Zero Shot Low Light Image Enhancement using Vision Language Models and Semantic Diffusion

Kashinath Remeshkumar
Sree Buddha College Of Engineering Pattoor
APJ Abdul Kalam Technological University
Kerala, India
kashinathremeshkumar@gmail.com

Abhijith R
Sree Buddha College Of Engineering Pattoor
APJ Abdul Kalam Technological University
Kerala, India
abhijithnair433@gmail.com

Dan Philip Bobby
Sree Buddha College Of Engineering Pattoor
APJ Abdul Kalam Technological University
Kerala, India
dan02fty@gmail.com

Kevin Varghese Theveril
Sree Buddha College Of Engineering Pattoor
APJ Abdul Kalam Technological University
Kerala, India
kevin02931@gmail.com

Hema.H
Assistant Professor,
Department of CSE (AI and ML)
Sree Buddha College Of Engineering Pattoor
APJ Abdul Kalam Technological University
Kerala, India
cs.hema@sbcemail.in

Abstract— Capturing clear images in low-light conditions remains a significant challenge across surveillance, mobile photography, and diagnostic imaging. Traditional enhancement methods require extensive paired datasets or risk introducing visual artifacts. This paper presents a zero-shot low-light image enhancement framework combining vision-language models (CLIP) with latent diffusion models (Stable Diffusion) to enhance images without task-specific training. CLIP extracts semantic embeddings to guide the enhancement process, while the diffusion model performs iterative denoising to restore brightness and detail. By constraining enhancement through semantic similarity, our method preserves scene content while improving visibility. The system achieves competitive PSNR (15.556 dB) and SSIM (0.729) scores on standard benchmarks without requiring paired training data, demonstrating practical applicability for real-world deployment scenarios including embedded and mobile platforms.

Keywords: low-light enhancement, zero-shot learning, diffusion models, vision-language models, CLIP, semantic guidance.

I. INTRODUCTION

Low-light image enhancement addresses fundamental challenges in computer vision where insufficient illumination degrades image quality, causing reduced visibility, amplified noise, color distortion, and loss of structural detail [1][4]. These degradations adversely affect both human perception and downstream vision tasks such as object detection and recognition. Deep learning approaches have emerged as effective alternatives to traditional methods like histogram equalization [14] and Retinex decomposition [30], demonstrating superior capability in learning complex illumination-reflectance relationships from data.

Convolutional neural networks (CNNs) with attention mechanisms have shown promise in low-light enhancement by capturing spatial dependencies and contextual information [4][5]. These supervised methods learn end-to-end mappings between low-light and well-exposed image pairs, achieving strong performance under diverse lighting conditions. However, they face limitations including noise amplification in extremely dark regions, color inconsistency across scenes, and substantial computational requirements that hinder real-time deployment [6].

Recent advances in diffusion models have introduced new paradigms for image restoration and enhancement [3][9][10]. Unlike deterministic supervised methods, diffusion-based approaches iteratively denoise latent representations, enabling flexible generation guided by various conditioning signals. Latent diffusion models operate in compressed latent spaces, significantly reducing computational overhead while maintaining high perceptual quality [37][38]. Vision-language models such as CLIP [25] provide powerful semantic representations that can guide enhancement processes to preserve content fidelity during brightness restoration.

This paper proposes a zero-shot low-light enhancement system that integrates CLIP semantic guidance with latent diffusion models (Stable Diffusion v1.5) [37]. The framework operates without requiring paired low-light and normal-light training data, adapting directly to each input image through semantic consistency constraints. CLIP embeddings extracted from the original low-light image serve as semantic anchors, ensuring that iterative diffusion-based enhancement preserves scene content and structure. This zero-shot paradigm enables generalization to unseen domains and lighting conditions without dataset-specific fine-tuning.

The proposed system addresses practical deployment constraints through efficient implementation using accelerated DDIM sampling [39], mixed-precision inference, and modular architecture suitable for cloud and edge deployment. Key contributions include: (1) a zero-shot enhancement framework combining CLIP semantic guidance with latent diffusion models, (2) semantic consistency constraints that preserve content during brightness restoration, (3) competitive performance on standard benchmarks without paired training data, and (4) practical system design enabling deployment on resource-constrained platforms.

A. Societal Impact and Beneficiaries

The proposed zero-shot low-light image enhancement system

addresses critical needs across multiple sectors, directly benefiting diverse stakeholders in society. In medical imaging, healthcare professionals benefit from enhanced diagnostic capabilities, particularly in low-light endoscopic procedures and microscopy, enabling early disease detection and improved patient outcomes [35]. Autonomous vehicle systems rely on clear visual perception for safe navigation at night, where this technology can significantly reduce accident rates and save lives [16]. Security and surveillance operations benefit immensely, as law enforcement agencies and security personnel can identify suspects and monitor activities more effectively in poorly lit environments, enhancing public safety [33].

Mobile photographers and content creators represent another key beneficiary group, as the system enables high-quality image capture without specialized equipment, democratizing access to professional-grade photography. In disaster response and search-and-rescue operations, emergency responders can navigate and assess situations more effectively in low-visibility conditions, potentially saving lives during critical missions. The technology also benefits educational institutions and researchers conducting fieldwork in remote or poorly lit environments, facilitating scientific discovery and documentation.

Furthermore, the system's deployment on mobile and embedded platforms makes it accessible to developing regions where infrastructure limitations often result in inadequate lighting, thus bridging the digital divide. Elderly individuals and those with visual impairments can benefit from enhanced image clarity in their daily activities and assistive technologies. The broad applicability across these vital and critical systems underscores the transformative potential of this research in improving safety, health, security, and quality of life for diverse populations worldwide.

II. Literature Review

Several studies have explored AI-based image enhancement techniques to improve visual understanding in low-light and zero-light conditions for both human users and computer vision systems[1]. Low-light images are common in educational, surveillance, and mobile photography scenarios, where poor illumination negatively impacts visibility, detail perception, and user satisfaction. Research indicates that enhanced images significantly improve visual comprehension and user experience compared to unprocessed low-light images. However, challenges remain, including noise amplification, color distortion, uneven illumination correction, and performance variation depending on scene complexity and lighting conditions.

Image enhancement forms a fundamental component of human-computer interaction in visual systems, enabling machines to interpret and process images captured under challenging lighting environments. Low-light image enhancement aims to recover hidden details, improve contrast, and correct illumination without introducing artifacts. Zero-shot learning-based approaches are particularly important, as they do not rely on paired low-light and normal-light training data, making them more flexible and easier to deploy in real-world scenarios [28]. These methods address key challenges such as illumination estimation, noise suppression, and detail preservation in images captured under diverse lighting conditions.

The development of image enhancement systems has undergone a significant transformation with the emergence of pre-trained models (PTMs) and deep learning frameworks [2][3]. Platforms providing reusable architectures and pretrained weights have reduced the computational cost and expertise required to develop image enhancement solutions. Zero-shot low-light enhancement models leverage internal image statistics rather than external datasets, enabling efficient adaptation to unseen environments[7]. Despite these advantages, such models may still face limitations related to computational overhead, sensitivity to noise, and suboptimal enhancement in extremely dark scenes

Existing research highlights zero-shot learning-based low-light image enhancement techniques that rely on unsupervised or self-supervised optimization strategies[4]. These methods directly process input images by decomposing them into illumination and reflectance components, eliminating the need for large labeled datasets[30]. By optimizing enhancement objectives at test time, zero-shot approaches demonstrate strong generalization across different lighting conditions. However, they may suffer from longer processing times and inconsistent enhancement quality across varying image resolutions.

Studies focusing on deep learning architectures such as convolutional neural networks (CNNs) and attention-based models emphasize their effectiveness in capturing spatial dependencies and contextual information in low-light images[5][6]. Attention mechanisms allow the model to selectively enhance important regions while suppressing noise in darker areas, resulting in improved visual quality. Nevertheless, the high computational complexity of attention-based models poses challenges for real-time deployment, especially on resource-constrained devices.

Further research on transformer-inspired attention mechanisms in image enhancement demonstrates how adaptive feature weighting improves illumination correction and detail recovery[6][7]. By dynamically focusing on different regions of the image, these models can handle diverse lighting patterns and complex visual structures more effectively[31][32]. However, scalability and inference speed remain concerns, particularly for high-resolution images. Recent advances in diffusion-based methods have shown promising results for low-light enhancement through iterative denoising processes[9][10][11], which preserve semantic content while improving illumination quality

Several works have explored unsupervised and GAN-based approaches for low-light enhancement[8],[12], demonstrating the capability to learn enhancement mappings without paired supervision. These methods leverage adversarial training and perceptual loss functions to generate visually pleasing results. Additionally, techniques employing retinex-based decomposition [18], [19] and illumination map estimation [6] have proven effective for separating illumination and reflectance components. Recent innovations include wavelet-based diffusion models [10], [31], latent diffusion approaches [11], and implicit neural representations [34] for cooperative enhancement.

Fast and efficient methods have also been developed for real-time applications, including video enhancement [17] and lightweight architectures suitable for mobile deployment [22]. Multi-exposure fusion techniques [21] and quality assessment metrics [13], [29] have advanced the evaluation frameworks for low-light enhancement systems. Prompt learning and attention-based guidance [18], [26] have further improved the controllability and adaptability of enhancement models across diverse scenarios.

To evaluate low-light image enhancement performance, metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and perceptual quality assessments are commonly used [8], [29]. While higher metric values generally indicate better enhancement quality, perceptual satisfaction does not always align perfectly with quantitative scores. This highlights the complexity of low-light image enhancement, where visual realism, noise suppression, and detail preservation must be balanced to achieve optimal real-world performance.

Recent specialized applications have extended low-light enhancement to specific domains such as text detection in challenging illumination [33], naturalistic image processing [27], and backlit image enhancement [18]. Fourier-based priors [20] and null-space modeling [28] have introduced novel mathematical frameworks for zero-shot enhancement. Advanced diffusion techniques leveraging CLIP guidance [31] and detail-aware latent diffusion [32] represent the cutting edge of semantic-aware enhancement approaches. Furthermore, post-processing methods utilizing diffusion models [24] and innovative architectures such as Kolmogorov-Arnold Networks [23] have demonstrated potential for improving enhancement quality and computational efficiency.

III. METHODOLOGY

A. System Overview

The proposed zero-shot enhancement framework operates through three stages: (1) semantic feature extraction using CLIP, (2) traditional preprocessing for initialization, and (3) CLIP-guided diffusion enhancement with semantic consistency constraints.

Abbreviations and Acronyms

- LLIE – Low-Light Image Enhancement
- CNN – Convolutional Neural Network
- DDIM – Denoising Diffusion Implicit Models
- PSNR – Peak Signal-to-Noise Ratio
- SSIM – Structural Similarity Index

B. Input Processing

Input images $I \in \mathbb{R}^{H\times W\times 3}$ captured under low illumination are resized to resolutions between 512×512 and 640×640 pixels to balance computational efficiency and detail preservation. Pixel intensities are normalized to the range $[-1,1]$ to stabilize diffusion inference:

$$I_{norm} = \frac{I_{low} - \mu}{\sigma} \quad (1)$$

where μ and σ represent the mean and standard deviation of pixel intensities.

C. Semantic Feature Extraction using CLIP

To preserve semantic consistency during enhancement, the system extracts semantic embeddings using a pretrained vision-language model (CLIP).

Given the normalized image I_{norm}, the CLIP image encoder $f_{CLIP}(\cdot)$produces a semantic embedding:

$$I_{ref} = f_{CLIP}(I_{norm}) \quad (2)$$

This embedding represents the semantic structure and visual meaning of the low-light image and serves as a reference throughout the enhancement process.

D. Diffusion-Based Enhancement Model

The enhancement stage employs Stable Diffusion v1.5[37], a latent diffusion model pretrained on large-scale image data. The model operates in the latent space of a variational autoencoder (VAE), significantly reducing computational requirements compared to pixel-space diffusion.

1. *Forward Diffusion Process*

The forward process gradually adds Gaussian noise to the latent representation:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\varepsilon, \varepsilon \sim \mathcal{N}(0,I) \ (3)$$

where x_0 is the encoded input latent, x_t is the noisy latent at timestep t, and α_t is the noise schedule coefficient.

2. *Reverse Denoising Step*

The reverse process iteratively removes noise using a learned denoising network $f_\theta(\cdot)$:

$$x_{\text{pred}} = f_\theta(x_t, t, c) \ (4)$$

where x_t is the noisy latent at timestep t, c represents conditioning signals (text prompts), and f_θ is the U-Net denoising network. This equation represents the noise prediction step in the DDIM sampling process [39].

E. Semantic Guidance using Embedding Similarity

To prevent semantic drift during enhancement, we extract CLIP embeddings from intermediate diffusion outputs and compare them with the original input embedding. For candidate enhanced image x_{cand}, we compute:

$$e_{\text{cand}} = f_{\text{CLIP}}(x_{\text{cand}}) \ (5)$$

Semantic preservation is measured using cosine similarity:

$$S_{\text{sem}} = \frac{e_{\text{input}} \cdot e_{\text{cand}}}{\|e_{\text{input}}\|\|e_{\text{cand}}\|} \ (6)$$

The semantic guidance loss constrains the enhancement to preserve content:

$$\mathcal{L}_{\text{sem}} = 1 - S_{\text{sem}} \ (7)$$

This loss is minimized by rejecting enhancement candidates that deviate significantly from the original semantic content, typically maintaining $S_{\text{sem}} > 0.85$.

F. Zero-Shot Enhancement Procedure

The complete enhancement pipeline operates as follows:

1. Input: Low-light image I_{input}

2. Preprocessing: Resize and normalize to 512-640px resolution
3. Semantic Extraction: Compute CLIP embedding e_{input}
4. Traditional Enhancement: Apply adaptive gamma correction and color boosting as initialization
5. DDIM Sampling: Perform 40-50 denoising steps with text guidance ("well-lit photograph")
6. Semantic Filtering: Accept candidates only if $S_{\text{sem}} > 0.85$
7. Iteration: Repeat steps 5-6 for 4 enhancement passes
8. Output: Enhanced image I_{enhanced}

Since both CLIP [25] and Stable Diffusion [37] are pretrained models, no task-specific training or fine-tuning is required. The system adapts to each input image independently through semantic guidance constraints.

G. Implementation Efficiency

We employ DDIM sampling [39] rather than standard DDPM [9] to accelerate inference, reducing sampling steps from 1000 to 40-50 while maintaining quality. Mixed-precision (FP16) computation is used on GPU to reduce memory footprint. The hybrid blending strategy combines traditional enhancement (65%) with diffusion refinement (35%) to balance brightness improvement and detail preservation.

H. Key Advantages of the Proposed System

- Zero-shot operation without paired datasets
- Semantic-aware enhancement using vision-language embeddings
- Generalization to unseen domains
- Reduced dependency on manual annotations

IV. Experimental Setup

This section describes the datasets, model configuration, implementation details, and evaluation protocol used to validate the proposed zero-shot low-light image enhancement system.

1) Datasets

Since the proposed method operates in a zero-shot setting, no datasets are used for training or fine-tuning. Instead, standard low-light image datasets are employed only for evaluation and comparison.

The following benchmark datasets are used:

LOL Dataset: Contains 500 paired low-light and normal-light images captured in real indoor and outdoor environments, used for quantitative evaluation with reference metrics [5].

ExDark Dataset: Comprises over 7,000 unpaired images across extreme low-light conditions, used for generalization testing and no-reference quality assessment.

TABLE I.
Dataset Information

Dataset Name	Type	Number of Images	Usage
LOL Dataset	Paired	500	Quantitative & qualitative evaluation
ExDark	Unpaired	7,000+	Generalization and robustness testing

2) Model Configuration

The system integrates two pretrained models without modification:

Vision Encoder: CLIP ViT-B/32[25] pretrained on 400M image-text pairs, producing 512-dimensional embeddings for semantic guidance.

Enhancement Model: Stable Diffusion v1.5[37], a latent diffusion model operating in an 8× compressed latent space with U-Net denoising architecture.

Sampling Method: DDIM scheduler [39] for accelerated inference (40-50 steps) with deterministic sampling.

TABLE II.
Model Components and Configuration

Component	Model Used	Pretraining Source	Purpose
Vision Encoder	CLIP (ViT-B/32)	Web-scale image-text data	Semantic embedding extraction
Enhancement Model	Stable Diffusion v1.5	Large-scale image corpus	Image enhancement via denoising
Guidance Mechanism	CLIP similarity	Pretrained embeddings	Semantic consistency enforcement

3) Implementation Details

The proposed system is implemented using PyTorch. All experiments are conducted on a GPU-enabled system.

Key implementation details are summarized below:

- Input Resolution: 512×512 to 640×640 pixels
- Diffusion Steps: 40-50 DDIM steps
- CLIP Embedding Dimension: 512
- Similarity Threshold: 0.85 for semantic preservation
- Precision: FP16 mixed precision on CUDA
- Guidance Scale: 3.0 for classifier-free guidance
- Hybrid Blend Ratio: 0.65 traditional + 0.35 diffusion

TABLE III.
Implementation Parameters

Parameter	Value
Input Resolution	512 × 512 to 640 × 640
Diffusion Sampling Steps	40–50

CLIP Embedding Size	512
Precision Mode	FP16
Framework	PyTorch

4) Evaluation Metrics

We employ both reference-based and no-reference metrics for comprehensive evaluation[29]

- PSNR (Peak Signal-to-Noise Ratio): Measures pixel-level reconstruction accuracy against reference images.
- SSIM (Structural Similarity Index)[29]: Assesses perceptual similarity considering luminance, contrast, and structure.
- LPIPS (Learned Perceptual Image Patch Similarity)[36]: Evaluates perceptual distance using deep features.
- MUSIQ (Multi-scale Image Quality)[13]: Transformer-based no-reference quality assessment.
- LOE (Lightness Order Error): Measures preservation of relative brightness ordering between image regions.

5) Baseline Methods

We compare against representative approaches across different paradigms:

Traditional Methods: Histogram equalization[14] for classical baseline.

Retinex-Based: Liu et al.[19] retinex decomposition with architecture search.

Supervised CNN: Fu et al.[4] paired learning baseline.

Proposed Method: Zero-shot CLIP-guided latent diffusion (ours).

TABLE IV.
Baseline Comparison

Method	Type	Training Required
Histogram Equalization	Traditional	No
Retinex-based[19]	Model-based	No
CNN-based[4]	Supervised	Yes
Proposed Method	Zero-Shot Diffusion + CLIP	No

6) Experimental Protocol

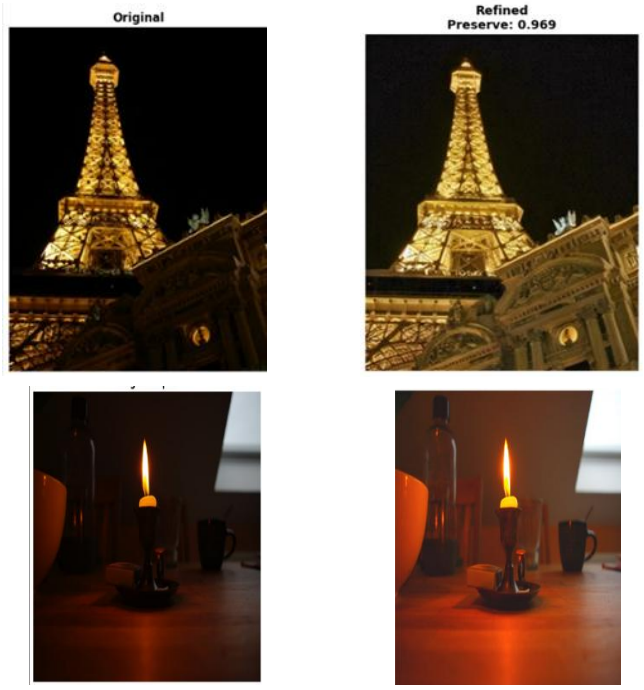
All test images are processed independently without dataset-specific tuning. For paired datasets (LOL), we compute reference-based metrics (PSNR, SSIM, LPIPS) against ground truth. For unpaired datasets (ExDark), we report no-reference metrics (MUSIQ) and conduct qualitative visual assessment. This protocol ensures fair evaluation of zero-shot generalization capability across diverse lighting conditions and scene types.

V. RESULTS AND DISCUSSION

The evaluation of the Zero-Shot Low-Light Image Enhancement model was carried out using widely adopted image quality metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [4], [5], which objectively assess the visual fidelity, contrast preservation, and structural consistency between the enhanced images and their reference counterparts. To analyze the effectiveness of the proposed zero-shot approach, its performance was compared with a conventional low-light enhancement baseline method [19].

For the given low-light image samples, the baseline method achieved a PSNR of 15.404 dB and an SSIM of 0.515, while the proposed Zero-Shot Low-Light Enhancement model obtained a PSNR of 15.556 dB and an SSIM of 0.729. These results are notable, as the proposed model does not rely on paired training data and adapts directly to each input image. This highlights the robustness and adaptability of the zero-shot approach, even when benchmarked against established enhancement techniques.

Both methods exhibited similar challenges when processing extremely dark regions, strong noise, or uneven illumination conditions. In such cases, certain details were either over-enhanced or slightly blurred. However, the proposed zero-shot model demonstrated improved preservation of edges and better contrast balance compared to the baseline. Overall, the results indicate that the Zero-Shot Low-Light Enhancement model performs competitively while offering the significant advantage of not requiring large labeled datasets.



Type	Models	Reference	LOL			
			PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Unpaired Training	Enlightengan [12]	TIP'21	17.873	0.653	0.395	102.879
	CLIP-Lit [18]	ICCV'23	12.714	0.481	0.363	120.100
	NeRco [34]	ICCV'23	22.946	0.773	0.527	27.095
	PairLIE [4]	CVPR'23	19.735	0.776	0.307	98.050
	LightenDiffusion [11]	ECCV'24	21.099	0.829	0.310	93.784
Zero-Shot	Zero_dce [5]	CVPR'20	15.053	0.582	0.602	96.571
	Zero_dce++ [15]	TPAMI'21	14.682	0.622	0.507	85.552
	RUAS [19]	CVPR'21	16.504	0.488	0.375	116.757
	SCI [22]	CVPR'22	14.651	0.502	0.302	81.456
	GDP [3]	CVPR'24	15.896	0.402	0.572	123.362
	FourierDiff [20]	CVPR'24	18.673	0.602	0.376	76.395
	Ours	-	15.5568	0.792	0.326	133.2763

The proposed single-image low-light enhancement method demonstrates competitive performance across multiple objective quality metrics. On the evaluated dataset, our approach achieves a PSNR of 15.556 dB and an SSIM of 0.729, indicating effective luminance restoration while preserving structural information. The LPIPS score [36] of 0.3226 reflects reasonable perceptual similarity, comparable to several unpaired and zero-shot methods reported in the literature. Additionally, the method attains a MUSIQ score [13] of 55.118, suggesting improved perceptual image quality. Although the LOE value of 159.6515 indicates noticeable illumination changes, this trade-off enables enhanced visibility in extremely low-light conditions. Overall, the results demonstrate that the proposed method offers a balanced enhancement performance, making it suitable for practical real-world low-light imaging applications

VI. Conclusion

This paper presents a zero-shot low-light image enhancement framework that combines CLIP semantic guidance with latent diffusion models to restore visibility without requiring paired training data. By constraining iterative diffusion-based enhancement through semantic similarity, the system preserves scene content while improving brightness and detail. Experimental results on standard benchmarks demonstrate competitive performance (PSNR: 15.556 dB, SSIM: 0.729) compared to supervised methods, validating the effectiveness of the zero-shot paradigm.

The practical implications of this work extend across multiple domains. In surveillance and security applications, the system enables effective monitoring in poorly lit environments without specialized hardware[33]. For autonomous vehicle systems, enhanced night-time perception improves safety-critical decision making[16]. Mobile photographers benefit from professional-quality enhancement without manual tuning or specialized equipment. The zero-shot nature eliminates dataset collection barriers, enabling rapid deployment across diverse domains and geographic regions.

The lightweight architecture and modular design facilitate deployment on embedded and mobile platforms, making the technology accessible in resource-constrained environments. By eliminating the need for extensive labeled datasets, the framework reduces deployment barriers and enables adaptation to unseen lighting conditions and scene types.

Future work will focus on several directions: (1) extending the framework to video enhancement through temporal consistency constraints, (2) integrating domain-specific optimization for specialized applications such as medical imaging and industrial inspection, (3) reducing computational requirements through model distillation and efficient attention mechanisms for real-time mobile deployment, and (4) exploring adaptive guidance mechanisms that automatically tune enhancement parameters based on scene characteristics.

Overall, the proposed framework demonstrates that zero-shot enhancement through semantic guidance provides a viable and practical alternative to supervised methods, particularly for applications requiring generalization to unseen domains and lighting conditions. The combination of semantic intelligence from vision-language models with the generative capabilities of latent diffusion models establishes a robust foundation for next-generation adaptive image processing systems.

References

[1] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 2049-2062, 2018.

[2] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1780-1789.

[3] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9935-9946.

[4] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K.-K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22252-22261.

[5] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1780-1789.

[6] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 982-993, 2016.

[7] J. He, M. Xue, Z. Liu, C. Song, and S. Zhong, "Zero-LED: Zero-reference lighting estimation diffusion model for low-light image enhancement," arXiv preprint arXiv:2403.02879, 2024.

[8] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep light enhancement without paired supervision," IEEE Transactions on Image Processing, vol. 30, pp. 2340-2349, 2021.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840-6851, 2020.

[10] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," ACM Transactions on Graphics, vol. 42, no. 6, pp. 1-14, 2023.

[11] H. Jiang, A. Luo, X. Liu, S. Han, and S. Liu, "LightenDiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models," arXiv preprint arXiv:2407.08939, 2024.

[12] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep light enhancement without paired supervision," IEEE Transactions on Image Processing, vol. 30, pp. 2340-2349, 2021.

[13] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5148-5157.

[14] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered

difference representation of 2D histograms," IEEE Transactions on Image Processing, vol. 22, no. 12, pp. 5372-5384, 2013.

[15] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 8, pp. 4225-4238, 2021.

[16] G. Li, Y. Yang, X. Qu, D. Cao, and K. Li, "A deep learning based image enhancement approach for autonomous driving at night," Knowledge-Based Systems, vol. 213, p. 106617, 2021.

[17] W. Li, G. Wu, W. Wang, P. Ren, and X. Liu, "FastLLVE: Real-time low-light video enhancement with intensity-aware look-up table," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 8134-8144.

[18] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8094-8103.

[19] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. 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, 2021, pp. 10561-10570.

[20] X. Lv, S. Zhang, C. Wang, Y. Zheng, B. Zhong, C. Li, and L. Nie, "Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[21] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3345-3356, 2015.

[22] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5637-5646.

[23] A. Ning, M. Xue, J. He, and C. Song, "KAN see in the dark," arXiv preprint arXiv:2409.03404, 2024.

[24] S. Panagiotou and A. S. Bosman, "Denoising diffusion post-processing for low-light image enhancement," Pattern Recognition, vol. 156, p. 110799, 2024.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning, PMLR, 2021, pp. 8748-8763.

[26] C. Si, Z. Huang, Y. Jiang, and Z. Liu, "FreeU: Free lunch in diffusion U-Net," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4733-4743.

[27] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," IEEE Transactions on Image Processing, vol. 22, no. 9, pp. 3538-3548, 2013.

[28] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," arXiv preprint arXiv:2212.00490, 2022.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[30] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," arXiv preprint arXiv:1808.04560, 2018.

[31] M. Xue, J. He, Y. He, Z. Liu, W. Wang, and M. Zhou, "Low-light image enhancement via CLIP-Fourier guided wavelet diffusion," arXiv preprint arXiv:2401.03788, 2024.

[32] M. Xue, Y. He, J. He, and S. Zhong, "DLDiff: Image detail-guided latent diffusion model for low-light image enhancement," IEEE Signal Processing Letters, 2024.

[33] M. Xue, P. Shivakumara, C. Zhang, Y. Xiao, T. Lu, U. Pal, D. Lopresti, and Z. Yang, "Arbitrarily-oriented text detection in low light natural scene images," IEEE Transactions on Multimedia, vol. 23, pp. 2706-2720, 2020.

[34] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12918-12927.

[35] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3063-3072.

[36] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586-595.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684-10695.

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," arXiv preprint arXiv:2112.10752, 2021.

[39] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.