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.

NeurIPS 2026

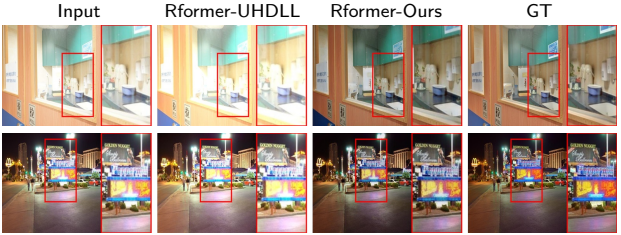
Code & data: `https://github.com/PolyU-VCLab/MESS`

Outline

1. Motivation: why do enhancement models fail to generalize?
2. Existing data construction strategies
3. MESS: Multi-Exposure Sequence Synthesis
   - RAW-guided training data
   - Frequency-based enhancement · Diffusion-based detail enhancement · Exposure-conditioned rendering
4. Building a 500K paired dataset
5. Experiments: cross-dataset results, scale, backbones, ablations
6. Conclusion & limitations

Motivation: Enhancement Models Generalize Poorly

- Adverse illumination tasks are studied in isolation:
  - Low-light enhancement (LLIE)
  - Exposure correction (EC)
  - Backlit enhancement (BIE)
- Trained on narrow data, models break on other tasks and on mixed under/over-exposure.
- Larger models or more complex designs alone do not solve it.



Retinexformer trained on UHD-LL vs. on MESS-generated data (tested on SICE).

The Real Bottleneck: Training Data

Task	Dataset	#images	#scenes
LLIE	LOLv2Real	789	263
LLIE	LSRW	5,650	4,157
LLIE	UHD-LL	2,150	724
EC	MSEC	24,330	5,000
EC	LCDP	1,700	1,700
EC	SICE	4,800	589
BIE	BAID	3,000	1,782
BIE	Backlit300	305	305
All	**MESS (ours)**	**500,000**	**500,000**

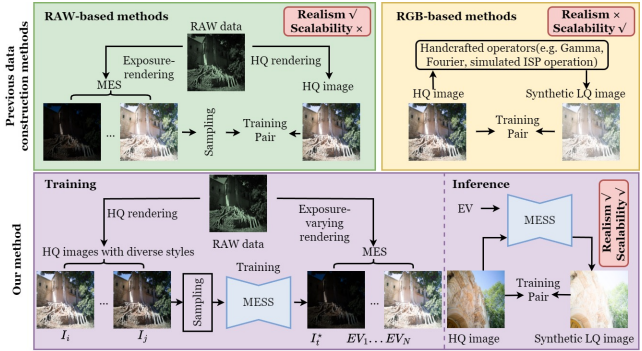
Scenes counted by CLIP similarity > 0.9 when not provided.

- Real pairs need manual capture & precise alignment $\Rightarrow$ expensive, hard to scale.
- Existing sets: a few hundred to a few thousand scenes.
- Limited coverage of scenes and degradation types.

Goal

Synthesize **realistic** *and* **scalable** paired data for illumination enhancement in the wild.

Existing Data Construction vs. MESS



RAW-based (MSEC, RCD, UNICE)
ISP exposure rendering of RAW data
realistic / not scalable

RGB-based (Gamma, Fourier, InvISP)
Handcrafted ops on clipped 8-bit RGB
scalable / less realistic

MESS (ours)
Learn RGB → MES from RAW supervision, then apply to large RGB collections
realistic & scalable

Key Idea and Contributions

Key idea

Use **RAW images** to supervise realistic exposure variation, and train an **RGB-to-MES generator** that transfers it to **in-the-wild RGB images** at scale.

1. **MESS framework**: scalable, realistic training-data synthesis for generalizable illumination enhancement.
2. **MES synthesis pipeline**:
   - frequency-based enhancement → style-robust scene representation
   - diffusion-based synthesis → completes details lost in 8-bit RGB
   - exposure-conditioned rendering → realistic, EV-controllable output
3. **500K paired dataset**: a *lightweight* network trained on it generalizes across LLIE, EC and BIE benchmarks.

RAW-guided Training Data

From each high-bit RAW image X, build a triplet $\{I_i, I_j, I_t^\star\}$:

- $I_i = \mathcal{T}_{\text{sty}}(X; s_i)$, $I_j = \mathcal{T}_{\text{sty}}(X; s_j)$: two style-varied renderings (MEF + tone mapping of a 7-EV stack, $-3 \cdots +3$ EV) $\rightarrow$ learn a style-robust representation
- $I_t^\star = \mathcal{T}_{\text{cam}}(X; e_t)$: camera-consistent rendering at target exposure $e_t \in [0, 1]$ $\rightarrow$ supervise exposure-controlled synthesis
- **39,642** RAW scenes from AODRAW, FiveK, HDRP, PASCAL-RAW, PPR10K, RAISE, RAW-NOD

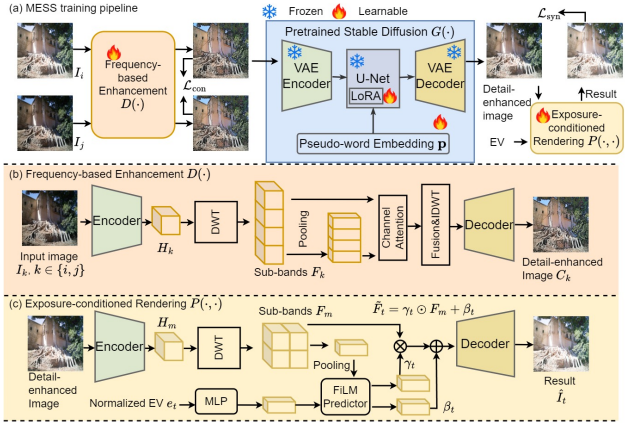
Style-varied rendering



Renderings with different EVs



MESS Framework Overview



Given source I_i and normalized target EV e_t:

$$\hat{I}_t = P\Big(G(D(I_i)),\, e_t\Big)$$

- D: frequency-based enhancement, $C_i = D(I_i)$
- G: diffusion detail enhancement, $I_m = G(C_i)$
- P: exposure-conditioned rendering, $\hat{I}_t = P(I_m, e_t)$

Varying e_t yields a full multi-exposure sequence (MES).

(1) Frequency-based Enhancement $D(\cdot)$

Why frequency? ISP-induced appearance changes (exposure, contrast) live mostly in low frequencies; structure & texture live in high frequencies.

Why DWT, not FFT? FFT lacks spatial localization; Haar DWT is simple, localized, with low coefficient correlation.

Pipeline

1. Encoder: conv + 4 ResBlocks $\rightarrow H_k \in \mathbb{R}^{64\times H\times W}$
2. Haar DWT $\rightarrow F_k \in \mathbb{R}^{256\times \frac{H}{2}\times \frac{W}{2}}$ (LL, LH, HL, HH)
3. Channel attention (avg + max pooling, shared MLP): $\tilde{F}_k = A_k \odot F_k$
4. IDWT + decoder $\rightarrow C_k$

Representation consistency

I_i, I_j come from the *same* RAW but differ in style:

$$\mathcal{L}_{\text{con}} = \left\| D(I_i) - D(I_j) \right\|_1$$

$\Rightarrow$ style-invariant, content-consistent representation.

(2) Diffusion-based Detail Enhancement $G(\cdot)$

Problem: 8-bit RGB has clipped / compressed dynamic range $\Rightarrow$ missing content must be synthesized, not just re-mapped.

Solution: one-step diffusion prior (SD-Turbo)

$$G(C_i) = \mathrm{Dec}_{\mathrm{VAE}}\Big(\mathcal{G}_\theta\big(\mathrm{Enc}_{\mathrm{VAE}}(C_i),\, \mathbf{p}\big)\Big)$$

- Frozen VAE encoder / decoder
- LoRA-adapted U-Net (q, k, v, out projections; $r = \alpha = 8$)
- Learnable shared pseudo-word embedding $\mathbf{p}$ (single token) replaces text prompts

Trainable parameters

Only LoRA layers + $\mathbf{p}$

LoRA projection

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

W frozen; A, B low-rank, trainable.

(3) Exposure-conditioned Rendering $P(\cdot)$

Adjust brightness, contrast and global tone while keeping recovered structure.

1. Encoder + Haar DWT: $I_m \to F_m \in \mathbb{R}^{256\times \frac{H}{2}\times \frac{W}{2}}$
2. Content descriptors: avg + max pooling of F_m (2×256-d)
3. Exposure embedding: MLP$(e_t) \to 128$-d
4. FiLM predictor on the 640-d concat $\to \gamma_t, \beta_t \in \mathbb{R}^{256}$
5. Modulate, IDWT, decode $\to \hat{I}_t$

Wavelet-domain FiLM

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

Residual form $1 + \gamma_t$ stabilizes modulation.

DWT separates illumination/contrast responses into sub-bands; affine modulation gives continuous control over e_t.

Training Objective & Implementation

Loss

$$\mathcal{L}_{\text{con}} = \|C_i - C_j\|_1$$
$$\mathcal{L}_{\text{syn}} = \|\hat{I}_t - I_t^\star\|_2^2$$
$$\mathcal{L}_{\text{total}} = \lambda\,\mathcal{L}_{\text{con}} + \mathcal{L}_{\text{syn}}, \quad \lambda = 1$$

MESS training

- 512×512 crops, 200K iterations
- AdamW, batch 64, 8×A100
- LR $10^{-4} \to 10^{-6}$ (cosine)
- Target EV $\sim \mathcal{U}[-3, 3]$

Downstream models

- 512×512, AdamW, LR 2×10^{-4}
- Batch 64, cosine, ℓ_1 loss, 300K iterations
- Identical protocol across data sources

From MESS to a 500K Paired Dataset

Source selection ($\sim$20M $\rightarrow$ 500K)

- Unsplash, SA-1B, Open Images
- Filter blur, artifacts, noise, bad exposure, near-duplicates
- Rank by NR-IQA ensemble (ARNIQA, MUSIQ, MANIQA, TOPIQ)

Degraded input synthesis

- EV in $[-3, +3]$, avoid near-0; under/over equally likely, biased to large shifts
- Reject near-identical, clipped or artifacted outputs

Pair = (MESS output, original HQ image)



MES synthesized by MESS from a single RGB image.

Experimental Setup

Downstream backbones (only the data changes)

- LLIE: FourLLIE, RetinexFormer, Multinex
- EC: ExSLOT, CSEC
- BIE: RAVE

Test sets (no fine-tuning)
UHDLL, LSRW, SICE, MSEC, BAID

Metrics

- FR: PSNR, SSIM, LPIPS
- NR: ARNIQA, MUSIQ, MANIQA

Training data compared

- **Real / in-domain**: UHDLL, LSRW, SICE, BAID
- **RAW-synthesized**: MSEC, RCD
- **RGB-synthesized**: Gamma, CSNorm-Fourier, UEC-based
- **Ours**: MESS-generated (500K)

Main Results (FourLLIE): Averages over 5 Test Sets

Training data	Type	PSNR↑	SSIM↑	LPIPS↓	ARNIQA↑	MUSIQ↑	MANIQA↑
UHDLL	Real	19.678	0.813	0.142	0.684	53.170	0.442
LSRW	Real	19.663	0.816	0.149	0.682	53.508	0.444
SICE	Real	19.965	0.816	0.149	0.684	53.752	0.444
MSEC	RAW-Syn	20.079	0.816	0.147	0.692	53.920	0.450
BAID	Real	18.662	0.774	0.182	0.657	52.223	0.427
RCD	RAW-Syn	19.101	0.798	0.154	0.678	53.155	0.439
Gamma	RGB-Syn	16.920	0.703	0.244	0.650	51.856	0.414
CSNorm-Fourier	RGB-Syn	18.399	0.755	0.206	0.659	52.551	0.428
UEC-based	RGB-Syn	19.645	0.795	0.169	0.681	53.317	0.441
MESS-generated	RGB-Syn	**21.923**	**0.846**	**0.121**	**0.709**	**55.510**	**0.467**

Red: best; blue: second best. MESS improves average PSNR by **+1.84 dB** over the strongest baseline (MSEC).

Cross-dataset PSNR (FourLLIE)

Training data	Type	UHDLL	LSRW	SICE	MSEC	BAID	Avg.
UHDLL	Real	**23.694**	19.357	16.152	18.728	20.461	19.678
LSRW	Real	21.383	**20.849**	17.246	18.964	19.875	19.663
SICE	Real	18.947	18.061	**22.385**	20.329	20.103	19.965
MSEC	RAW-Syn	19.348	18.214	20.367	**22.116**	20.352	20.079
BAID	Real	17.684	16.935	18.276	18.742	21.674	18.662
RCD	RAW-Syn	19.702	18.242	17.478	19.924	20.159	19.101
UEC-based	RGB-Syn	19.137	17.841	20.228	20.553	20.467	19.645
MESS-generated	RGB-Syn	23.416	20.514	21.769	21.982	**21.935**	**21.923**

Gamma and CSNorm-Fourier rows omitted for space (both below MESS on every test set).

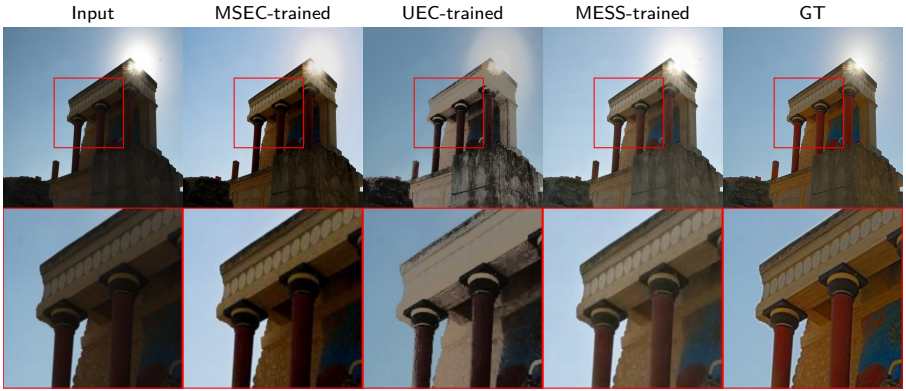
- Task-specific data wins only in-domain and drops sharply elsewhere (e.g. UHDLL → SICE: 16.15 dB).
- MESS is best or second best on every test set without any in-domain data.

Qualitative Comparison: Low-light Enhancement



- FourLLIE trained on different data, tested on LSRW (zoom-ins in the bottom row).
- UHDLL-trained amplifies JPEG artifacts; LSRW-trained (in-domain!) still blurs details.
- MESS-trained avoids artifacts and keeps sharper structures.

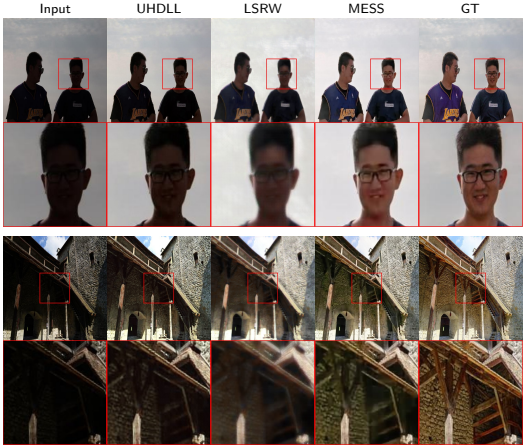
Qualitative Comparison: Exposure Correction



- FourLLIE trained on different data, tested on SICE.
- RAW-based MSEC data: ineffective correction; RGB-based UEC data: unnatural colors.
- MESS-trained: natural exposure, better color fidelity and texture preservation.

More Qualitative Results

BIE / EC – vs. real-data training



LLIE / EC – vs. synthetic training data



MESS-trained models give more stable exposure, more natural colors and better-preserved structures.

Is It Quality or Scale? Both.

Data scale (FourLLIE, 5-set average)

Training data	PSNR↑	SSIM↑	LPIPS↓
LSRW	19.663	0.816	0.149
SICE	19.965	0.816	0.149
MESS-5K	20.220	0.822	0.143
MESS-10K	20.590	0.830	0.137
MESS-50K	21.199	0.839	0.129
MESS-500K	**21.923**	**0.846**	**0.121**

At a similar size (5K), MESS already beats real data.

RAW-only vs. RGB expansion

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

- Matched size: better HQ RGB targets than RAW MEF targets.
- Full set: extra gain from scale.

Architecture-agnostic Gains

Average PSNR over UHDLL, LSRW, SICE, MSEC, BAID (dB)

Task	Backbone	Best baseline (MSEC)	**MESS-generated**	Gain
LLIE	FourLLIE	20.079	**21.923**	+1.84
LLIE	RetinexFormer	19.736	**21.626**	+1.89
LLIE	Multinex	18.331	**20.172**	+1.84
EC	ExSLOT	17.251	**19.212**	+1.96
EC	CSEC	16.766	**18.737**	+1.97
BIE	RAVE	17.082	**18.935**	+1.85

Averages computed from the per-test-set results in the paper (main text and Appendix C).

- Same architecture and protocol; only the training data changes.
- Consistent $\sim$1.8–2.0 dB gain across LLIE, EC and BIE backbones; NR metrics show a similar trend.

Ablation Study (FourLLIE, 5-set average)

Components

Variant	PSNR↑	SSIM↑	LPIPS↓
w/o $\mathcal{L}_{\text{con}}$	21.354	0.835	0.132
w/o frequency enhancement	20.936	0.828	0.140
w/o diffusion detail enh.	21.112	0.832	0.136
w/o exposure-cond. rendering	20.781	0.821	0.145
FFT instead of DWT	21.447	0.837	0.130
Full MESS	**21.923**	**0.846**	**0.121**

Exposure injection in $P(\cdot)$

Renderer design	PSNR↑	LPIPS↓
No exposure condition	20.781	0.145
Concat EV + RGB features	21.082	0.137
Global FiLM on RGB features	21.324	0.132
DWT-FiLM w/o content desc.	21.541	0.128
DWT-FiLM + content desc.	**21.923**	**0.121**

- Every component helps; exposure-conditioned rendering matters most.
- DWT > FFT: spatially localized frequency modeling is beneficial.
- Frequency-domain, content-adaptive modulation gives the most accurate exposure synthesis.

Conclusion & Limitations

Conclusion

- **MESS** transfers RAW-derived exposure variation to in-the-wild RGB images, combining RAW realism with RGB scalability.
- Frequency-based enhancement + diffusion detail compensation + exposure-conditioned rendering.
- A 500K-pair dataset lets a *lightweight* model generalize across LLIE, EC and BIE benchmarks, consistently across six backbones.

Limitations & future work

- Exposure priors come from a fixed set of RAW datasets $\rightarrow$ expand the RAW pool.
- Image selection relies on NR-IQA metrics $\rightarrow$ use VLMs for more robust data selection.

Thank you!

Questions?

Paper & code: https://github.com/PolyU-VCLab/MESS

Ruodai Cui ruodai.cui@connect.polyu.hk