UNIVERSITY OF CALIFORNIA, MERCED

Reference-Guided Image Enhancement with Retinex Priors and Diffusion Models

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Electrical Engineering and Computer Science

by

Siddharth Seth

Committee in charge:

Professor Ming-Hsuan Yang, Chair
Professor Shawn Newsam
Professor Meng Tang

2025

The thesis of Siddharth Seth is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Professor Shawn Newsam

Professor Meng Tang

Professor Ming-Hsuan Yang Chair

University of California, Merced

2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Prof. Ming-Hsuan Yang, whose mentorship, unwavering support, and thoughtful guidance have shaped both the trajectory of my research and the way I approach problems. His patience, availability, and keen insight have made an indelible impact on my growth, both as a researcher and as a person.

I am also immensely thankful to my labmates for the countless brainstorming sessions and collaborative spirit that helped make research fun. I sincerely appreciate the time and effort my committee members, Prof. Shawn Newsam and Prof. Meng Tang, devoted to evaluating my work and offering valuable feedback.

Finally, I would like to thank my family and friends for being my anchor throughout this journey. Their unconditional love, encouragement during the toughest days, and belief in me—even when I doubted myself—have been my greatest source of strength. Whether through a comforting call, a shared laugh, or simply their presence, they have supported me in more ways than I can ever fully express.

ABSTRACT OF THE THESIS

Reference-Guided Image Enhancement with Retinex Priors and Diffusion Models

by

Siddharth Seth
Master of Science in Electrical Engineering and Computer Science
University of California Merced, 2025
Professor Ming-Hsuan Yang, Chair

Image restoration, particularly low-light and compression artifact removal, remains a challenging problem due to severe loss of illumination, color distortion, and high-frequency detail degradation. Traditional methods often struggle to preserve semantic content, while existing learning-based approaches require large annotated datasets and are limited in handling extreme underexposure. In this paper, we propose a novel reference-based image enhancement framework, ReGIE, that integrates Retinex priors, reference image conditioning, and a diffusion model guided via ControlNet. Our method leverages a Retinex-based decomposition of the degraded image to provide structural illumination priors and conditions the generative process using a semantically similar reference image retrieved from a multi-view dataset. To mitigate the over-smoothing effects common in latent diffusion models, we introduce a high-frequency fidelity constraint using Discrete Fourier Transform and Sobel filtering, encouraging edge-aware restoration. Extensive experiments on multi-view datasets demonstrate that our model achieves state-of-the-art performance in both quantitative metrics (PSNR, SSIM, DreamSim) and qualitative perceptual quality. Our results highlight the potential of combining physical priors, generative modeling, and reference-based conditioning for controllable and high-fidelity image enhancement in extremely challenging lighting conditions and compression artifacts.

Chapter 1

Introduction

Image restoration is a fundamental task in computer vision that focuses on recovering high-quality images from degraded observations affected by various distortions such as noise, blur, low resolution, compression artifacts, and exposure issues [11]. These degradations can arise due to limitations in imaging sensors, environmental conditions, motion, or lossy transmission and storage processes. The goal of image restoration is to reconstruct visually appealing and perceptually accurate images while preserving fine details, structures, and textures.

Image compression is indispensable in modern imaging pipelines, enabling efficient storage and transmission of large volumes of visual data. However, lossy compression methods such as JPEG, while widely used due to their simplicity and efficiency, introduce a range of artifacts—blocking, ringing, blurring, and color distortion that significantly degrade visual quality and hinder the performance of downstream vision applications [47, 5]. Consequently, developing robust restoration techniques that can effectively mitigate these artifacts is essential for both perceptual enhancement and reliable machine vision performance.

Capturing high-quality images under diverse and challenging lighting conditions is a critical problem in modern computer vision [4, 48]. Lighting plays a fundamental role in image formation, influencing the visibility, color balance, contrast, and overall perceptual quality of captured images. While well-lit environments allow for clearer, more detailed captures, extreme lighting conditions—such as low-light (under-exposure) introduce significant degradations that compromise both human visual perception and automated com-

puter vision algorithms.

In under-exposed (low-light) conditions, the lack of sufficient illumination results in dark, low-contrast images, making it difficult to discern objects and details [29]. This is particularly challenging for vision systems operating at night, in dimly lit indoor environments, or under adverse weather conditions like fog or heavy rain. Low-light images frequently suffer from:

- Noise amplification – Due to increased sensor gain (ISO) settings, images exhibit excessive noise, graininess, and speckle distortions [40].
- Reduced contrast and visibility – Shadows and dark regions become indistinguishable, leading to loss of critical scene information [52].
- Color distortion – Under low-light, cameras struggle to accurately capture colors, often resulting in unnatural color shifts or washed-out tones [50].
- Detail loss and blur – Motion blur increases in low-light conditions as cameras use longer exposure times to compensate for darkness [4].

Conversely, in over-exposed (high-brightness) conditions, such as direct sunlight, high-glare environments, or reflective surfaces, images suffer from:

- Saturation and highlight clipping – Overexposed regions lose all detail, turning into bright white patches where information is irreversibly lost [14].
- Dynamic range limitations – Standard imaging sensors struggle to balance bright and dark areas simultaneously [55].
- Lens flare and reflection artifacts – Harsh lighting can introduce unwanted reflections and flares, further degrading the image quality [44].
- Color fading and unnatural tones – High brightness levels wash out colors, resulting in low contrast and loss of vibrancy [62].

These problems are not merely aesthetic concerns—they have severe implications for downstream computer vision tasks. Many automated vision systems rely on high-quality

images to extract meaningful features for decision-making, and degraded images can severely impact the performance of:

- Object Detection and Recognition – Poor lighting or compression artifacts can make it difficult for AI models to distinguish objects from the background, especially in surveillance, security, and retail applications [25].
- Tracking and Scene Understanding – Motion blur and extreme lighting conditions can disrupt object tracking algorithms, affecting real-time applications such as autonomous driving and robotics [58].
- Medical and Healthcare Imaging – In radiology, endoscopy, and microscopy, incorrect exposure or compression artifacts can lead to misinterpretation of critical details [65].
- Digital Photography and Computational Aesthetics – Over/underexposed images fail to capture the intended artistic vision, requiring extensive post-processing to restore their quality [18].

Given these widespread challenges, the development of robust image enhancement and exposure correction techniques is essential. Modern image restoration methods, including deep learning-based approaches [59, 4], physics-driven modeling [12], and hybrid enhancement frameworks, are continuously evolving to address these issues. The goal is to restore natural-looking images that preserve details, maintain accurate colors, and optimize visibility for both human observers and computer vision algorithms.

1.1 Challenges in Exposure Correction

Correcting exposure and enhancing low-light images is inherently an ill-posed problem, as different lighting conditions introduce various types of degradations that require context-aware restoration strategies. We next describe the key challenges.

Visibility Degradation and Color Distortion. Low-light images often suffer from intensity attenuation and color distortion due to sensor limitations [52]. Conventional enhancement methods tend to over-amplify certain color channels, leading to unnatural tones

and chromatic shifts. Over-exposed images, on the other hand, may lose important texture details, especially in highlighted or saturated regions, making it difficult to recover lost information.

Noise Amplification and Detail Preservation. Under-exposed images captured in low-light conditions frequently contain photon noise, sensor noise, and compression artifacts [40], which get amplified during enhancement if not handled properly. Some methods, particularly aggressive contrast enhancement approaches, tend to smooth out fine textures, making images appear overly artificial [29].

Limitations of Traditional Methods. Traditional approaches such as histogram equalization [11] can improve contrast but often introduce artifacts and do not consider local exposure variations. Gamma correction works well in some cases but assumes a fixed non-linearity, which does not adapt to different exposure conditions. Retinex-based methods [20, 12], inspired by human visual perception, decompose an image into reflectance and illumination components, but often fail to handle noise and color distortions in complex scenes.

1.2 Compression Artifacts

Mitigating compression artifacts has been an active area of research. Traditional methods, including filtering [51], dictionary learning [34], and sparse coding [8], often lack robustness and generalizability. The advent of deep learning has revolutionized artifact removal, with early CNN-based approaches like ARCNN [7] and DnCNN [60] demonstrating notable improvements in restoring compressed images. More recently, generative models and diffusion-based methods have emerged as powerful tools for image restoration. Diffusion models, originally designed for high-fidelity image generation, have shown promise in inverse problems like denoising, super-resolution, and artifact removal [45, 43]. However, most existing diffusion models rely on multi-step sampling, leading to significant computational overhead and making them unsuitable for real-time or resource-constrained settings.

To overcome these limitations, Guo et al. propose CODiff, a compression-aware one-step diffusion model for JPEG artifact removal [13]. CODiff introduces a Compression-

aware Visual Embedder (CaVE) that integrates JPEG quantization priors into a powerful feature extractor. This feature representation is used to condition the one-step denoising process, drastically reducing inference time. Furthermore, CODiff employs a dual learning strategy combining explicit quality prediction and implicit reconstruction objectives. This dual guidance enables the model to better understand compression levels and recover visually pleasing images even under heavy degradation.

In parallel, other architectures such as Swin2SR [26, 3] leverage transformer-based backbones to capture long-range dependencies and achieve state-of-the-art results in compressed image super-resolution. Similarly, DAGN [63] explores the synergy between compression-sensitive and compression-invariant features to enhance restoration. Despite these advancements, achieving high-quality restoration across diverse compression levels while maintaining real-time efficiency remains an open problem. In this work, we aim to address these challenges using our ControlNet diffusion model.

1.3 Scope of This Work

In this study, we present a novel framework for image restoration, ReGIE, particularly low-light enhancement and image compression artifacts mitigation that combines:

Retinex Theory: This works on the theory of decomposing an image into its illumination and reflectance components. This helps in better handling exposure variations while minimizing artifacts and noise.

Diffusion-Based Restoration Mechanisms: For realistic exposure enhancement through iterative refinements and better generalizability, leveraging diffusion priors obtained from millions of real-world images.

This work aims to correct under-exposed images in a unified framework, suppressing noise without sacrificing detail richness. Using a Retinex model, we can ensure real-time efficiency, making the method practical for autonomous driving, medical imaging, and mobile photography applications, i.e., deployment-friendly, while using a diffusion model helps generalize to varying environments. By addressing the fundamental challenges and leveraging recent advancements, we seek to push the boundaries of automated exposure correction and compression artifact removal, making it more adaptive, efficient, and gener-

alizable to diverse real-world conditions.

Chapter 2

Related Work

Recent advances in image enhancement tackle challenges in both low-light and compression-degraded scenarios. Methods like Zero-DCE [12] and LLFlow [49] enhance low-light images without paired data, while transformer-based models such as Retinexformer [2] improve structure preservation. For compression artifact removal, approaches like SwinIR [26] and DAGN [63] leverage attention mechanisms and dual-path learning. CODiff [13] further introduces a one-step diffusion model with JPEG priors for fast and high-fidelity restoration.

2.1 Low-Light Image Enhancement

Low-light image enhancement (LLIE) is a longstanding problem in computer vision, aiming to improve the visibility and perceptual quality of images captured under suboptimal illumination. Traditional approaches rely on hand-crafted priors and physical imaging models. Retinex theory [22] forms the foundation for many classical methods, where an image is decomposed into reflectance and illumination components to simulate human visual perception. Numerous variants [20, 12, 52] have been proposed to address illumination estimation and reflectance preservation, although they often struggle with noise amplification and color distortion in extremely dark regions.

Learning-based methods, particularly those driven by convolutional neural networks (CNNs), have gained prominence in recent years. Supervised models such as EnlightenGAN [19], KinD [62], and Zero-DCE [12] have shown substantial improvements by learn-

ing illumination correction directly from paired or unpaired data. However, these models often require large-scale annotated datasets, and their ability to generalize to real-world low-light conditions remains limited.

2.2 Reference-Based Image Restoration

Reference-based image restoration introduces auxiliary images as guidance to improve the restoration of a degraded target. These auxiliary inputs—referred to as reference images—can provide structural, contextual, or appearance-based cues that are otherwise absent in the corrupted image. This strategy has been explored in super-resolution [64, 57], denoising [9], and video frame interpolation [35]. In low-light settings, reference images can serve as semantic anchors, compensating for the severe loss of detail, texture, and illumination cues present in the target image.

Unlike multi-frame or stereo-based enhancement techniques, reference-based restoration does not necessarily assume temporal continuity or geometric alignment. Instead, the goal is to exploit visual similarity and semantic overlap between the target and reference. Approaches like SGZ [54], SMORE [33], and TTSR [56] employ reference encoders to extract transferable appearance features, which are then fused with the target's structural information using attention or correlation modules. Our method builds on this idea but leverages a diffusion-based framework with explicit control pathways to disentangle structure and appearance through structured priors and reference image embeddings.

2.3 Diffusion Models for Image Restoration

Denoising diffusion probabilistic models (DDPMs) [16, 46] have emerged as a powerful class of generative models capable of synthesizing high-quality images through iterative denoising of Gaussian noise. Recent advancements have extended their application to image-to-image translation tasks, such as super-resolution [43], inpainting [31], and blind restoration [24]. The core advantage of diffusion models lies in their capacity to model complex image distributions without adversarial training, while naturally supporting iterative refinement.

However, vanilla diffusion models are often slow to converge and are not inherently guided by spatial priors. Conditional variants, such as guided diffusion [6], classifier-free guidance [17], and latent diffusion models (LDMs) [41], address this by incorporating conditioning signals into the reverse process. In particular, ControlNet [61] has introduced a flexible and scalable framework for injecting spatial control via zero-initialized convolutional branches. This allows pretrained diffusion models to respond to structure-preserving cues like edge maps, depth maps, or semantic masks. Our work adapts ControlNet to condition on Retinex priors and reference image features, enabling spatial and appearance-aware low-light enhancement.

2.4 High-Frequency Preservation in Restoration

A major challenge in image restoration, especially under strong degradations such as low-light conditions, is the recovery of high-frequency details. These include edges, textures, and fine structures that are often lost during encoding or diffusion. Previous works have attempted to address this through perceptual losses [21], adversarial training [23], and frequency-domain constraints [53]. In the context of VAEs and diffusion models, VAE-induced blurring is a common issue due to the smooth latent representations.

To mitigate this, recent studies have proposed injecting high-pass filtered components, using Discrete Fourier Transform (DFT) [36], Wavelet transforms [27], or edge-based regularizers such as Sobel operators [38]. Our method integrates high-frequency fidelity constraints that operate in both the frequency and gradient domains. These constraints are applied at inference time to encourage structural consistency between the output and the ground truth, even in fine-grained detail.

In contrast to prior works, our approach combines Retinex theory, reference-based conditioning, and high-frequency constraints within a unified ControlNet-based diffusion framework. By doing so, we enable structure-aware, appearance-guided enhancement that is especially effective in extreme low-light conditions. Our design facilitates controllable, high-fidelity restoration while maintaining flexibility across varying lighting and scene complexities.

Chapter 3

Approach

We approach the problem by formulating it as a reference-based image restoration task. In this setting, the goal is to recover a high-quality target image from a degraded or noisy input, using an additional reference image that provides complementary information. The reference image typically shares semantic or structural similarity with the target, such as being from the same scene, subject, or identity, but captured under different conditions (e.g., pose, lighting, or resolution). By leveraging the correlation between the input and the reference, the model can more accurately infer missing details, reduce artifacts, and improve the fidelity of the restored output. This formulation allows us to exploit both low-level pixel-wise cues and high-level contextual information from the reference, leading to more robust and visually coherent restoration performance.

Our work specifically focuses on low-light and compression artifact images enhancement within the reference-based image restoration framework. Low-light conditions often lead to images with poor visibility, elevated noise levels, and significant loss of detail, particularly in darker regions. By leveraging a well-lit reference image that shares semantic or structural similarity with the low-light input, our method aims to guide the enhancement process more effectively. Leveraging a multi-view scene setup, the reference view provides additional context to compensate for missing information, helping the model to recover fine textures, correct color and brightness distortions, and suppress noise. This targeted formulation allows us to address the inherent ambiguity in low-light scenes by grounding the restoration in real, high-quality visual cues drawn from the reference image.

To tackle the task of low-light and compression artifact image enhancement, our pro-

posed framework, ReGIE, introduces several key innovations:

- We incorporate structural cues derived from Retinex decomposition to guide the enhancement process, enabling effective recovery of global illumination in severely underexposed images by explicitly modeling noise.
- By leveraging semantic and stylistic information from reference images, the model restores fine-grained details that are typically lost in degraded conditions.
- We build on a ControlNet-augmented diffusion architecture to enable precise, spatially aligned conditioning at multiple scales throughout the generative process, allowing for disentangled and controllable image synthesis.
- To suppress VAE-induced blurring and preserve edge sharpness, we introduce a loss term that combines Discrete Fourier Transform (DFT) and Sobel filtering in the frequency domain.

Together, these components form a robust enhancement pipeline capable of addressing both global illumination recovery and detail preservation in challenging low-light and compression artifact-affected scenes.

3.1 Preliminaries

Diffusion models. are a class of generative models that rely on a two-phase process: a *forward (diffusion)* process and a *reverse (denoising)* process. These models are inspired by nonequilibrium thermodynamics and are capable of generating complex data, such as images, by learning to gradually corrupt and then reconstruct data.

In the forward process, also known as the noising process, a clean input image x_0 is progressively corrupted by adding Gaussian noise over a series of discrete time steps $t \in \{0, 1, \ldots, T\}$. At each step, the input is transformed into a noisier version x_t according to the following equation:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon_t$$

Here, $\bar{\alpha}_t$ is a variance scheduling parameter that determines the proportion of the original signal preserved at time step t, and $\epsilon_t \sim \mathcal{N}(0, I)$ represents isotropic Gaussian noise. As t

increases, the signal degrades progressively until x_T is approximately pure Gaussian noise.

The reverse process aims to invert the corruption and reconstruct the clean image x_0 starting from the noisy image x_T. This is achieved by learning a denoising function parameterized by a neural network ϵ_θ, which estimates the noise component ϵ_t added at each step.

The model is trained to predict ϵ_t using a noise prediction objective. A commonly used loss function is the mean squared error (MSE) between the true noise and the predicted noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,t,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

By minimizing this objective, the network learns to denoise step-by-step, enabling the generation of high-quality samples during inference by reversing the diffusion trajectory from pure Gaussian noise back to a structured image.

Inputs Assumption. We aim to tackle the task of restoring images degraded due to under-saturation, with a specific focus on correcting exposure in low-light conditions. To address this, we assume access to ground-truth (GT) paired training data consisting of a clean, well-exposed image $I_{\text{gt}} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding degraded, low-light version $I_a \in \mathbb{R}^{H\times W\times 3}$. The artifact-affected image I_a exhibits significant under-saturation due to low illumination conditions at capture time, while I_{gt} serves as the target of the same image captured under ideal lighting. We also assume the availability of a corresponding reference image, $I_r \in \mathbb{R}^{H\times W\times 3}$, which provides complementary visual information to guide the restoration process. The reference image is assumed to be semantically or structurally related to the target scene, potentially sharing viewpoint, content, or context, but may differ in appearance due to variations in pose, lighting, or other factors.

While the paired supervision allows the model to learn a direct mapping between low-light and well-lit imagery, we provide the restoration network I_r to capture the underlying semantics, textures, and illumination patterns necessary for effective enhancement at a scene level. Additionally, this setting enables the network to disentangle true scene content from noise and lighting artifacts, thus enabling more faithful restoration.

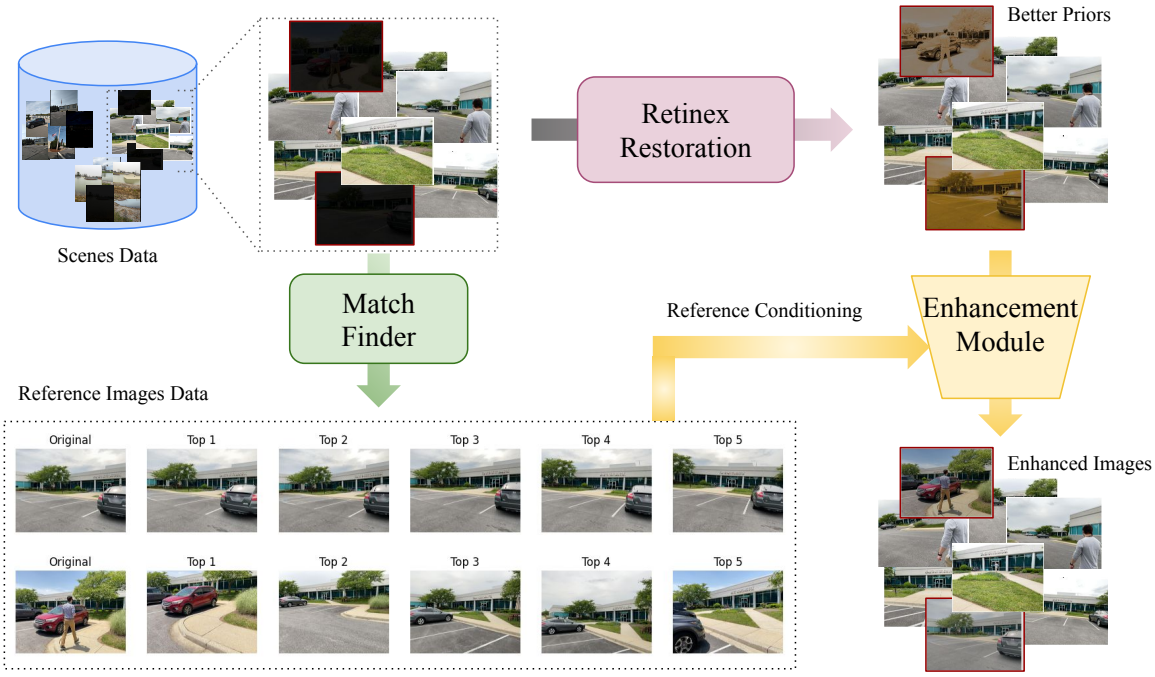

Figure 3.1: Overall pipeline of our proposed approach for low-light image enhancement.

3.2 Overall Method

Our reference-based image restoration framework, ReGIE, operates under the assumption that, for each training sample, we have access to an artifact-degraded image along with its corresponding clean target and a reference image. However, rather than requiring manual collection of reference images for each artifact image, we leverage *multi-view scene data*, which inherently contains multiple perspectives and lighting variations of the same scene. This allows us to extract suitable reference-clean pairs directly from the existing dataset, significantly reducing the need for additional reference image acquisition. The selected reference image shares semantic or structural similarity with the target image, offering complementary information for effective restoration.

To provide a strong inductive bias for the downstream generative model, we preprocess the degraded (low-light) artifact images using a *Retinex-based decomposition module*. This module enhances visibility by separating the illumination and reflectance components of the image, effectively suppressing degradation introduced by low-light conditions. The enhanced output serves as a refined prior input to the diffusion model, emphasizing critical

structural and content features in the corrupted image.

Our core generative model builds upon a *diffusion architecture with a ControlNet backbone*, enabling conditional generation guided by both structural cues and reference-based appearance information. The model takes as input the artifact image I_a, the corresponding clean target I_{gt}, and the reference image I_r, and generates a restored output $\hat{I}_a$. The network is trained to preserve the structural and spatial layout of the artifact image while borrowing appearance features such as texture, color, and tone from the reference image, resulting in perceptually coherent and visually pleasing restorations.

To further improve reconstruction quality, we address the common issue of *VAE-induced blurriness* in diffusion models, which often results in the loss of fine details. We incorporate a Fourier transform-based module to retain high-frequency information during the generative process. This frequency-domain representation helps to enhance edge sharpness and texture fidelity by fusing back fine-grained spectral components into the final output, thus mitigating the smoothing effects of latent compression and producing sharper, more detailed images. We show the overall approach in Fig. 3.1.

Since our proposed method remains the same for low-light image enhancement and compression artifact removal, we simply describe each of the components forming the proposed low-light image enhancement approach in detail.

3.3 Retinex-based Exposure Correction

According to the Retinex theory, an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ can be decomposed into a reflectance image $\mathbf{R} \in \mathbb{R}^{H\times W\times 3}$ and an illumination map $\mathbf{L} \in \mathbb{R}^{H\times W}$ as

$$\mathbf{I} = \mathbf{R} \odot \mathbf{L}, \tag{3.1}$$

where $\odot$ denotes the element-wise multiplication. This Retinex model assumes $\mathbf{I}$ is corruption-free, which is inconsistent with the real under-exposed or over-exposed scenes. We identify that the corruptions predominantly arise from two sources. First, the use of high-ISO and long-exposure settings in dark scenes naturally leads to the introduction of noise and artifacts. Second, the process of brightening these images may tend to not only amplify existing noise and artifacts but may also result in incorrectly exposed images and color distortions.

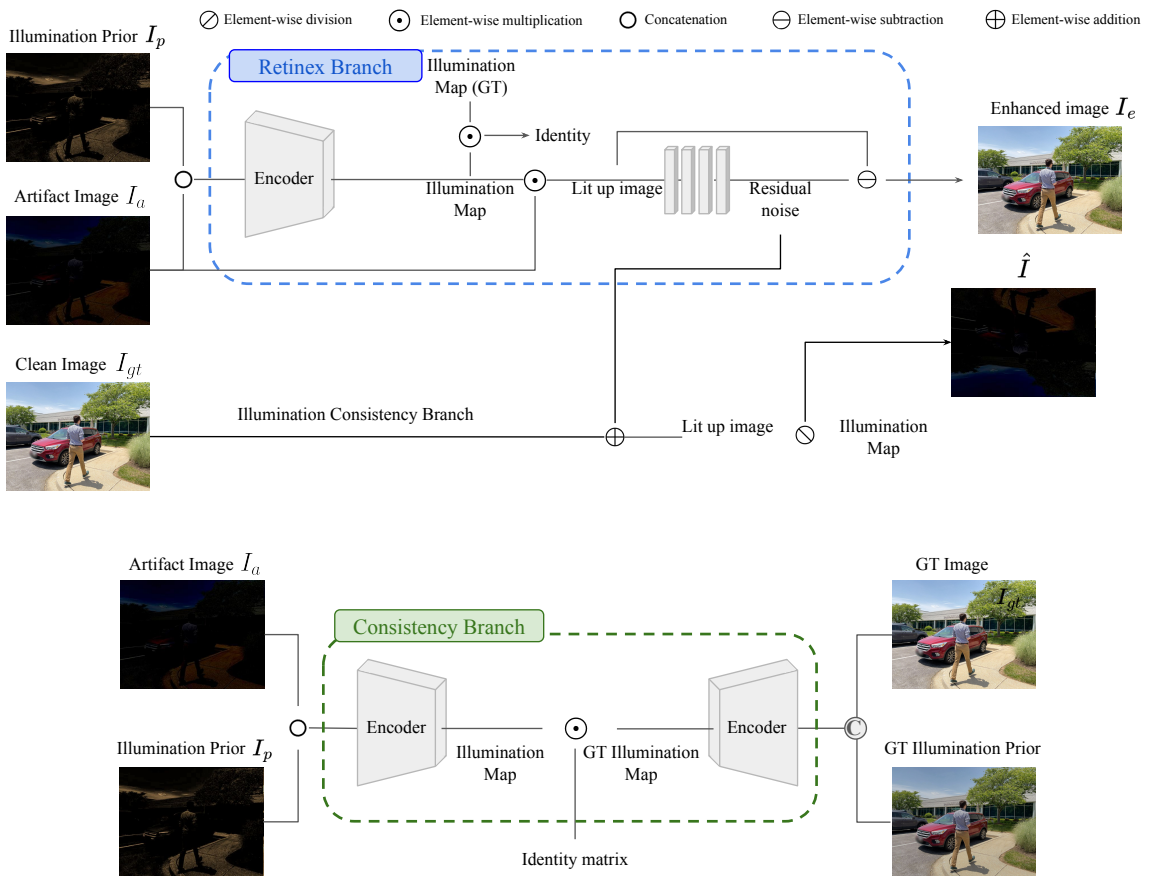

Figure 3.2: Retinex module to process highly undersaturated images and provide a strong inductive bias for the generative model.

To account for the corruptions, the following equation models the perturbations in $\mathbf{R}$ and $\mathbf{L}$:

$$\begin{aligned}\mathbf{I} &= (\mathbf{R} + \hat{\mathbf{R}}) \odot (\mathbf{L} + \hat{\mathbf{L}}) \\ &= \mathbf{R} \odot \mathbf{L} + \mathbf{R} \odot \hat{\mathbf{L}} + \hat{\mathbf{R}} \odot (\mathbf{L} + \hat{\mathbf{L}}),\end{aligned} \tag{3.2}$$

where $\hat{\mathbf{R}} \in \mathbb{R}^{H \times W \times 3}$ and $\hat{\mathbf{L}} \in \mathbb{R}^{H \times W}$ denote the perturbations. Similar to [2], we regard $\mathbf{R}$ as a well-exposed image. To light up $\mathbf{I}$, we element-wise multiply the two sides of Eq. (3.2) by a light-up map $\bar{\mathbf{L}}$ such that $\bar{\mathbf{L}} \odot \mathbf{L} = \mathbf{1}$ as

$$\mathbf{I} \odot \bar{\mathbf{L}} = \mathbf{R} + \mathbf{R} \odot (\hat{\mathbf{L}} \odot \bar{\mathbf{L}}) + (\hat{\mathbf{R}} \odot (\mathbf{L} + \hat{\mathbf{L}})) \odot \bar{\mathbf{L}}, \tag{3.3}$$

where $\hat{\mathbf{R}} \odot (\mathbf{L} + \hat{\mathbf{L}})$ represents the noise and artifacts hidden in the dark scenes and are amplified by $\bar{\mathbf{L}}$. $\mathbf{R} \odot (\hat{\mathbf{L}} \odot \bar{\mathbf{L}})$ indicates the under-/over-exposure and color distortion caused

by the light-up process. We simplify Eq. (3.3) as

$$\mathbf{I}_{lu} = \mathbf{I} \odot \bar{\mathbf{L}} = \mathbf{R} + \mathbf{C}, \tag{3.4}$$

where $\mathbf{I}_{lu} \in \mathbb{R}^{H \times W \times 3}$ represents the lit-up image and $\mathbf{C} \in \mathbb{R}^{H \times W \times 3}$ indicates the overall corruption term.

We show a simplified diagram of Retinexformer in Fig. 3.2. We then formulate the exposure correction loss function between the retinex-mitigated image I_e and the ground truth clean image I_{gt} as

$$\mathcal{L}_{ec}(I_{gt}, I_e) = \frac{1}{N} \sum_{i=1}^{N} (I_{gt}^i - I_e^i)^2 \tag{3.5}$$

3.4 Reference dataset collection

To construct our reference-based image restoration dataset, we leverage multi-view scene data that inherently captures the same scene from various perspectives and under varying lighting conditions. For each scene, we iterate over individual images and apply augmentation techniques to simulate underexposure or degradation, resulting in the artifact image, I_a. A clean image, I_{gt} of the same view, is selected to serve as the ground-truth target, and an additional reference image, I_r is identified from within the same scene based on semantic or structural similarity. This process yields a triplet consisting of the clean image, its augmented version - the artifact image, and the reference image. To enhance reference selection, we perform a similarity matching step, ranking candidate reference images using a matching function and selecting the top-k most relevant ones (e.g., rank-1, rank-2, rank-5). Furthermore, we evaluate the impact of different downscaling factors (e.g., 0.5, 0.25, 0.125, 0.0625) on both the runtime and accuracy of similarity estimation. This allows us to assess the trade-off between computational efficiency and matching performance, guiding the choice of scale factor for large-scale data generation. The resulting dataset provides rich, diverse triplets for training and evaluating reference-guided image enhancement models. The algorithm is shown in Algorithm 1.

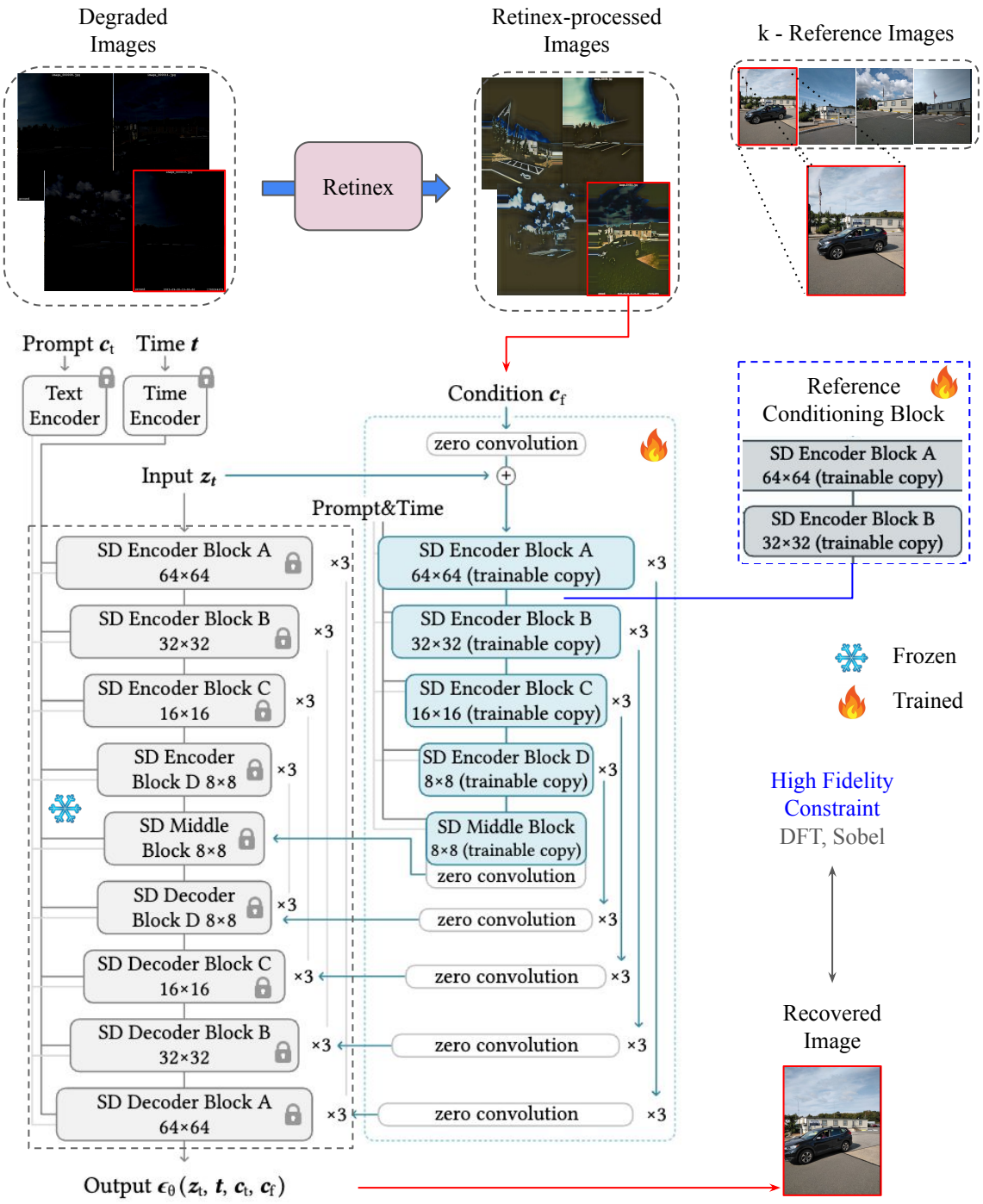

Figure 3.3: ControlNet with Retinex-based pre-processing for reference-based image restoration. A diffusion model with induced priors on exposure correction produces the corresponding mitigated image.

Algorithm 1: Reference Triplet Dataset Construction

Require: Multi-view scene dataset $\mathcal{S}$, similarity matching function $\mathcal{M}$, scale factors $\mathcal{F} = \{0.5, 0.25, 0.125, 0.0625\}$

Ensure: Triplet dataset $\mathcal{D}$ of form $\{(I_{a\_i}, I_{\text{gt\_i}}, I_{r\_i})\}$

for each scene $s \in \mathcal{S}$ **do**
  for each image $I_{gt\_i} \in s$[images] **do**
    Apply data augmentation to I_{gt_i} to obtain $I_{\text{a\_i}}$
    Retrieve a clean, similar-view image I_r from the same scene
    Append triplet $(I_{gt\_i}, I_{\text{a\_i}}, I_{r\_i})$ to dataset $\mathcal{D}$
  end for
end for

// Similarity Matching (Triplet Selection)
for each image I_i in scene s **do**
  Compute similarity scores $\mathcal{M}(I_i, I_j)$ for all $I_j \in s$, where $j \neq i$
  Rank the images by similarity score
  Select top-k matches (e.g., rank-1, rank-2, rank-5) as candidate references I_r
end for

// Scale Factor Evaluation (Optional for Trade-off Analysis)
for each scale factor $f \in \mathcal{F}$ **do**
  Downscale images by factor f
  Evaluate similarity matching performance and runtime
  Record trade-off between speed and accuracy
end for
return Triplet dataset $\mathcal{D}$ and optional similarity-ranking metrics

3.5 Diffusion Enhancement Module

At the core of our restoration framework is a diffusion-based generative model enhanced with a ControlNet backbone, enabling conditional image generation guided by multiple structured inputs. Unlike conventional diffusion models that are conditioned solely

on time-step and textual prompts, ControlNet introduces auxiliary conditioning branches, allowing the integration of spatial priors and external appearance cues. This design significantly improves generation quality, especially under severe degradation scenarios such as low-light imaging. We design it to incorporate both structural priors and reference-based appearance cues into the denoising process. The objective is to reconstruct a high-quality image $\hat{I}_a$ from a degraded input I_a, leveraging a corresponding clean target I_{gt} to predict the noisa and a reference image I_r as guidance.

At each diffusion step t, the model receives a noisy latent variable z_t, which is obtained by perturbing I_a through a forward diffusion process. The generation is conditioned on three additional signals. A timestep and prompt embedding c_t, typically encoded via sinusoidal or learned positional embeddings is the usual signal given in the controller architecture. A structural prior $c_l = f_{\text{retinex}}(I_a)$ extracted from the low-light input using a Retinex decomposition module captures illumination and edge-aware features. Finally, an appearance prior $c_r = f_{\text{ref}}(I_r)$ is computed by encoding the reference image I_r through a dedicated, trainable encoder. As we want the diffusion model to focus on collecting the appearance information from the reference image, we keep the text prompts null.

These conditioning signals are injected into the diffusion U-Net using zero-convolution layers within the ControlNet architecture, allowing for precise spatial alignment without disrupting the pretrained weights of the diffusion backbone. Specifically, the noise prediction network ϵ_θ operates as:

$$\epsilon_\theta(z_t, t, c_t, c_l, c_r),$$

where z_t is denoised with the guidance of both c_l and c_r. The Retinex prior c_l biases the generation toward preserving structural fidelity in $\hat{I}_a$, while the reference embedding c_r contributes high-level appearance attributes such as color, tone, and texture.

The ControlNet backbone follows a U-Net structure with encoder and decoder blocks at multiple spatial resolutions (e.g., 64×64, 32×32). Each resolution stage processes the latent input z_t and integrates the conditional features via feature-wise addition and normalization. The encoded latent is then progressively decoded to reconstruct the denoised image. The model is trained using the standard noise prediction objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{I_a, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_t, c_l, c_r)\|_2^2 \right].$$

Through this multi-branch conditioning strategy, the ControlNet-enhanced diffusion model

learns to synthesize structurally consistent and visually plausible restorations, grounded in both the corrupted input I_a and the semantically aligned reference I_r.

3.6 High-frequency Preservation

In image restoration and generation tasks, preserving fine-grained semantic and spatial details is critical for producing perceptually convincing outputs. However, the standard VAE-based reconstructions often fail to retain high-frequency components such as textures, edges, and especially text, which are crucial for semantic fidelity. To address this limitation, we explicitly enforce high-frequency consistency between the generated image and its ground truth counterpart, we introduce a fidelity constraint that compares their responses to high-pass filtering operations. We utilize two complementary operators: a frequency-domain operator $\mathcal{F}(\cdot)$, implemented via the Discrete Fourier Transform (DFT) [36], and an edge-domain operator $\mathcal{S}(\cdot)$, implemented via the Sobel filter [38]. The Fourier-based component transitions an image into the frequency domain, isolates high-frequency components using a high-pass filter, and reconstructs them using the inverse transform. Simultaneously, the Sobel operator captures prominent gradient information to highlight edge structures.

Given a ground truth image I_{gt} and a generated image $D_\theta(\mathbf{z}_{t\to 0})$ decoded from the estimated latent representation at time step t, the fidelity loss $\mathcal{L}_f$ is computed as:

$$\begin{aligned}\mathcal{L}_f(I_{\text{gt}}, D_\theta(\mathbf{z}_{t\to 0})) = \|\mathcal{F}(I_{\text{gt}}) - \mathcal{F}(D_\theta(\mathbf{z}_{t\to 0}))\|_2^2 \\ + \|\mathcal{S}(I_{\text{gt}}) - \mathcal{S}(D_\theta(\mathbf{z}_{t\to 0}))\|_2^2\end{aligned} \tag{3.6}$$

To obtain the latent $\mathbf{z}_{t\to 0}$, we use the predicted noise ϵ from the diffusion model H to perform deterministic denoising from step t to step 0.

The final training objective is a weighted combination:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}},$$

where we have expanded $\mathcal{L}_f = \mathcal{L}_{\text{freq}} + \mathcal{L}_{\text{edge}}$ and λ_{freq} and λ_{edge} are hyperparameters controlling the contribution of the auxiliary losses.

Chapter 4

Experiments

In this section, we describe the dataset used for training and evaluation, baselines compared with, evaluation metrics, and implementation details.

4.1 Datasets and Evaluation Metrics

To evaluate our proposed reference-based image enhancement framework, ReGIE, for low-light and compression artifacts, we conduct experiments on the WRIVA challenge dataset [1]. The dataset is a multi-view image dataset with multiple scenes captured in different outdoor scenarios. The multi-view setup helps us formulate the image restoration problem as a reference-based image restoration problem. The dataset also features different cameras for capturing the scenes. The diversity in the outdoor environment makes the dataset challenging to train on. We use a subset of the dataset, using cameras A08 and M07. Each of the cameras captures several outdoor scenes. Since the dataset is originally clean, we simulate low-light conditions and compression artifacts using the albumentations library. The library provides different parameters that help us generate varying illumination conditions and levels of compression artifacts. We use the evaluation set present in the dataset. Since the evaluation set consists of images with different artifact types, we only choose the ones affected by low exposure and compression artifact scenarios.

For reference image construction, we first compute visual similarity using a traditional feature-matching method such as SIFT [30]. To improve robustness, we further refine the ranking using a learned similarity function based on CLIP features [39]. From this ranked

Figure 4.1: Qualitative comparison of our proposed approach with other low-light enhancement methods on the M07 set. Our method consistently shows better retention of structural details and color information. Best seen in color.

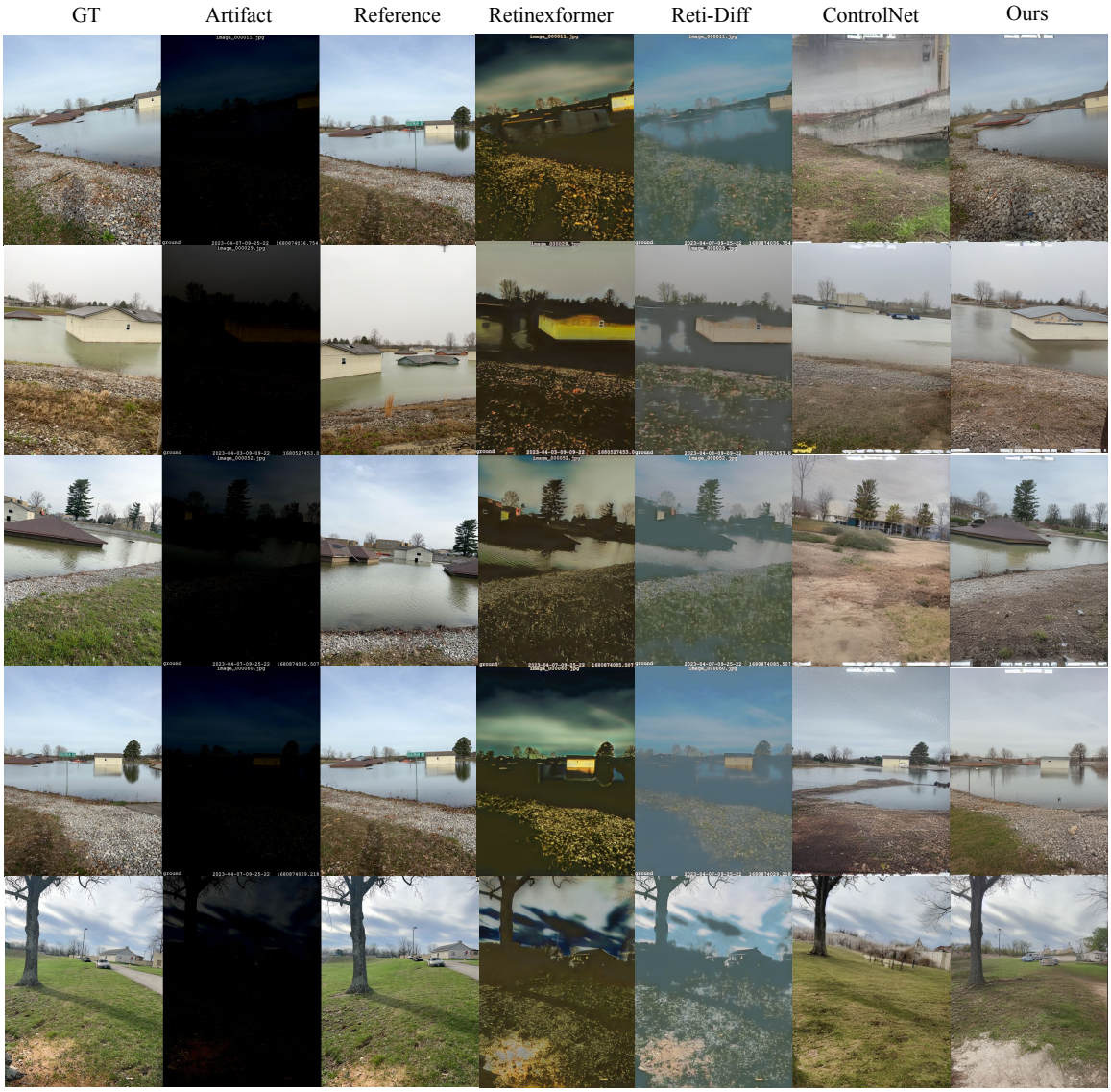

Figure 4.2: Qualitative comparison of our proposed approach with other low-light enhancement methods on the A08 set. Our method consistently shows better retention of structural details and color information, while Retinexformer and Reti-Diff show washed-out visual results. Best seen in color.

list, we select the top-k similar candidates and choose the most relevant reference image for each target image within a scene.

We follow [42] to employ widely adopted non-reference perceptual metrics DreamSim

score [10]. It bridges the gap between low-level metrics (e.g. LPIPS, PSNR, SSIM) and high-level measures (e.g. CLIP).

4.2 Qualitative Evaluation

Since we propose the first reference-based low-light and compression artifact image correction, we choose to compare our results with the baseline - ControlNet that takes multiple conditionings as input, i.e., the artifact image and the corresponding reference image. We also compare ReGIE with single-image low-light/restoration methods for a broader discussion about the advantages of using a reference image for image restoration.

Fig. 4.1 and Fig. 4.2 showcases visual comparisons with competing methods. Our approach, ReGIE, produces images with more natural illumination, better texture recovery, and fewer artifacts, especially in challenging dark regions. The reference conditioning helps recover fine details (e.g., text, signage) that other methods fail to restore, while the Retinex prior contributes to balanced exposure correction.

4.3 Quantitative Results

We compare our proposed method against state-of-the-art low-light enhancement methods, including Reti-Diff [15], Retinexformer [2], DiffPlugIn [28], and ControlNet [61].

Table 4.1: Quantitative comparison for low-light image enhancement on A08 subset.

Method	**PSNR↑**	**SSIM↑**	**DreamSim↓**
Retinexformer	11.23	0.31	0.69
DiffPlugIn	12.1	0.46	0.55
Reti-Diff	12.8	**0.54**	0.61
ControlNet	13.21	0.39	0.46
ReGIE (Ours)	**14.97**	0.49	**0.30**

Our method achieves the best performance across the DreamSim and PSNR metrics on both A08 and M07 sets, demonstrating both high fidelity and perceptual quality. We achieve the best performance on the A08 set for the SSIM metric and are close to Reti-Diff on the M07 set. We show the results in Table 4.1 and Table 4.2.

Table 4.2: Quantitative comparison for low-light image enhancement on M07 subset.

Method	**PSNR↑**	**SSIM↑**	**DreamSim↓**
Retinexformer	12.46	0.43	0.65
DiffPlugIn	13.56	0.51	0.57
Reti-Diff	13.32	**0.52**	0.60
ControlNet	14.20	0.41	0.37
ReGIE (Ours)	**16.91**	0.49	**0.17**

Table 4.3: Quantitative comparison for compression artifact removal on M07 subset.

Method	**PSNR↑**	**SSIM↑**	**DreamSim↓**
PromptIR	18.86	0.56	0.41
DA-CLIP	19.91	0.69	0.35
ControlNet	20.41	0.77	0.23
ReGIE (Ours)	**21.14**	**0.81**	**0.12**

We also compare our proposed method against state-of-the-art compression artifact enhancement methods, including DA-CLIP [32], PromptIR [37], and our own baseline ControlNet [61].

Our method achieves the best performance across the DreamSim and PSNR metrics on both A08 and M07 sets, demonstrating both high fidelity and perceptual quality. We achieve the best performance on the A08 set for the SSIM metric and on the M07 set. We show the results in Table 4.3 and Table 4.4.

4.4 Discussion

Our results highlight the effectiveness of combining Retinex-based illumination priors with reference image conditioning in a diffusion-based generative setting. The integration of ControlNet allows for precise injection of structural and semantic cues, enabling high-quality enhancement without relying on extensive fine-tuning or adversarial losses. Moreover, the frequency-domain fidelity constraint (DFT + Sobel) helps preserve edge sharpness and mitigates VAE-induced blurring typically observed in latent diffusion models. The reference-guided enhancement strategy proves particularly advantageous in scenes with fine details, where a purely image-to-image mapping is insufficient.

Table 4.4: Quantitative comparison for compression artifact removal on A08 subset.

Method	**PSNR↑**	**SSIM↑**	**DreamSim↓**
PromptIR	18.60	0.64	0.51
DA-CLIP	19.30	0.62	0.42
ControlNet	20.20	0.71	0.27
ReGIE (Ours)	**20.76**	**0.77**	**0.15**

Table 4.5: Ablation study on A08 subset for compressed image enhancement.

Method	**PSNR↑**	**SSIM↑**	**DreamSim↓**
ControlNet	20.20	0.71	0.27
Ours w/o Retinex priors	18.7	0.69	0.24
Ours w/o HFP	19.45	0.74	0.19
ReGIE (Ours)	**20.76**	**0.77**	**0.15**

In the context of image compression artifact mitigation, this framework also demonstrates strong potential. Compression artifacts often manifest as blockiness, ringing, and loss of high-frequency details—issues that conventional mitigation models struggle to restore due to their limited inductive bias toward spatial fidelity. By incorporating both frequency-domain constraints and reference-based conditioning, our model effectively recovers structural coherence and restores perceptual quality, even in highly degraded inputs. The illumination prior aids in decoupling compression-induced tonal distortions from semantic content, while ControlNet ensures spatial alignment with the reference image. Together, these components create a robust system capable of enhancing not only poorly illuminated images but also those compromised by aggressive compression.

4.4.1 Ablation Study

To better understand the contributions of each component in our framework, we conduct an ablation study by incrementally removing key modules and observing their impact on performance. We evaluate PSNR (↑), SSIM (↑), and DreamSim (↓) across various configurations. The results are summarized in Table 4.5.

Effect of Retinex priors. Removing Retinex-based illumination priors leads to a noticeable drop in PSNR and SSIM, suggesting their importance in providing global illumination cues and structural guidance. Despite a slight improvement in DreamSim, the overall

perceptual quality degrades, indicating the Retinex priors' role in enhancing photometric consistency.

Effect of high-frequency fidelity (HFP). Constraint Eliminating the frequency-domain fidelity loss (DFT + Sobel) results in reduced PSNR and perceptual sharpness, as reflected in the higher DreamSim score. This demonstrates that the HFP constraint is crucial for edge preservation and mitigating VAE-induced blurring artifacts.

Comparison with ControlNet baseline. While the ControlNet-only baseline benefits from reference-based spatial conditioning, it lacks both the illumination-aware Retinex guidance and the high-frequency constraint. Our full model outperforms the baseline by a significant margin across all metrics, validating the complementary benefits of our design components.

4.4.2 Reference Condition Block

To inject semantic and structural priors using reference conditioning, the model leverages a set of reference images k. These are processed through a Reference Conditioning Block, composed of trainable copies of the shallow Stable Diffusion (SD) encoder blocks (specifically the 64×64 and 32×32 resolutions). These blocks extract multi-scale features from the references and align them with the degraded inputs. The diffusion model backbone follows the U-Net architecture used in Stable Diffusion, consisting of multi-resolution encoder and decoder stages: The encoder comprises SD Encoder Blocks A through D, progressively reducing spatial dimensions (from 64×64 to 8×8). The bottleneck is handled by a frozen SD Middle Block (8×8). The decoder reconstructs the image through symmetrical blocks, culminating in the output at the original spatial resolution.

4.4.3 Comparison to ControlNet

The default ControlNet architecture takes in one conditional input, for example, a single degraded image. However, a straightforward extension to our reference-based problem setup is to have two ControlNets. We can then give the degraded image as one conditional signal and the reference image as another.

ReGIE differs with directly ControlNet in several points. First, we use a Retinex-based

image enhancement module as the pre-processing step. This decomposes the input image into illumination and reflectance components. The enhanced degraded images better retain structural details. Second, we modified the ControlNet architecture as we found that conditioning on all ControlNet blocks is inefficient. Instead of using two ControlNets, we use one ControlNet for the degraded image, and for conditioning the reference image, we use only the first block. This helps semantically map information from the reference image to the degraded image. Lastly, we apply a high-frequency preservation loss. We use FFT and Sobel operators to minimize high-frequency losses, as the output of diffusion models often fails to retain high-frequency information due to VAE compression.

Figure 4.3: Qualitative comparison of our proposed approach with other compression artifact removal methods and baselines on the A08 set. Our method consistently shows better retention of structural details and color information, while DA-CLIP struggles to remove the artifact as seen in the visual results. Best seen in color.

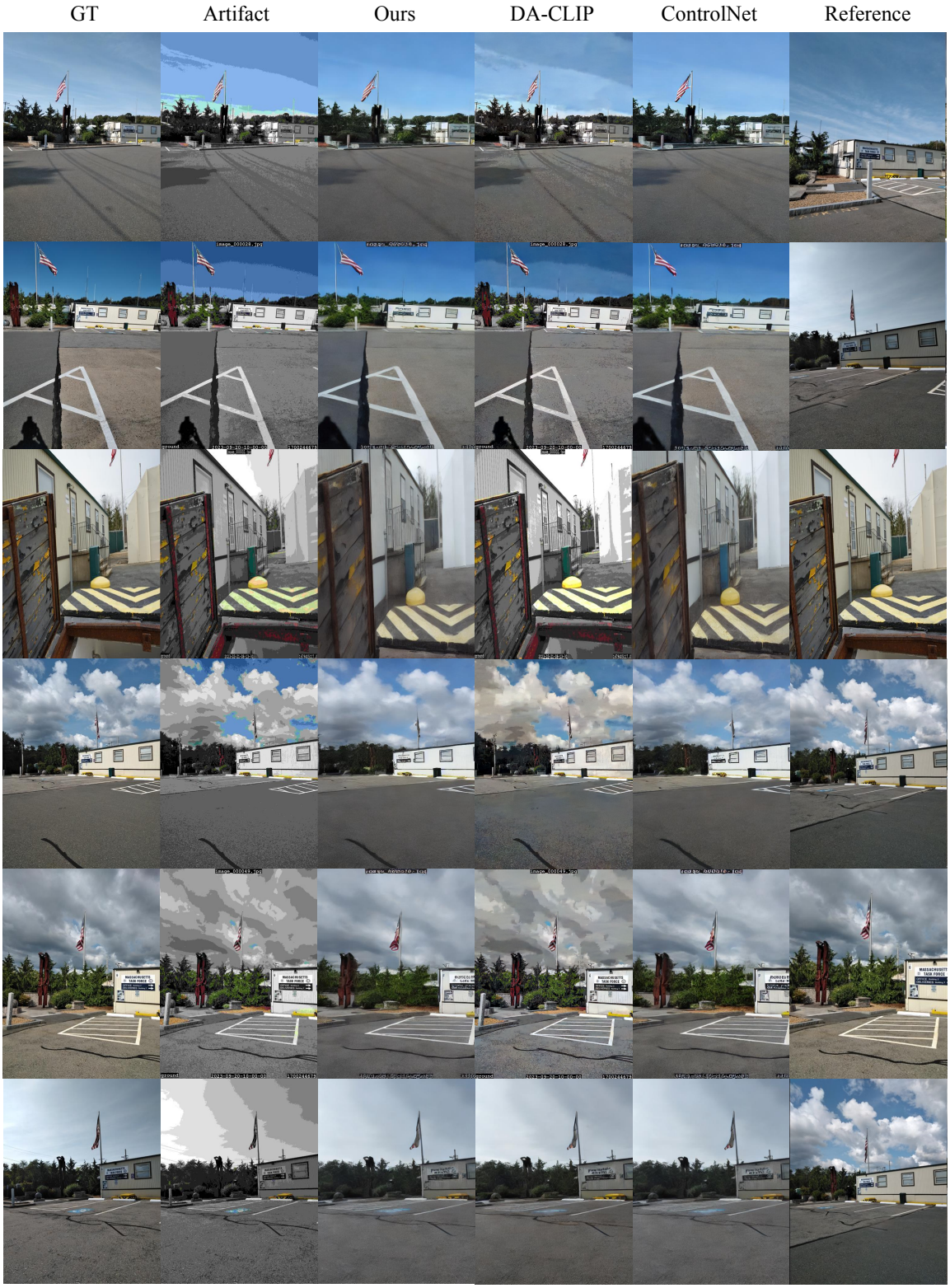

Figure 4.4: Qualitative comparison of our proposed approach with other compression artifact removal methods and baselines on the M07 set. Our method consistently shows better retention of structural details and color information, while DA-CLIP struggles to remove the artifact, as seen in the visual results. Best seen in color.

Chapter 5

Conclusion and Future Work

5.1 Conclusion and Future Work

In this paper, we proposed a framework, ReGIE, for reference-based low-light and compressed image enhancement that integrates Retinex-based illumination priors, reference-guided conditioning, and a ControlNet-based diffusion model. Our method is specifically designed to address the challenges of semantic degradation and detail loss in severely underexposed and compressed images. By leveraging the structural information extracted from Retinex decomposition, generative priors from a ControlNet-based diffusion model, and the semantic and stylistic cues provided by a reference image, the model is able to recover both global illumination and fine-grained details. The use of ControlNet enables precise and spatially aligned conditioning at multiple stages of the diffusion process, facilitating a disentangled and controllable generation pathway. Furthermore, we introduced a high-frequency fidelity constraint, incorporating both Discrete Fourier Transform and Sobel operators, to enhance edge sharpness and suppress the blurring artifacts typically introduced by VAE-based diffusion models.

Our experiments, conducted on a multi-view scene dataset, demonstrate that our approach consistently outperforms prior methods across a range of quantitative metrics, including PSNR, SSIM, and most importantly, DreamSim. Qualitative comparisons highlight the model's ability to produce perceptually pleasing results with enhanced texture, better exposure correction, and fewer artifacts.

Although our current model focuses on single-reference and fully supervised settings,

there are several promising directions for future work. One natural extension would be to support multi-reference fusion, where information from multiple retrieved images is adaptively combined to provide more diverse and comprehensive guidance. Additionally, moving beyond the need for paired supervision by exploring unpaired or semi-supervised training strategies—potentially leveraging generative adversarial losses or cycle-consistency objectives—would enhance the model's applicability to real-world, uncurated datasets.

Another important avenue is the extension of our framework to video sequences. Introducing temporal priors or motion-aware reference alignment mechanisms could support temporally consistent low-light video enhancement, which remains a challenging task. Efficiency is another critical aspect, and future work could investigate faster sampling techniques or model distillation methods to reduce the computational overhead associated with diffusion-based inference, making the model viable for real-time applications. Lastly, incorporating scene-aware retrieval strategies, such as those informed by depth estimation or semantic segmentation, could improve the relevance and alignment of selected reference images, further boosting restoration quality in complex environments.

Overall, this work demonstrates the potential of combining physics-based priors, reference-guided generation, and conditional diffusion modeling for high-quality restoration under extreme low-light conditions. We hope this framework lays the groundwork for future innovations at the intersection of generative modeling and extreme low-light and compressed image enhancement.

Bibliography

[1] M. Brown, M. Chan, and M. Twardowski. Wriva public data, 2024.

[2] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023.

[3] J. Cao, K. Zhang, R. Timofte, and L. V. Gool. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. *arXiv preprint arXiv:2209.11345*, 2022.

[4] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. *ACM Transactions on Graphics (TOG)*, 2018.

[5] A. L. Cunha, T. F. Chan, and J. Dong. Image restoration: A review of recent advances. *IEEE Signal Processing Magazine*, 2022.

[6] P. Dhariwal and A. Q. Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[7] C. Dong, Y. Deng, C. C. Loy, and X. Tang. Compression artifacts reduction by a deep convolutional network. In *IEEE International Conference on Computer Vision (ICCV)*, pages 576–584, 2015.

[8] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[9] H. Fan, Z. Zhang, and J. Xu. Briefly supervised learning for low-light image enhancement. In *CVPR*, 2019.

[10] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2024.

[11] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson Education, 2008.

[12] C. Guo, Y. Li, J. Guo, C. C. Loy, J. Hou, S. S. Kwong, and R. Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, 2020.

[13] J. Guo, Z. Chen, W. Li, Y. Guo, and Y. Zhang. Codiff: Compression-aware one-step diffusion model for artifact removal. *arXiv preprint arXiv:2502.09873*, 2025.

[14] S. W. Hasinoff, D. Sharlet, F. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, 2016.

[15] C. He, C. Fang, Y. Zhang, K. Li, L. Tang, C. You, F. Xiao, Z. Guo, and X. Li. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. In *ICLR*, 2025.

[16] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[17] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2022.

[18] A. Ignatov, L. Van Gool, and R. Timofte. Rendering natural camera raw images. *IEEE Access*, 2020.

[19] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, and P. Zhou. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 2021.

[20] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. In *IEEE Transactions on Image Processing*, 1997.

[21] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[22] E. H. Land. The retinex theory of color vision. *Scientific American*, 1977.

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[24] H. Li, Z. Zhang, B. Dai, and C. C. Loy. Blind image restoration via neural implicit priors. *arXiv preprint arXiv:2203.11844*, 2022.

[25] Y. Li, X. Sun, Q. Li, Y. Wang, G. Yu, and R. Jin. Lightweight and accurate face detection for mobile devices. In *CVPR*, 2019.

[26] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021.

[27] P. Liu, H. Zhang, B. He, K. Li, Z. Wang, and Z. Zhou. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 773–782, 2018.

[28] Y. Liu, Z. Ke, F. Liu, N. Zhao, and R. W. Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *CVPR*, 2024.

[29] K. G. Lore, A. Akintayo, and S. Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

[30] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[31] A. Lugmayr, M. Danelljan, and R. Timofte. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

[32] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön. Controlling vision-language models for universal image restoration. In *ICLR*, 2024.

[33] Z. Luo, Y. Lu, J. Ma, R. Timofte, and Z. Liu. Structure-consistent reference-based image super-resolution. In *ECCV*, 2022.

[34] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(4):791–804, 2012.

[35] S. Niklaus, L. Mai, and F. Liu. Softmax splatting for video frame interpolation. In *CVPR*, 2020.

[36] H. J. Nussbaumer. *Fast Fourier Transform and Convolution Algorithms*. Springer, 1982.

[37] V. Potlapalli, S. W. Zamir, S. Khan, and F. Khan. Promptir: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[38] W. K. Pratt. *Digital Image Processing: PIKS Scientific Inside*. John Wiley & Sons, 2007.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[40] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang. Lr-gan: Low-light image enhancement using real low-light images. *Pattern Recognition*, 2020.

[41] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[43] C. Saharia, W. Chan, S. Saxena, L. Li, D. J. Fleet, M. Norouzi, and T. Salimans. Image super-resolution via iterative refinement. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[44] X. Shen, J. Cao, L. J. Karam, Y.-H. Hsieh, and J.-B. Huang. Modeling flare for hdr imaging. In *ECCV*, 2020.

[45] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Dhariwal, M. Abbeel, and I. Sutskever. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[46] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[47] G. K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, 1992.

[48] Y. Wang, X. Gao, C. Shen, and M. Sun. Deep upe: A deep underexposed photo enhancer. In *Proceedings of the ACM Multimedia*, 2019.

[49] Y. Wang, Z. Huang, R. T. Tan, and L. Zhang. Low-light image enhancement via real low-light images. In *CVPR*, 2022.

[50] Y. Wang, Y. Lin, C. Dong, and C. C. Loy. Seeing in the dark with zero-shot implicit alignment. In *CVPR*, 2021.

[51] Z. Wang, A. Bovik, and B. L. Evans. Blind measurement of blocking artifacts in images. In *IEEE International Conference on Image Processing (ICIP)*, 2000.

[52] C. Wei, W. Wang, W. Yang, and J. Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018.

[53] H. Xu, W. Yang, Q. Qi, R. Xu, X. Wang, W. Xu, and J. Liu. Learning to restore low-light images via decomposition-and-enhancement. In *NeurIPS*, 2020.

[54] Q. Xu, C. Li, and D. Tao. Semantic-guided zero-shot image super-resolution. In *CVPR*, 2020.

[55] Q. Yan, D. Zhang, D. Wang, B. Dai, and D. Lin. Attention-guided network for low-light image enhancement. In *CVPR*, 2019.

[56] F. Yang, H. Liu, Q. Zhang, and H. Yang. Learning texture transformer network for image super-resolution. In *CVPR*, 2020.

[57] F. Yang, H. Liu, Q. Zhang, and H. Yang. Learning texture transformer network for image super-resolution. *TPAMI*, 2021.

[58] Z. Yu, T. Lin, C. Shen, and Y. Liu. Path aggregation network for instance segmentation. *Neurocomputing*, 2021.

[59] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

[60] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[61] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[62] Y. Zhang, J. Zhang, and C. Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM Multimedia*, 2019.

[63] X. Zhao, Y. Chen, J. Ren, and W. Zuo. Dagn: Dual awareness guidance network for jpeg artifact removal. *arXiv preprint arXiv:2405.09291*, 2024.

[64] H. Zheng, L. Zhang, and Y. Wang. Semi-supervised learning for image super-resolution via generative adversarial networks. In *CVPR*, 2018.

[65] Y. Zhou, Z. Xie, and Y. Liu. Models and techniques for image enhancement in medical imaging: A review. *IEEE Access*, 2021.