SPACE: Structure-Preserving Cross-Modal Image Enhancement for Extreme Low-Light Conditions

Yue Zhang[1] , **Zhiliang Wu**[2] , **Yuxuan Hou**[1] and **Hehe Fan**[1*]
[1]Zhejiang University
[2]Nanyang Technological University

Abstract

Low-Light Image Enhancement (LLIE) remains challenging due to severe noise, color artifacts, and structural degradation in extreme low-light environments. Existing methods primarily rely on photometric cues and often fail to preserve geometric consistency, resulting in blurred edges and distorted textures. To address this limitation, we introduce **LAMP**, a large-scale dataset of **L**ow-light **A**ligned **M**ultimodal **P**airs for LLIE. LAMP contains over 18k high-quality aligned pairs and consists of two complementary subsets. `LAMP-Real` is captured using LiDAR-based depth sensors under physically controlled illumination, while `LAMP-Synthetic` is generated through a physics-calibrated degradation pipeline. Both subsets provide precisely aligned RGB images, depth maps, and pixel-wise semantic annotations. Building upon LAMP, we propose **SPACE**, a **S**tructure **P**reservation **A**ware **C**ross-modal **E**nhancement framework that explicitly leverages geometric priors. SPACE introduces a Depth-Adaptive HVI Transformation to decouple luminance and chrominance under depth guidance, effectively suppressing color-space noise. Furthermore, a Depth-Manifold Modulated Attention mechanism constrains feature interactions within a learned depth manifold, ensuring structural coherence during enhancement. Extensive experiments demonstrate that SPACE consistently outperforms state-of-the-art methods in visual quality and structural fidelity. The code and data are available at https://github.com/YueCheong/SPACE.

1 Introduction

Low-Light Image Enhancement (LLIE) is a fundamental task for the reliable operation of indoor robotic systems and intelligent agents. In complex indoor environments, these systems rely heavily on visual sensors for navigation, object interaction, and scene understanding [Zhang *et al.*, 2025b]. Restoring visibility and structural details from under-exposed images is therefore essential to maintain the robustness of these agents under challenging illumination conditions.

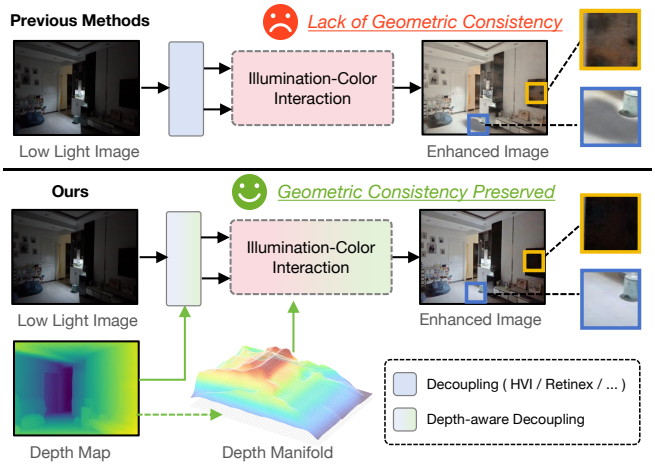

Figure 1: Comparison between previous photometric-only LLIE methods (top) and our geometry-aware approach (bottom), which explicitly leverages depth priors to preserve geometric consistency.

Existing LLIE methods such as heuristic physical modeling [Stark, 2000], discriminative feature mapping [Lin *et al.*, 2025a], and distributional generative reconstruction [Liu *et al.*, 2026] have made significant progress in iluminance restoration. However, despite these advances, current frameworks face a fundamental challenge in preserving geometric consistency. As illustrated in Fig. 1, traditional photometric-only frameworks frequently over-smooth dark structures and introduce texture artifacts, leading to blurred edges and distorted scene geometry. This limitation is further exacerbated by the scarcity of large-scale multimodal datasets providing aligned depth information in extreme low-light scenarios.

To address these issues, we first introduce **LAMP**, a large-scale dataset of **L**ow-light **A**ligned **M**ultimodal **P**airs. Unlike previous RGB-only benchmarks, LAMP provides over 18k pairs of precisely aligned low/normal-light RGB-D images and pixel-wise semantic annotations. This dataset is composed of two complementary subsets: `LAMP-Real`, captured via LiDAR sensors to ensure physical authenticity, and `LAMP-Synthetic`, generated through a physics-calibrated pipeline to ensure scalability. Building upon this benchmark,

*Corresponding author.

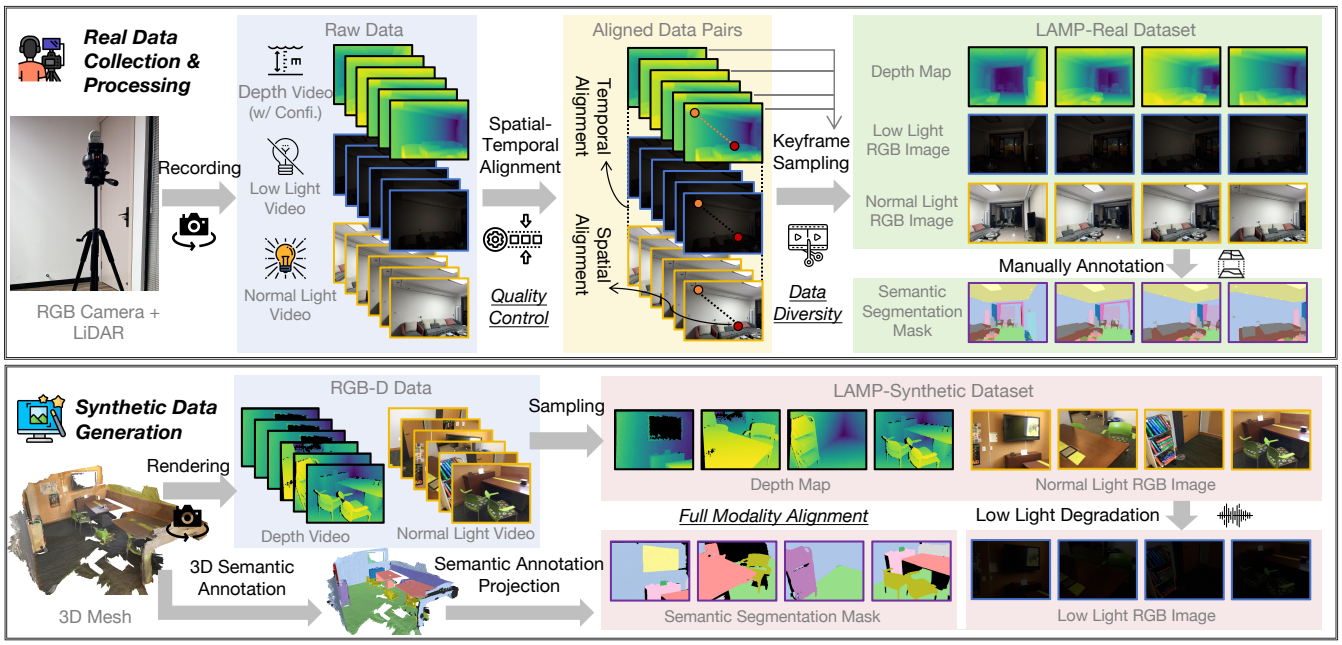

Figure 2: Overview of the LAMP dataset construction pipeline. (Top) LAMP-Real: Real-world low-light RGB-D data are captured using LiDAR-equipped devices under physically controlled illumination, followed by spatial–temporal alignment, keyframe sampling, and manual semantic annotation. (Bottom) LAMP-Synthetic: Large-scale synthetic RGB-D data are generated from 3D scenes via rendering, semantic projection, and physics-calibrated low-light degradation, ensuring full modality alignment and scalable data generation.

we propose **SPACE**, a **S**tructure-**P**reservation-**A**ware **C**ross-modal **E**nhancement framework that explicitly leverages geometric priors through two core designs. First, we introduce a Depth-Adaptive HVI Transformation (DAHT) to decouple luminance and chrominance under depth guidance, ensuring stable color decomposition in severely low-light regions. Second, we propose a Depth-Manifold Modulated Attention (DMMA) mechanism that constrains feature interactions within a learned depth manifold, thereby preserving structural coherence and sharp object boundaries during reconstruction. The contributions of this paper are threefold:

- **Dataset.** We introduce LAMP, the first large-scale multimodal low-light benchmark that provides high-fidelity RGB-D data and semantic annotations, facilitating geometry-aware low-light research.
- **Framework.** We propose the SPACE framework, which incorporates DAHT and DMMA to explicitly enforce geometric constraints, significantly improving the structural fidelity of enhanced images.
- **Experiment.** Extensive experiments demonstrate that SPACE achieves competitive performance on multiple benchmarks and exhibits a superior balance between model complexity and restoration quality.

2 Related Work

2.1 Low-Light Image Enhancement

Low-Light Image Enhancement (LLIE) methods can be categorized into Traditional Cognition (TC), Feed-Forward Enhancement (FFE), and Generative Reconstruction (GR) paradigms. TC methods [Wei *et al.*, 2018] rely on histogram equalization or Retinex-based decomposition; while interpretable and training-free, they often amplify noise. FFE methods [Bai *et al.*, 2024; Chen *et al.*, 2025] utilize deep networks to learn direct mappings or illumination-reflectance corrections, improving global coherence but frequently oversmoothing textures or introducing artifacts [Guo and Hu, 2023]. GR methods [Lin *et al.*, 2025b] leverage GANs or diffusion models [Wang *et al.*, 2025; Shang *et al.*, 2024] to produce high-quality visual synthesis, yet often suffer from heavy computational costs and reduced fidelity to the input. Unlike these approaches that primarily optimize photometric appearance, we introduce depth-aware guidance [Zhang *et al.*, 2023] to ensure geometric consistency in complex scenes.

2.2 Color Space Modeling for LLIE

Color-space modeling aims to decouple exposure from color to minimize chromatic distortion. Early methods rely on hand-crafted color spaces (e.g., HSV [Hu *et al.*, 2023], YUV [Wang *et al.*, 2007], and YCbCr [Lee *et al.*, 2012]) or Retinex-style decomposition [Fan *et al.*, 2025], but frequently encounter color casts or unstable illumination estimation in extreme darkness. While recent HVI-based models [Yan *et al.*, 2025] improve intensity continuity and stability, they remain limited by a purely photometric perspective. Drawing inspiration from physics-based models [Wei *et al.*, 2021; Xu *et al.*, 2023; Zhou *et al.*, 2025], our work incorporates structural cues to modulate the decoupled transformation, achieving more stable and faithful enhancement in extremely low-light conditions.

3 The LAMP Datasets

3.1 Dataset Overview

To investigate the role of geometric cues in low-light perception, we introduce **LAMP** (**L**ow-light **A**ligned **M**ultimodal **P**airs), a new benchmark comprising two complementary subsets, `LAMP-Real` and `LAMP-Synthetic` (Fig. 2), balancing physical realism and scalable coverage. `LAMP-Real` contains 1,530 paired low- and normal-light images from 165 real indoor scenes, while `LAMP-Synthetic` provides 16,968 aligned pairs rendered from 707 indoor scenes for large-scale controllable data generation. Each sample includes strictly aligned multimodal annotations, consisting of paired RGB images under different illumination, depth maps with optional confidence, and pixel-wise semantic labels over 30 categories. LAMP supports systematic evaluation across multiple low-light perception tasks, including enhancement, semantic segmentation, and depth estimation.

3.2 LAMP-Real: Real-world Data Capture

`LAMP-Real` is captured using an iPhone 15 Pro Max equipped with a LiDAR sensor, mounted on a stabilized gimbal to ensure spatial consistency. For each scene, we record synchronized normal-light RGB, low-light RGB, and LiDAR-based depth streams, where low-light conditions are created by physically reducing ambient illumination rather than altering camera exposure. As shown in Fig. 2 (top), spatial-temporal alignment is applied to synchronize multimodal streams. Keyframes are extracted every 15° from 110° rotation sequences, yielding 1,530 aligned pairs, split into 1,380 training and 150 validation samples. All keyframes are manually annotated with pixel-level semantic labels covering 30 categories.

3.3 LAMP-Synthetic: Data Simulation

To enable large-scale aligned data, we construct `LAMP-Synthetic` using 707 3D indoor scenes from ScanNet [Dai *et al.*, 2017]. As shown in Fig. 2 (bottom), high-fidelity rendering generates normal-light RGB images and depth maps, while semantic labels are obtained via projection of 3D annotations to ensure pixel-accurate alignment. To reduce the domain gap, we apply a physics-calibrated degradation pipeline by matching Y-channel luminance statistics of real low-light images and injecting sensor-specific noise derived from `LAMP-Real`. The resulting 16,968 pairs are split into 13,560 training and 3,408 validation samples, forming an independent evaluation track complementary to real-world data.

3.4 Characteristics of LAMP

LAMP redefines multimodal low-light benchmarks by addressing the data scarcity and modality gaps in existing datasets (Table 1). Its core advantages are as follows:

- **Complete Multimodal Alignment:** While prior LLIE datasets are predominantly limited to RGB pairs, LAMP provides strictly aligned RGB, high-fidelity depth, and 30-class semantic masks. This integration allows models to resolve visual ambiguities in extreme darkness by leveraging geometric and structural priors.

Dataset	Real/Syn.	Amb.	Depth	Sem.	Video	Scale
LLNet [Lore *et al.*, 2017]	Syn.	–	✗	✗	✗	169
LOL-V1 [Wei *et al.*, 2018]	Real	✗	✗	✗	✗	1,000
SID [Chen *et al.*, 2018]	Real	✗	✗	✗	✗	5,518
SICE [Cai *et al.*, 2018]	Real+Syn.	✗	✗	✗	✓	4,413
DRV [Chen *et al.*, 2019]	Real	✗	✗	✗	✓	–
LOL-V2 [Yang *et al.*, 2021]	Real+Syn.	✗	✗	✗	✗	3578
SDSD [Wang *et al.*, 2021]	Real	✗	✗	✗	✓	–
TM-DIED [Vonikakis, 2021]	Real	✓	✗	✗	✗	222
LSRW [Hai *et al.*, 2023]	Real	✗	✗	✗	✗	11,300
LED [Wang *et al.*, 2024b]	Real	✗	✓	✗	✗	2730
LOM [Cui *et al.*, 2024]	Real	✗	✗	✗	✓	650
LAMP (Ours)	Real+Syn.	✓	✓	✓	✓	36,996

Table 1: Comparison of LAMP with existing low-light datasets. Real/Syn. indicates whether low-light data are captured under real illumination or synthetically generated. Amb. denotes whether low-light conditions arise from physical ambient illumination changes (✓) or are simulated by camera parameter adjustment (✗). Depth and Sem. indicate available depth and semantic annotations. Scale reports the total number of low and normal-light images.

- **Physics-Grounded Authenticity:** By physically manipulating ambient light rather than camera exposure, `LAMP-Real` preserves genuine photon-starved statistics. This physical fidelity is extended to the synthetic data through sensor-specific noise modeling, ensuring high statistical consistency and effectively bridging the sim-to-real gap.

4 The SPACE Method

As illustrated in Fig. 3, the SPACE framework restores low-light images by integrating geometric priors from depth maps across three stages: (1) Depth-Adaptive HVI Transformation (DAHT) for luminance-chrominance decoupling; (2) Multi-Expert Feature Extraction for depth-manifold anchoring; and (3) Depth-Manifold Modulated Attention (DMMA) for geometry-constrained refinement.

4.1 Depth-Adaptive HVI Transformation

Depth-Adaptive HVI Transformation (DAHT) decouples luminance from chrominance while suppressing color noise. We map the input $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ into the HVI space [Yan *et al.*, 2025], which consists of Intensity ($\boldsymbol{I}$) and orthogonal chromatic coordinates ($\boldsymbol{H}, \boldsymbol{V}$). While $\boldsymbol{H}$ and $\boldsymbol{V}$ are projected from the Hue ($\mathbf{h}$) and Saturation ($\mathbf{s}$) of the HSV space [Hu *et al.*, 2023], the Intensity $\boldsymbol{I}$ is calculated via Max-RGB [Land, 1977]:

$$\boldsymbol{I}(u,v) = \max_{c\in R,G,B} \boldsymbol{X}_c(u,v). \tag{1}$$

To eliminate artifacts, we introduce a depth-aware modulation coefficient $\alpha_{\mathbf{k}}$:

$$\alpha_{\mathbf{k}}(u,v) = \left(\sin\left(\frac{\pi \boldsymbol{I}(u,v)}{2}\right) + \epsilon\right)^{\mathcal{A}\cdot K}, \tag{2}$$

where $\mathcal{A}$ is a depth-adaptive modulation map derived from depth $\mathbf{D}$ through the DA-Kernel, and K is a learnable scaling factor. The final chromatic components $(\boldsymbol{H}, \boldsymbol{V})$ are:

$$\boldsymbol{H}(u,v), \boldsymbol{V}(u,v) = \alpha_{\mathbf{k}}(u,v)\cdot \mathbf{s}(u,v)\cdot \left(\cos\frac{\pi \mathbf{h}}{3}, \sin\frac{\pi \mathbf{h}}{3}\right). \tag{3}$$

This allows $(\boldsymbol{H}, \boldsymbol{V}, \boldsymbol{I})$ to provide a geometrically consistent representation for feature extraction.

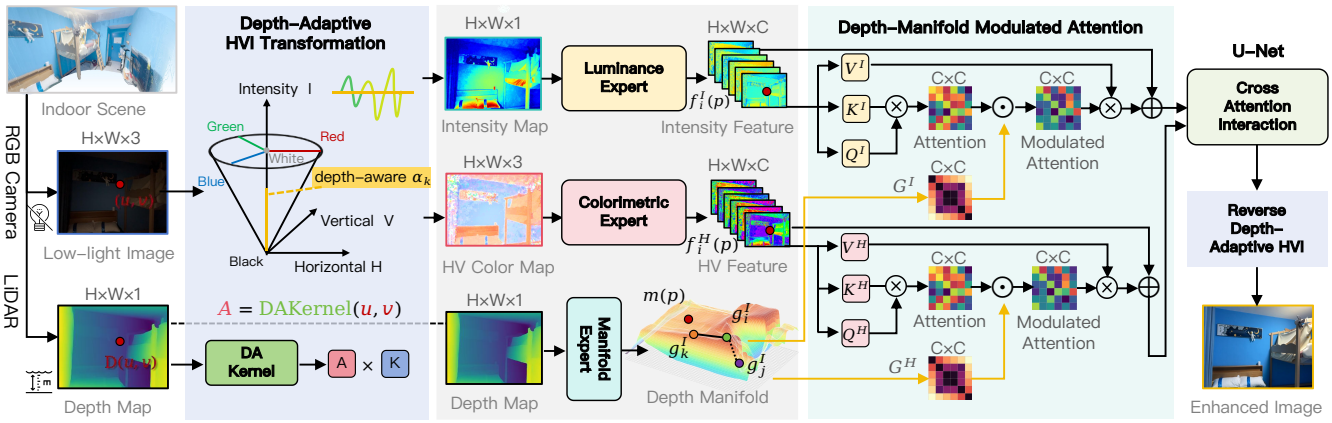

Figure 3: Overview of the proposed SPACE framework. The architecture systematically integrates geometric priors through four stages: (1) Depth-Adaptive HVI Transformation (DAHT) decouples luminance and chrominance under depth guidance; (2) Multi-Expert Feature Extraction grounds appearance features in a learned high-dimensional depth manifold; (3) Depth-Manifold Modulated Attention (DMMA) enforces explicit geometric affinity constraints on feature interactions; and (4) the cross-attention module enables bidirectional interaction between luminance and chrominance features, followed by final reconstruction via a reverse DAHT process.

4.2 Multi-Expert Feature Extraction

This module anchors visual features onto a high-dimensional depth manifold. Specifically, the Luminance expert Φ_L, Colorimetric expert Φ_C, and Manifold expert Φ_M extract features from the intensity $\boldsymbol{I}$, chromatic components $\boldsymbol{H}$ and $\boldsymbol{V}$, and depth $\boldsymbol{D}$ as follows:

$$\mathbf{f}^I = \Phi_L(\boldsymbol{I}), \quad \mathbf{f}^H = \Phi_C(\boldsymbol{H}, \boldsymbol{V}), \quad \mathbf{m} = \Phi_M(\boldsymbol{D}). \tag{4}$$

In this formulation, $\mathbf{f}^I$ and $\mathbf{f}^H \in \mathbb{R}^{\mathcal{H}\times\mathcal{W}\times\mathcal{C}}$ represent appearance feature maps, while $\mathbf{m} \in \mathbb{R}^{\mathcal{H}\times\mathcal{W}\times\mathcal{C}_m}$ denotes the learned geometric manifold. To link latent semantics to physical space, each channel i is assigned a geometric centroid $\mathbf{g}_i \in \mathbb{R}^{\mathcal{C}_m}$ defined as the spatial expectation on $\mathbf{m}$:

$$\mathbf{g}_i = \mathbb{E}p \in \Omega[m(p)|\mathbf{f}_i] = \frac{\sum_{p\in\Omega} \phi(\mathbf{f}_i(p)) \cdot \mathbf{m}(p)}{\sum_{p\in\Omega} \phi(\mathbf{f}_i(p))}, \tag{5}$$

where p denotes the pixel coordinate within the spatial domain Ω, $\mathbf{f}_i(p)$ is the activation value of the i-th channel, and $\phi(\cdot)$ is a non-linear activation for spatial weight normalization. This formulation assigns a unique Expected Physical Identity to each channel by independently computing the intensity-branch centroids $\mathbf{g}_i^I$ and HV-branch centroids $\mathbf{g}_i^H$. Ultimately, these physically anchored representations ensure structural consistency across HVI components and provide the geometric prerequisite for subsequent modulated attention, effectively shielding the enhancement process from illumination and sensor noise.

4.3 Depth-Manifold Modulated Attention

Depth-Manifold Modulated Attention (DMMA) refines feature representations by imposing geometric constraints on self-attention to preserve structural coherence in extreme low-light reconstruction. Unlike standard attention that relies solely on visual similarity and is sensitive to noise, DMMA leverages the physical identities derived in Sec. 4.2 to construct a geometric affinity matrix $\mathbf{G}^\star \in \mathbb{R}^{\mathcal{C}\times\mathcal{C}}$, which encodes inter-channel physical proximity into the attention weights. The index $\star \in \{I, H\}$ specifies the respective intensity and chrominance branches within the dual-stream architecture. Specifically, for any two feature channels i and j within a given branch, the affinity $\mathbf{G}_{ij}^\star$ is computed as:

$$\mathbf{G}_{ij}^\star = \exp\left(-\frac{\|\mathbf{g}_i - \mathbf{g}_j\|_2^2}{\sigma}\right), \tag{6}$$

where $\mathbf{g}_i^\star$ and $\mathbf{g}_j^\star \in \mathbb{R}^{\mathcal{C}_m}$ denote the geometric centroids in the manifold space and σ is a temperature hyperparameter regulating the geometric constraint. By defining $\mathbf{G}^\star$ over the channel dimension $\mathcal{C}$, the module ensures that interactions are modulated by the expected physical depth of the underlying semantic concepts.

During the refinement stage, for an input feature map $\mathbf{f}^\star \in \{\mathbf{f}^I, \mathbf{f}^H\}$, the query $\mathbf{Q}^\star$, key $\mathbf{K}^\star$, and value $\mathbf{V}^\star$ matrices are generated via linear projections. The final modulated attention is formulated by integrating the geometric affinity into the attention map:

$$\hat{\mathbf{f}}^\star = \mathrm{Softmax}\left(\frac{\mathbf{Q}^\star \mathbf{K}^{\star T}}{\sqrt{d_k}}\right) \odot \mathbf{G}^\star \cdot \mathbf{V}^\star + \mathbf{f}^\star, \tag{7}$$

where $\odot$ denotes the channel-wise modulation operation and d_k is the scaling factor. By integrating these geometric constraints, the DMMA mechanism compels the framework to prioritize feature exchange between semantic components anchored to the same manifold region, which effectively suppresses noise propagation across object boundaries. Consequently, DMMA formulates the enhancement task as a geometry-constrained optimization that ensures structural stability even in near-zero contrast regions. By leveraging these channel-wise physical identities, the SPACE framework successfully preserves sharp edges and accurate scene geometry under extreme illumination.

Method	Complexity		LOL-V1			LAMP-Real			LAMP-Synthetic		
	Params (M)	FLOPs (G)	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
KinD [Zhang *et al.*, 2019]	8.020	34.990	23.018	0.843	0.156	12.400	0.585	0.534	18.421	0.732	0.315
KinD++ [Zhang *et al.*, 2021]	8.270	2970.500	23.107	0.805	0.144	12.488	0.568	0.581	18.654	0.725	0.328
SNR-aware [Xu *et al.*, 2022]	39.120	26.350	26.716	0.851	0.152	14.663	0.587	0.614	21.320	0.794	0.218
QuadPrior [Wang *et al.*, 2024a]	1252.750	1103.200	22.849	0.800	0.201	13.452	0.588	0.674	20.105	0.758	0.284
Bread [Guo and Hu, 2023]	2.020	19.850	25.299	0.847	0.155	13.629	0.601	0.626	21.047	0.782	0.245
RetinexNet [Wei *et al.*, 2018]	0.840	587.470	18.915	0.427	0.470	15.337	0.614	0.780	17.214	0.601	0.512
RetinexFormer [Cai *et al.*, 2023]	1.530	15.850	27.140	0.850	0.129	14.621	0.600	0.521	22.856	0.834	0.165
RetinexMamba [Bai *et al.*, 2024]	3.590	34.760	25.758	0.823	0.167	14.626	0.612	0.490	22.147	0.815	0.188
LLFormer [Wang *et al.*, 2023]	24.550	22.520	25.758	0.823	0.167	14.689	0.587	0.734	22.210	0.806	0.224
LLFlow [Wang *et al.*, 2022]	17.420	358.400	24.998	0.871	0.117	13.729	0.612	0.618	21.842	0.821	0.197
EnlightenGAN [Jiang *et al.*, 2021]	114.350	61.010	20.003	0.691	0.317	11.445	0.483	0.713	16.541	0.614	0.456
NeRCo [Yang *et al.*, 2023]	12.540	48.210	22.450	0.812	0.214	14.110	0.593	0.612	20.365	0.762	0.271
LYT-Net [Brateanu *et al.*, 2024]	3.490	0.045	27.230	0.853	0.089	14.669	0.629	0.623	23.104	0.842	0.142
GSAD [Hou *et al.*, 2023]	17.360	442.020	27.605	0.876	0.092	14.448	0.618	0.576	23.412	0.855	0.136
CWNet [Zhang *et al.*, 2025a]	1.230	11.300	23.600	0.850	0.065	14.431	0.605	0.665	20.874	0.812	0.198
CIDNet [Yan *et al.*, 2025]	1.880	7.570	28.201	0.889	0.079	15.265	0.636	0.494	24.156	0.872	0.115
SPACE (ours)	1.930	10.560	28.400*	0.903*	0.060*	16.207	0.661	0.477	25.243	0.894	0.088

Table 2: Quantitative comparison on LOL-V1 [Wei *et al.*, 2018], LAMP-Real and LAMP-Synthetic datasets. $*$ indicates depth maps estimated by Depth-Anything [Yang *et al.*, 2024] as ground-truth depth is unavailable in LOL-V1. The FLOPs is tested on a single 256 × 256 image. Red, Blue, and Green indicate the best, second-best, and third-best performance.

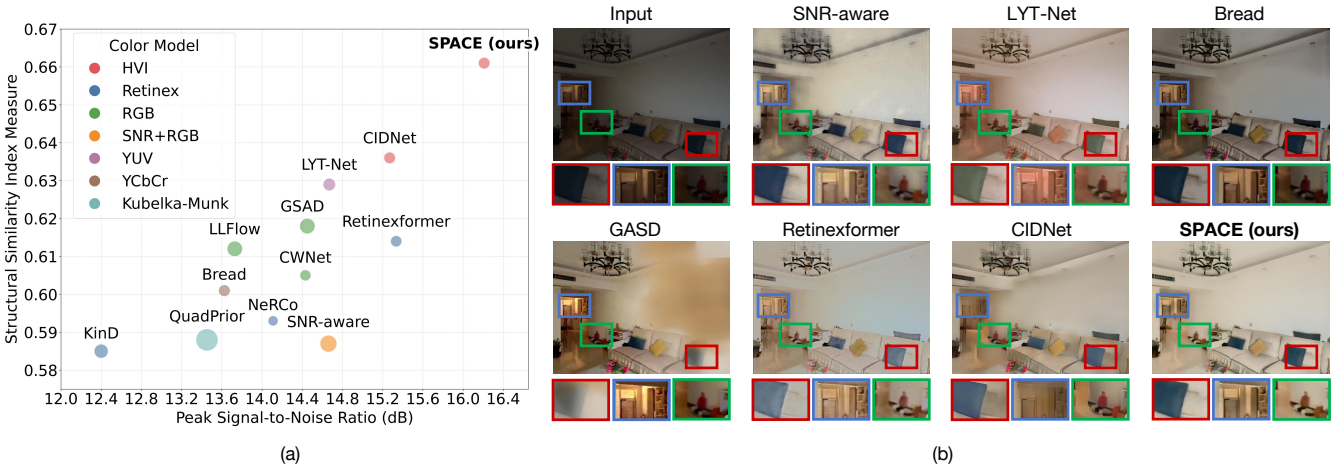

Figure 4: Comparison with advanced methods on the LAMP-Real dataset. (a) Performance Trade-off: Scatter plot of PSNR versus SSIM, where bubble color denotes the utilized color space and bubble size corresponds to the model parameter count. SPACE achieves the state-of-the-art balance. (b) Qualitative Evaluation: Visual comparison on a representative scene. Zoomed-in patches (red, blue, and green boxes) highlight SPACE's superior ability to restore sharp structural edges and suppress noise artifacts.

4.4 Image Reconstruction

The U-Net serves as the structural backbone for joint restoration, where DMMA and cross-attention modules are integrated to perform alternating feature refinement. While DMMA enforces geometric constraints within each branch to preserve structural coherence, the Cross-Attention facilitates bidirectional interaction between the luminance and chrominance domains. This alternating mechanism enables the reciprocal refinement of structural details and color consistency by leveraging the complementary information anchored to the geometric manifold. Finally, the processed features are decoded into restored HVI components and projected back to the sRGB space via a Reverse DAHT process, ensuring that the geometric stability and noise-free representations are successfully preserved in the final Enhanced Image.

5 Experiments

5.1 Experiment Settings

Datasets. Besides the proposed LAMP-Real and LAMP-Synthetic datasets, experiments are also conducted on the widely adopted LOL-V1 [Wei *et al.*, 2018], SICE [Cai *et al.*, 2018], and SID [Chen *et al.*, 2018] benchmarks.

Evaluation Metrics. Following [Yan *et al.*, 2025], we quantitatively evaluate performance using Peak Signal-to-Noise Ratio (PSNR↑), Structural Similarity Index (SSIM↑), and Learned Perceptual Image Patch Similarity (LPIPS↓). We additionally report model complexity in terms of the number of parameters and FLOPs.

Implementation Details. Our framework is trained using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 1×10^{-4} and steadily decreased

Method		Performance		
Color Space Modeling	w/ depth	PSNR↑	SSIM↑	LPIPS↓
sRGB	✗	$14.850_{(-1.357)}$	$0.621_{(-0.040)}$	$0.542_{(+0.065)}$
Retinex	✗	$15.215_{(-0.992)}$	$0.630_{(-0.031)}$	$0.528_{(+0.051)}$
HSV	✗	$15.482_{(-0.725)}$	$0.638_{(-0.023)}$	$0.510_{(+0.033)}$
HVI (w/o DAHT)	✗	$15.816_{(-0.391)}$	$0.649_{(-0.012)}$	$0.491_{(+0.014)}$
HVI (w/ DAHT)	✓	**16.207**	**0.661**	**0.477**
Network Architecture	w/ dual	PSNR↑	SSIM↑	LPIPS↓
UNet Baseline	✗	$12.315_{(-3.892)}$	$0.478_{(-0.183)}$	$0.695_{(+0.218)}$
w/ Vanilla SA	✗	$13.684_{(-2.523)}$	$0.552_{(-0.109)}$	$0.621_{(+0.144)}$
w/ Vanilla SA + CA	✓	$14.852_{(-1.355)}$	$0.615_{(-0.046)}$	$0.538_{(+0.061)}$
w/ DMMA	✗	$15.936_{(-0.271)}$	$0.653_{(-0.008)}$	$0.486_{(+0.009)}$
w/ DMMA + CA (full)	✓	**16.207**	**0.661**	**0.477**

Table 3: Ablation study of color space modeling and architectural components on LAMP-Real. "SA" indicates self-attention [Zamir *et al.*, 2022], "CA" denotes lighten Cross-Attention [Yan *et al.*, 2025], and "dual" represents the dual-branch design.

to 1×10^{-7} using a cosine annealing scheme. Training is conducted on a single NVIDIA RTX 4090 GPU. Additional training details are provided in the Appendix.

5.2 Main Results on LLIE

Quantitative Comparison. As shown in Table 2 and Fig. 4(a), SPACE achieves superior performance across all three benchmarks. Specifically, it achieves PSNR scores of 16.207 dB and 25.243 dB on LAMP-Real and LAMP-Synthetic, respectively, outperforming the second-best method, CIDNet [Yan *et al.*, 2025], by margins of 0.942 dB and 1.087 dB. Furthermore, on the LOL-V1 [Wei *et al.*, 2018] dataset, our method achieves a PSNR of 28.400 dB. This result highlights that competitive reconstruction performance can be achieved even when relying on estimated depth, as ground-truth depth is unavailable for this benchmark.

Qualitative Evaluation. Results in Fig. 4(b) confirm that SPACE excels at recovering structural details and maintaining color fidelity in extreme low-light environments. While baseline methods such as Retinexformer [Cai *et al.*, 2023] or GSAD [Hou *et al.*, 2023] often produce blurred boundaries and residual artifacts, our framework preserves sharp edges and realistic textures. By leveraging geometric centroids to ground semantic features, the model effectively distinguishes physical surfaces from noise, ensuring high perceptual quality and geometric consistency across diverse indoor scenes.

Computational Efficiency. As shown in the performance–complexity analysis in Fig. 4(a), SPACE exhibits a favorable trade-off between restoration quality and model complexity compared to existing methods. While many prior approaches improve performance by substantially increasing model size, SPACE achieves strong PSNR and SSIM with a compact design of 1.610M parameters and 15.570G FLOPs.

5.3 Ablation Study

We conduct a thorough ablation study to verify the effectiveness of our proposed color space modeling and architectural components, with results summarized in Table 3.

Ablation of Color Space Modeling. Table 3 demonstrates that our proposed DAHT significantly outperforms standard color space representations by achieving the highest PSNR of 16.207 dB. Compared to sRGB, Retinex, and HSV modeling, our DAHT provides substantial PSNR gains of 1.357 dB, 0.992 dB, and 0.725 dB, respectively. Notably, removing the depth-adaptive mechanism (w/o DAHT) results in a 0.391 dB decline, confirming that the proposed transformation effectively refines the HVI-based color modeling.

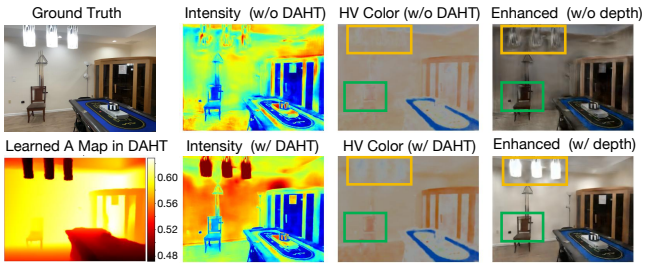

Figure 5: Visualization of the effectiveness of the DAHT module.

Method	Arch. Type	PSNR↑	SSIM↑	LPIPS↓	Time (s)↓
LEDNet	CNN	24.152 (+3.456)	0.835 (+0.010)	0.116 (-0.005)	0.032
SNR-aware	Transformer	23.324 (+0.771)	0.842 (-0.009)	0.128 (-0.054)	0.041
LLFormer	Transformer	24.218 (+2.615)	0.854 (+0.060)	0.115 (-0.094)	0.082
CIDNet	Transformer	24.785 (+0.629)	0.878 (+0.006)	0.102 (-0.013)	**0.031**
GSAD	Diffusion	**25.356** (+3.562)	0.878 (+0.030)	0.104 (-0.010)	0.285
DiffLight	CNN+Diffusion	25.120 (+1.364)	0.865 (+0.003)	0.098 (-0.012)	0.442
SPACE (ours)	Transformer	25.243	**0.894**	**0.088**	0.038

Table 4: Results of applying DAHT as a plug-in to various LLIE methods [Zhou *et al.*, 2022; Xu *et al.*, 2022; Wang *et al.*, 2023; Hou *et al.*, 2024; Feng *et al.*, 2024] on the LAMP-Synthetic dataset. Red values in parentheses indicate the performance gain relative to the original baseline color space. Time denotes the average GPU inference time per image.

Ablation of Network Architecture. The structural contribution of each module is verified in Table 3 by comparing the vanilla UNet baseline (12.315 dB) against our progressively refined framework. While integrating vanilla Self-Attention (SA) improves the PSNR to 13.684 dB, our DMMA achieves a far superior PSNR of 15.936 dB, demonstrating the efficacy of anchoring attention weights to the geometric manifold. Furthermore, the integration of lighten Cross-Attention (CA) within the dual-branch design enables robust information exchange between HVI components, resulting in the optimal PSNR of 16.207 dB for the full model.

5.4 Further Analysis and Discussion

Effectiveness of DAHT Module. DAHT leverages depth-aware guidance to modulate the transformation into the HVI space. As illustrated in Fig. 5, the learned modulation map $\mathcal{A}$ adaptively steers this process, refining both intensity and colorimetric representations. Specifically, the HV Color map under DAHT exhibits significantly superior structural integrity compared to the baseline, ensuring that the final enhanced output preserves accurate textures and color fidelity while aligning closely with the ground truth.

Generalizability of DAHT. As shown in Table 4, we evaluate DAHT as a plug-and-play module across a variety of state-of-the-art low-light enhancement architectures. The integration of DAHT consistently improves performance across all models, yielding significant PSNR gains such as 3.456 dB for LEDNet [Zhou *et al.*, 2022], 2.615 dB for LLFormer [Wang *et al.*, 2023], and 1.364 dB for DiffLight [Feng

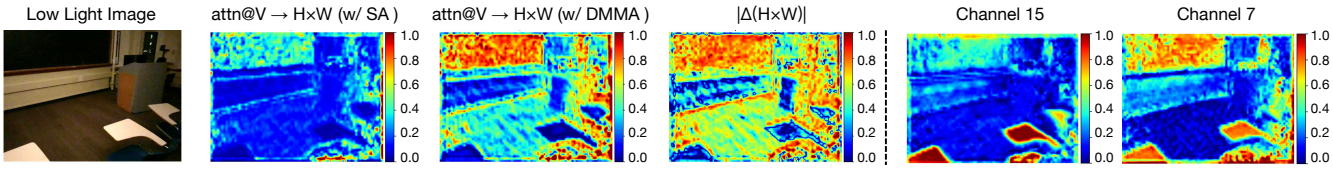

Figure 6: Visual analysis of the proposed DMMA module. (Left) Comparison of aggregated attention-weighted features (attn@V) between standard Self-Attention (SA) and DMMA. The difference map $|\Delta|$ highlights how the depth manifold guides structural enhancement in low-light regions. (Right) High similarity between Channels 15 and 7 in capturing consistent geometric structures, demonstrating the cross-channel feature alignment enforced by depth-manifold modulation.

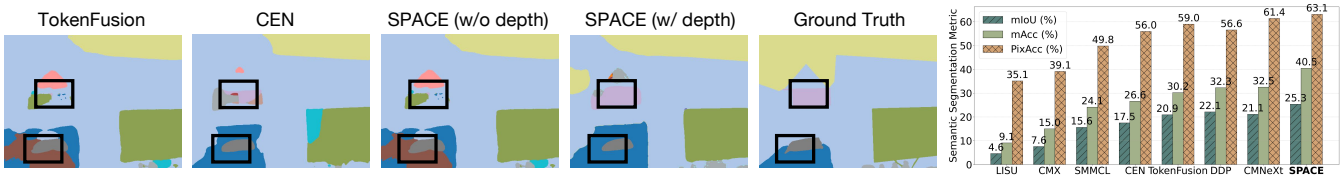

Figure 7: Semantic segmentation results on LAMP-Real. Qualitative comparison (left) shows SPACE (w/ depth) better preserves structural details. Quantitative results (right) confirm SPACE achieves superior performance over advanced methods.

Method	LAMP-Real		SICE		SID	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Compared Baselines w/o Depth Information						
RetinexNet	15.337	0.614	12.424	0.613	15.695	0.395
LLFlow	13.729	0.612	12.737	0.617	16.226	0.367
CIDNet	15.265	0.636	13.435	0.642	22.904	0.676
SPACE w/ Different Depth Estimation Methods						
MiDaS v3	15.756	0.640	13.512	0.646	23.284	0.682
ZoeDepth	15.942	0.651	13.628	0.650	23.592	0.690
Marigold	16.122	0.660	13.745	0.656	23.852	0.699
DepthAnything	16.084	0.657	13.821	0.653	23.845	0.706
Real Depth	16.207	0.661	–*	–*	–*	–*

Table 5: Comparison of depth-free baselines and our SPACE using various depth sources. Utilizing monocular depth estimates [Ranftl *et al.*, 2020; Bhat *et al.*, 2022; Ke *et al.*, 2025; Yang *et al.*, 2024] validates the practical utility and robustness of our framework in the absence of ground-truth (GT) depth. Red and Blue indicate the best and second-best performance. "*" indicates datasets where GT depth is unavailable.

et al., 2024]. These results highlight the universal applicability of our depth-adaptive transformation, effectively enhancing both CNN-based and Transformer-based frameworks.

Effectiveness of DMMA Module. DMMA uses depth-manifold to enhance structural awareness in challenging low-light environments. As shown in Fig. 6 (Left), DMMA achieves more coherent and intense feature responses than standard Self-Attention (SA), with the difference map $|\Delta|$ highlighting the structural gains provided by depth-manifold modulation. Crucially, the modulated attention mechanism induces high inter-channel correlation between Channels 15 and 7, resulting in their synchronized focus on consistent geometric structures (Fig. 6, Right). This phenomenon demonstrates that the depth manifold acts as a strong geometric constraint that enforces cross-channel feature alignment, enabling the model to reliably extract stable and physically meaningful spatial layouts despite severe visual noise.

Robustness to Depth Sources. Table 5 evaluates the robustness of SPACE when ground-truth depth is unavailable, by using estimated depth maps from monocular models such as Marigold [Ke *et al.*, 2025] and DepthAnything [Yang *et al.*, 2024]. Despite the lack of precise LiDAR data, SPACE maintains a PSNR of 16.084 dB using DepthAnything estimates, which closely matches the 16.207 dB achieved with real depth. This performance stability validates the practical utility of our framework in real-world environments where ground-truth geometric priors are unavailable.

Extension to Low-light Semantic Segmentation. We further extend the SPACE architecture to low-light semantic segmentation on the LAMP dataset. Fig. 7 demonstrates SPACE's quantitative superiority and qualitative structural consistency over representative baselines. Additional details and results are provided in the Appendix.

6 Conclusion

In this paper, we introduce the LAMP dataset and SPACE framework to systematically address geometric consistency challenges in extreme low-light conditions. LAMP provides a large-scale multimodal benchmark with over 18k precisely aligned low/normal-light RGB-D pairs and pixel-wise semantic annotations, helping fill the gap in geometric modalities in existing low-light datasets. Built on LAMP, SPACE leverages the proposed Depth-Adaptive HVI Transformation (DAHT) to decouple luminance and chrominance, and Depth-Manifold Modulated Attention (DMMA) to enforce explicit geometric constraints, leading to improved structural fidelity in restored images. Extensive experiments show that SPACE achieves competitive performance across multiple benchmarks while maintaining a favorable trade-off between model complexity and reconstruction quality. Future work will explore the application of geometric priors to additional downstream low-light visual perception tasks, such as video enhancement and semantic segmentation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62472381) and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

[Bai *et al.*, 2024] Jiesong Bai, Yuhao Yin, and Qiyuan He. Retinexmamba: Retinex-based mamba for low-light image enhancement. *arXiv preprint arXiv:2405.03349*, 2024.

[Bhat *et al.*, 2022] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: combining relative and metric depth (official implementation). *IEEE TPAMI*, 2022.

[Brateanu *et al.*, 2024] Alexandru Brateanu, Raul Balmez, Adrian Avram, and CC Orhei. Lyt-net: Lightweight yuv transformer-based network for low-light image enhancement. *arXiv preprint arXiv:2401.15204*, 2024.

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

[Cai *et al.*, 2023] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, pages 12504–12513, 2023.

[Chen *et al.*, 2018] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, pages 3291–3300, 2018.

[Chen *et al.*, 2019] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, pages 3185–3194, 2019.

[Chen *et al.*, 2025] Kerui Chen, Zhiliang Wu, Wenjin Hou, Kun Li, Hehe Fan, and Yi Yang. Prompt-aware controllable shadow removal. *arXiv preprint arXiv:2501.15043*, 2025.

[Cui *et al.*, 2024] Ziteng Cui, Lin Gu, Xiao Sun, Xianzheng Ma, Yu Qiao, and Tatsuya Harada. Aleth-nerf: Illumination adaptive nerf with concealing field assumption. In *AAAI*, 2024.

[Dai *et al.*, 2017] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017.

[Fan *et al.*, 2025] Guodong Fan, Zishu Yao, Guang-Yong Chen, Jian-Nan Su, and Min Gan. Iniretinex: Rethinking retinex-type low-light image enhancer via initialization perspective. In *AAAI*, volume 39, pages 2834–2842, 2025.

[Feng *et al.*, 2024] 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.

[Guo and Hu, 2023] Xiaojie Guo and Qiming Hu. Low-light image enhancement via breaking down the darkness. *IJCV*, 131(1):48–66, 2023.

[Hai *et al.*, 2023] 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. *JVCIR*, 90:103712, 2023.

[Hou *et al.*, 2023] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *NeurIPS*, 36:79734–79747, 2023.

[Hou *et al.*, 2024] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *NeurIPS*, 36, 2024.

[Hu *et al.*, 2023] Changhui Hu, Lintao Xu, Yanyong Guo, Xiaoyuan Jing, Xiaobo Lu, and Pan Liu. Hsv-3s and 2d-gda for high-saturation low-light image enhancement in night traffic monitoring. *IEEE TITS*, 24(12):15190–15206, 2023.

[Jiang *et al.*, 2021] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 30:2340–2349, 2021.

[Ke *et al.*, 2025] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025.

[Land, 1977] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

[Lee *et al.*, 2012] Sooyeon Lee, Youngshin Kwak, Youn Jin Kim, Sehyeok Park, and Jaehyun Kim. Contrast-preserved chroma enhancement technique using ycbcr color space. *IEEE Transactions on Consumer Electronics*, 58(2):641–645, 2012.

[Lin *et al.*, 2025a] Yingqi Lin, Xiaogang Xu, Jiafei Wu, Yan Han, and Zhe Liu. Geometric-aware low-light image and video enhancement via depth guidance. *IEEE TIP*, 2025.

[Lin *et al.*, 2025b] Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaohu Xing, Wenxue Li, Lei Zhu, and Xinghao Ding. Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. In *AAAI*, volume 39, pages 5307–5315, 2025.

[Liu *et al.*, 2026] Yuetong Liu, Yunqiu Xu, Yang Wei, Xiuli Bi, and Bin Xiao. Clear nights ahead: Towards multi-weather nighttime image restoration. In *AAAI*, volume 40, pages 7395–7403, 2026.

[Lore *et al.*, 2017] 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.

[Ranftl *et al.*, 2020] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2020.

[Shang *et al.*, 2024] Kai Shang, Mingwen Shao, Chao Wang, Yuanshuo Cheng, and Shuigen Wang. Multi-domain multi-scale diffusion model for low-light image enhancement. In *AAAI*, volume 38, pages 4722–4730, 2024.

[Stark, 2000] Jalex Stark. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE TIP*, 9(5):889–896, 2000.

[Vonikakis, 2021] Vasileios Vonikakis. Tm-died: The most difficult image enhancement dataset, 2021.

[Wang *et al.*, 2007] Lingxue Wang, Yuanmeng Zhao, Weiqi Jin, Shiming Shi, and Shengxiang Wang. Real-time color transfer system for low-light level visible and infrared images in yuv color space. In *Signal Processing, Sensor Fusion, and Target Recognition XVI*, volume 6567, pages 539–546. SPIE, 2007.

[Wang *et al.*, 2021] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *ICCV*, pages 9700–9709, 2021.

[Wang *et al.*, 2022] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *AAAI*, volume 36, pages 2604–2612, 2022.

[Wang *et al.*, 2023] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *AAAI*, volume 37, pages 2654–2662, 2023.

[Wang *et al.*, 2024a] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *CVPR*, pages 26057–26066, 2024.

[Wang *et al.*, 2024b] Zhen Wang, Dongyuan Li, Guang Li, Ziqing Zhang, and Renhe Jiang. Multimodal low-light image enhancement with depth information. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4976–4985, 2024.

[Wang *et al.*, 2025] Tao Wang, Kaihao Zhang, Yong Zhang, Wenhan Luo, Björn Stenger, Tong Lu, Tae-Kyun Kim, and Wei Liu. Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement. *Pattern Recognition*, 166:111628, 2025.

[Wei *et al.*, 2018] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[Wei *et al.*, 2021] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE TPAMI*, 44(11):8520–8537, 2021.

[Xu *et al.*, 2022] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *CVPR*, pages 17714–17724, 2022.

[Xu *et al.*, 2023] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *CVPR*, pages 9893–9903, 2023.

[Yan *et al.*, 2025] Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *CVPR*, pages 5678–5687, 2025.

[Yang *et al.*, 2021] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE TIP*, 30:2072–2086, 2021.

[Yang *et al.*, 2023] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *ICCV*, pages 12918–12927, 2023.

[Yang *et al.*, 2024] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022.

[Zhang *et al.*, 2019] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, pages 1632–1640, 2019.

[Zhang *et al.*, 2021] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *IJCV*, 129:1013–1037, 2021.

[Zhang *et al.*, 2023] Yue Zhang, Hehe Fan, Yi Yang, and Mohan Kankanhalli. Dpmix: Mixture of depth and point cloud video experts for 4d action segmentation. *arXiv preprint arXiv:2307.16803*, 2023.

[Zhang *et al.*, 2025a] 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 *ICCV*, pages 8789–8799, 2025.

[Zhang *et al.*, 2025b] Yue Zhang, Yingzhao Jian, Hehe Fan, Yi Yang, and Roger Zimmermann. Uni3d-moe: Scalable multimodal 3d scene understanding via mixture of experts. *arXiv preprint arXiv:2505.21079*, 2025.

[Zhou *et al.*, 2022] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, pages 573–589. Springer, 2022.

[Zhou *et al.*, 2025] Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. *arXiv preprint arXiv:2502.04370*, 2025.