the target exposure condition and performs exposure-aware feature modulation, producing outputs that better match realistic camera characteristics.

Based on MESS, we curate 500K high-quality in-the-wild RGB images and synthesize exposure-perturbed counterparts to automatically construct a large-scale paired training set. For each source image, one synthesized exposure variant is sampled from the generated MES as the degraded input, while the original high-quality image is used as the target. To validate the effectiveness of the synthesized data, we train a lightweight enhancement network on the resulting dataset. Despite its compact model size, the trained model generalizes better than existing methods across multiple real-world illumination enhancement benchmarks.

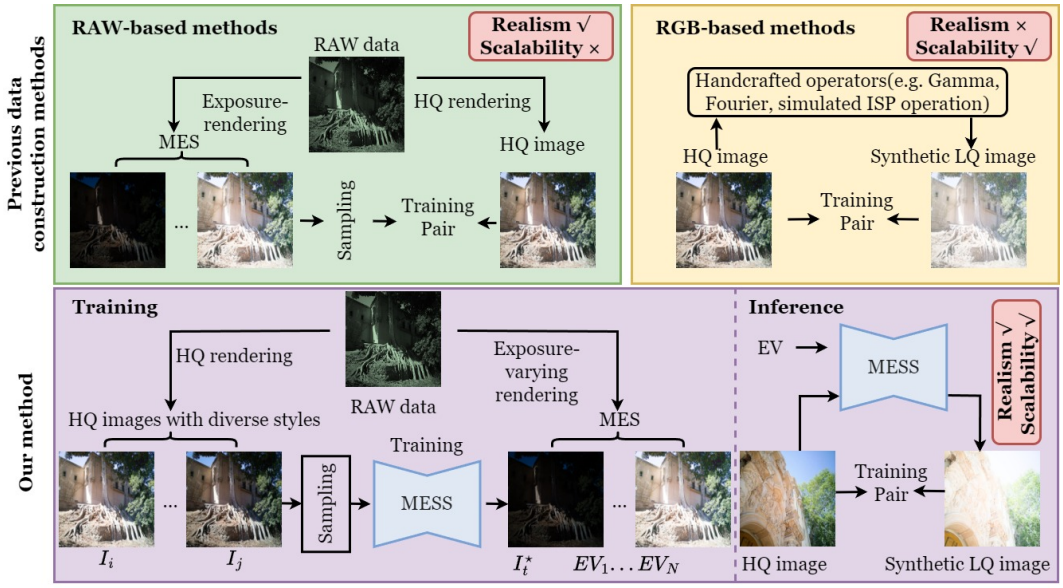

Figure 2: **Comparison of different data construction strategies for image illumination enhancement.** RAW-based methods (top left) produce realistic exposure variations but lack scalability, while RGB-based methods (top right) are scalable but often less realistic. In contrast, our proposed **MESS** (bottom) transfers RAW-derived exposure priors to in-the-wild RGB images, enabling both realistic and scalable synthesis of training pairs. Notions are defined in §3.

Our contributions are summarized as follows. (1) We propose **MESS**, a new framework for scalable and realistic training data synthesis for generalizable image enhancement by transferring RAW-derived exposure variations to in-the-wild RGB images. (2) We develop an MES synthesis pipeline, including frequency-based enhancement for style-robust scene representation, diffusion-based intermediate synthesis for detail completion, and exposure-conditioned rendering for realistic exposure-aware output generation. (3) We build a large-scale synthetic dataset of 500K paired samples with MESS, and show that even a lightweight enhancement model trained on this dataset can achieve strong generalization across multiple real-world image enhancement tasks.

2 Related Work

Image Illumination Enhancement Methods and Datasets. Existing work on illumination enhancement can be broadly categorized into three related but distinct tasks: low-light image enhancement (LLIE) [22, 27, 36, 35], exposure correction (EC) [38, 12, 25, 1, 26], and backlit image enhancement (BIE) [42, 48, 59]. Many methods can achieve strong performance on their target benchmarks by using specialized architectures such as CNNs, Transformers, and diffusion models [27, 68, 16, 38]. However, most of them are developed for a single task and trained on relatively narrow data distributions, limiting the generalization to real-world images with mixed degradations. This limitation is related to the problems of existing datasets. Popular benchmarks such as LOLv2Real [75], UHD-LL [36], LSRW [22], BAID [48], BAID [48] and SICE [7] are limited in scale, scene diversity and degradation coverage. As shown in Table 1, LOLv1 and LOLv2Real contain only a few hundred images, while larger datasets such as LSRW [22], UHD-LL [36], BAID [48], and SICE [7] still contain only a few thousand images. This limited scale makes it difficult to achieve strong generalization across diverse real-world scenarios. Some studies have explored ISP-like operations in the RAW domain, including MSEC [1] and LCDP [62], which partially alleviates this issue by enabling more physically grounded exposure simulation through ISP rendering [1, 62]. However, these methods are still constrained by the availability of high-quality RAW datasets.

Synthetic Data for Illumination Enhancement. Synthetic data provide a practical alternative to the costly collection of real paired data for image enhancement. For example, LLNet [47] generates degraded samples through gamma-based perturbations of high-quality reference images, whereas CSNorm [76] adopts a Fourier-transform-based degradation strategy. However, such approaches often

Table 1: Summary of existing datasets for image enhancement. "#images" denotes the number of training images, while "#scenes" denotes the number of unique scenes. For datasets LOLv2Real [75], UHD-LL [36], LSRW [22], and BAID [48], the exact number of scenes is not provided. We treat images with a CLIP similarity score greater than 0.9 as duplicated images from the same scene.

Task	Dataset	#images	#scenes	Task	Dataset	#images	#scenes
LLIE	LOLv2Real [75]	789	263	EC	LCDP [62]	1,700	1,700
LLIE	LSRW [22]	5,650	4,157	EC	SICE [7]	4,800	589
LLIE	UHD-LL [36]	2,150	724	BIE	BAID [48]	3,000	1,782
EC	MSEC [1]	24,330	5,000	BIE	Backlit300 [42]	305	305

struggle to produce physically realistic degradations. Since illumination variation is approximately linear in the RAW domain, mapping sRGB images back to RAW space enables more physically grounded simulation of illumination degradations. Motivated by this observation, several ISP-aware synthesis pipelines have been proposed. For instance, Unprocessing [5] and CycleISP [77] show that explicit modeling of the camera ISP can generate more realistic paired data and improve generalization to real-world images. InvISP [73] formulates the ISP as an invertible pipeline to reconstruct RAW images from sRGB inputs, while ParamISP [31] further incorporates camera metadata, such as ISO and exposure time, into the ISP model. Following this line of research, recent LLIE-oriented synthesis pipelines [46, 69] first map sRGB images back to RAW space, then simulate low-light degradations and ISP variations, and finally render the results back to sRGB for training. Nevertheless, the mapping from RAW to RGB is inherently irreversible, as the dynamic range is clipped or compressed during the imaging pipeline. Hence, simplified ISP operators alone are insufficient to faithfully recover RAW from sRGB images, which limits the realism and fidelity of the synthesized data.

3 Multi-Exposure Sequence Synthesis: Methodology

This section describes the proposed MESS framework and its training strategy. We first present the data preparation procedure and overall training pipeline in §3.1, including RAW-guided supervision for learning camera-consistent exposure variation and RGB-based synthesis for large-scale paired data construction. We then introduce the three main components: frequency-based enhancement in §3.2, diffusion-based detail enhancement in §3.3, and exposure-conditioned rendering in §3.4. Finally, we summarize the training objectives in §3.5.

3.1 Training Framework Overview

Training Data Preparation. As shown in the bottom left of Fig. 2, we construct RAW-guided triplets $\{I_i, I_j, I_t^\star\}$ from each high-bit RAW image X. Specifically, $I_i = \mathcal{T}_{\text{sty}}(X; s_i)$ and $I_j = \mathcal{T}_{\text{sty}}(X; s_j)$, where $\mathcal{T}_{\text{sty}}$ denotes style-varied rendering, and s_i and s_j are rendering-style parameters. Meanwhile, $I_t^\star = \mathcal{T}_{\text{cam}}(X; e_t)$ is the image rendered under the target exposure using a camera-consistent ISP pipeline, where $\mathcal{T}_{\text{cam}}$ denotes camera-consistent exposure rendering and $e_t \in [0, 1]$ is the normalized exposure value. The image pair $\{I_i, I_j\}$ is used to learn a style-robust scene representation, while the image $I_t^\star$ supervises exposure-controlled synthesis. More details of the training data construction process can be found in **Appendix A**.

Training Pipeline. MESS aims to synthesize realistic multi-exposure sequences (MES) from high-quality RGB images in-the-wild, enabling scalable paired training data construction. Given a source RGB image I_i and a normalized target exposure value e_t, MESS predicts an exposure-controlled output $\hat{I}_t = \text{MESS}(I_i, e_t)$. Both $\hat{I}_t$ and the RAW-rendered reference $I_t^\star$ correspond to the same target exposure. By varying e_t, the model generates an MES for the same input image.

As shown in Fig. 3(a), MESS consists of three sequentially connected modules. The *frequency-based enhancement module* $D(\cdot)$ extracts a style-robust scene representation $C_i = D(I_i)$. The *diffusion-based detail enhancement module* $G(\cdot)$ enriches C_i with plausible structural and textural details, producing an intermediate image $I_m = G(C_i)$. The *exposure-conditioned rendering module* $P(\cdot)$ injects the normalized target exposure value e_t into I_m and generates the final exposure-aware RGB output $\hat{I}_t = P(I_m, e_t)$. Once the MESS model is trained, it is then applied to high-quality in-the-wild RGB images to synthesize MES. As shown in the bottom right of Fig. 2, for each RGB image, one

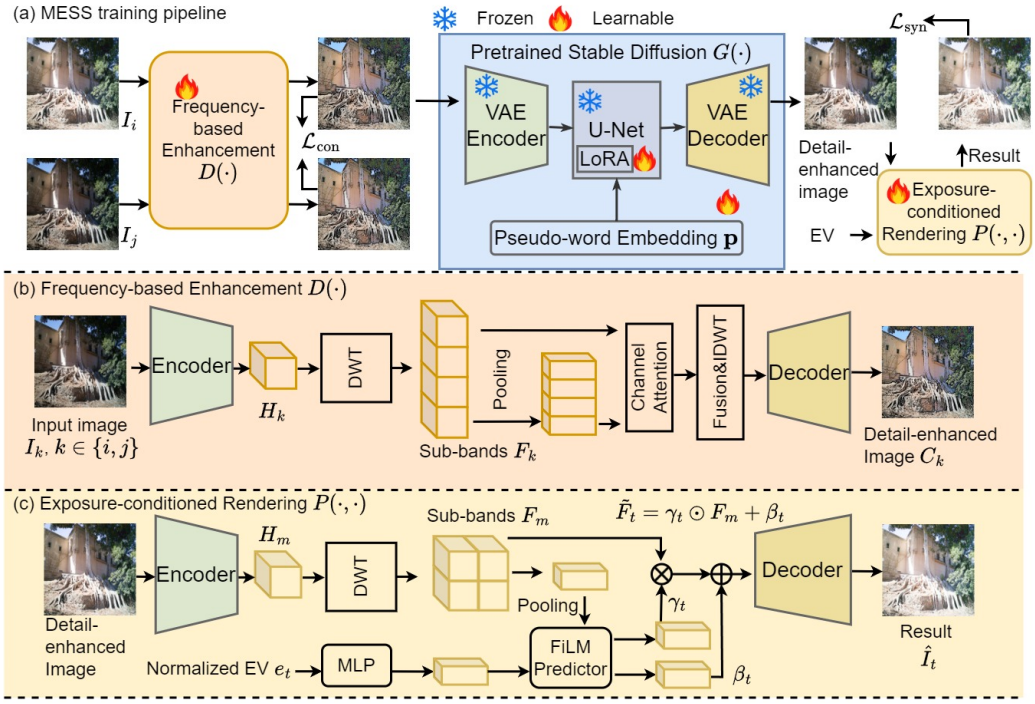

Figure 3: **MESS Training Framework.** (a) The training pipeline: Given a single RGB image and a target exposure value, MESS generates the corresponding exposure-adjusted image through three modules: frequency-based enhancement, diffusion-based synthesis, and exposure-conditioned rendering. (b) Structure of the frequency-based enhancement module. (c) Structure of the exposure-conditioned rendering module.

synthesized exposure variant is sampled as the degraded input, while the original RGB image serves as the target. The resulting pairs can be used to train a lightweight downstream enhancement network.

3.2 Frequency-based Enhancement

To reduce rendering-style variations caused by camera ISP pipelines while preserving shared scene content, MESS uses a frequency-based enhancement module $D(\cdot)$, as illustrated in Fig. 3(b). This design is motivated by the fact that ISP-induced appearance variations, e.g., exposure and contrast changes, are mainly captured by low-frequency responses, whereas high-frequency components preserve stable structures and textures [36, 25, 61]. Given an input image I_k, $k \in \{i, j\}$, this module first encodes it into feature maps H_k, then decomposes the features into different frequency bands, enhances them, and finally decodes them back to the image domain.

Many frequency-based enhancement methods employ Fast Fourier Transform (FFT) for the feature decomposition. However, FFT lacks spatial localization capability and is limited in modeling local variations [61, 36, 25]. We instead adopt Discrete Wavelet Transform (DWT), in particular the Haar wavelet transform due to its simplicity and reduced coefficient correlation [54, 57, 3, 79], to perform the decomposition. Haar DWT is applied to each channel of H_k, decomposing it into one low-frequency sub-band and three high-frequency sub-bands, collectively denoted as F_k. To adaptively fuse the wavelet sub-bands, we apply average pooling and max pooling to F_k and feed the pooled descriptors into a channel attention block. The attention-enhanced representation is computed as $\tilde{F}_k = A_k \odot F_k$, where A_k denotes the learned channel weights and $\odot$ denotes channel-wise multiplication. The Inverse DWT (IDWT) is then applied to the processed features, which are finally decoded as the enhanced image representation C_k.

During training, the style-varied renderings I_i and I_j are generated from the same RAW image. We impose the representation consistency loss in §3.5 on $C_i = D(I_i)$ and $C_j = D(I_j)$, encouraging $D(\cdot)$ to produce style-invariant and content-consistent features.

3.3 Diffusion-based Detail Enhancement

After frequency-based enhancement, the resulting representation C_i is passed to a diffusion-based generator G to synthesize the missing details due to dynamic-range clipping or compression. The intermediate output is $I_m = G(C_i)$. We adopt the accelerated Stable Diffusion backbone (SD-Turbo) [56] to perform this task for its strong generative prior and high efficiency. The diffusion module G consists of a frozen VAE encoder, a LoRA-adapted U-Net denoising function $\mathcal{G}_\theta$, and a frozen VAE decoder. Since SD-Turbo is originally a text-to-image model, text conditioning needs to be provided. Inspired by previous works [18, 51, 43, 39], we introduce a learnable shared pseudo-word embedding $\mathbf{p}$ to adapt the diffusion prior to the target synthesis task.

Given C_i, we first obtain the latent code $z = \mathrm{Enc}_{\mathrm{VAE}}(C_i)$ using the frozen VAE encoder. We then perform denoising as $z' = \mathcal{G}_\theta(z, \mathbf{p})$, where $\mathcal{G}_\theta$ is parameterized by the LoRA-adapted U-Net and $\mathbf{p}$ is the learnable shared pseudo-word embedding. The frozen VAE decoder maps z' to the diffusion output $I_m = \mathrm{Dec}_{\mathrm{VAE}}(z')$, that is, $G(C_i) = \mathrm{Dec}_{\mathrm{VAE}}(\mathcal{G}_\theta(\mathrm{Enc}_{\mathrm{VAE}}(C_i), \mathbf{p}))$. Following previous diffusion-based super-resolution works [71, 72, 58], we freeze the VAE and part of the U-Net, optimizing only the LoRA layers and pseudo-word embedding $\mathbf{p}$.

3.4 Exposure-conditioned Rendering

As shown in Fig. 3(c), the exposure-conditioned renderer $P(\cdot)$ receives the diffusion-enhanced image I_m and normalized target exposure value e_t to output the final RGB image $\hat{I}_t = P(I_m, e_t)$. It aims to adjust exposure-dependent appearance, including brightness, contrast, and global tone, while preserving the structure and textures recovered by the diffusion module.

Since such appearance changes are mainly reflected in low-frequency components, we adopt the DWT-based modulation design in §3.2. Specifically, the renderer first extracts feature maps H_m from I_m and applies Haar DWT to obtain wavelet sub-band features F_m. For each sub-band, global average pooling and max pooling are used to obtain compact content descriptors. The normalized target exposure value $e_t \in [0, 1]$ is encoded by a lightweight MLP into an exposure embedding. The content descriptors and exposure embedding are concatenated and fed into a FiLM [53] predictor, which outputs channel-wise affine parameters γ_t and β_t. The exposure-conditioned sub-band features are computed as $\tilde{F}_t = \gamma_t \odot F_m + \beta_t$, where $\odot$ denotes channel-wise multiplication, followed by IDWT and decoding to produce $\hat{I}_t$. Our formulation provides an efficient exposure-control mechanism: DWT decomposes illumination- and contrast-related responses into multi-resolution sub-bands, while exposure-conditioned affine modulation enables continuous rendering control with respect to e_t.

3.5 Training Loss

MESS is trained with two objectives: representation consistency and synthesis reconstruction. The representation consistency loss is applied to the style-varied renderings I_i and I_j. Since they share the same scene content, the enhanced representations $C_i = D(I_i)$ and $C_j = D(I_j)$ are encouraged to be close, yielding $\mathcal{L}_{\mathrm{con}} = \|C_i - C_j\|_1$. The synthesis reconstruction loss supervises the final synthesized image $\hat{I}_t$ with the RAW-rendered reference $I_t^\star$ under the same target exposure, which is defined as $\mathcal{L}_{\mathrm{syn}} = \|\hat{I}_t - I_t^\star\|_2^2$.

The final training loss is $\mathcal{L}_{\mathrm{total}} = \lambda \mathcal{L}_{\mathrm{con}} + \mathcal{L}_{\mathrm{syn}}$, where λ balances the two objectives.

4 Experiments

4.1 Experimental Setup

We evaluate MESS as a training data generator for generalizable image illumination enhancement. MESS is first trained with RAW-guided supervision to learn camera-consistent exposure variations. After training, we apply it to high-quality in-the-wild RGB images to synthesize paired training data for downstream image enhancement tasks (i.e., LLIE, EC and BIE). Details of RAW datasets, RGB image collection, and downstream paired training data generation are provided in **Appendix B**.

Implementation Details. MESS is trained on random 512×512 crops for 200K iterations using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) on 8 NVIDIA A100 GPUs. The total batch size is 64. The learning

Table 2: The results of full-reference scores using FourLLIE backbone [61]. The model trained on each data source are evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID	Average
UHDLL [36]	Real	PSNR ↑	23.694	19.357	16.152	18.728	20.461	19.678
LSRW [22]	Real	PSNR ↑	21.383	20.849	17.246	18.964	19.875	19.663
SICE [7]	Real	PSNR ↑	18.947	18.061	22.385	20.329	20.103	19.965
MSEC [1]	Raw-Syn	PSNR ↑	19.348	18.214	20.367	22.116	20.352	20.079
BAID [48]	Real	PSNR ↑	17.684	16.935	18.276	18.742	21.674	18.662
RCD [12]	Raw-Syn	PSNR ↑	19.702	18.242	17.478	19.924	20.159	19.101
Gamma [47]	RGB-Syn	PSNR ↑	16.845	14.692	17.248	17.681	18.136	16.920
CSNorm-Fourier [76]	RGB-Syn	PSNR ↑	18.079	16.456	18.843	19.224	19.391	18.399
UEC-based [12]	RGB-Syn	PSNR ↑	19.137	17.841	20.228	20.553	20.467	19.645
MESS-generated	RGB-Syn	PSNR ↑	23.416	20.514	21.769	21.982	21.935	21.923
UHDLL [36]	Real	SSIM ↑	0.914	0.701	0.762	0.819	0.871	0.813
LSRW [22]	Real	SSIM ↑	0.897	0.731	0.786	0.814	0.853	0.816
SICE [7]	Real	SSIM ↑	0.825	0.689	0.864	0.842	0.858	0.816
MSEC [1]	Raw-Syn	SSIM ↑	0.832	0.697	0.816	0.881	0.855	0.816
BAID [48]	Real	SSIM ↑	0.782	0.641	0.764	0.792	0.893	0.774
RCD [12]	Raw-Syn	SSIM ↑	0.836	0.694	0.781	0.822	0.856	0.798
Gamma [47]	RGB-Syn	SSIM ↑	0.745	0.568	0.712	0.729	0.763	0.703
CSNorm-Fourier [76]	RGB-Syn	SSIM ↑	0.798	0.614	0.767	0.781	0.816	0.755
UEC-based [12]	RGB-Syn	SSIM ↑	0.833	0.659	0.804	0.827	0.852	0.795
MESS-generated	RGB-Syn	SSIM ↑	0.910	0.718	0.851	0.869	0.884	0.846
UHDLL [36]	Real	LPIPS ↓	0.084	0.177	0.202	0.158	0.091	0.142
LSRW [22]	Real	LPIPS ↓	0.113	0.149	0.196	0.164	0.125	0.149
SICE [7]	Real	LPIPS ↓	0.157	0.181	0.145	0.139	0.123	0.149
MSEC [1]	Raw-Syn	LPIPS ↓	0.142	0.178	0.166	0.131	0.117	0.147
BAID [48]	Real	LPIPS ↓	0.184	0.226	0.206	0.187	0.105	0.182
RCD [12]	Raw-Syn	LPIPS ↓	0.135	0.180	0.183	0.151	0.121	0.154
Gamma [47]	RGB-Syn	LPIPS ↓	0.235	0.292	0.258	0.231	0.206	0.244
CSNorm-Fourier [76]	RGB-Syn	LPIPS ↓	0.198	0.244	0.217	0.203	0.169	0.206
UEC-based [12]	RGB-Syn	LPIPS ↓	0.153	0.208	0.174	0.179	0.132	0.169
MESS-generated	RGB-Syn	LPIPS ↓	0.091	0.156	0.138	0.134	0.087	0.121

rate is initialized as 1×10^{-4} and decayed to 1×10^{-6} with a cosine schedule. The target EV is uniformly sampled from $[-3, 3]$ EV. The weight λ in the final loss is set to 1. For downstream image enhancement models, training images are cropped to 512×512. We use AdamW with an initial learning rate of 2×10^{-4}, batch size 64, cosine learning rate decay, and an ℓ_1 reconstruction loss. Each downstream model is trained for 300K iterations. Unless otherwise specified, architecture, resolution, optimizer, batch size, schedule, and training iterations are kept fixed across data sources.

Evaluation Protocol. To comprehensively evaluate the effectiveness of MESS, we select representative backbone models from each downstream task, and train them on MESS-generated data. Specifically, we employ FourLLIE [61], RetinexFormer [8], and Multinex [4] from LLIE, ExSlot [29] and CSEC [38]from EC, and RAVE [17] from BIE. We then evaluate these models on five commonly used test sets for different image enhancement tasks, including UHDLL [36], LSRW [22], SICE [7], MSEC [1], and BAID [48], to evaluate their generalization performance.

For comparison, we also train these models on three types of datasets to demonstrate the superiority of our MESS-generated data. (1) In-domain datasets, i.e., the associated training sets of the six employed test sets. (2) Datasets synthesized by RGB-based methods, including Gamma correction [47], UEC-based synthesis [12], and CSNorm-Fourier [76]. (3) Datasets synthesized by RAW-based methods, including MSEC [1] and RCD [12]. Note that MSEC is also an in-domain dataset in our experiment.

We report PSNR, SSIM and LPIPS [78] as full-reference (FR) metrics, and ARNIQA [2], MUSIQ [30], MANIQA [74], and TOPIQ [9] as no-reference (NR) metrics.

4.2 Main Results

Due to page limit, in the main paper we can only report the results using the backbone model FourLLIE [61]. The results using the other backbone models are provided in **Appendix C**.

Table 3: The results of no-reference scores using FourLLIE backbone [61]. The model trained on each data source are evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID	Average
UHDLL [36]	Real	ARNIQA ↑	0.681	0.650	0.694	0.693	0.703	0.684
LSRW [22]	Real	ARNIQA ↑	0.662	0.671	0.688	0.691	0.699	0.682
SICE [7]	Real	ARNIQA ↑	0.666	0.630	0.720	0.711	0.693	0.684
MSEC [1]	Raw-Syn	ARNIQA ↑	0.663	0.648	0.719	0.720	0.708	0.692
BAID [48]	Real	ARNIQA ↑	0.628	0.611	0.652	0.667	0.727	0.657
RCD [12]	Raw-Syn	ARNIQA ↑	0.663	0.640	0.693	0.697	0.699	0.678
Gamma [47]	RGB-Syn	ARNIQA ↑	0.625	0.604	0.662	0.673	0.685	0.650
CSNorm-Fourier [76]	RGB-Syn	ARNIQA ↑	0.631	0.611	0.676	0.681	0.695	0.659
UEC-based [12]	RGB-Syn	ARNIQA ↑	0.655	0.639	0.694	0.700	0.717	0.681
MESS-generated	RGB-Syn	ARNIQA ↑	0.698	0.661	0.726	0.721	0.738	0.709
UHDLL [36]	Real	MUSIQ ↑	57.904	50.639	50.855	52.439	54.013	53.170
LSRW [22]	Real	MUSIQ ↑	56.741	52.180	51.263	53.064	54.291	53.508
SICE [7]	Real	MUSIQ ↑	55.483	50.421	54.966	54.134	53.757	53.752
MSEC [1]	Raw-Syn	MUSIQ ↑	56.225	50.762	52.685	55.065	54.863	53.920
BAID [48]	Real	MUSIQ ↑	52.876	49.214	50.936	51.872	56.218	52.223
RCD [12]	Raw-Syn	MUSIQ ↑	56.191	50.623	51.706	53.255	54.001	53.155
Gamma [47]	RGB-Syn	MUSIQ ↑	53.218	48.973	51.495	52.320	53.273	51.856
CSNorm-Fourier [76]	RGB-Syn	MUSIQ ↑	54.130	49.435	52.216	53.054	53.921	52.551
UEC-based [12]	RGB-Syn	MUSIQ ↑	55.073	49.939	53.045	53.783	54.747	53.317
MESS-generated	RGB-Syn	MUSIQ ↑	58.136	52.087	55.672	55.214	56.442	55.510
UHDLL [36]	Real	MANIQA ↑	0.463	0.418	0.438	0.454	0.439	0.442
LSRW [22]	Real	MANIQA ↑	0.466	0.429	0.446	0.445	0.433	0.444
SICE [7]	Real	MANIQA ↑	0.452	0.408	0.467	0.461	0.430	0.444
MSEC [1]	Raw-Syn	MANIQA ↑	0.451	0.412	0.462	0.476	0.449	0.450
BAID [48]	Real	MANIQA ↑	0.421	0.397	0.426	0.432	0.457	0.427
RCD [12]	Raw-Syn	MANIQA ↑	0.454	0.411	0.441	0.451	0.436	0.439
Gamma [47]	RGB-Syn	MANIQA ↑	0.418	0.394	0.425	0.421	0.414	0.414
CSNorm-Fourier [76]	RGB-Syn	MANIQA ↑	0.436	0.402	0.437	0.440	0.424	0.428
UEC-based [12]	RGB-Syn	MANIQA ↑	0.448	0.422	0.448	0.451	0.434	0.441
MESS-generated	RGB-Syn	MANIQA ↑	0.478	0.439	0.473	0.482	0.465	0.467

Quantitative Results. Tables 2 and 3 report full-reference and no-reference results, respectively, from which we see that the training data are critical for model generalization. First, models trained on task-specific paired datasets perform competitively on their corresponding tasks but generalize poorly to other enhancement tasks. For example, UHDLL and LSRW mainly cover underexposure, while SICE [7], MSEC [1] and BAID [48] focus on exposure correction and backlit enhancement, respectively. Second, existing RGB-based synthesis improves scalability but over-simplifies degradation assumptions. Gamma correction [47] mainly models global intensity changes, while UEC-based [12] and CSNorm-Fourier [76] lack camera-consistent exposure variations. Third, RAW-based synthesis improves a little the generalization but remains limited in scalability. In contrast, MESS-generated data consistently improves both fidelity and perceptual quality across LLIE, EC, and BIE benchmarks, boosting significantly the generalization performance of the lightweight FourLLIE backbone model, which is originally designed for the LLIE task only.

Qualitative Results. Figure 4 shows qualitative comparisons for the LLIE task under cross-dataset settings. All models use the same architecture but are trained on different data sources, and are tested on LSRW. Although UHDLL [36] and LSRW [22] are both real-captured LLIE datasets, models trained on them still exhibit visible artifacts when tested across datasets. The UHDLL-trained model amplifies JPEG artifacts, while the LSRW-trained model, despite being trained and tested on LSRW, still produces blurred details. In contrast, the model trained on MESS-generated data avoids these artifacts and preserves sharper structures. Figure 5 shows another example of cross-dataset test using the EC task. All models use the same architecture but are trained on different data sources, and are tested on SICE. The RAW-based MSEC [1] training data fail to provide effective enhancement in this case. The RGB-based UEC [12] training data improve exposure but lead to unnatural color appearance.In contrast, MESS-generated data yield more natural exposure correction with better color fidelity and texture preservation. More visual comparisons can be found in **Appendix D**.

Ablation Study. We also perform ablation studies on the frequency-based enhancement module, diffusion-based detail enhancement module, and exposure-conditioned rendering module. The results

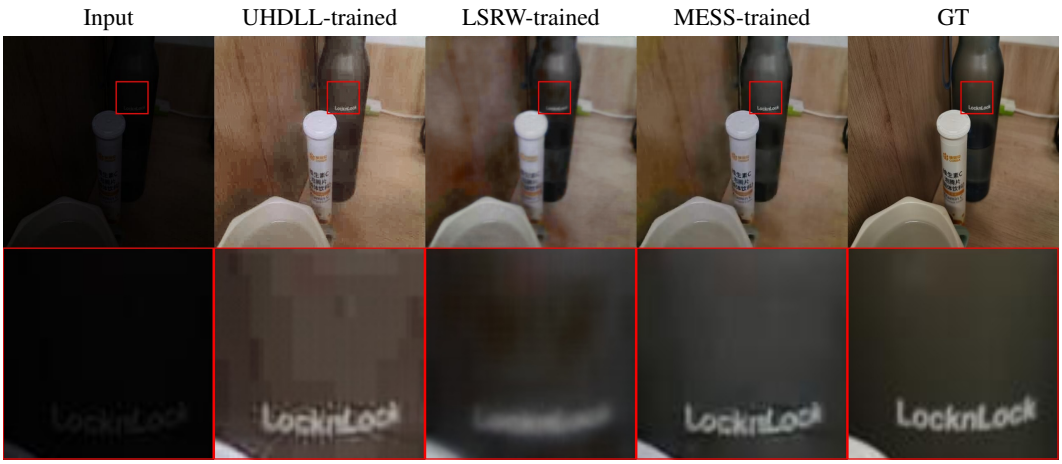

Figure 4: Qualitative comparison of the FourLLIE models trained on UHDLL, LSRW and our MESS generated datasets. The test image is from the LSRW dataset.

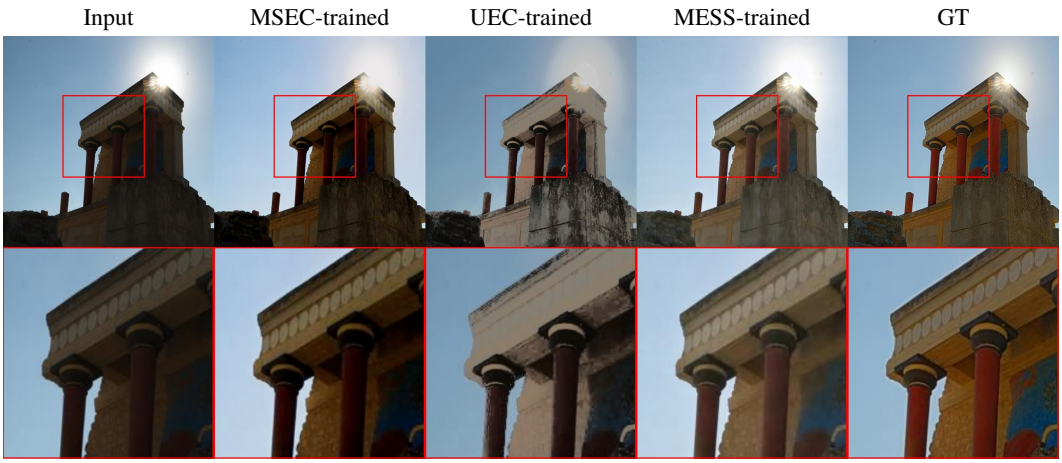

Figure 5: Qualitative comparison of the FourLLIE models trained on MSEC, UEC and MESS generated datasets. The test image is from the SICE dataset.

verify that each component contributes to the quality of synthesized data and the generalization ability of the downstream model. The details can be found in **Appendix E**.

Effect of Scale of Training Data. We evaluate the effect of training data scale by randomly sampling three subsets of MESS, containing 5K, 10K, and 50K pairs, respectively, and comparing them with the full 500K dataset. All settings use the FourLLIE backbone [61] and the same downstream training protocol described in the experimental setup. MESS-5K is approximately comparable in size to LSRW [22] and SICE [7]. Table 4 reports macro-average results over UHDLL [36], LSRW [22], SICE [7], MSEC [1], and BAID [48]. MESS-5K achieves better PSNR, SSIM and LPIPS scores than both the real-data baselines, possibly benefiting from its high-quality ground truth. Further gains with larger training sets highlight the benefits of both data quality and scale.

RAW-only versus RGB Expansion. We compare RAW-only training data with matched-size and full-size MESS-RGB datasets under identical downstream architectures and training protocols, using FourLLIE [61] as the backbone. The RAW-only dataset contains 39,642 scenes and training pairs, with exposure-perturbed inputs and MEF-generated targets obtained using an approach similar to UNICE [11]. The matched-size MESS-RGB dataset contains the same number of RGB scenes and pairs, using curated high-quality RGB images as targets. Table 5 shows that MESS-RGB dataset improves in all three macro-average full-reference metrics at matched size. The advantages reflect the better target quality of MESS-RGB data over RAW-only MEF targets. Further gains with the full 500K set suggest an additional benefit from increased data scale.

Table 4: Dataset-scale comparison using FourLLIE. Scores are macro-averaged over the five test sets. MESS-5K is approximately comparable in size to LSRW and SICE. Values in red and blue indicate the best and second-best results, respectively.

Training data	PSNR ↑	SSIM ↑	LPIPS ↓	ARNIQA ↑	MANIQA ↑	MUSIQ ↑
LSRW	19.663	0.816	0.149	0.682	0.444	53.508
SICE	19.965	0.816	0.149	0.684	0.444	53.752
MESS-5K	20.220	0.822	0.143	0.686	0.451	54.050
MESS-10K	20.590	0.830	0.137	0.692	0.456	54.410
MESS-50K	21.199	0.839	0.129	0.701	0.463	54.988
MESS-500K	21.923	0.846	0.121	0.709	0.467	55.510

Table 5: RAW-only and MESS-RGB comparison. Results are macro-averaged over UHDLL, LSRW, SICE, MSEC, and BAID. Values in red and blue indicate the best and second-best results, respectively.

Training data	Unique scenes	Pairs	PSNR ↑	SSIM ↑	LPIPS ↓
RAW-only	39,642	39,642	20.764	0.829	0.139
MESS-RGB, matched size	39,642	39,642	21.083	0.837	0.131
MESS-RGB, full set	500,000	500,000	21.923	0.846	0.121



Figure 6: Visualization of MESS synthesis. The source RGB image is followed by synthesized exposure variants from MESS at -3, -1, $+1$, and $+3$ EV, respectively.

Visualization of MESS Synthesis Outputs. Figure 6 shows exposure-controlled synthesis from a single in-the-wild RGB image. Varying the target EV produces under- and overexposed variants of the same source image.

5 Conclusion

In this paper, we proposed MESS, a scalable data synthesis framework for generalizable image illumination enhancement. MESS integrated frequency-based enhancement, diffusion-based detail compensation, and exposure-conditioned rendering. By transferring RAW-derived exposure variations to large-scale in-the-wild RGB images, MESS combined RAW-rendering realism with RGB-data scalability, enabling realistic and scalable training data synthesis. Using MESS, we constructed 500K training samples from high-quality in-the-wild RGB images, and trained a lightweight model on this dataset, which demonstrated strong generalization performance across diverse real-world illumination enhancement benchmarks. It is anticipated that MESS can advance scalable data synthesis for robust image enhancement in the wild.

Limitations. Despite the promising results, MESS has two main limitations. First, it currently relies on a fixed set of RAW datasets to derive exposure priors; expanding the RAW data pool could further improve the realism and coverage of the synthesized data. Second, MESS currently selects the best-rendered image based on several NR-IQA metrics; incorporating powerful Vision Language Models (VLMs) into this process could yield more robust data selection.

References

[1] Mahmoud Afifi, Konstantinos G. Derpanis, Björn Ommer, and Michael S. Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9153–9163, 2021.

[2] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 189–198, 2024.

[3] Elissavet Batziou, Konstantinos Ioannidis, Ioannis Patras, Stefanos Vrochidis, and Ioannis Kompatsiaris. Low-light image enhancement based on u-net and haar wavelet pooling. In *MultiMedia Modeling*, volume 13834 of *Lecture Notes in Computer Science*, pages 510–522. Springer, 2023. doi: 10.1007/978-3-031-27818-1_42.

[4] Alexandru Brateanu, Tingting Mu, Codruta O. Ancuti, and Cosmin Ancuti. Multinex: Lightweight low-light image enhancement via multi-prior retinex. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. doi: 10.1109/CVPR.2011.5995413.

[7] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4), 2018. doi: 10.1109/TIP.2018.2794218.

[8] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12504–12513, 2023.

[9] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024. doi: 10.1109/TIP.2024.3378466.

[10] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300, 2018.

[11] Ruodai Cui and Lei Zhang. Unice: Training a universal image contrast enhancer, 2025. URL `https://arxiv.org/abs/2507.17157`.

[12] Ruodai Cui, Li Niu, and Guosheng Hu. Unsupervised exposure correction. In *Computer Vision – ECCV 2024*, volume 15064 of *Lecture Notes in Computer Science*, pages 252–268. Springer, 2024. doi: 10.1007/978-3-031-72658-3_15.

[13] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 219–224, 2015.

[14] Frederic Drago, Karol Myszkowski, Thomas Annen, and Norishige Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum*, 22(3):419–426, 2003. doi: 10.1111/1467-8659.00689.

[15] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics*, 21(3):257–266, 2002.

[16] Yixu Feng, Shuo Hou, Haotian Lin, Yu Zhu, Peng Wu, Wei Dong, Jinqiu Sun, Qingsen Yan, and Yanning Zhang. Difflight: Integrating content and detail for low-light image enhancement. In *CVPR*, pages 6143–6152, 2024.

[17] Tatiana Gaintseva, Martin Benning, and Gregory Slabaugh. Rave: Residual vector embedding for clip-guided backlit image enhancement, 2024. URL `https://arxiv.org/abs/2404.01889`.

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023.

[19] Ardeshir A. Goshtasby. Fusion of multi-exposure images. *Image and Vision Computing*, 23(6):611–618, 2005.

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. doi: 10.1109/TIP.2016.2639450.

[22] 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. *JVCI*, 90:103712, 2023.

[23] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):192:1–192:12, 2016. doi: 10.1145/2980179.2980254.

[24] Jie Huang, Yajing Liu, Xueyang Fu, Man Zhou, Yang Wang, Feng Zhao, and Zhiwei Xiong. Exposure normalization and compensation for multiple-exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *ECCV*, pages 163–180, 2022.

[26] Jie Huang, Feng Zhao, Man Zhou, Jie Xiao, Naishan Zheng, Kaiwen Zheng, and Zhiwei Xiong. Learning sample relationship for exposure correction. In *CVPR*, pages 9904–9913, 2023.

[27] Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *ECCV*, pages 161–179, 2024.

[28] 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 Transactions on Image Processing*, 30:2340–2349, 2021. doi: 10.1109/TIP.2021.3051462.

[29] D. Jung, D. Kim, G. Wang, et al. Exposure-slot: Exposure-centric representations learning with slot-in-slot attention for region-aware exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17892–17901, 2025.

[30] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5148–5157, 2021.

[31] Woohyeok Kim, Geonu Kim, Junyong Lee, Seungyong Lee, Seung-Hwan Baek, and Sunghyun Cho. Paramisp: Learned forward and inverse isps using camera parameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26067–26076, 2024.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2020. doi: 10.1007/s11263-020-01316-z.

[34] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3487–3497, 2021.

[35] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *TPAMI*, 2021. doi: 10.1109/TPAMI.2021.3063604.

[36] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. In *International Conference on Learning Representations*, 2023.

[37] Hui Li, Kede Ma, Hongwei Yong, and Lei Zhang. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 2020.

[38] Yiyu Li, Ke Xu, Gerhard Petrus Hancke, and Rynson W.H. Lau. Color shift estimation-and-correction for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25389–25398, 2024.

[39] Zhe Li, Lei Zhang, Zheren Fu, Kun Zhang, and Zhendong Mao. Hierarchy-aware pseudo word learning with text adaptation for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24319–24329, 2025.

[40] Zhong-Yu Li, Xin Jin, Boyuan Sun, Chun-Le Guo, and Ming-Ming Cheng. Towards raw object detection in diverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[41] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 653–661, 2021.

[42] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8094–8103, 2023.

[43] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. Fine-grained textual inversion network for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 240–250, 2024. doi: 10.1145/3626772.3657831.

[44] Jiaying Liu, Dejia Xu, Wenhan Yang, Minhao Fan, and Haofeng Huang. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 129:1153–1184, 2021. doi: 10.1007/s11263-020-01418-8.

[45] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[46] Yu Long, Qinghua Lin, Zhihua Wang, Kai Zhang, Jianguo Zhang, and Yuming Fang. Enhancing low-light images: A synthetic data perspective on practical and generalizable solutions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5784–5792, 2025. doi: 10.1609/aaai.v39i6.32617.

[47] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. doi: 10.1016/j.patcog.2016.06.008.

[48] Xiaoqian Lv, Shengping Zhang, Qinglin Liu, Haozhe Xie, Bineng Zhong, and Huiyu Zhou. Backlitnet: A dataset and network for backlit image enhancement. *Computer Vision and Image Understanding*, 218: 103403, 2022. doi: 10.1016/j.cviu.2022.103403.

[49] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Pacific Conference on Computer Graphics and Applications*, 2007.

[50] Igor Morawski, Yu-An Chen, Yu-Sheng Lin, Shusil Dangi, Kai He, and Winston H. Hsu. Genisp: Neural isp for low-light machine cognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 630–639, 2022.

[51] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023. doi: 10.1145/3581783.3612137.

[52] A. Omid-Zohoor, C. Young, D. Ta, and B. Murmann. Towards always-on mobile object detection: Energy vs. performance tradeoffs for embedded hog feature extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–14, 2017. doi: 10.1109/TCSVT.2017.2653187.

[53] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3942–3951, 2018. doi: 10.1609/AAAI.V32I1.11671.

[54] Piotr Porwik and Agnieszka Lisowska. The haar-wavelet transform in digital image processing: Its status and achievements. *Machine Graphics and Vision*, 13(1/2):79–98, 2004.

[55] Erik Reinhard, Michael M. Stark, Peter Shirley, and James A. Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002. doi: 10.1145/566654.566575.

[56] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15144 of *Lecture Notes in Computer Science*, pages 87–103, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73016-0. doi: 10.1007/978-3-031-73016-0_6.

[57] Radomir S. Stanković and Bogdan J. Falkowski. The haar wavelet transform: Its status and achievements. *Computers & Electrical Engineering*, 29(1):25–44, 2003. doi: 10.1016/S0045-7906(01)00011-8.

[58] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2333–2343, 2025. doi: 10.1109/CVPR52734.2025.00223.

[59] Yoshiaki Ueda, Daiki Moriyama, Takanori Koga, and Noriaki Suetake. Histogram specification-based image enhancement for backlit image. In *ICIP*, pages 958–962, 2020.

[60] Unsplash. Unsplash full dataset 1.3.0. `https://unsplash.com/data`, 2025. Official dataset release for research and machine learning.

[61] Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *ACMMM*, pages 7459–7469, 2023.

[62] Haoyuan Wang, Ke Xu, and Rynson W. H. Lau. Local color distributions prior for image enhancement. In *Computer Vision – ECCV 2022*, volume 13678 of *Lecture Notes in Computer Science*, pages 343–359. Springer, 2022. doi: 10.1007/978-3-031-19797-0_20.

[63] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6849–6857, 2019.

[64] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22(9):3538–3548, 2013. doi: 10.1109/TIP.2013.2261309.

[65] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26057–26066, 2024.

[66] Yang Wang, Long Peng, Liang Li, Yang Cao, and Zheng-Jun Zha. Decoupling-and-aggregating for image exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[67] Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, and Shuaicheng Liu. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[68] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C. Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *ICCV*, pages 12438–12448, 2023.

[69] Zhihua Wang, Yu Long, Qinghua Lin, Kai Zhang, Yazhu Zhang, Yuming Fang, Li Liu, and Xiaochun Cao. Towards realistic low-light image enhancement via isp-driven data modeling. *arXiv preprint arXiv:2504.12204*, 2025.

[70] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[71] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2938.

[72] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[73] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[74] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1191–1200, 2022.

[75] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. doi: 10.1109/TIP.2021.3050850.

[76] Mingde Yao, Jie Huang, Xin Jin, Ruikang Xu, Shenglong Zhou, Man Zhou, and Zhiwei Xiong. Generalized lightness adaptation with channel selective normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10668–10679, 2023.

[77] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2696–2705, 2020.

[78] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[79] Tongshun Zhang, Pingping Liu, Yubing Lu, Mengen Cai, Zijian Zhang, Zhe Zhang, and Qiuzhan Zhou. Cwnet: Causal wavelet network for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Appendix

This appendix provides additional details and results to support the main paper. Sec. A presents implementation details, including the construction of RAW-guided training data and the network architecture of MESS. Sec. B describes how we use the trained MESS generator to construct the large-scale paired RGB training set. Sec. C reports additional quantitative results with other downstream enhancement backbones. Sec. D provides additional qualitative comparisons across low-light enhancement, exposure correction, and backlit enhancement benchmarks. Finally, Sec. E provides ablation studies that analyze the contribution of each component in MESS.

A Additional Implementation Details

A.1 Details of RAW-guided Training Data Construction

This section provides more details for the data construction process described in Sec. 3.1. We collect high-bit RAW images from publicly available datasets, including AODRAW [40], FiveK [6], HDRP [23], PASCAL-RAW [52], PPR10K [41], RAISE [13], and RAW-NOD [50]. Images with severe artifacts, invalid metadata, calibration charts, or extremely poor visual quality are removed before rendering. For each RAW image X, we first render a seven-exposure sequence by adjusting the exposure value to

$$\{-3, -2, -1, 0, +1, +2, +3\}\,\mathrm{EV}.$$

This sequence is used to construct two types of rendered images. We first render exposure-specific references with a camera-consistent ISP setting. The target exposure value is randomly sampled and normalized to $e_t \in [0, 1]$, yielding the reference image $I_t^\star$ for exposure supervision.

We then construct a style-varied rendering pool from the same RAW image. In addition to directly rendered exposure variants, we apply multiple multi-exposure fusion (MEF) [49, 19, 37] and tone mapping [55, 14, 15] algorithms to the seven-exposure sequence. These operations produce images with diverse tone, contrast, brightness, saturation, and local detail. During training, two images are randomly sampled from this pool to form I_i and I_j.

After training MESS on RAW-guided triplets, we apply it to high-quality RGB images collected from in-the-wild datasets. For each RGB image, we sample one or more target exposure values and synthesize exposure-varied outputs. One synthesized image is randomly selected as the input, while the original RGB image is used as the target. This process produces paired data for downstream enhancement training.

The final dataset contains 39,642 groups after manual filtering. Each group contains synthesized exposure-varied images associated with the same high-quality target image. The dataset statistics are summarized in Tab. 6, and representative examples are shown in Fig. 7.

A.2 Network Architecture Details

This section describes with more details the architecture of the proposed MESS network. As shown in Figure 2 in the main paper, MESS consists of three sequential modules: the frequency-based enhancement module $D(\cdot)$, the diffusion-based detail enhancement module $G(\cdot)$, and the exposure-conditioned rendering module $P(\cdot)$. All convolutional layers use reflection padding unless otherwise specified. LeakyReLU with a negative slope of 0.2 is used in all convolutional blocks.

A.2.1 Frequency-based Enhancement Module

The frequency-based enhancement module $D(\cdot)$ adopts an encoder–wavelet attention–decoder structure. Given an RGB input image $I_k \in \mathbb{R}^{3\times H\times W}$, where $k \in \{i, j\}$, the encoder $E_D(\cdot)$ maps it to a 64-channel feature map $H_k \in \mathbb{R}^{64\times H\times W}$.

The encoder consists of one 3×3 convolutional layer followed by four residual blocks. The first convolution maps the RGB input to 64 channels. Each residual block contains two 3×3 convolutional layers with 64 channels and a residual skip connection. The encoder does not use spatial downsampling.

Dataset	AODRAW	PASCAL-RAW	RAISE	FiveK
# Groups	4396	3971	6992	3651
Dataset	HDRP	PPR10K	RAW-NOD	Total
# Groups	3155	9790	7687	39642

Table 6: Statistics of the constructed training dataset after manual filtering. Each group contains synthesized exposure-varied images associated with the same high-quality target image.



Figure 7: Representative examples of the constructed dataset. The style-varied renderings show diverse tones, contrast, and local details, while the exposure-varied renderings correspond to different EV settings of the same scene.

A single-level Haar DWT is applied to H_k. For each feature channel, the transform produces one low-frequency sub-band and three high-frequency sub-bands, denoted by LL, LH, HL, and HH. The four sub-bands are concatenated along the channel dimension, resulting in a wavelet feature $F_k \in \mathbb{R}^{256\times H/2\times W/2}$.

A wavelet-domain channel attention block adaptively reweights the concatenated sub-band feature. Specifically, global average pooling and global max pooling are applied to F_k, producing two 256-dimensional channel descriptors. The two descriptors are passed through a shared two-layer MLP with hidden dimension 64 and output dimension 256. The outputs of the two branches are summed and activated by a sigmoid function to obtain the channel attention weights A_k. The attended wavelet feature is computed as $\tilde{F}_k = A_k \odot F_k$, where A_k is broadcast along the spatial dimensions.

The attended wavelet feature is transformed back by inverse Haar DWT, producing $\tilde{H}_k \in \mathbb{R}^{64\times H\times W}$. The decoder $Q_D(\cdot)$ then maps $\tilde{H}_k$ to the enhanced representation C_k. The decoder contains four residual blocks with 64 channels and a final 3×3 convolutional layer that maps 64 channels to 3 channels. A sigmoid activation is applied at the end to constrain the output to $[0, 1]$.

A.2.2 Diffusion-based Detail Enhancement Module

The diffusion-based detail enhancement module $G(\cdot)$ is built on SD-Turbo and operates in the latent space. It consists of a frozen VAE encoder, a LoRA-adapted U-Net, a learnable pseudo-word embedding, and a frozen VAE decoder.

Given the enhanced representation C_i, we first normalize it from $[0, 1]$ to $[-1, 1]$ and feed it into the frozen VAE encoder to obtain the latent code z. The latent code is then processed by the SD-Turbo U-Net conditioned on a learnable pseudo-word embedding $\mathbf{p}$. The output latent z' is decoded by the frozen VAE decoder and rescaled to $[0, 1]$, producing the intermediate image I_m.

The pseudo-word embedding $\mathbf{p}$ consists of a single learnable token and is shared across all training samples. No hand-crafted text prompt is used.

LoRA adapters are inserted into the attention projection layers of the U-Net, including the query, key, value, and output projections in both self-attention and cross-attention blocks. For a frozen projection matrix W, the LoRA-adapted projection is defined as

$$y = Wx + \frac{\alpha}{r} BAx, \tag{1}$$

where A and B are low-rank trainable matrices. We use LoRA rank $r = 8$ and scaling factor $\alpha = 8$. Only the LoRA parameters and the pseudo-word embedding are learnable in this module, while the VAE encoder, VAE decoder, and original U-Net weights remain frozen.

A.2.3 Exposure-conditioned Rendering Module

The exposure-conditioned rendering module $P(\cdot)$ contains a convolutional encoder, a wavelet-domain FiLM modulation block, and a decoder. Given the diffusion-enhanced image $I_m \in \mathbb{R}^{3\times H\times W}$, the encoder $E_P(\cdot)$ extracts a 64-channel feature map $H_m \in \mathbb{R}^{64\times H\times W}$.

The encoder $E_P(\cdot)$ has the same structure as $E_D(\cdot)$, consisting of one 3×3 convolutional layer followed by four residual blocks. The first convolution maps the RGB input to 64 channels, and each residual block contains two 3×3 convolutional layers with a residual skip connection.

A single-level Haar DWT is applied to H_m. The resulting LL, LH, HL, and HH sub-bands are concatenated along the channel dimension, producing a wavelet feature $F_m \in \mathbb{R}^{256\times H/2\times W/2}$.

The scalar target exposure value $e_t \in [0, 1]$ is encoded by a two-layer MLP. The exposure encoder maps the one-dimensional input to a 128-dimensional embedding through a 64-dimensional hidden layer, with a ReLU activation after the first linear layer.

To predict content-adaptive modulation parameters, global average pooling and global max pooling are applied to F_m, producing two 256-dimensional descriptors. These descriptors are concatenated with the 128-dimensional exposure embedding, resulting in a 640-dimensional vector. A two-layer FiLM predictor maps this vector to a 512-dimensional output, which is split into two 256-dimensional vectors, γ_t and β_t. The wavelet feature is modulated by

$$\tilde{F}_t = (1 + \gamma_t) \odot F_m + \beta_t, \tag{2}$$

where γ_t and β_t are broadcast along the spatial dimensions. The residual form $1 + \gamma_t$ is used to stabilize affine modulation.

The modulated feature $\tilde{F}_t$ is transformed back by inverse Haar DWT, yielding a 64-channel feature map. The decoder $Q_P(\cdot)$ contains four residual blocks with 64 channels and a final 3×3 convolutional layer that maps 64 channels to 3 channels. A sigmoid activation is applied to produce the final RGB output $\hat{I}_t \in [0, 1]^{3\times H\times W}$.

B Construction of Large-scale Paired RGB Training Data

After training on RAW-guided triplets, MESS is used as an exposure-variation generator to construct paired RGB data at scale. Given a selected high-quality RGB image I_{hq}, we treat it as the normal-exposure reference and synthesize an exposure-degraded counterpart $\tilde{I}$ by feeding I_{hq} and a sampled exposure condition into MESS. The resulting pair $(\tilde{I}, I_{\mathrm{hq}})$ is used for downstream enhancement training, where $\tilde{I}$ serves as the degraded input and I_{hq} serves as the restoration target. Although MESS can generate multiple exposure variants for each source image, we store one synthesized degraded image for each target image in the final downstream training set, resulting in 500K paired RGB samples.

B.1 Source Image Selection

We start from approximately 20 million in-the-wild RGB images collected from public image sources, including Unsplash [60], SA-1B [32], and Open Images [33]. Before quality ranking, we remove images that are unsuitable as restoration targets, such as images with obvious motion blur, strong compression artifacts, severe sensor noise, extremely low contrast, or large saturated or under-exposed regions. Near-duplicate images are also removed using perceptual hashing and visual feature similarity to improve scene diversity. For the remaining candidates, we evaluate perceptual quality using multiple no-reference IQA models, including ARNIQA [2], MUSIQ [30], MANIQA [74], and TOPIQ [9]. Since different IQA models have different score ranges and preference directions, their outputs are normalized before computing an ensemble quality score. The top 500K images according to the ensemble score are retained as high-quality RGB targets. This strategy reduces the bias of any single quality metric and favors images that are consistently sharp, natural, well-exposed, and visually pleasing.

B.2 Exposure-degraded Input Synthesis

For each selected target image, we sample one exposure condition within the same exposure range used for MESS training, corresponding to $[-3, +3]$ EV before normalization. Exposure values close to normal exposure are avoided, since they usually produce trivial pairs with limited restoration difficulty. Under-exposure and over-exposure cases are synthesized with equal probability. To emphasize challenging degradations, the sampling distribution is biased toward larger exposure shifts, while a smaller portion of moderate exposure shifts is retained to improve training stability. After synthesis, each generated pair is checked by lightweight validity tests. We reject a synthesized input if it is nearly identical to the target, contains excessive black or saturated pixels, or shows visible structural artifacts produced by MESS. For rejected samples, a new exposure condition is sampled and the synthesis process is repeated.

B.3 Final Paired Dataset

The final RGB training set contains 500K pairs. Each pair consists of one MESS-synthesized exposure-degraded image and the corresponding high-quality RGB target. Because the degraded images are generated from diverse in-the-wild scenes and cover both under-exposure and over-exposure conditions, the resulting dataset provides broad scene and degradation diversity for training a generalizable enhancement network.

C Additional Results with Other Downstream Backbones

In the main paper, we report the results using FourLLIE [61] as the representative downstream backbone due to space limitation. In this section, we provide additional experiments with more downstream image enhancement models to further verify the generality of MESS-generated training data. Specifically, we evaluate RetinexFormer [8] and Multinex [4] for low-light image enhancement, ExSLOT [29] and CSEC [38] for exposure correction, and RAVE [17] for backlit image enhancement.

For each downstream architecture, we keep the network architecture, training resolution, optimizer, batch size, learning rate schedule, loss function, and number of training iterations fixed. Only the training data source is changed. All trained models are evaluated on the same five test sets, including UHDLL [36], LSRW [22], SICE [7], MSEC [1], and BAID [48]. We report full-reference metrics, including PSNR, SSIM, and LPIPS [78], and no-reference metrics, including ARNIQA [2], MUSIQ [30], and MANIQA [74].

C.1 Results with Low-Light Image Enhancement Backbones

Evaluation with etinexFormer. Tables 7 and 8 show the full-reference and no-reference results using RetinexFormer [8] as the downstream backbone. The results demonstrate that MESS-generated data consistently provides strong cross-dataset generalization. Compared with task-specific paired datasets, MESS-generated data achieves more balanced performance across low-light, exposure correction, and backlit enhancement benchmarks. Compared with RGB-based synthetic datasets, MESS-generated data yields better perceptual quality and stronger robustness to diverse illumination degradations.

Evaluation with Multinex. Tables 9 and 10 show the results using Multinex [4] as the downstream backbone. MESS-generated data again achieves competitive or superior performance across different test sets. This indicates that the benefit of MESS is not tied to a specific LLIE architecture, but comes from the improved diversity and camera-consistent exposure supervision of the generated data.

C.2 Results with Exposure Correction Backbones

Evaluation with ExSLOT. Tables 11 and 12 report the results using ExSLOT [29] as the downstream backbone. Only the training data source is changed, while the backbone and training setting remain fixed. The results verify whether MESS-generated data can improve the generalization ability of an exposure correction architecture across LLIE, EC, and BIE benchmarks.

Evaluation with CSEC. Tables 13 and 14 report the results using CSEC [38] as the downstream backbone. Similar to other backbones, the architecture and training protocol are fixed, and only the data source is changed. These experiments further examine whether MESS-generated training data provides architecture-agnostic improvements.

C.3 Results with Backlit Image Enhancement Backbone

Evaluation with RAVE. Tables 15 and 16 report the results using RAVE [17] as the downstream backbone. The same RAVE architecture is trained with different data sources and evaluated on all test sets without fine-tuning. This setting examines whether MESS-generated data can also benefit a backlit image enhancement architecture under cross-task evaluation.

D Additional Qualitative Results

In this section, we provide additional qualitative results to further analyze the generalization ability of MESS-generated training data. Figs. 8, 9, and 10 compare MESS with models trained on real-captured task-specific datasets, i.e., UHDLL and LSRW. These results show that although real paired datasets are physically valid, their limited scene and degradation diversity makes the trained models less robust under cross-dataset and cross-task evaluation. In particular, models trained on UHDLL or LSRW often over-brighten sky and building regions, desaturate vegetation, blur window or facade structures, and amplify local artifacts.

Figs. 11, 12, and 13 compare MESS with different data synthesis strategies, including RAW-based rendering with MSEC and RGB-based rendering with UEC. As shown in these qualitative comparisons, the MESS-trained model achieves more stable visual quality, more natural color reproduction, and better preservation of object structures across low-light enhancement, exposure correction, and backlit enhancement scenarios.

E Ablation Study

We conduct ablation studies to evaluate the contribution of each component in MESS. All variants use the same FourLLIE backbone to synthesize paired training data, with identical downstream training and evaluation protocols. The results are averaged over UHDLL, LSRW, SICE, MSEC, and BAID.

As shown in Table 17, removing any major component consistently degrades performance. The representation consistency loss improves the robustness of the learned scene representation against rendering-style variations. The frequency-based enhancement module helps separate structural information from exposure-related appearance changes, leading to more reliable synthesized training pairs. The diffusion-based detail enhancement module mainly benefits perceptual quality by compensating for missing local details. The exposure-conditioned rendering module is crucial for generating controllable exposure variants; without explicit exposure conditioning, the synthesized pairs contain less accurate exposure relationships and lead to weaker downstream generalization. Replacing DWT with FFT performs better than removing frequency modeling, but remains inferior to the full model, confirming the advantage of spatially localized wavelet representations.

Table 18 compares different exposure injection strategies. The renderer without exposure condition performs the worst, while directly concatenating the exposure value with RGB features brings only limited improvement. RGB-domain FiLM is more effective, but still lacks explicit frequency decomposition. DWT-FiLM further improves the results by modulating multi-resolution frequency components. With additional content descriptors, the full renderer achieves the best performance across all metrics, demonstrating the effectiveness of content-adaptive exposure modulation for realistic exposure synthesis.

Table 7: Full-reference results using RetinexFormer [8] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	PSNR ↑	22.509	19.024	15.687	18.745	20.025
LSRW [22]	Real	PSNR ↑	20.175	20.496	16.745	18.980	19.440
SICE [7]	Real	PSNR ↑	17.770	17.711	21.899	20.300	19.645
MSEC [1]	Raw-Syn	PSNR ↑	18.162	17.890	19.910	22.809	19.910
BAID [48]	Real	PSNR ↑	16.508	16.591	17.782	18.750	21.471
RCD [12]	Raw-Syn	PSNR ↑	18.701	17.855	18.641	20.768	20.243
Gamma [47]	RGB-Syn	PSNR ↑	15.659	14.311	16.759	17.671	17.666
CSNorm-Fourier [76]	RGB-Syn	PSNR ↑	16.885	16.090	18.370	19.209	18.950
UEC-based [12]	RGB-Syn	PSNR ↑	17.976	17.473	19.731	20.547	19.981
MESS-generated	RGB-Syn	PSNR ↑	22.231	20.161	21.591	22.936	21.210
UHDLL [36]	Real	SSIM ↑	0.901	0.715	0.763	0.819	0.892
LSRW [22]	Real	SSIM ↑	0.876	0.746	0.784	0.816	0.878
SICE [7]	Real	SSIM ↑	0.812	0.702	0.864	0.841	0.881
MSEC [1]	Raw-Syn	SSIM ↑	0.821	0.714	0.816	0.886	0.878
BAID [48]	Real	SSIM ↑	0.771	0.657	0.764	0.793	0.917
RCD [12]	Raw-Syn	SSIM ↑	0.824	0.704	0.792	0.852	0.884
Gamma [47]	RGB-Syn	SSIM ↑	0.731	0.584	0.714	0.728	0.789
CSNorm-Fourier [76]	RGB-Syn	SSIM ↑	0.783	0.630	0.766	0.783	0.838
UEC-based [12]	RGB-Syn	SSIM ↑	0.822	0.672	0.803	0.827	0.876
MESS-generated	RGB-Syn	SSIM ↑	0.897	0.733	0.851	0.882	0.908
UHDLL [36]	Real	LPIPS ↓	0.068	0.172	0.192	0.158	0.083
LSRW [22]	Real	LPIPS ↓	0.095	0.143	0.189	0.164	0.103
SICE [7]	Real	LPIPS ↓	0.141	0.176	0.130	0.143	0.101
MSEC [1]	Raw-Syn	LPIPS ↓	0.124	0.172	0.159	0.132	0.094
BAID [48]	Real	LPIPS ↓	0.168	0.222	0.199	0.188	0.066
RCD [12]	Raw-Syn	LPIPS ↓	0.122	0.185	0.178	0.149	0.082
Gamma [47]	RGB-Syn	LPIPS ↓	0.219	0.286	0.250	0.202	0.148
CSNorm-Fourier [76]	RGB-Syn	LPIPS ↓	0.181	0.238	0.208	0.230	0.185
UEC-based [12]	RGB-Syn	LPIPS ↓	0.137	0.200	0.167	0.180	0.110
MESS-generated	RGB-Syn	LPIPS ↓	0.075	0.150	0.137	0.129	0.070

Table 8: No-reference results using RetinexFormer [8] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	ARNIQA ↑	0.667	0.657	0.693	0.694	0.680
LSRW [22]	Real	ARNIQA ↑	0.648	0.685	0.685	0.692	0.678
SICE [7]	Real	ARNIQA ↑	0.651	0.643	0.717	0.711	0.671
MSEC [1]	Raw-Syn	ARNIQA ↑	0.647	0.658	0.707	0.713	0.684
BAID [48]	Real	ARNIQA ↑	0.613	0.626	0.650	0.666	0.705
RCD [12]	Raw-Syn	ARNIQA ↑	0.641	0.651	0.691	0.708	0.681
Gamma [47]	RGB-Syn	ARNIQA ↑	0.611	0.616	0.658	0.673	0.662
CSNorm-Fourier [76]	RGB-Syn	ARNIQA ↑	0.616	0.624	0.671	0.682	0.672
UEC-based [12]	RGB-Syn	ARNIQA ↑	0.643	0.653	0.690	0.699	0.695
MESS-generated	RGB-Syn	ARNIQA ↑	0.684	0.675	0.723	0.720	0.716
UHDLL [36]	Real	MUSIQ ↑	57.696	50.842	50.615	52.444	53.977
LSRW [22]	Real	MUSIQ ↑	56.335	52.359	50.991	53.050	54.254
SICE [7]	Real	MUSIQ ↑	55.283	50.596	54.710	54.120	53.720
MSEC [1]	Raw-Syn	MUSIQ ↑	55.990	50.936	52.401	54.951	53.831
BAID [48]	Real	MUSIQ ↑	52.681	49.143	50.675	51.855	56.205
RCD [12]	Raw-Syn	MUSIQ ↑	55.661	50.558	51.696	53.987	54.331
Gamma [47]	RGB-Syn	MUSIQ ↑	53.000	49.417	51.267	52.325	53.268
CSNorm-Fourier [76]	RGB-Syn	MUSIQ ↑	53.921	49.604	51.952	53.055	53.892
UEC-based [12]	RGB-Syn	MUSIQ ↑	54.843	50.089	52.785	53.791	54.757
MESS-generated	RGB-Syn	MUSIQ ↑	57.928	52.266	55.416	55.112	56.429
UHDLL [36]	Real	MANIQA ↑	0.465	0.407	0.432	0.455	0.422
LSRW [22]	Real	MANIQA ↑	0.461	0.419	0.437	0.445	0.418
SICE [7]	Real	MANIQA ↑	0.449	0.398	0.460	0.460	0.414
MSEC [1]	Raw-Syn	MANIQA ↑	0.450	0.402	0.455	0.466	0.425
BAID [48]	Real	MANIQA ↑	0.419	0.387	0.417	0.431	0.440
RCD [12]	Raw-Syn	MANIQA ↑	0.443	0.404	0.445	0.464	0.422
Gamma [47]	RGB-Syn	MANIQA ↑	0.418	0.383	0.420	0.419	0.396
CSNorm-Fourier [76]	RGB-Syn	MANIQA ↑	0.436	0.392	0.431	0.439	0.409
UEC-based [12]	RGB-Syn	MANIQA ↑	0.448	0.412	0.439	0.450	0.415
MESS-generated	RGB-Syn	MANIQA ↑	0.477	0.429	0.466	0.462	0.448

Table 9: Full-reference results using Multinex [4] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	PSNR ↑	20.622	17.579	14.033	18.743	18.473
LSRW [22]	Real	PSNR ↑	18.295	19.061	15.103	18.971	17.878
SICE [7]	Real	PSNR ↑	15.878	16.274	20.073	20.302	18.084
MSEC [1]	Raw-Syn	PSNR ↑	16.274	16.448	18.265	22.322	18.345
BAID [48]	Real	PSNR ↑	14.611	15.146	16.149	18.748	19.918
RCD [12]	Raw-Syn	PSNR ↑	16.818	16.417	16.995	20.769	18.685
Gamma [47]	RGB-Syn	PSNR ↑	13.768	12.886	15.123	17.669	16.110
CSNorm-Fourier [76]	RGB-Syn	PSNR ↑	14.994	14.662	16.731	19.209	17.382
UEC-based [12]	RGB-Syn	PSNR ↑	16.087	16.038	18.092	20.551	18.433
MESS-generated	RGB-Syn	PSNR ↑	20.343	18.727	19.950	22.181	19.657
UHDLL [36]	Real	SSIM ↑	0.868	0.689	0.734	0.819	0.869
LSRW [22]	Real	SSIM ↑	0.844	0.720	0.756	0.815	0.855
SICE [7]	Real	SSIM ↑	0.778	0.676	0.827	0.841	0.859
MSEC [1]	Raw-Syn	SSIM ↑	0.787	0.688	0.787	0.875	0.855
BAID [48]	Real	SSIM ↑	0.738	0.631	0.736	0.792	0.894
RCD [12]	Raw-Syn	SSIM ↑	0.799	0.674	0.762	0.857	0.863
Gamma [47]	RGB-Syn	SSIM ↑	0.699	0.557	0.685	0.744	0.765
CSNorm-Fourier [76]	RGB-Syn	SSIM ↑	0.751	0.603	0.738	0.766	0.816
UEC-based [12]	RGB-Syn	SSIM ↑	0.788	0.646	0.776	0.827	0.853
MESS-generated	RGB-Syn	SSIM ↑	0.864	0.707	0.832	0.879	0.885
UHDLL [36]	Real	LPIPS ↓	0.118	0.200	0.224	0.159	0.124
LSRW [22]	Real	LPIPS ↓	0.146	0.170	0.221	0.164	0.143
SICE [7]	Real	LPIPS ↓	0.191	0.203	0.162	0.143	0.142
MSEC [1]	Raw-Syn	LPIPS ↓	0.175	0.200	0.190	0.135	0.135
BAID [48]	Real	LPIPS ↓	0.218	0.248	0.231	0.188	0.106
RCD [12]	Raw-Syn	LPIPS ↓	0.179	0.216	0.206	0.144	0.125
Gamma [47]	RGB-Syn	LPIPS ↓	0.269	0.314	0.282	0.210	0.200
CSNorm-Fourier [76]	RGB-Syn	LPIPS ↓	0.231	0.266	0.240	0.222	0.214
UEC-based [12]	RGB-Syn	LPIPS ↓	0.187	0.229	0.199	0.179	0.151
MESS-generated	RGB-Syn	LPIPS ↓	0.125	0.177	0.169	0.132	0.111

Table 10: No-reference results using Multinex [4] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	ARNIQA ↑	0.614	0.622	0.672	0.692	0.646
LSRW [22]	Real	ARNIQA ↑	0.595	0.645	0.666	0.689	0.639
SICE [7]	Real	ARNIQA ↑	0.598	0.605	0.699	0.710	0.637
MSEC [1]	Raw-Syn	ARNIQA ↑	0.612	0.618	0.690	0.702	0.647
BAID [48]	Real	ARNIQA ↑	0.559	0.585	0.631	0.675	0.669
RCD [12]	Raw-Syn	ARNIQA ↑	0.602	0.615	0.672	0.703	0.653
Gamma [47]	RGB-Syn	ARNIQA ↑	0.560	0.577	0.641	0.666	0.627
CSNorm-Fourier [76]	RGB-Syn	ARNIQA ↑	0.564	0.585	0.656	0.680	0.636
UEC-based [12]	RGB-Syn	ARNIQA ↑	0.586	0.614	0.673	0.701	0.659
MESS-generated	RGB-Syn	ARNIQA ↑	0.631	0.635	0.705	0.708	0.680
UHDLL [36]	Real	MUSIQ ↑	58.031	51.432	55.099	52.412	58.954
LSRW [22]	Real	MUSIQ ↑	56.670	52.985	55.401	53.058	59.207
SICE [7]	Real	MUSIQ ↑	55.588	51.224	59.133	54.147	58.732
MSEC [1]	Raw-Syn	MUSIQ ↑	56.328	51.563	56.852	54.650	58.789
BAID [48]	Real	MUSIQ ↑	53.023	50.013	55.005	51.868	61.163
RCD [12]	Raw-Syn	MUSIQ ↑	56.001	51.229	56.137	53.982	59.295
Gamma [47]	RGB-Syn	MUSIQ ↑	53.318	49.751	55.660	52.316	58.237
CSNorm-Fourier [76]	RGB-Syn	MUSIQ ↑	54.259	50.246	56.391	53.079	58.858
UEC-based [12]	RGB-Syn	MUSIQ ↑	55.227	50.753	57.241	53.786	59.709
MESS-generated	RGB-Syn	MUSIQ ↑	58.263	52.892	59.839	54.507	61.387
UHDLL [36]	Real	MANIQA ↑	0.417	0.372	0.407	0.454	0.377
LSRW [22]	Real	MANIQA ↑	0.412	0.382	0.416	0.446	0.368
SICE [7]	Real	MANIQA ↑	0.402	0.360	0.437	0.460	0.368
MSEC [1]	Raw-Syn	MANIQA ↑	0.400	0.367	0.431	0.455	0.378
BAID [48]	Real	MANIQA ↑	0.370	0.351	0.396	0.434	0.402
RCD [12]	Raw-Syn	MANIQA ↑	0.397	0.361	0.412	0.461	0.387
Gamma [47]	RGB-Syn	MANIQA ↑	0.368	0.346	0.396	0.421	0.353
CSNorm-Fourier [76]	RGB-Syn	MANIQA ↑	0.385	0.353	0.406	0.441	0.362
UEC-based [12]	RGB-Syn	MANIQA ↑	0.395	0.375	0.418	0.451	0.370
MESS-generated	RGB-Syn	MANIQA ↑	0.427	0.392	0.442	0.462	0.394

Table 11: Full-reference results using ExSLOT [29] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	PSNR ↑	19.254	16.508	12.332	17.656	16.346
LSRW [22]	Real	PSNR ↑	16.957	17.977	13.427	18.935	15.748
SICE [7]	Real	PSNR ↑	14.518	15.173	17.934	20.345	16.013
MSEC [1]	Raw-Syn	PSNR ↑	14.927	15.361	16.544	23.200	16.222
BAID [48]	Real	PSNR ↑	13.270	14.047	14.439	18.762	17.823
RCD [12]	Raw-Syn	PSNR ↑	15.469	15.322	15.283	20.558	16.564
Gamma [47]	RGB-Syn	PSNR ↑	12.381	11.825	13.404	18.721	14.002
CSNorm-Fourier [76]	RGB-Syn	PSNR ↑	13.646	13.585	14.983	19.243	15.261
UEC-based [12]	RGB-Syn	PSNR ↑	14.697	14.989	16.393	20.578	16.332
MESS-generated	RGB-Syn	PSNR ↑	18.976	17.642	18.550	23.329	17.562
UHDLL [36]	Real	SSIM ↑	0.859	0.673	0.660	0.821	0.857
LSRW [22]	Real	SSIM ↑	0.837	0.702	0.733	0.813	0.843
SICE [7]	Real	SSIM ↑	0.769	0.660	0.811	0.841	0.848
MSEC [1]	Raw-Syn	SSIM ↑	0.776	0.666	0.762	0.890	0.842
BAID [48]	Real	SSIM ↑	0.725	0.613	0.711	0.793	0.882
RCD [12]	Raw-Syn	SSIM ↑	0.782	0.654	0.735	0.856	0.856
Gamma [47]	RGB-Syn	SSIM ↑	0.689	0.539	0.708	0.729	0.752
CSNorm-Fourier [76]	RGB-Syn	SSIM ↑	0.744	0.585	0.714	0.782	0.804
UEC-based [12]	RGB-Syn	SSIM ↑	0.779	0.631	0.752	0.828	0.841
MESS-generated	RGB-Syn	SSIM ↑	0.855	0.689	0.798	0.894	0.873
UHDLL [36]	Real	LPIPS ↓	0.127	0.218	0.240	0.158	0.138
LSRW [22]	Real	LPIPS ↓	0.154	0.190	0.235	0.165	0.158
SICE [7]	Real	LPIPS ↓	0.202	0.224	0.177	0.143	0.156
MSEC [1]	Raw-Syn	LPIPS ↓	0.184	0.218	0.206	0.121	0.153
BAID [48]	Real	LPIPS ↓	0.229	0.267	0.247	0.187	0.121
RCD [12]	Raw-Syn	LPIPS ↓	0.185	0.229	0.228	0.145	0.145
Gamma [47]	RGB-Syn	LPIPS ↓	0.240	0.331	0.298	0.230	0.241
CSNorm-Fourier [76]	RGB-Syn	LPIPS ↓	0.278	0.284	0.257	0.205	0.201
UEC-based [12]	RGB-Syn	LPIPS ↓	0.198	0.250	0.211	0.178	0.165
MESS-generated	RGB-Syn	LPIPS ↓	0.134	0.197	0.184	0.125	0.125

Table 12: No-reference results using ExSLOT [29] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	ARNIQA ↑	0.661	0.650	0.693	0.692	0.680
LSRW [22]	Real	ARNIQA ↑	0.640	0.673	0.690	0.690	0.675
SICE [7]	Real	ARNIQA ↑	0.646	0.633	0.727	0.709	0.669
MSEC [1]	Raw-Syn	ARNIQA ↑	0.641	0.652	0.720	0.728	0.687
BAID [48]	Real	ARNIQA ↑	0.610	0.613	0.654	0.669	0.705
RCD [12]	Raw-Syn	ARNIQA ↑	0.639	0.649	0.707	0.707	0.682
Gamma [47]	RGB-Syn	ARNIQA ↑	0.604	0.607	0.664	0.673	0.662
CSNorm-Fourier [76]	RGB-Syn	ARNIQA ↑	0.610	0.615	0.678	0.682	0.675
UEC-based [12]	RGB-Syn	ARNIQA ↑	0.633	0.640	0.694	0.700	0.695
MESS-generated	RGB-Syn	ARNIQA ↑	0.678	0.663	0.721	0.725	0.716
UHDLL [36]	Real	MUSIQ ↑	57.446	50.672	52.614	52.452	56.011
LSRW [22]	Real	MUSIQ ↑	56.085	52.205	52.996	53.064	56.266
SICE [7]	Real	MUSIQ ↑	55.039	50.460	56.700	54.161	55.777
MSEC [1]	Raw-Syn	MUSIQ ↑	55.782	50.815	54.446	55.648	55.843
BAID [48]	Real	MUSIQ ↑	52.403	49.264	52.691	51.901	58.217
RCD [12]	Raw-Syn	MUSIQ ↑	55.431	50.475	53.725	53.998	56.359
Gamma [47]	RGB-Syn	MUSIQ ↑	52.746	48.979	53.211	52.339	55.288
CSNorm-Fourier [76]	RGB-Syn	MUSIQ ↑	53.669	49.489	53.941	53.037	55.923
UEC-based [12]	RGB-Syn	MUSIQ ↑	54.643	49.963	54.779	53.801	56.759
MESS-generated	RGB-Syn	MUSIQ ↑	57.678	52.112	57.406	55.814	58.441
UHDLL [36]	Real	MANIQA ↑	0.444	0.387	0.423	0.455	0.410
LSRW [22]	Real	MANIQA ↑	0.440	0.397	0.430	0.446	0.401
SICE [7]	Real	MANIQA ↑	0.428	0.378	0.452	0.460	0.399
MSEC [1]	Raw-Syn	MANIQA ↑	0.446	0.380	0.447	0.486	0.408
BAID [48]	Real	MANIQA ↑	0.400	0.365	0.413	0.434	0.426
RCD [12]	Raw-Syn	MANIQA ↑	0.433	0.373	0.431	0.468	0.415
Gamma [47]	RGB-Syn	MANIQA ↑	0.396	0.362	0.412	0.420	0.385
CSNorm-Fourier [76]	RGB-Syn	MANIQA ↑	0.412	0.370	0.420	0.441	0.394
UEC-based [12]	RGB-Syn	MANIQA ↑	0.427	0.388	0.434	0.450	0.402
MESS-generated	RGB-Syn	MANIQA ↑	0.456	0.407	0.453	0.493	0.434

Table 13: Full-reference results using CSEC [38] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	PSNR ↑	19.145	16.386	11.687	18.706	15.269
LSRW [22]	Real	PSNR ↑	16.860	17.850	12.807	18.961	14.696
SICE [7]	Real	PSNR ↑	14.421	15.039	17.321	20.334	14.905
MSEC [1]	Raw-Syn	PSNR ↑	14.824	15.197	15.910	22.728	15.169
BAID [48]	Real	PSNR ↑	12.278	13.928	13.841	18.745	16.763
RCD [12]	Raw-Syn	PSNR ↑	15.177	15.184	14.654	20.761	15.508
Gamma [47]	RGB-Syn	PSNR ↑	13.139	11.701	12.778	17.687	12.958
CSNorm-Fourier [76]	RGB-Syn	PSNR ↑	13.526	13.433	14.397	19.204	14.217
UEC-based [12]	RGB-Syn	PSNR ↑	14.564	14.845	15.787	20.580	15.306
MESS-generated	RGB-Syn	PSNR ↑	18.867	17.515	17.937	22.866	16.502
UHDLL [36]	Real	SSIM ↑	0.873	0.672	0.701	0.819	0.857
LSRW [22]	Real	SSIM ↑	0.852	0.702	0.725	0.814	0.846
SICE [7]	Real	SSIM ↑	0.782	0.661	0.805	0.842	0.850
MSEC [1]	Raw-Syn	SSIM ↑	0.789	0.666	0.756	0.873	0.847
BAID [48]	Real	SSIM ↑	0.706	0.611	0.706	0.791	0.885
RCD [12]	Raw-Syn	SSIM ↑	0.784	0.659	0.736	0.859	0.854
Gamma [47]	RGB-Syn	SSIM ↑	0.740	0.539	0.653	0.727	0.807
CSNorm-Fourier [76]	RGB-Syn	SSIM ↑	0.756	0.587	0.710	0.781	0.756
UEC-based [12]	RGB-Syn	SSIM ↑	0.793	0.630	0.744	0.828	0.843
MESS-generated	RGB-Syn	SSIM ↑	0.869	0.689	0.792	0.868	0.876
UHDLL [36]	Real	LPIPS ↓	0.103	0.218	0.235	0.158	0.127
LSRW [22]	Real	LPIPS ↓	0.131	0.190	0.227	0.163	0.147
SICE [7]	Real	LPIPS ↓	0.174	0.221	0.170	0.145	0.145
MSEC [1]	Raw-Syn	LPIPS ↓	0.162	0.220	0.196	0.128	0.141
BAID [48]	Real	LPIPS ↓	0.202	0.267	0.240	0.189	0.110
RCD [12]	Raw-Syn	LPIPS ↓	0.154	0.226	0.218	0.148	0.134
Gamma [47]	RGB-Syn	LPIPS ↓	0.255	0.333	0.290	0.233	0.229
CSNorm-Fourier [76]	RGB-Syn	LPIPS ↓	0.218	0.283	0.251	0.203	0.194
UEC-based [12]	RGB-Syn	LPIPS ↓	0.174	0.250	0.205	0.180	0.154
MESS-generated	RGB-Syn	LPIPS ↓	0.110	0.197	0.177	0.126	0.114

Table 14: No-reference results using CSEC [38] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	ARNIQA ↑	0.682	0.659	0.708	0.692	0.700
LSRW [22]	Real	ARNIQA ↑	0.661	0.682	0.701	0.692	0.695
SICE [7]	Real	ARNIQA ↑	0.669	0.643	0.734	0.709	0.690
MSEC [1]	Raw-Syn	ARNIQA ↑	0.664	0.657	0.725	0.722	0.701
BAID [48]	Real	ARNIQA ↑	0.629	0.623	0.668	0.667	0.734
RCD [12]	Raw-Syn	ARNIQA ↑	0.664	0.655	0.719	0.706	0.705
Gamma [47]	RGB-Syn	ARNIQA ↑	0.627	0.614	0.674	0.674	0.682
CSNorm-Fourier [76]	RGB-Syn	ARNIQA ↑	0.633	0.623	0.689	0.680	0.691
UEC-based [12]	RGB-Syn	ARNIQA ↑	0.654	0.649	0.710	0.701	0.713
MESS-generated	RGB-Syn	ARNIQA ↑	0.699	0.672	0.740	0.729	0.723
UHDLL [36]	Real	MUSIQ ↑	57.028	50.210	52.349	52.464	55.537
LSRW [22]	Real	MUSIQ ↑	55.667	51.738	52.746	53.064	55.811
SICE [7]	Real	MUSIQ ↑	54.582	49.989	56.478	54.142	55.266
MSEC [1]	Raw-Syn	MUSIQ ↑	55.360	50.302	54.223	55.400	55.376
BAID [48]	Real	MUSIQ ↑	51.984	48.770	52.446	51.863	57.750
RCD [12]	Raw-Syn	MUSIQ ↑	55.017	49.972	53.494	53.997	55.884
Gamma [47]	RGB-Syn	MUSIQ ↑	52.359	48.506	53.002	52.338	54.791
CSNorm-Fourier [76]	RGB-Syn	MUSIQ ↑	53.274	48.989	53.752	53.060	55.466
UEC-based [12]	RGB-Syn	MUSIQ ↑	54.173	49.521	54.570	53.769	56.253
MESS-generated	RGB-Syn	MUSIQ ↑	57.260	51.645	57.184	55.554	57.974
UHDLL [36]	Real	MANIQA ↑	0.448	0.384	0.427	0.455	0.417
LSRW [22]	Real	MANIQA ↑	0.447	0.393	0.433	0.446	0.411
SICE [7]	Real	MANIQA ↑	0.432	0.374	0.455	0.460	0.408
MSEC [1]	Raw-Syn	MANIQA ↑	0.433	0.378	0.452	0.479	0.419
BAID [48]	Real	MANIQA ↑	0.403	0.360	0.415	0.434	0.435
RCD [12]	Raw-Syn	MANIQA ↑	0.437	0.373	0.443	0.461	0.421
Gamma [47]	RGB-Syn	MANIQA ↑	0.400	0.360	0.414	0.422	0.394
CSNorm-Fourier [76]	RGB-Syn	MANIQA ↑	0.418	0.366	0.427	0.440	0.401
UEC-based [12]	RGB-Syn	MANIQA ↑	0.431	0.388	0.434	0.450	0.414
MESS-generated	RGB-Syn	MANIQA ↑	0.460	0.403	0.461	0.486	0.443

Table 15: Full-reference results using RAVE [17] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	PSNR ↑	19.276	16.488	12.454	18.718	16.573
LSRW [22]	Real	PSNR ↑	16.983	18.003	13.557	18.981	15.976
SICE [7]	Real	PSNR ↑	14.521	15.214	18.673	20.311	16.182
MSEC [1]	Raw-Syn	PSNR ↑	14.922	15.362	16.650	22.001	16.473
BAID [48]	Real	PSNR ↑	13.249	14.094	14.574	18.772	18.035
RCD [12]	Raw-Syn	PSNR ↑	15.459	15.332	15.398	20.761	16.803
Gamma [47]	RGB-Syn	PSNR ↑	12.441	13.639	13.513	17.681	14.222
CSNorm-Fourier [76]	RGB-Syn	PSNR ↑	13.668	11.857	15.140	19.249	15.470
UEC-based [12]	RGB-Syn	PSNR ↑	14.709	15.006	16.531	20.536	16.586
MESS-generated	RGB-Syn	PSNR ↑	18.998	17.668	18.365	21.869	17.774
UHDLL [36]	Real	SSIM ↑	0.857	0.671	0.709	0.820	0.855
LSRW [22]	Real	SSIM ↑	0.833	0.702	0.735	0.813	0.841
SICE [7]	Real	SSIM ↑	0.769	0.661	0.799	0.828	0.845
MSEC [1]	Raw-Syn	SSIM ↑	0.775	0.670	0.763	0.870	0.845
BAID [48]	Real	SSIM ↑	0.724	0.614	0.712	0.791	0.881
RCD [12]	Raw-Syn	SSIM ↑	0.785	0.651	0.743	0.855	0.855
Gamma [47]	RGB-Syn	SSIM ↑	0.689	0.538	0.661	0.781	0.752
CSNorm-Fourier [76]	RGB-Syn	SSIM ↑	0.743	0.584	0.716	0.729	0.804
UEC-based [12]	RGB-Syn	SSIM ↑	0.775	0.630	0.754	0.829	0.840
MESS-generated	RGB-Syn	SSIM ↑	0.853	0.689	0.812	0.874	0.872
UHDLL [36]	Real	LPIPS ↓	0.132	0.219	0.243	0.157	0.144
LSRW [22]	Real	LPIPS ↓	0.161	0.189	0.235	0.164	0.160
SICE [7]	Real	LPIPS ↓	0.206	0.222	0.178	0.143	0.161
MSEC [1]	Raw-Syn	LPIPS ↓	0.191	0.217	0.206	0.141	0.152
BAID [48]	Real	LPIPS ↓	0.230	0.266	0.248	0.186	0.124
RCD [12]	Raw-Syn	LPIPS ↓	0.181	0.227	0.227	0.144	0.147
Gamma [47]	RGB-Syn	LPIPS ↓	0.283	0.334	0.299	0.232	0.243
CSNorm-Fourier [76]	RGB-Syn	LPIPS ↓	0.247	0.284	0.258	0.205	0.205
UEC-based [12]	RGB-Syn	LPIPS ↓	0.203	0.246	0.214	0.180	0.168
MESS-generated	RGB-Syn	LPIPS ↓	0.139	0.196	0.185	0.138	0.128

Table 16: No-reference results using RAVE [17] backbone. The model trained on each data source is evaluated on all test sets without fine-tuning. Values in red and blue indicate the best and second-best results, respectively.

Training Data	Type	Metric	UHDLL	LSRW	SICE	MSEC	BAID
UHDLL [36]	Real	ARNIQA ↑	0.657	0.650	0.692	0.693	0.679
LSRW [22]	Real	ARNIQA ↑	0.637	0.672	0.687	0.692	0.672
SICE [7]	Real	ARNIQA ↑	0.642	0.633	0.718	0.699	0.667
MSEC [1]	Raw-Syn	ARNIQA ↑	0.638	0.645	0.708	0.697	0.680
BAID [48]	Real	ARNIQA ↑	0.603	0.613	0.649	0.668	0.701
RCD [12]	Raw-Syn	ARNIQA ↑	0.632	0.643	0.691	0.702	0.689
Gamma [47]	RGB-Syn	ARNIQA ↑	0.601	0.604	0.660	0.671	0.658
CSNorm-Fourier [76]	RGB-Syn	ARNIQA ↑	0.608	0.614	0.675	0.680	0.667
UEC-based [12]	RGB-Syn	ARNIQA ↑	0.629	0.641	0.691	0.700	0.691
MESS-generated	RGB-Syn	ARNIQA ↑	0.674	0.662	0.724	0.704	0.712
UHDLL [36]	Real	MUSIQ ↑	57.529	50.730	52.641	52.449	56.127
LSRW [22]	Real	MUSIQ ↑	56.168	52.299	53.067	53.058	56.357
SICE [7]	Real	MUSIQ ↑	55.096	50.518	56.744	54.119	55.867
MSEC [1]	Raw-Syn	MUSIQ ↑	55.830	50.856	54.476	54.413	55.936
BAID [48]	Real	MUSIQ ↑	52.473	49.331	52.705	51.879	58.310
RCD [12]	Raw-Syn	MUSIQ ↑	55.492	50.526	53.759	53.994	56.447
Gamma [47]	RGB-Syn	MUSIQ ↑	52.846	49.091	53.247	52.308	55.360
CSNorm-Fourier [76]	RGB-Syn	MUSIQ ↑	53.735	49.582	54.003	53.045	55.989
UEC-based [12]	RGB-Syn	MUSIQ ↑	54.678	50.033	54.834	53.760	56.816
MESS-generated	RGB-Syn	MUSIQ ↑	57.761	52.206	57.450	54.573	58.534
UHDLL [36]	Real	MANIQA ↑	0.443	0.388	0.422	0.452	0.406
LSRW [22]	Real	MANIQA ↑	0.438	0.398	0.430	0.446	0.400
SICE [7]	Real	MANIQA ↑	0.430	0.378	0.451	0.460	0.399
MSEC [1]	Raw-Syn	MANIQA ↑	0.426	0.381	0.448	0.449	0.408
BAID [48]	Real	MANIQA ↑	0.395	0.365	0.412	0.433	0.424
RCD [12]	Raw-Syn	MANIQA ↑	0.423	0.379	0.436	0.469	0.419
Gamma [47]	RGB-Syn	MANIQA ↑	0.397	0.363	0.409	0.422	0.380
CSNorm-Fourier [76]	RGB-Syn	MANIQA ↑	0.414	0.370	0.420	0.440	0.389
UEC-based [12]	RGB-Syn	MANIQA ↑	0.425	0.392	0.433	0.452	0.402
MESS-generated	RGB-Syn	MANIQA ↑	0.455	0.408	0.457	0.445	0.432

Table 17: Ablation study of the proposed MESS framework. All variants are used to generate training data for the same FourLLIE backbone, and the trained models are evaluated on five test sets without fine-tuning. The reported results are averaged over UHDLL, LSRW, SICE, MSEC, and BAID.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓	ARNIQA ↑	MUSIQ ↑	MANIQA ↑
w/o representation consistency loss	21.354	0.835	0.132	0.698	54.821	0.456
w/o frequency-based enhancement	20.936	0.828	0.140	0.692	54.307	0.449
w/o diffusion-based detail enhancement	21.112	0.832	0.136	0.695	54.563	0.453
w/o exposure-conditioned rendering	20.781	0.821	0.145	0.688	54.018	0.445
FFT instead of DWT	21.447	0.837	0.130	0.700	54.963	0.458
Full MESS	**21.923**	**0.846**	**0.121**	**0.709**	**55.510**	**0.467**

Table 18: Ablation study on the design of the exposure-conditioned rendering module. All variants use the same frequency-based enhancement module and diffusion-based detail enhancement module. Only the way of injecting the target exposure condition is changed.

Renderer Design	PSNR ↑	SSIM ↑	LPIPS ↓	ARNIQA ↑	MUSIQ ↑	MANIQA ↑
No exposure condition	20.781	0.821	0.145	0.688	54.018	0.445
Concatenate EV with RGB features	21.082	0.831	0.137	0.695	54.576	0.452
Global FiLM on RGB features	21.324	0.836	0.132	0.701	54.905	0.459
DWT-FiLM w/o content descriptor	21.541	0.840	0.128	0.704	55.102	0.462
DWT-FiLM with content descriptor	**21.923**	**0.846**	**0.121**	**0.709**	**55.510**	**0.467**

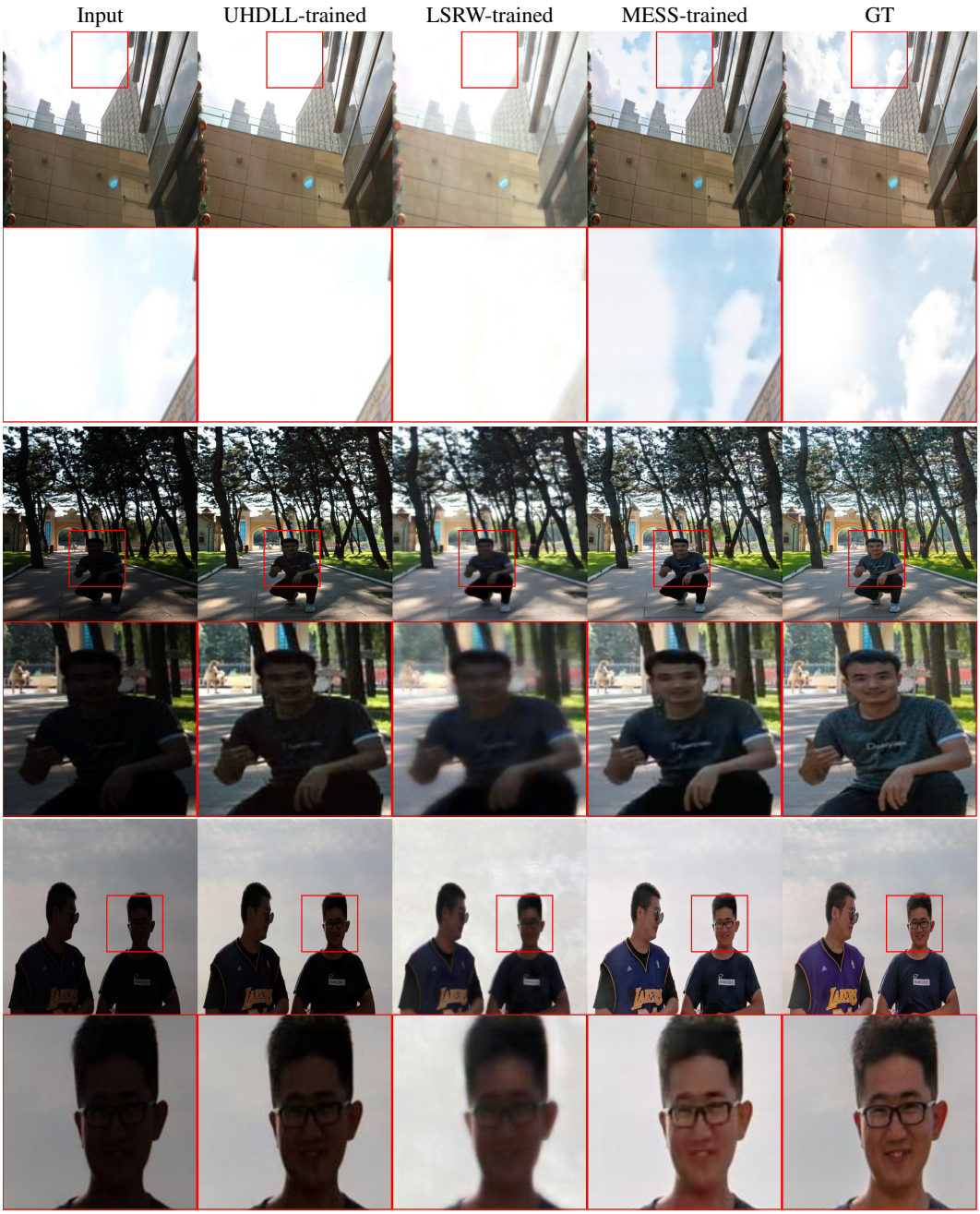

Figure 8: Qualitative comparison of the FourLLIE models trained on UHDLL, LSRW, and our MESS-generated datasets. The test images are from the BAID dataset.

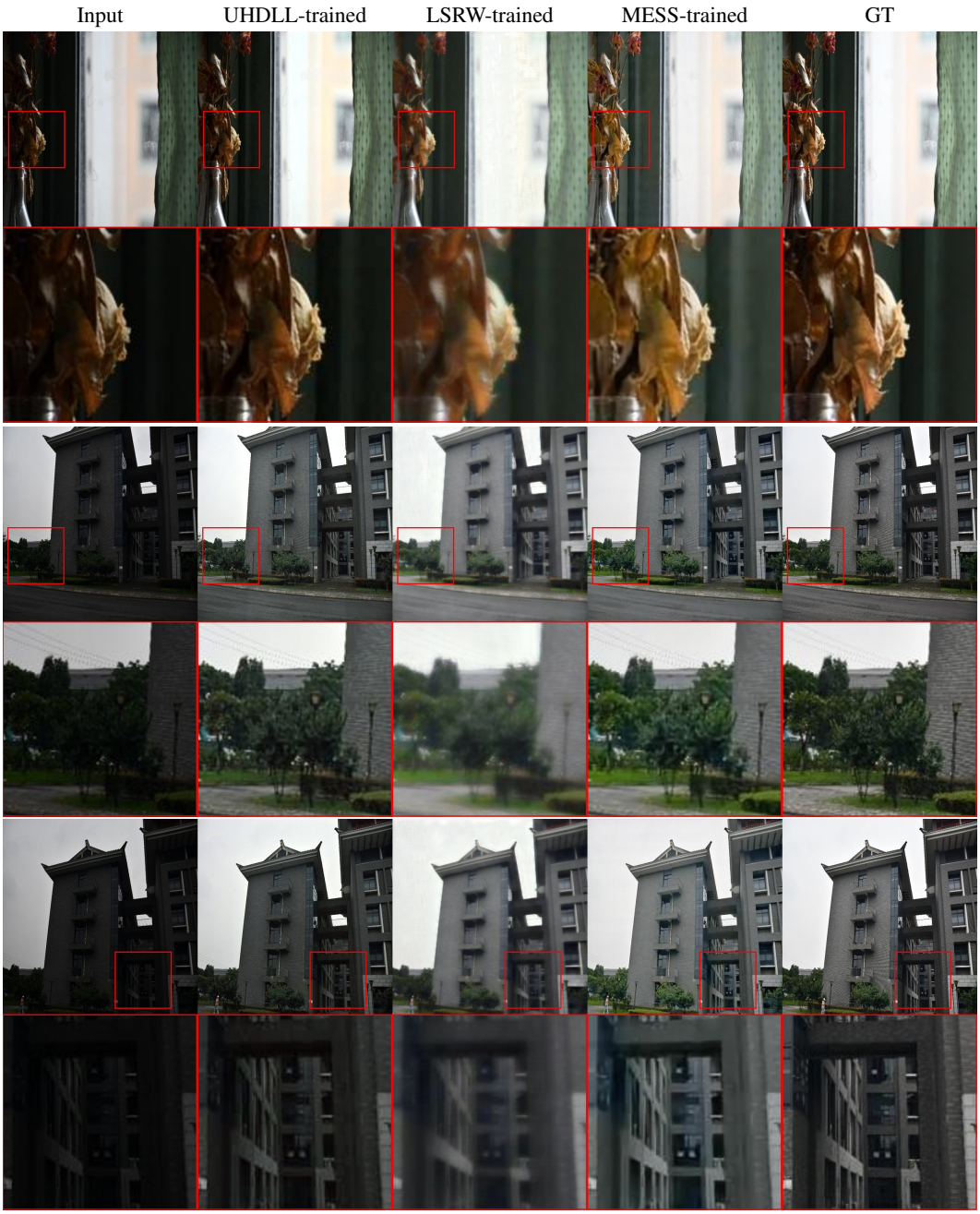

Figure 9: Qualitative comparison of the FourLLIE models trained on UHDLL, LSRW, and our MESS-generated datasets. The test images are from the LSRW dataset.



Figure 10: Qualitative comparison of the FourLLIE models trained on UHDLL, LSRW, and our MESS-generated datasets. The test images are from the SICE dataset.

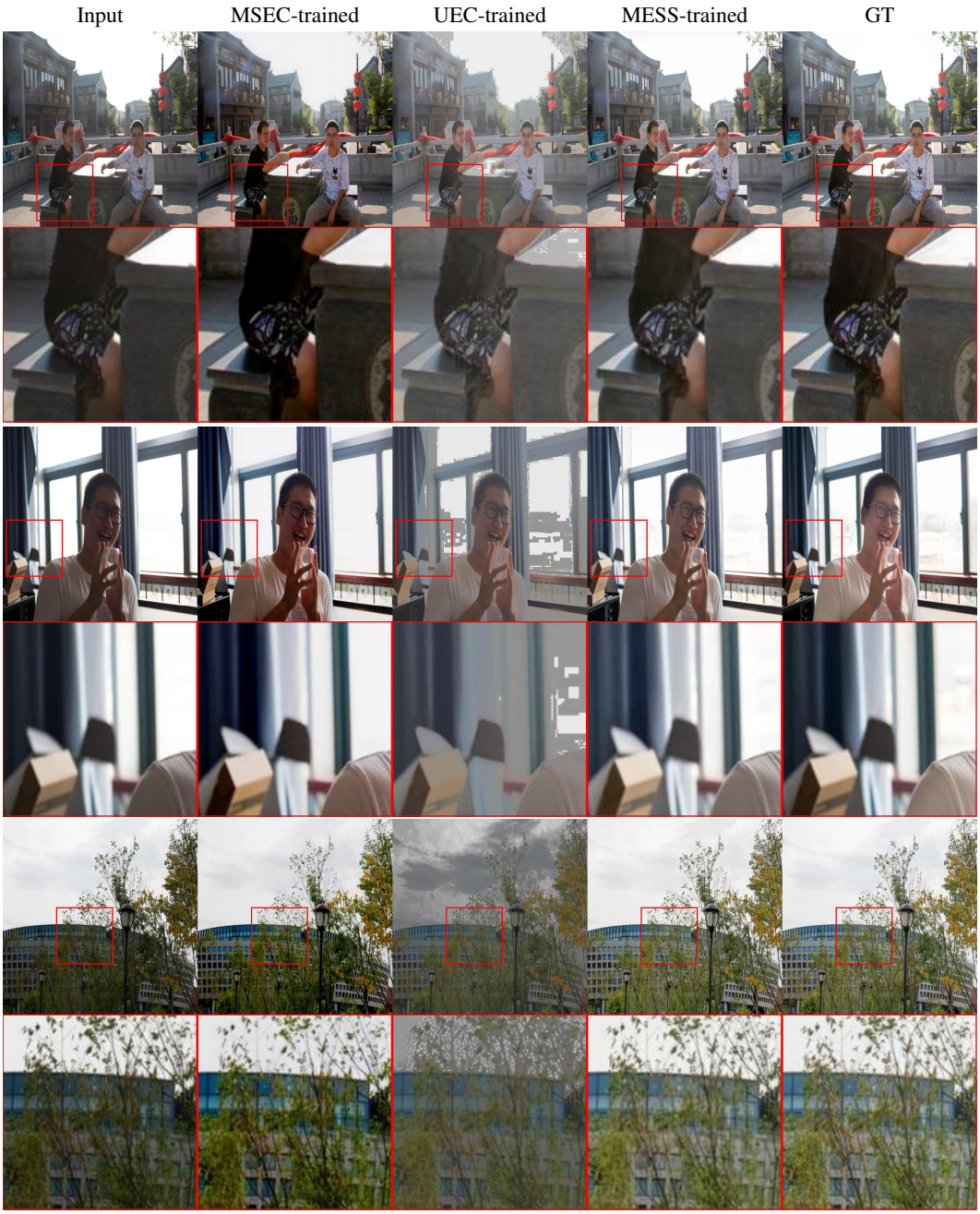

Figure 11: Qualitative comparison of the FourLLIE models trained on MSEC, UEC, and MESS-generated datasets. The test images are from the BAID dataset.



Figure 12: Qualitative comparison of the FourLLIE models trained on MSEC, UEC, and MESS-generated datasets. The test images are from the SICE dataset.

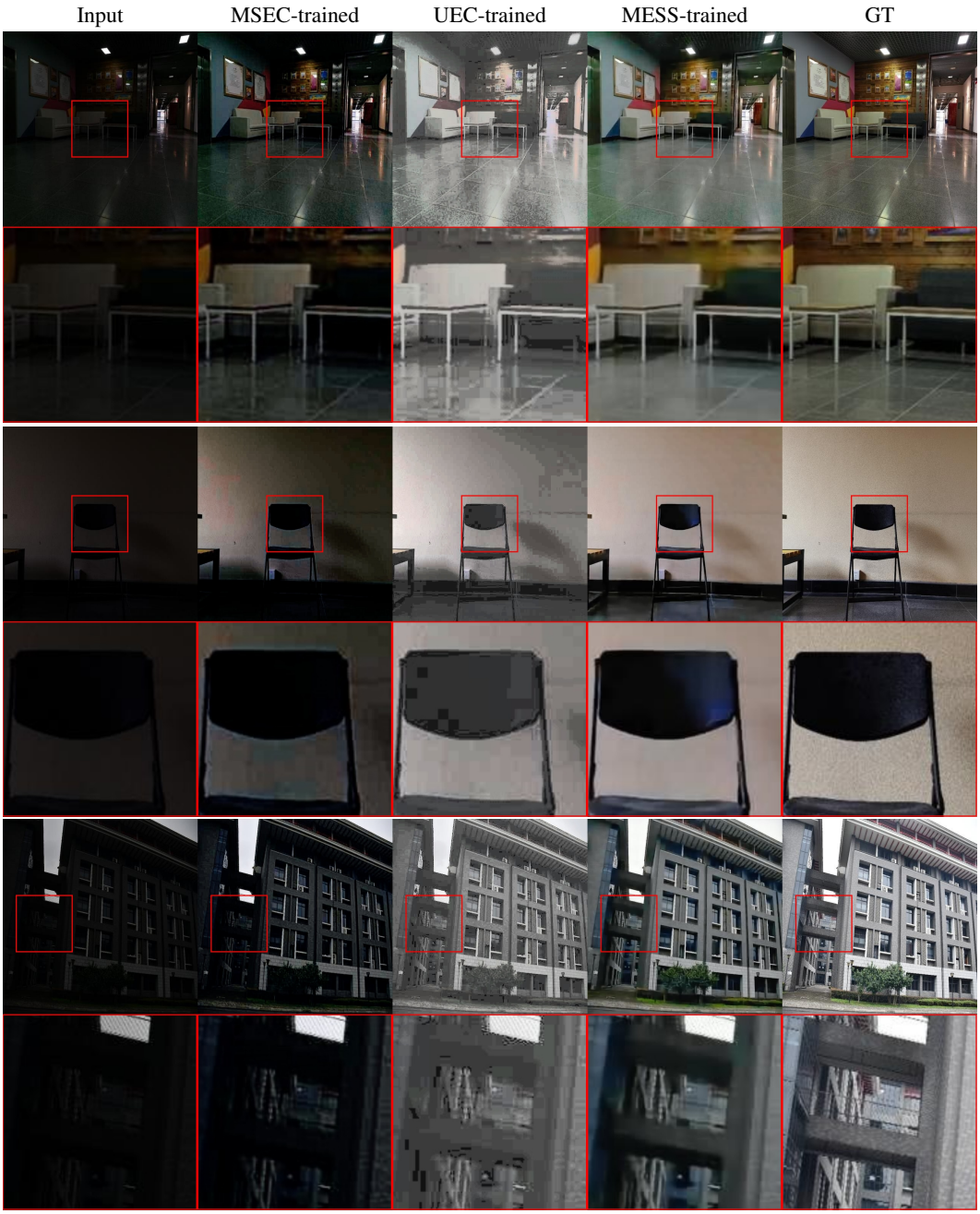

Figure 13: Qualitative comparison of the FourLLIE models trained on MSEC, UEC, and MESS-generated datasets. The test images are from the LSRW dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: Yes.

   Justification: The abstract and introduction state that the paper proposes MESS, a scalable multi-exposure sequence synthesis framework that transfers RAW-derived exposure variations to in-the-wild RGB images, constructs a 500K paired training set, and improves generalization across multiple image illumination enhancement tasks. These claims are supported by the method description, dataset construction procedure, main experiments, additional backbone evaluations, and ablation studies.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: Yes.

   Justification: The paper includes a dedicated limitations paragraph. It states that MESS currently relies on a fixed set of RAW datasets to derive exposure priors, and that the current best-rendered image selection strategy is based on several no-reference IQA metrics. The paper also suggests that expanding the RAW data pool and incorporating stronger vision-language models for data selection may further improve the framework.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: N/A.

   Justification: The paper does not present formal theoretical results, theorems, lemmas, or mathematical proofs. It is primarily a methodological and empirical study on data synthesis for image illumination enhancement.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: Yes.

   Justification: The paper provides the main information needed to reproduce the experimental comparisons in principle, including the RAW-guided training data construction, source RGB image selection pipeline, exposure-degraded input synthesis procedure, downstream paired dataset construction, network architecture details, training settings, evaluation benchmarks, metrics, and ablation variants. Some minor implementation details and exact filtering thresholds are not fully specified, but the information disclosed is sufficient to understand and approximately reproduce the main experimental conclusions.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: No.

   Justification: The paper states that the code, dataset, and models will be released, but the current manuscript does not provide an active repository, download links, release instructions, or complete reproduction scripts. Therefore, open access is promised but not yet provided in the submitted paper.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: Yes.

Justification: The paper specifies the main experimental settings, including crop size, number of iterations, optimizer, AdamW parameters, batch size, learning rate, learning-rate schedule, loss functions, exposure sampling range, GPU type for MESS training, downstream training protocol, test datasets, compared data sources, and evaluation metrics. The appendix further provides network architecture details and dataset construction details. Although not every hyperparameter selection rationale is discussed, the reported settings are sufficient to understand how the experiments were conducted.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: No.

   Justification: The quantitative results are reported as point estimates over benchmark datasets. The paper does not report error bars, confidence intervals, standard deviations over multiple random seeds, or formal statistical significance tests.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: No.

   Justification: The paper reports that MESS is trained using 8 NVIDIA A100 GPUs, but it does not provide sufficient compute-resource information for all experiments, such as GPU memory, training time, inference time, total computational cost, or detailed resource requirements for each downstream model and ablation.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: Yes.

   Justification: The work focuses on image illumination enhancement and data synthesis using public datasets and generated exposure variations. The paper does not describe human-subject experiments, sensitive personal data collection, or deployment in high-risk decision-making settings. Based on the information provided, the research appears to conform to standard ethical research practice.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: No.

    Justification: The paper does not include a dedicated broader-impact discussion. While the proposed method may benefit robust image enhancement in real-world visual applications, the manuscript does not explicitly discuss potential positive societal impacts, possible negative consequences, misuse risks, or limitations in downstream deployment.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: N/A.

    Justification: The paper releases, or plans to release, an image enhancement data synthesis framework, paired training data, and enhancement models rather than a high-risk generative model such as a text-to-image generator, large language model, or scraped dataset designed for identity-sensitive applications. The work does not appear to require specialized misuse safeguards beyond standard dataset and model release documentation.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: No.

Justification: The paper cites the existing datasets, models, and methods used in the study, including RAW datasets, RGB image sources, IQA models, downstream backbones, and SD-Turbo. However, the manuscript does not explicitly list the license names, terms of use, or version-specific usage conditions for all existing assets. Therefore, the crediting is partially addressed, but license documentation is incomplete.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: Yes.

Justification: The paper introduces a new MESS-generated paired RGB training dataset containing 500K samples. The manuscript and appendix describe how the data are generated, including source image selection, quality filtering, exposure-condition sampling, validity checks, and final paired-data construction. The paper also states that the dataset, code, and models will be released. Although release links are not yet provided, the new asset itself is documented in the paper.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: N/A.

Justification: The paper does not involve crowdsourcing, user studies, human annotation tasks, participant instructions, or compensation.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: N/A.

Justification: The paper does not conduct human-subject experiments or collect data directly from study participants. Therefore, IRB approval or an equivalent human-subjects review is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: N/A.

Justification: The paper does not use LLMs as an important, original, or non-standard component of the core research methodology. Although the paper mentions vision-language models as a possible future improvement for data selection, they are not part of the proposed method or experiments.