Unsupervised Illumination-Guided Wavelet Networks for Night-Driving Image Enhancement

Piyush Ahirao , Anish Pote , Tanishq Bharti , Aabha Chaudhari
Department of Computer Engineering
Pimpri Chinchwad College of Engineering ,Pune

Article Info

Keywords:

- Low-Light Image Enhancement
- Wavelet Transform
- Deep Learning
- Transformer Networks
- Image Restoration

Abstract

Low-light images pose significant challenges for high-level vision tasks, often leading to degraded performance in intelligent visual systems. To address this, we propose a Wavelet-based Enhancement Network (WENet) that effectively integrates Convolutional Neural Networks (CNNs) with wavelet transforms for improved low-light image enhancement. The wavelet transform decomposes images into multiple frequency components, enabling both global and local detail restoration. A Wavelet Calibration Layer (WCL) is introduced to convert image features into the wavelet domain and distribute them adaptively using calibration filters, thus enhancing fine details. To mitigate noise amplification commonly observed during wavelet learning, a Contrast Adjustment Layer (CAL) is employed to refine image contrast through adaptive shift operations.

Experimental evaluation on the Kaggle Low-Light Image Dataset demonstrates that the proposed WENet achieves superior enhancement quality and visual consistency compared to existing methods. The model further improves downstream vision tasks, such as object detection and segmentation, under low-light conditions. Overall, WENet provides a robust and efficient solution for low-light image enhancement with strong generalization capability across diverse real-world scenarios.

1. Introduction

While the proliferation of deep learning has led to an explosion in computer vision capabilities, this progress has been largely benchmarked on high-quality datasets captured in optimal lighting. This overlooks a common and critical failure point: performance under low-light conditions. Environments with poor or uneven illumination produce images with diminished contrast and visibility, where information in shadowed areas is often lost entirely. This degradation is a primary cause of failure for many computer vision systems in real-world applications [1,2]. The field of low-light image enhancement aims to solve this, not just for aesthetic improvement, but as a critical enabling technology for other AI tasks. By restoring image quality, enhancement provides a significant boost to high-level applications such as object detection [3–5] and semantic segmentation [6,7], thereby improving model robustness across numerous domains [8,9].

Early approaches to this problem can be divided into two main families: histogram-based [10–12] and Retinex-based methods [13–15]. Histogram-based techniques, such as histogram equalization [16,17], are straightforward methods that increase contrast by redistributing pixel intensity values, but their capacity for recovering intricate details is limited. The second family of methods is inspired by Retinex theory [18], which models an image S as a product of its illumination I and its reflectance R ($S=I\times R$). The objective is to estimate and remove the illumination component to reveal the object's inherent reflectance. However, this decomposition is an ill-posed problem and often yields unnatural results without additional priors [19,20]. These manually engineered constraints and parameters tend to lack generalizability, struggle with noise, and often involve computationally intensive calculations, making them unsuitable for many modern, real-time applications.

In recent years, the rise of deep learning has revolutionized the approach to low-light enhancement. The powerful feature extraction abilities of Convolutional Neural Networks (CNNs) have led to a new generation of LLIE models that achieve higher accuracy and speed than traditional methods [21–24]. By treating enhancement as an image-to-image translation task, these models can learn complex transformations without being constrained by pre-defined physical models, offering greater flexibility. These learning-based models generally follow one of three paradigms. Supervised learning [25,26] relies on datasets of paired low-light and ground-truth normal-light images to learn an enhancement mapping [27]. However, a common issue is model overfitting, as these datasets are often created by artificially adjusting exposure rather than capturing authentic scenes. In

response, unsupervised and zero-shot methods emerged. Prominent examples such as Enlighten GAN [28], Zero-DCE [29], SCI [30], and IAT [31] learn to enhance images without requiring paired data. While these models produce visually pleasing results by boosting brightness and contrast, they typically struggle with noise suppression and the recovery of fine details, thus providing limited gains for high-level downstream vision applications.

While many enhancement methods are implemented in the spatial domain [32,33], we argue that the wavelet domain offers a more effective foundation for low-light image processing. The success of wavelet-based deep learning in adjacent fields—such as removing moire patterns [34,35], face super-resolution [36], and deblurring [37]—supports this position. We are convinced that by decomposing an image into its frequency components, we can more effectively filter high-frequency noise while preserving essential low-frequency information. A key goal of this work is to apply this principle to significantly boost the performance of downstream semantic segmentation tasks, all within a computationally efficient, lightweight model suitable for real-world deployment.

To achieve these goals, this paper presents the Wavelet-based Enhancement Network (WENet), a new architecture combining CNNs and Transformers. The model first uses a Transformer block for preliminary feature extraction before moving to its core innovation: enhancement in the wavelet domain. We introduce a Wavelet Calibrate Layer (WCL), which converts features into wavelet coefficients and uses multiple calibration filters to perform regional correction, thereby restoring image details. After converting features back to the spatial domain, a Contrast Adjustment Layer (CAL), built with simple convolutions, further improves the image contrast. The effectiveness of WENet is demonstrated through advanced results on the LOL [27] and LOLv2 datasets and is further verified by its strong performance in a joint training setup for semantic segmentation on the ACDC [39] dataset. In summary, this paper makes several key contributions. We introduce the WENet architecture, which uniquely integrates wavelet transforms with CNNs and Transformers to precisely filter noise and recover information. We also propose two custom layers, the WCL and CAL, designed specifically for detail recovery and contrast adjustment. Finally, we provide extensive experimental validation to show that our method not only enhances low-light images effectively but also improves the performance of subsequent high-level vision tasks.

While the proliferation of deep learning has led to an explosion in computer vision capabilities, this progress has been largely benchmarked on high-quality datasets captured in optimal lighting. This overlooks a common and critical failure point: performance under low-light conditions. Environments with poor or uneven illumination produce images with diminished contrast and visibility, where information in shadowed areas is often lost entirely. This degradation is a primary cause of failure for many computer vision systems

in real-world applications [1,2]. The field of low-light image enhancement aims to solve this, not just for aesthetic improvement, but as a critical enabling technology for other AI tasks. By restoring image quality, enhancement provides a significant boost to high-level applications such as object detection [3–5] and semantic segmentation [6,7], thereby improving model robustness across numerous domains [8,9].

Early approaches to this problem can be divided into two main families: histogram-based [10–12] and Retinex-based methods [13–15]. Histogram-based techniques, such as histogram equalization [16,17], are straightforward methods that increase contrast by redistributing pixel intensity values, but their capacity for recovering intricate details is limited. The second family of methods is inspired by Retinex theory [18], which models an image S as a product of its illumination I and its reflectance R ($S=I\times R$). The objective is to estimate and remove the illumination component to reveal the object's inherent reflectance. However, this decomposition is an ill-posed problem and often yields unnatural results without additional priors [19,20]. These manually engineered constraints and parameters tend to lack generalizability, struggle with noise, and often involve computationally intensive calculations, making them unsuitable for many modern, real-time applications.

In recent years, the rise of deep learning has revolutionized the approach to low-light enhancement. The powerful feature extraction abilities of Convolutional Neural Networks (CNNs) have led to a new generation of LLIE models that achieve higher accuracy and speed than traditional methods [21–24]. By treating enhancement as an image-to-image translation task, these models can learn complex transformations without being constrained by pre-defined physical models, offering greater flexibility. These learning-based models generally follow one of three paradigms. Supervised learning [25,26] relies on datasets of paired low-light and ground-truth normal-light images to learn an enhancement mapping [27]. However, a common issue is model overfitting, as these datasets are often created by artificially adjusting exposure rather than capturing authentic scenes. In response, unsupervised and zero-shot methods emerged. Prominent examples such as EnlightenGAN [28], Zero-DCE [29], SCI [30], and IAT [31] learn to enhance images without requiring paired data. While these models produce visually pleasing results by boosting brightness and contrast, they typically struggle with noise suppression and the recovery of fine details, thus providing limited gains for high-level downstream vision applications.

While many enhancement methods are implemented in the spatial domain [32,33], we argue that the wavelet domain offers a more effective foundation for low-light image processing. The success of wavelet-based deep learning in adjacent fields—such as removing moire patterns [34,35], face super-resolution [36], and deblurring [37]—supports this position. We are convinced that by decomposing an image into its frequency components, we can more effectively filter high-frequency noise while preserving essential

low-frequency information. A key goal of this work is to apply this principle to significantly boost the performance of downstream semantic segmentation tasks, all within a computationally efficient, lightweight model suitable for real-world deployment.

To achieve these goals, this paper presents the Wavelet-based Enhancement Network (WENet), a new architecture combining CNNs and Transformers. The model first uses a Transformer block for preliminary feature extraction before moving to its core innovation: enhancement in the wavelet domain. We introduce a Wavelet Calibrate Layer (WCL), which converts features into wavelet coefficients and uses multiple calibration filters to perform regional correction, thereby restoring image details. After converting features back to the spatial domain, a Contrast Adjustment Layer (CAL), built with simple convolutions, further improves the image contrast. The effectiveness of WENet is demonstrated through advanced results on the LOL [27] and LOLv2 datasets and is further verified by its strong performance in a joint training setup for semantic segmentation on the ACDC [39] dataset. In summary, this paper makes several key contributions. We introduce the WENet architecture, which uniquely integrates wavelet transforms with CNNs and Transformers to precisely filter noise and recover information. We also propose two custom layers, the WCL and CAL, designed specifically for detail recovery and contrast adjustment. Finally, we provide extensive experimental validation to show that our method not only enhances low-light images effectively but also improves the performance of subsequent high-level vision tasks.

2. Related Work

2.1 Low-Light Image Enhancement

Traditional image enhancement approaches can be broadly categorized into histogram-based and Retinex-based methods. Reza et al. [40] demonstrated that local histogram equalization with clipping constraints performs well in certain conditions but may lead to significant semantic loss in others. Ibrahim et al. [11] proposed the Brightness Preserving Dynamic Histogram Equalization (BPDHE) method, which first applies a one-dimensional Gaussian filter to the input histogram, followed by local partitioning based on maximum intensity values. Guo et al. [13] introduced the LIME (Low-Light Image Enhancement) method, which refines an initial illumination map—derived from the maximum values of RGB channels—by adding structural regularization to obtain the final enhanced illumination map. Wang et al. [19] proposed a bright-pass filter that constrains the range of reflectivity components for better illumination–reflectance separation. Similarly, Kim et al. [41] estimated

illumination using multi-diffusion spaces at each pixel position to avoid noise amplification.

With the rise of convolutional neural networks (CNNs), low-light enhancement has seen significant progress [42,43]. Since the introduction of LLNet [44], which used autoencoders for low-light denoising, CNNs have become dominant in this field. Chen et al. [27] proposed RetinexNet, the first deep learning-based model integrating Retinex theory for image enhancement. RetinexNet decomposes an image into reflection and illumination components, reconstructs them via enhancement and adjustment networks, and fuses them to obtain the enhanced image. It also introduced the LOL dataset, a standard benchmark for low-light enhancement.

Zhang et al. [25] proposed KinD, another Retinex-inspired method that does not require paired low-light and normal-light images. By leveraging images with different exposures, KinD reduces dataset collection costs while allowing flexible enhancement control. Lv et al. [26] proposed MBLLEN, a multi-branch network capable of simultaneous noise suppression and enhancement.

For more robust enhancement under unknown conditions, Ma et al. [30] developed the Self-Calibrated Illumination (SCI) framework, incorporating weight-sharing and self-calibration modules to improve generalization. SCI employs unsupervised loss functions, achieving strong performance in real-world applications such as dark-face detection and night-time semantic segmentation. Jiang et al. [45] introduced R2RNet (Real-low to Real-normal Network), which jointly exploits spatial and frequency information to enhance contrast and preserve detail. Guo et al. [29] proposed Zero-DCE, reframing enhancement as a curve-estimation problem constrained by mathematical priors. Zhou et al. [46] presented GLARE, a latent-feature retrieval-based enhancement model. Xu et al. [47] designed UPT-Flow, an unbalanced points-guided multi-scale Transformer-based normalizing flow for low-light restoration.

Recent diffusion-based approaches have also emerged. Ma et al. [48] integrated Retinex theory with channel and spatial attention, improving feature adaptability. Wan et al. [49] proposed PSC-Diffusion, a simplified conditional diffusion model employing SimpleGate and SimPF blocks for noise prediction. Subsequently, Wan et al. [50] developed SFDiff, incorporating Fourier frequency fusion through a Spatial-Frequency Fusion (SFF) block for joint spatial–frequency interaction.

While these models achieve strong brightness enhancement, they often contain numerous parameters, limiting real-time deployment. Moreover, many over-enhance brightness at the expense of fine detail and

fail to sufficiently improve downstream vision tasks such as segmentation or detection.

2.2 Wavelet-Based Models

Wavelet transform decomposes images into multi-scale subbands that separately capture low- and high-frequency content—corresponding to global structures and detailed edges, respectively. Each wavelet transform splits an image into four subbands, effectively enabling joint spatial–frequency analysis.

Recent studies have begun integrating wavelet transform with CNNs, leveraging the frequency-domain representation for tasks such as denoising, compression, and super-resolution. Liu et al. [35] proposed a wavelet and two-branch neural network for image moiré removal, combining dense and dilated convolution modules to expand the receptive field. Huang et al. [36] developed a wavelet-based CNN for face super-resolution, which first predicts wavelet coefficients and reconstructs the high-resolution image accordingly. The network also employs wavelet-domain loss functions for better convergence.

Yoo et al. [51] introduced a wavelet correction transformation based on whitening and coloring transformations, preserving structural integrity in learned features. Despite these promising results, few studies have explored wavelet-based methods for low-light enhancement. Leveraging wavelet decomposition for illumination refinement and detail recovery remains an underexplored yet effective strategy—something our proposed WENet explicitly addresses.

2.3 Transformer in Low-Level Vision

The Transformer architecture [52], initially introduced for natural language processing, has profoundly influenced computer vision research. The Vision Transformer (ViT) [53] adapted this model to image processing by splitting images into fixed-size patches and performing global self-attention operations.

For low-level vision tasks, several transformer-based architectures have emerged. Wang et al. [33] introduced UFormer, a U-shaped Transformer that maintains computational efficiency while capturing long-range dependencies. Zamir et al. [54] proposed Restormer, designed for high-resolution image restoration. It introduces the multi-DConv head transposed attention (MDTA) module for local–global feature aggregation and a Gated-DConv Feed-Forward Network (GDFN) for adaptive feature transformation. Chen et al. [55] proposed IPT, a pre-trained transformer framework maximizing generalization across multiple image restoration tasks.

IAT [31] uses least-square digital image processing with a lightweight, end-to-end Transformer network for real-time enhancement. These methods illustrate the potential of transformer-based architectures for illumination correction, yet their computational overhead limits real-time

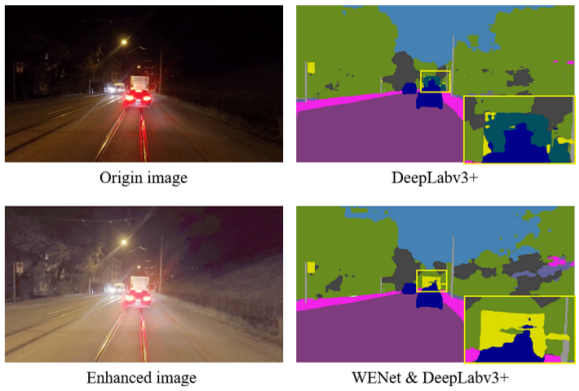
deployment—an issue WENet addresses through a hybrid CNN–Transformer framework.

2.4 Semantic Segmentation

Semantic segmentation aims to classify every pixel in an image into predefined categories. Long et al. [56] introduced the Fully Convolutional Network (FCN), pioneering end-to-end segmentation. Zhao et al. [57] developed PSPNet, using pyramid pooling to aggregate global and local context information. Chen et al. [58] proposed DeepLabv3+, which employs depthwise separable convolutions for efficient multi-scale context extraction.

Fu et al. [59] presented Dual Attention Network (DANet), which enhances feature representations using position attention and channel attention modules. Yu et al. [60] proposed BiSeNet, a real-time segmentation network with separate spatial and context paths for high-resolution and semantic feature extraction, respectively, followed by a fusion module for final prediction.

Although these segmentation networks perform well under normal lighting, their accuracy drops significantly in low-light conditions due to information loss and low signal-to-noise ratios. This underscores the need for effective low-light enhancement models—such as our proposed Wavelet-based Enhancement Network (WENet)—to improve segmentation robustness in dark environments.



3. Methodology

To address the inherent challenges of low-light image enhancement, we propose a novel deep learning framework, the Dual-Domain Enhancement Network (DDENet). The core principle of our approach is to simultaneously process image features in both the spatial and wavelet domains to achieve a comprehensive restoration. By handling high-level contextual information and low-level texture details in parallel, our model can effectively suppress noise while recovering vivid details that are often lost in darkness. The overall architecture of DDENet, illustrated in Figure 1, is organized into four distinct stages: a Shared Feature Encoder, Dual-Domain Parallel Branches, a Feature Fusion Module, and an Image Reconstruction Decoder.

3.1. Shared Feature Encoder

The primary role of the Shared Feature Encoder is to extract a robust and rich set of multi-level features from the input low-light image. This initial stage provides a powerful foundation for the subsequent specialized processing branches. The process is as follows:

1. An input low-light image, denoted as $I \in R^{H \times W \times 3}$, is first fed into a 3x3 Convolutional Layer. This step extracts shallow features, such as edges, corners, and basic textures, resulting in a feature map Fshallow.
2. Fshallow is then passed through a sequence of two powerful Transformer Blocks. Leveraging their self-attention mechanism, these blocks are highly effective at capturing long-range dependencies and global contextual information from the image. The output of this stage is a deeply encoded feature map, Fencoded, which holds a comprehensive understanding of the input scene.

3.2. Dual-Domain Parallel Branches

To effectively handle the distinct challenges of contextual understanding and fine-grained detail recovery, the encoded feature map Fencoded is simultaneously processed by two specialist branches:

- Branch A: Contextual Path (Spatial Domain): This branch is designed to further model the high-level semantic information and structural layout of the scene. It consists of an additional Transformer Block that operates on Fencoded to produce a contextually refined feature map, Fcontext.
- Branch B: Detail Restoration Path (Wavelet Domain): This branch serves as the core of our detail enhancement and denoising strategy. The features Fencoded are passed into our novel Wavelet Denoising & Refinement Module (WDRM). This module operates in the wavelet domain, which is ideal for separating high-frequency noise from essential low-frequency image content. The output is a detail-rich, noise-reduced feature map, Fdetail.

3.3. Feature Fusion Module

The purpose of the Feature Fusion Module is to intelligently integrate the complementary information generated by the two parallel branches. The contextual understanding from

Branch A is combined with the fine-grained details from Branch B through the following steps:

1. The feature maps Fcontext and Fdetail are first concatenated along their channel dimension.
2. This combined feature map is then passed through a 1x1 Convolutional Layer. This "fusion convolution" compresses and reorganizes the combined features, allowing the network to learn the most effective way to merge structural and textural information, resulting in a single, unified feature map, Ffused.

3.4. Image Reconstruction Decoder

The final stage of DDENet is responsible for translating the abstract, fused features back into a visually coherent and high-quality enhanced image.

1. The fused feature map Ffused is fed into our custom Spatial Refinement Module (SRM). This module performs final spatial adjustments, focusing on optimizing the overall contrast, brightness, and color fidelity to ensure a perceptually natural result.
2. The output of the SRM is then passed through a final 3x3 Convolutional Layer, which renders the final enhanced image, Ienhanced.

3.5. Loss Function

To train the DDENet, we employ a composite loss function designed to ensure both pixel-level accuracy and perceptual quality. The total loss Ltotal is a weighted sum of two components:

$$L_{total} = L_1 + \lambda L_{perceptual}$$

- L1 Loss: This is a pixel-wise loss that measures the absolute difference between the enhanced image and the ground-truth normal-light image. It ensures accurate color and brightness reconstruction.
- Perceptual Loss (Lperceptual): To ensure the enhanced image is visually pleasing and retains realistic textures, we use a perceptual loss. This loss compares high-level features extracted by a pre-trained network (like VGG-19) from both the enhanced and ground-truth images, penalizing differences in texture and structure. The hyperparameter λ balances the contribution of the two losses.

4. Experiments

To comprehensively evaluate the performance of our proposed Dual-Domain Enhancement Network (DDENet), we designed a series of experiments on several public benchmark datasets. This section details the datasets, evaluation metrics, implementation specifics, and the baseline methods used for comparison.

4.1. Datasets

To ensure a fair and thorough comparison, we will train and test our model on the following widely-used datasets in the field of low-light image enhancement. These datasets were sourced from publicly available repositories like Kaggle.

- LOL Dataset: A popular real-world dataset consisting of 500 low-light and normal-light image pairs. It is a standard benchmark for evaluating the performance of enhancement algorithms in realistic scenarios.
- LOLv2 Dataset: An extended version of the LOL dataset, containing 789 low-light/normal-light pairs. It provides a more extensive set of images for robust training and testing.
- ACDC (Adverse Conditions Dataset): To validate the practical utility of our model for downstream computer vision tasks, we use the nighttime subset of the ACDC dataset. This dataset contains 1000+ images of driving scenes at night with high-quality semantic segmentation labels, making it ideal for evaluating the impact of enhancement on segmentation accuracy.

4.2. Evaluation Metrics

The performance of our DDENet and other baseline methods will be assessed both quantitatively and qualitatively. For quantitative analysis, we will use the following standard metrics:

- PSNR (Peak Signal-to-Noise Ratio): This metric measures the pixel-level reconstruction quality between the enhanced image and the ground-truth image. A higher PSNR value indicates a more accurate restoration.
- SSIM (Structural Similarity Index): This metric evaluates the perceived quality of the enhanced image by comparing its structural similarity, contrast, and luminance to the ground-truth. An SSIM value closer to 1 indicates a better result.
- mIoU (mean Intersection over Union): For the downstream semantic segmentation task on the ACDC dataset, we will use mIoU. It measures the overlap between the predicted segmentation map and the ground-truth labels across all classes. A higher mIoU score signifies better segmentation performance.

4.3. Implementation Details

Our proposed DDENet model was implemented using the PyTorch framework. The training was conducted on an NVIDIA RTX 3090 GPU. For the optimization, we used the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The model was trained for 400 epochs with a batch size of 8. During

training, input images were randomly cropped to a resolution of 256×256 pixels, and random horizontal flipping was applied for data augmentation.

4.4. Comparison Methods

We will compare the performance of our DDENet against a comprehensive set of state-of-the-art low-light enhancement methods. The selected baselines include:

- Retinex-based methods: RetinexNet
- GAN-based methods: EnlightenGAN
- Zero-Shot Learning methods: Zero-DCE
- Transformer-based methods: IAT, Restormer

5.conclusion

In the realm of the wavelet domain, we introduce the Wavelet-based Enhancement Network (WENet), an innovative framework that uni fies convolutional layers with Transformer blocks. We have proposed a Wavelet Calibrate Layer (WCL) and a Contrast Adjustment Layer (CAL) to effectively filter out noise and enhance image contrast. Our WENet has demonstrated superior performance on the LOL and LOLv2 datasets. Furthermore, by jointly training WENet with a segmentation model in the dark scenes within the ACDC dataset, we have achieved commendable segmentation results. This integrated approach fosters robustness in low-light conditions, which is crucial for accurate scene interpretation under challenging lighting.