

Article

DepthLux: Employing Depthwise Separable Convolutions for Low-Light Image Enhancement [†]

Raul Balmez ¹, Alexandru Brateanu ¹, Ciprian Orhei ², Codruta O. Ancuti ^{2,*} and Cosmin Ancuti ^{2,*}¹ Department of Computer Science, University of Manchester, Manchester M13 9PL, UK;

raul.balmez@student.manchester.ac.uk (R.B.); alexandru.bratanu@student.manchester.ac.uk (A.B.)

² Faculty of Electronics, Telecommunications and Information Technologies, Polytechnic University Timisoara, 300223 Timisoara, Romania; ciprian.orhei@upt.ro

* Correspondence: codruta-o.ancuti@upt.ro (C.O.A.); cosmin.ancuti@upt.ro (C.A.)

[†] IEEE International Symposium on Electronics and Telecommunications (ISETC), Timisoara, Romania, 7–8 November 2024.

Abstract: Low-light image enhancement is an important task in computer vision, often made challenging by the limitations of image sensors, such as noise, low contrast, and color distortion. These challenges are further exacerbated by the computational demands of processing spatial dependencies under such conditions. We present a novel transformer-based framework that enhances efficiency by utilizing depthwise separable convolutions instead of conventional approaches. Additionally, an original feed-forward network design reduces the computational overhead while maintaining high performance. Experimental results demonstrate that this method achieves competitive results, providing a practical and effective solution for enhancing images captured in low-light environments.

Keywords: image sensor restoration; low-light enhancement; vision transformer



Academic Editor: Alessandro Artusi

Received: 29 January 2025

Revised: 22 February 2025

Accepted: 27 February 2025

Published: 1 March 2025

Citation: Balmez, R.; Brateanu, A.; Orhei, C.; Ancuti, C.O.; Ancuti, C.

DepthLux: Employing Depthwise Separable Convolutions for Low-Light Image Enhancement. *Sensors* **2025**, *25*, 1530. <https://doi.org/10.3390/s25051530>

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the Creative Commons

Attribution (CC BY) license

[\(https://creativecommons.org/licenses/by/4.0/\)](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Image sensors frequently encounter difficulties in low-light environments, including increased noise and diminished contrast, both of which negatively impact image quality. Low-light image enhancement (LLIE) plays an important role in computer vision (CV), focusing on enhancing the illumination—as shown in Figure 1, contrast, quality, and clarity of images affected by limited visibility and distortions in challenging illuminated settings [1]. Under low-light conditions, image sensors often struggle with noise and diminished contrast, which compromise overall image quality. This degradation not only leads to visually unappealing results but also negatively impacts the functionality of many CV systems.

Low-light environments are characterized by insufficient illumination levels, which often introduce other issues like artifacts and noise. These visual defects are determined by the strong relationship between the colors and the brightness, where the latter can strongly influence the representation of the chromatic element. As the generic low-light environment is not a uniform space in terms of the light distribution, multiple methods have been attempted to demonstrate their ability to enhance images captured under very different conditions [2,3].

Well-established LLIE techniques, such as gamma correction and histogram equalization [1], enhance the illumination factor in images but do not mitigate the other deficiencies introduced by the absence of light. Other classical CV methods attempt to improve these techniques by incorporating illumination factors but often result in unwanted artifacts

and imbalances in the enhanced images [1]. A widely adopted approach leverages cognitive models inspired by Retinex theory [4]. The foundational Retinex algorithm is based on selecting an appropriate surround function to determine pixel weighting within the neighborhood of a target pixel, leading to the development of the single-scale Retinex algorithm [5]. This was later expanded upon with multiscale Retinex algorithms [6,7].

The Retinex theory was first mirrored in the deep learning (DL) approaches through the use of CNNs which provided the first well-established LLIE solutions. These methods often struggle to adapt to different light settings, as they rely on a one-to-one mapping in the enhancement process and tend to underfit in the absence of long training times. On the other hand, CNNs successfully manage to capture local contexts but struggle to learn the light patterns represented by long-range dependencies.

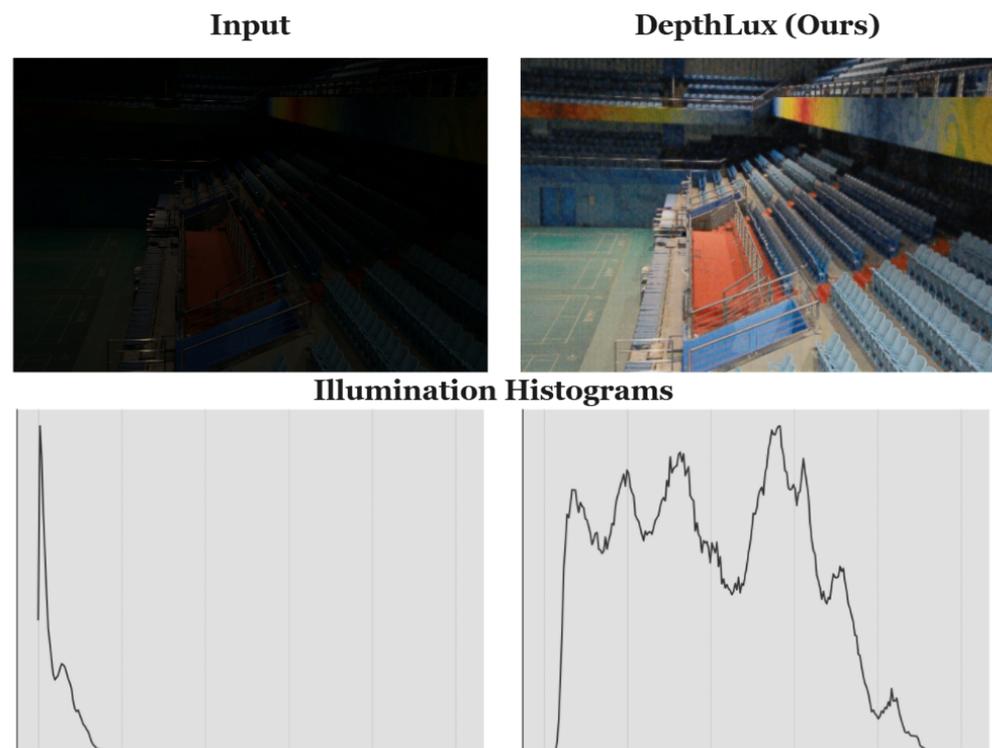


Figure 1. Visual representation of a low-light image and its enhanced counterpart, along with their illumination histograms. We assume illumination to be the maximum value across the channels of each pixel.

The Transformer’s self-attention mechanism [8], later adapted for Vision Transformers (ViTs) [9], offers a partial remedy to these limitations in image processing tasks. However, ViTs face notable drawbacks, including prohibitive computational overhead and resource inefficiency, stemming from the quadratic scaling of self-attention and the intensive use of dense and convolutional layers within feed-forward networks (FFNs).

Depthwise separable convolutions (DSCs) have emerged as an innovation design of CNNs, particularly for applications requiring computational efficiency and reduced model size. This technique decomposes the standard convolution operation into two distinct processes: depthwise convolution and pointwise convolution. The primary advantage of this separation is the significant reduction in the number of parameters and computational complexity, making it particularly suitable for deployment on resource-constrained devices such as mobile phones and embedded systems [10].

The challenge of addressing low-light deficiencies has often been met with computationally intensive solutions, characterized by large parameter counts and intricate

mechanisms. We propose a reimagination of the FFN, which traditionally relies on computationally expensive layers to enhance feature representations. Our work demonstrates that DSC can serve as an efficient alternative, maintaining performance integrity while drastically reducing complexity.

Building on the methods and insights presented in [11], we extend this to provide additional motivation for the proposed approach and contextualize the task further. In addition, we introduce a qualitative results section to highlight the perceptual appeal of the enhanced images and conduct an ablation study to justify architectural and training choices. The key innovations introduced by DepthLux can be summarized as follows:

- Reimagining the structure of the FFN, Depthwise Separable Convolutions (DSCs) are employed exclusively as a substitute for traditional convolutional and fully connected layers, significantly enhancing the model's scalability and computational cost.
- Competitive results are presented while having fewer model parameters than most of the current LLIE methods.
- To address the computational bottleneck of the ViT's self-attention mechanism, a contraction–expansion strategy is applied, compressing the feature space prior to attention map computation and restoring it afterward, thereby reducing the computational overhead.

The paper is organized as follows: Section 1 introduces the research and highlights its significance. Section 2, titled “Related Work”, provides an overview of the LLIE domain. In Section 3, “Methods”, we describe the proposed model and its key components. Section 4 presents the results obtained on the LOL dataset [12]. Finally, the paper concludes with an ablation study in Section 5 and a discussion of future research directions in Section 6.

2. Related Work

LLIE has progressed from traditional CV techniques to advanced DL frameworks, each focusing on different methods to solve the problem. The early methods relied on histogram equalization, gamma correction, and physical models that corrected illumination by estimating environmental light and guided filtering [13]. Retinex-based approaches [4,14–16] addressed illumination correction by decomposing images into two components: a high-frequency component (reflectance) and a low-frequency component (illumination). These methods often used mathematical frameworks [17,18], commonly known as Retinex-based variational methods. However, they frequently overlooked noise and artifacts present in low-light environments, limiting their practical effectiveness.

The introduction of DL fundamentally shifted the paradigm in LLIE. CNN-based approaches, exemplified by LLNet [19], pioneered this transformation, followed by advancements such as EnGAN [20], which employed generative models to directly produce normal-light images from low-light inputs. Unsupervised methods, such as Zero-DCE [21], were also introduced, which use differentiable curve estimation and remove the need for paired data but struggle with color fidelity under extreme conditions.

Wei et al. [12] extended Retinex theory to DL, proposing a decomposition framework to enhance illumination and reconstruct well-lit images. Despite their progress, these CNN-driven methods often require complex training procedures and substantial computational resources. Furthermore, their local receptive fields restrict their ability to model global interactions and to capture long-range dependencies, a critical requirement for handling diverse low-light scenarios.

In recent years, Transformers have emerged as a powerful alternative to CNNs, reshaping image restoration tasks [22–25]. Unlike CNNs, Transformers such as ViTs [9] utilize self-attention mechanisms to capture long-range dependencies, providing a significant advantage in LLIE. Transformer-based models like UFormer [26], Restormer [27], Retinex-

former [28], and LytNet [2] have demonstrated this capability effectively while also proving that the effectiveness of the Transformer architecture can also be found in computationally efficient methods. UFormer reimagines the U-Net architecture [29] by substituting convolutions with Transformer blocks while maintaining a hierarchical encoder–decoder structure. KAN- \mathcal{T} [30] introduces an innovative attention mechanism based on Kolmogorov–Arnold networks and seamlessly integrates it into a Transformer architecture.

Retinexformer further advances this field by leveraging illumination-based representations to model non-local interactions in regions with varying lighting conditions. The presented methods also offer lightweight solutions, with the total number of trainable parameters ranging from 1.61 million to 26 million. DepthLux, in particular, has 9.75 million parameters, striking a balance between efficiency and performance. This highlights the importance of prioritizing computational efficiency in the design of ViTs for image restoration, ensuring that the models remain practical for real-world applications without compromising quality.

Computational efficiency has also become a focal point in LLIE research. DSCs, for instance, offer an elegant solution to reduce computational complexity by splitting standard convolutions into depthwise and pointwise operations [10,31]. While the depthwise operation applies a filter per channel, the pointwise convolution combines these outputs using 1×1 convolutions, achieving a balance between efficiency and performance. This approach not only reduces the number of parameters but also accelerates inference, making it ideal for real-time applications.

Attention mechanisms have played a pivotal role in advancing LLIE by highlighting features corresponding to poorly lit regions. However, their high computational cost has presented challenges. Transformer-based methods address this issue through innovative strategies such as reducing the dimensions of the attention map to $C \times C$, where C is the feature depth [27], or employing feature contraction and expansion techniques around attention computations [2]. These approaches effectively balance performance with efficiency.

The evolution of LLIE methods reflects a continuous effort to bridge the gap between theoretical advances and practical applications. Traditional methods, while foundational, lack robustness in real-world scenarios. CNN-based approaches have marked a significant leap forward but remain constrained by their inability to capture global dependencies. Transformers, with their self-attention mechanisms, have emerged as a transformative force, addressing these limitations and paving the way for more sophisticated LLIE frameworks.

3. Methods

Our approach demonstrates that standard convolutions, commonly employed in most LLIE ViTs, can be efficiently replaced with depthwise separable convolutions (DSCs), yielding minimal performance degradation while significantly reducing model parameters. Based on a conventional transformer framework as illustrated in Figure 2, DepthLux represents a paradigm shift in the traditional design of feed-forward networks (FFNs) by introducing the Depthwise Separable Illumination Block (DSIB). Additionally, it aims to optimize the self-attention mechanism through the Performance-Optimized Attention Block (POAB).

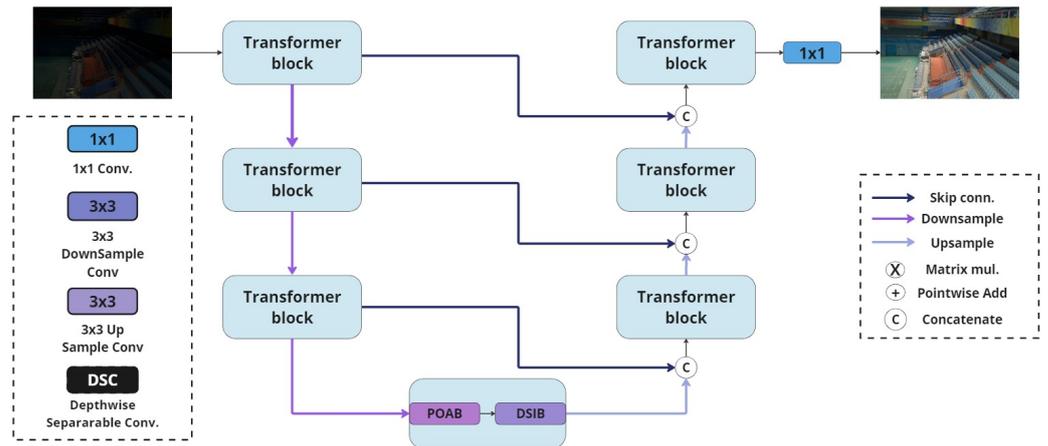


Figure 2. The architecture of DepthLux introducing a conventional Transformer architecture enhanced with the POAB and the DSIB.

3.1. Depthwise Separable Illumination Block

The innovation introduced in the Depthwise Separable Illumination Block (DSIB), which forms the core of DepthLux's FFNs, lies in its exclusive reliance on DSCs. This design choice plays a pivotal role in achieving the computational efficiency of the model. To illustrate, consider a convolution operation with a kernel of size $D_k \times D_k$, M input channels, and N output channels. A standard convolution requires $D_k \times D_k \times M \times N$ multiplications, as each output channel computes a weighted sum over the entire input space. In contrast, a depthwise separable convolution significantly reduces this complexity by factorizing the operation into two distinct steps: (1) a depthwise convolution, which applies a single convolutional filter to each input channel independently, and (2) a pointwise convolution, which combines these filtered outputs across channels. As a result, the total number of multiplications required by a DSC is reduced to $D_k \times D_k \times M + M \times N$, a substantial improvement over standard convolutions. Letting x_m represent the input channels, k the convolution kernel, $k^{\text{depthwise}}$ and $k^{\text{pointwise}}$ the depthwise and pointwise convolution kernels, and y_n the output, we can compare the functionality of the convolution and the DSC as follows:

$$\text{Convolution: } y_n = \sum_{m=1}^M (x_m * k), \quad \forall n \in \{1, \dots, N\}, \quad (1)$$

$$\text{DS Convolution: } y_n = \sum_{m=1}^M \left((x_m * k^{\text{depthwise}}) * k^{\text{pointwise}} \right), \quad \forall n \in \{1, \dots, N\}, \quad (2)$$

The integration of the 3×3 kernel, $k^{\text{depthwise}} \in \mathbb{R}^{3 \times 3}$ for depthwise convolutions (DWConv), alongside the 1×1 kernel, $k^{\text{pointwise}} \in \mathbb{R}^{1 \times 1}$, for pointwise convolutions (PWConv), enables a comprehensive extraction of features across varying spatial scales while maintaining computational efficiency. The parallel configuration of depthwise separable convolutions (DSCs) at the onset of the DSIB (illustrated in Figure 3) is explicitly designed to model diverse, multiscale illumination patterns. This architecture enables the extraction of rich illumination features across various scales, which are subsequently aggregated through the summation of the feature maps:

$$\mathbf{F}'_{\text{in}} = \sum_{i=1}^2 \text{DSC}^{(i)} \left(\text{DSC}^{(i)} (\mathbf{F}_{\text{in}}) \right), \quad \mathbf{F}'_{\text{in}} \in \mathbb{R}^{H \times W \times C}. \quad (3)$$

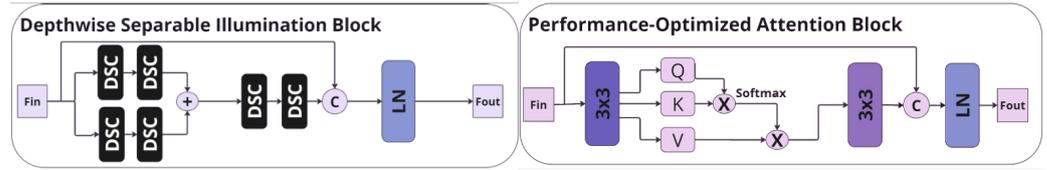


Figure 3. DSIB, which represents the feed-forward section of our proposed Transformer, exclusively uses DSC (left). POAB is built to reduce the computational complexity of the model by contracting and then expanding the feature space (right).

The fused map is further processed and refined through another set of DSC, which extracts the most relevant information. To transmit information to future layers, we employ a skip connection originating from $F_{in} \in \mathbb{R}^{H \times W \times C}$, which is concatenated (Equation (4)) with the output processed using multiple DSCs:

$$F_{out} = [F_{in}, DSC(DSC(F'_{in}))], \quad F_{out} \in \mathbb{R}^{H \times W \times 2C} \quad (4)$$

DepthLux introduces the innovative application of the Gaussian Error Linear Unit (GELU) activation following each stage of the DSC. By modulating inputs according to their magnitude and alignment with the Gaussian distribution mean, GELU ensures the preservation of nuanced features even in noise-contaminated settings, offering a distinct advantage over traditional activation functions:

$$DSC(F_{in}) = GELU(PWConv(GELU(DWConv(F_{in})))) \quad (5)$$

3.2. Performance-Optimized Attention Block

Figure 3 provides a schematic representation of the POAB block. The proposed methodology aims to establish that models with a significantly reduced computational overhead can achieve performance levels comparable to state-of-the-art approaches. This concept is exemplified within the attention module.

The input feature tensor $F_{in} \in \mathbb{R}^{H \times W \times C}$ is first reduced in resolution by a factor of four through a 3×3 convolution, producing a condensed representation:

$$F_{down} = conv3 \times 3_{down}(F_{in}), \quad F_{down} \in \mathbb{R}^{H/4 \times W/4 \times C/2}$$

Following this compression, the fundamental self-attention components—query (Q), key (K), and value (V)—are computed via fully connected layers as described in Equation (6):

$$Q = W_Q \cdot F_{down}; \quad K = W_K \cdot F_{down}; \quad V = W_V \cdot F_{down} \quad (6)$$

The attention mechanism responsible for capturing long-range dependencies can be formally expressed as follows, where dim represents the number of output channels used in the computation process of Q, K, and V:

$$Attn = \text{Softmax}\left(\frac{QK^T}{\sqrt{dim}}\right)V, \quad Attn \in \mathbb{R}^{H/4 \times W/4 \times C} \quad (7)$$

Upon calculating the attention map, the feature space is restored to its original dimensions using a 3×3 transposed convolution:

$$Attn_{out} = conv3 \times 3_{up}(Attn), \quad Attn_{out} \in \mathbb{R}^{H \times W \times C} \quad (8)$$

To preserve spatial information during the attention map computation, the input feature tensor $F_{in} \in \mathbb{R}^{H \times W \times C}$ is routed through a skip connection, ensuring its integration into deeper layers. This tensor is subsequently concatenated with the attention output $Attn_{out} \in \mathbb{R}^{H \times W \times C}$ which is expanded as formalized in Equation (8). The resulting feature map undergoes layer normalization, which aligns the feature distributions to ensure stability and consistency across channels.

The expansion–contraction strategy helps reduce the computational cost of the attention mechanism. By downsampling $F_{in} \in \mathbb{R}^{H \times W \times C}$ to $F_{down} \in \mathbb{R}^{H/4 \times W/4 \times C/2}$, we also reduce the depth of the features, optimizing the attention mechanism. While this leads to some information loss, fully mitigating it would require quadrupling the number of output channels to maintain the number of elements between the input feature map and the downsampled map. However, this would result in an additional 42 million parameters, which would significantly increase the model’s complexity.

4. Results

4.1. Implementation Details

The architecture adopts a standard transformer-based design, comprising three encoder–decoder pairs and a bottleneck block that serves as both a refinement module and a bridge between the encoding and decoding phases. The dimensional settings for the encoders and their respective decoders are specified as [64, 128, 256], where each value represents the embedding size, which determines the number of features in each layer. The bottleneck block operates with a fixed embedding size of 256, maintaining consistency in feature representation at the network’s deepest level. In addition, the transformer follows a U-Net structure: after each encoder, a downsampling operation reduces the spatial resolution to capture higher-level features, and before each decoder, an upsampling step is applied to restore the resolution. This arrangement ensures that the network effectively integrates the global context with local detail during both the encoding and decoding processes.

4.2. Training Details

DepthLux is tailored to handle a diverse range of images affected by insufficient illumination. To rigorously evaluate its performance, experiments are conducted on three benchmark datasets: LOL-v1, LOL-v2 real [32], and LOL-v2 synthetic [32]. Performance is measured using the Peak Signal-to-Noise Ratio (PSNR) and the Structure Similarity Index Measure (SSIM) [33], as shown in Table 1. These datasets represent the current standard for assessing state-of-the-art methods in LLIE. The framework used for the implementation is TensorFlow, and the training is performed on a Nvidia 3090 GPU using image patches sized 256×256 pixels. Data augmentation strategies, such as random cropping and flipping, are employed to enrich the training dataset and enhance model generalization.

The optimization process leverages the Adam optimizer, configured with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To further guide the training towards superior image restoration quality, a loss function based on PSNR as described in [2] is adopted. This loss function emphasizes the reconstruction of high-fidelity images and is detailed in Equation (9):

$$\mathcal{L}_{PSNR} = 40 - PSNR(\text{predicted}, \text{ground truth}) \quad (9)$$

Table 1. Results on LOL-v1 and LOL-v2 datasets and parameter counts of different models. **Red**, **blue**, and **green** metrics represent first, second, and third places, respectively.

Methods	Param (M)	LOL-v1		LOL-v2-R		LOL-v2-S	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SID [34]	7.76	14.35	0.436	13.24	0.442	15.04	0.610
DeepUPE [35]	1.02	14.38	0.446	13.27	0.452	15.08	0.623
RF [36]	21.54	15.23	0.452	14.05	0.458	15.97	0.632
DeepLPF [37]	1.77	15.28	0.473	14.10	0.480	16.02	0.587
UFormer [26]	5.29	16.36	0.771	18.82	0.771	19.66	0.871
RetinexNet [12]	0.62	16.77	0.462	18.37	0.723	17.13	0.798
IPT [23]	115.31	16.27	0.504	19.80	0.813	18.30	0.811
Sparse [32]	2.33	17.20	0.640	20.06	0.816	22.05	0.905
EnGAN [20]	8.64	17.48	0.652	18.64	0.677	16.57	0.734
RUAS [38]	0.003	18.23	0.720	18.37	0.723	16.55	0.652
FIDE [39]	8.62	18.27	0.665	16.85	0.678	15.20	0.612
DRBN [40]	2.21	19.86	0.834	20.13	0.830	23.22	0.927
KinD [41]	8.03	20.87	0.799	17.54	0.669	16.26	0.591
Restormer [27]	26.13	22.43	0.823	19.94	0.827	21.41	0.830
MIRNet [42]	5.90	24.14	0.842	20.36	0.782	21.94	0.846
SNR-Net [43]	4.01	24.61	0.842	21.48	0.849	24.14	0.928
LLFlow [44]	37.68	25.13	0.872	26.20	0.888	24.81	0.919
Retinexformer [28]	1.61	25.16	0.845	22.80	0.840	25.67	0.930
LLFormer [45]	24.55	25.76	0.823	26.20	0.819	28.01	0.927
LL-SKF [46]	39.91	26.80	0.879	28.45	0.905	29.11	0.953
DepthLux (Ours)	9.75	26.06	0.793	26.16	0.794	28.69	0.920

To address the challenges posed by local minima during the training of deep neural networks, a learning rate decay strategy is implemented. In particular, a cosine annealing schedule is employed, gradually reducing the learning rate from 2×10^{-4} to 1×10^{-6} over the course of training. This scheduling mechanism is designed to promote more stable convergence and improve the overall performance of the model.

4.3. Results Discussion

From a quantitative point of view, DepthLux achieves a competitive result which ranks the method second and third on the three tested datasets in terms of PSNR. We achieve an improvement of 9.7 dB against Uformer [26], 3.63 dB against Restormer [27], and 8.58 dB against EnGAN [20], while being 71%, 74%, and 62% more efficient than LLFormer [44], LLFlow [45], and Restormer [27], respectively, in terms of parameter count. However, some limitations in both performance and computational efficiency remain. These issues are highlighted by the gap with LL-SKF [46] across all three tested datasets, as well as the minimal 0.04 dB difference between DepthLux and both LLFormer [44] and LLFlow [45] on the LOL-v2 real datasets.

The qualitative results in Figure 4 demonstrate that DepthLux effectively enhances images without introducing artifacts or color distortions, unlike SNR and LLFlow. This highlights its ability to preserve natural image quality while improving visibility. However, due to the design trade-offs made in the POAB block to enhance computational efficiency, a slight graininess can be observed in the enhanced images. While this effect is minimal, it represents a compromise between performance and efficiency.

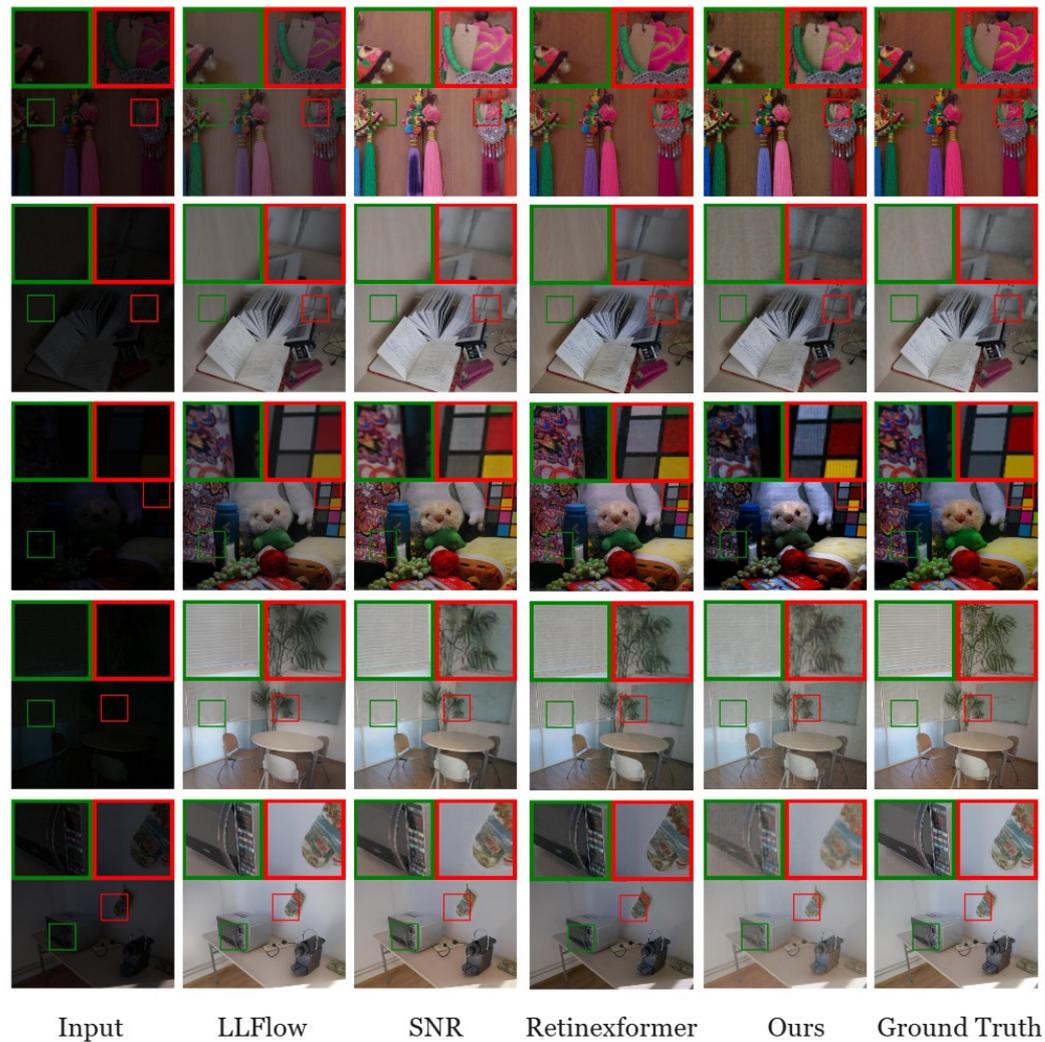


Figure 4. Qualitative results on the LOL dataset [12] comparing DepthLux with other SOTA LLIE methods.

4.4. Scaling Considerations

The results presented confirm that DepthLux achieves parameter efficiency by replacing conventional convolutions with DSC. In the current architecture, employing standard convolutions within the FFN yields a total of 26,402,980 trainable parameters. Substituting these with DSC reduces this count drastically to 9,754,780. To further support this assertion, we evaluate DepthLux’s scalability, examining how varying model sizes impact parameter count when using DSC versus traditional convolutions. This is accomplished by modulating a scaling factor that governs layer dimensions throughout the network.

The parameter growth trend for models utilizing standard convolutions follows a significantly steeper gradient compared to those employing DSC as depicted by their respective curves. This pronounced difference underscores the enhanced scalability of DepthLux.

5. Ablation Study

We conduct an ablation study on the LOL-v1 dataset to demonstrate the effectiveness of our proposed framework, and utilize PSNR to measure performance. The ablation study focuses on activation function and loss function variants. The results of the study can be found in Table 2.

Table 2. Ablation experiments on activation functions and loss functions. **Red** metrics represent the best results.

(a) The Effect of Different Activation Functions Used in the DSC.		(b) Impact of Different Loss Functions on Performance.	
Activation	PSNR	Loss Function	PSNR
No Activation	25.70	\mathcal{L}_1	25.71
ReLU	25.98	SSIM	25.86
PReLU	26.00	\mathcal{L}_2	26.01
GELU	26.06	$\mathcal{L}_{\text{PSNR}}$	26.06

To provide further justification for our selection of the loss function in training this model, we conduct a series of experiments comparing the performance of DepthLux when trained using different loss functions. Specifically, we evaluate the model's performance with three loss functions: \mathcal{L}_1 loss, \mathcal{L}_2 loss, and our proposed $\mathcal{L}_{\text{PSNR}}$ loss function.

\mathcal{L}_1 loss minimizes the absolute differences between the predicted and ground truth values, promoting robustness to outliers. \mathcal{L}_2 loss focuses on minimizing squared differences, penalizing larger errors more heavily. The $\mathcal{L}_{\text{SSIM}}$ loss, defined as

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(y_{\text{target}}, y_{\text{predicted}}) \quad (10)$$

$\mathcal{L}_{\text{SSIM}}$ focuses on preserving the structural similarity between the enhanced and ground truth images, improving the perceptual quality and structural fidelity. In contrast, our $\mathcal{L}_{\text{PSNR}}$ loss directly optimizes the PSNR by aligning the loss function with the evaluation criterion. The $\mathcal{L}_{\text{PSNR}}$ loss prioritizes features critical for high-quality reconstructions, leading to superior performance (0.35 dB compared to \mathcal{L}_1 , 0.2 dB compared to $\mathcal{L}_{\text{SSIM}}$, and 0.05 dB compared to \mathcal{L}_2).

GELU [47] outperforms no activation (by 0.37 dB), ReLU [48] (by 0.08 dB), and PReLU [49] (by 0.06 dB) when used in DSC due to its smooth and probabilistic formulation, which enables better feature discrimination and gradient flow. Unlike ReLU and PReLU, GELU transitions inputs non-linearly while avoiding abrupt changes or gradient instability. This smoothness allows DSC to model complex relationships more effectively, improving feature extraction and stability during training.

6. Conclusions and Future Work

This study establishes that ViTs can effectively utilize DSC as a viable alternative to traditional convolutional layers. By integrating DSC into the model architecture, the parameter count is significantly reduced, and computational efficiency is improved, leading to a more resource-conscious design without compromising performance in LLIE tasks. Our analysis from Figure 5 shows that DSC achieves an approximate 2.5× reduction in parameter count compared to standard convolutions, requiring only ~40% of the parameters needed for similar architectures. The reduced complexity introduced by DSC highlights its potential as a cornerstone for developing lightweight and scalable ViT architectures, paving the way for more efficient deployment in real-world applications.

The use of DSC not only reduces computational complexity but also demonstrates positive results on the LOL-v1 and LOL-v2 datasets as highlighted in the manuscript. To further expand our evaluation, we plan to include non-uniformly illuminated image datasets, such as the one provided in [50], in our future work.

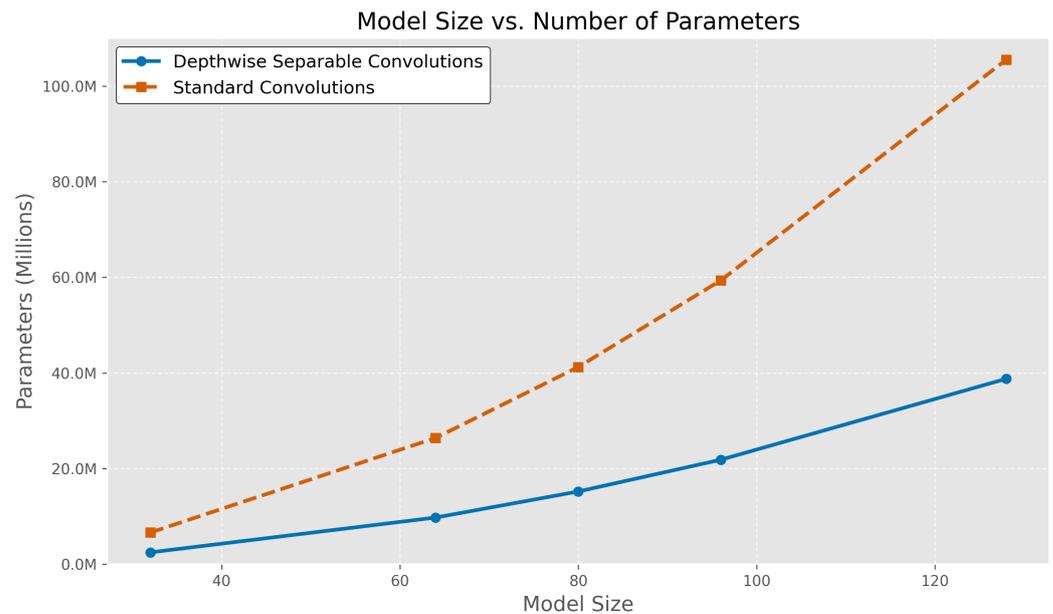


Figure 5. A comparison of the number of parameters and model sizes between models employing DSCs and those utilizing standard convolutions. The model size values correspond to the parameter that regulates the layer dimensions.

Future directions will aim to overcome the current framework’s limitations, focusing on optimizing the trade-off between model size and performance, particularly under extremely resource-constrained conditions. Research efforts will also explore the integration of complementary advancements, such as hybrid attention mechanisms that combine the strengths of different attention paradigms or adaptive depthwise operations tailored to varying input complexities. These innovations hold promise for further enhancing the capability and versatility of ViTs, ensuring they remain at the forefront of LLIE solutions and other computationally demanding tasks.

Author Contributions: Conceptualization, A.B., R.B., C.O., C.A. and C.O.A.; methodology, A.B., R.B., C.O., C.A. and C.O.A.; software, A.B. and R.B.; validation, A.B., R.B. and C.O.; formal analysis, A.B., R.B., C.O., C.A. and C.O.A.; investigation, A.B., R.B., C.O., C.A. and C.O.A.; resources, A.B., R.B., C.O., C.A. and C.O.A.; data curation, A.B., R.B., C.O., C.A. and C.O.A.; writing—original draft preparation, A.B., R.B., C.O., C.A. and C.O.A.; writing—review and editing, A.B., R.B., C.O., C.A. and C.O.A.; visualization, A.B., R.B., C.O., C.A. and C.O.A.; supervision, C.A. and C.O.A.; project administration, C.O., C.A. and C.O.A.; funding acquisition, C.A. and C.O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CV	Computer Vision
DW-Conv	Depthwise Convolutions
CNN	Convolutional Neural Network

DL	Deep Learning
DSC	Depthwise Separable Convolutions
DSIB	Depthwise Separable Illumination Block
FFN	Feed-Forward Network
GELU	Gaussian Error Linear Unit
K	Self Attention Key
LLIE	Low-light Image Enhancement
LN	Layer Normalization
POAB	Performance-Optimized Attention Block
PReLU	Parametric Rectified Linear Unit
PSNR	Peak Signal-to-Noise Ratio
PW-Conv	Depthwise Convolutions
ReLU	Rectified Linear Unit
Q	Self-Attention Query
SOTA	State of the Art
SSIM	Structural Similarity Index Measure
V	Self Attention Value
ViT	Vision Transformers

References

1. Wang, W.; Wu, X.; Yuan, X.; Gao, Z. An experiment-based review of low-light image enhancement methods. *IEEE Access* **2020**, *8*, 87884–87917. [[CrossRef](#)]
2. Brateanu, A.; Balmez, R.; Avram, A.; Orhei, C. LYT-Net: Lightweight YUV Transformer-based Network for Low-Light Image Enhancement. *arXiv* **2024**, arXiv:2401.15204.
3. Brateanu, A.; Balmez, R. Kolmogorov-Arnold Networks in Transformer Attention for Low-Light Image Enhancement. In Proceedings of the IEEE 2024 International Symposium on Electronics and Telecommunications (ISETC), Timisoara, Romania, 7–8 November 2024.
4. Land, E.H. The retinex theory of color vision. *Sci. Am.* **1977**, *237*, 108–129. [[CrossRef](#)] [[PubMed](#)]
5. Jobson, D.J.; Rahman, Z.u.; Woodell, G.A. Properties and performance of a center/surround retinex. *IEEE Trans. Image Process.* **1997**, *6*, 451–462. [[CrossRef](#)] [[PubMed](#)]
6. Rahman, Z.u.; Jobson, D.J.; Woodell, G.A. Multi-scale retinex for color image enhancement. In Proceedings of the 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland, 16–19 September 1996; IEEE: New York, NY, USA, 1996; Volume 3, pp. 1003–1006.
7. Jobson, D.J.; Rahman, Z.u.; Woodell, G.A. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **1997**, *6*, 965–976. [[CrossRef](#)]
8. Vaswani, A. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
10. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [[CrossRef](#)]
11. Balmez, R.; Brateanu, A. DepthLux: Depthwise Separable Convolution Transformer. In Proceedings of the IEEE 2024 International Symposium on Electronics and Telecommunications (ISETC), Timisoara, Romania, 7–8 November 2024; pp. 1–4.
12. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep retinex decomposition for low-light enhancement. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018.
13. Yu, S.Y.; Zhu, H. Low-Illumination Image Enhancement Algorithm Based on a Physical Lighting Model. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 28–37. [[CrossRef](#)]
14. Park, S.; Yu, S.; Moon, B.; Ko, S.; Paik, J. Low-light image enhancement using variational optimization-based retinex model. *IEEE Trans. Consum. Electron.* **2017**, *63*, 178–184. [[CrossRef](#)]
15. Gu, Z.; Li, F.; Fang, F.; Zhang, G. A novel retinex-based fractional order variational model for images with severely low light. *IEEE Trans. Image Process.* **2019**, *29*, 3239–3253. [[CrossRef](#)]
16. Cai, B.; Xu, X.; Guo, K.; Jia, K.; Hu, B.; Tao, D. A jointbintrinsic-extrinsic prior model for retinex. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

17. Liang, Z.; Liu, W.; Yao, R. Contrast Enhancement by Nonlinear Diffusion Filtering. *IEEE Trans. Image Process.* **2016**, *25*, 673–686. [[CrossRef](#)]
18. Lin, Y.H.; Lu, Y.C. Low-Light Enhancement Using a Plug-and-Play Retinex Model With Shrinkage Mapping for Illumination Estimation. *IEEE Trans. Image Process.* **2022**, *31*, 4897–4908. [[CrossRef](#)] [[PubMed](#)]
19. Lore, K.G.; Akintayo, A.; Sarkar, S. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognit.* **2017**, *61*, 650–662. [[CrossRef](#)]
20. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z. Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [[CrossRef](#)] [[PubMed](#)]
21. Guo, C.G.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 1780–1789.
22. Cai, Y.; Lin, J.; Wang, H.; Yuan, X.; Ding, H.; Zhang, Y.; Timofte, R.; Van Gool, L. Degradation-Aware Unfolding Half-Shuffle Transformer for Spectral Compressive Imaging. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022.
23. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-Trained Image Processing Transformer. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021.
24. Jang, S.I.; Pan, T.; Li, Y.; Heidari, P.; Chen, J.; Li, Q.; Gong, K. Spach Transformer: Spatial and channel-wise transformer based on local and global self-attentions for PET image denoising. *IEEE Trans. Med. Imaging* **2023**, *43*, 2036–2049. [[CrossRef](#)] [[PubMed](#)]
25. Liu, D.; Mao, Y.; Zuo, Y.; An, P.; Fang, Y. Light Field Angular Super-Resolution Network Based on Convolutional Transformer and Deep Deblurring. *IEEE Trans. Comput. Imaging* **2024**, *10*, 1736–1748. [[CrossRef](#)]
26. Wang, Z.; Cun, X.; Bao, J.; Liu, J. Uformer: A General U-Shaped Transformer for Image Restoration. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 28 November–9 December 2022.
27. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 28 November–9 December 2022.
28. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 12504–12513.
29. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015.
30. Brateanu, A.; Balmez, R.; Orhei, C.; Ancuti, C.; Ancuti, C. Enhancing Low-Light Images with Kolmogorov–Arnold Networks in Transformer Attention. *Sensors* **2025**, *25*, 327. [[CrossRef](#)]
31. Yang, H.; Jang, H.; Kim, T.; Lee, B. Non-temporal lightweight fire detection network for intelligent surveillance systems. *IEEE Access* **2019**, *7*, 169257–169266. [[CrossRef](#)]
32. Yang, W.; Wang, W.; Huang, H.; Wang, S.; Liu, J. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Trans. Image Process.* **2021**, *30*, 2072–2086. [[CrossRef](#)]
33. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)]
34. Chen, C.; Chen, Q.; Do, M.N.; Koltun, V. Seeing motion in the dark. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
35. Wang, R.; Zhang, Q.; Fu, C.W.; Shen, X.; Zheng, W.S.; Jia, J. Underexposed Photo Enhancement Using Deep Illumination Estimation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
36. Kosugi, S.; Yamasaki, T. Unpaired Image Enhancement Featuring Reinforcement-Learning-Controlled Image Editing Software. In Proceedings of the Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020.
37. Moran, S.; Marza, P.; McDonagh, S.; Parisot, S.; Slabaugh, G. DeepLpf: Deep local parametric filters for image enhancement. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
38. Liu, R.; Ma, L.; Zhang, J.; Fan, X.; Luo, Z. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021.
39. Xu, K.; Yang, X.; Yin, B.; Lau, R.W. Learning to restore low-light images via decomposition-and-enhancement. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
40. Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; Liu, J. Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. *IEEE Trans. Image Process.* **2021**, *30*, 3461–3473. [[CrossRef](#)]

41. Zhang, Y.; Zhang, J.; Guo, X. Kindling the darkness: A practical low-light image enhancer. In Proceedings of the ACM Multimedia, Nice, France, 21–25 October 2019.
42. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Learning enriched features for real image restoration and enhancement. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
43. Xu, X.; Wang, R.; Fu, C.W.; Jia, J. SNR-Aware Low-Light Image Enhancement. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 28 November–9 December 2022.
44. Wang, Y.; Wan, R.; Yang, W.; Li, H.; Chau, L.P.; Kot, A. Low-light image enhancement with normalizing flow. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 27 February–2 March 2022; Volume 36, pp. 2604–2612.
45. Wang, T.; Zhang, K.; Shen, T.; Luo, W.; Stenger, B.; Lu, T. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 2654–2662.
46. Yuhui, W.; Chen, P.; Guoqing, W.; Yang, Y.; Jiwei, W.; Chongyi, L.; Shen, H.T. Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
47. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
48. Agarap, A. Deep learning using rectified linear units (relu). *arXiv* **2018**, arXiv:1803.08375.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1026–1034.
50. Pu, T.; Zhu, Q. Non-Uniform Illumination Image Enhancement via a Retinal Mechanism Inspired Decomposition. *IEEE Trans. Consum. Electron.* **2024**, *70*, 747–756. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.