Dformer: Learning Efficient Image Restoration with Perceptual Guidance

Nodirkhuja Khudjaev, Roman Tsoy, S M A Sharif, Azamat Myrzabekov, Seongwan Kim, Jaeho Lee*
Opt-AI Inc.
LG Sciencepark, Seoul, South Korea
{nodir, roma, sharif, azamat, swan.kim, jaeho.lee}@opt-ai.kr

Abstract

Image restoration tasks incorporate widespread real-world application. Apart from its significant practicability, generic deep image restoration methods still fail to handle complex tasks, like shadow removal, low-light enhancement, etc. This paper addresses the limitations of existing image restoration methods by introducing a novel deep architecture. The proposed model incorporates illumination mapping inspired by the Retinex theory within a double encoder-decoder network. Additionally, it utilizes a multi-head cross-attention mechanism to correlate input and reconstructed images to generate plausible and refined images. The proposed model employs a perceptual optimization strategy to tackle intricate restoration tasks effectively. It surpasses state-of-the-art methods in demanding tasks such as shadow removal, low-light image enhancement, and blind compress image enhancement, all while utilizing fewer trainable parameters. Our method is selected among the top solutions in the New Trends in Image Restoration and Enhancement'24 (NTIRE) challenge for shadow removal, securing a top position without resorting to score-boosting techniques such as ensembling.

1. Introduction

Image restoration commonly refers to manipulating and enhancing low-quality images by correcting imperfections or artifacts that degrade their visual appearance [40]. These imperfections can include noise, blur, distortion, shadows, and other forms of degradation introduced during image acquisition, transmission, or processing. Image restoration techniques aim to recover or improve the original image by removing or reducing these imperfections while preserving salient details, structures, and color information. Notably, image restoration techniques such as shadow removal, low-light enhancement, denoising, and deblurring incorporate numerous applications in the real world, including photography [13], medical imaging [27], satellite imaging [3], autonomous driving [29], and forensic analysis [6].

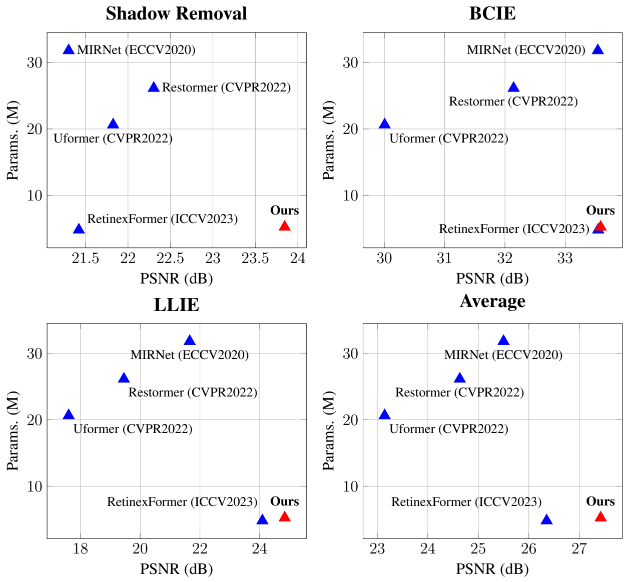

Figure 1. Comparison of models based on their PSNR (Peak Signal-to-Noise Ratio) and the number of parameters across SOTA image restoration tasks (a) shows the comparison for Shadow Removal [36]. (b) shows the comparison for Blind Compressed Image Enhancement (BCIE) [44]. (c) shows the comparison for Low Light Image Enhancement (LLIE) [23]. (d) shows the comparison of the average performance of each model across all tasks.

The widespread usability and the emergence of deep learning have inspired the vision community to address challenging image restoration tasks through their innovative approaches. Many studies [21, 35, 38, 51] from open research have focused on developing a single-network architecture that can efficiently handle multiple tasks rather than designing separate models for respective tasks. Several recent studies have used convolutional neural networks (CNNs) that allow for end-to-end learning of feature representations directly from the available data samples [15]. However, vanilla CNN networks have constraints and fail to

[*] Corresponding author

restore realistic images due to their limited ability to model long-range pixel dependencies[34, 48, 50].

To address the limitations, recent work from the image restoration domain has leveraged transformer-based deep architecture with self-attention (SA) [37] mechanisms to tackle image restoration effectively. Typically, these attention blocks play a crucial role in capturing long-range pixel interactions, allowing the deep architecture in parallelization and effective representation learning. It is worth noting that the SA in Transformers has adopted from the natural language tasks and illustrates state-of-the-art (SOTA) performance in high-level vision tasks. However, such a self-attention-based transformer model with large spatial window sizes increased the computational complexity of the network and failed to perform as expected on the dataset with limited samples, as shown in Fig. 1.

To tackle the issues of the existing image restoration methods, we introduced a novel deep network with dual encoder-decoder. Our proposed network leverages illumination mapping to boost the spatial luminance information inspired by the Retinex theory [16]. To the best concern, this is the first open literature work exploring the feasibility of illumination boosting for generic image restoration. In addition to that, our model learns to reconstruct an intermediate output by utilizing the first encoder-decoder. Later, we correlate our reconstructed image with the input using a cross-attention module [28]. Our second encoder-decoder block learns to refine the intermediate reconstruction from the correlated features to produce plausible enhanced images. Apart from the architecture, we introduce a perceptual optimization strategy, including a luminance-chrominance loss to maintain the color luminance consistency in the enhanced images. Experimental results illustrate that our proposed method can handle numerous image restoration tasks and outperform the existing methods with significantly less trainable parameters. Fig. 2 illustrates the performance of the proposed method in diverse image restoration tasks. Our main contributions are as follows:

- We proposed a novel Dformer that leverages illumination guidance with MCA in double encoder-decoder architecture. Our method is guided by a novel perceptual loss to perform evenly on image restoration tasks.
- We extensively studied the applicability of our method in numerous challenging image restoration tasks like shadow removal, LLIE and BCIE.
- We outperformed the SOTA image restoration tasks in fidelity score along with fewer trainable parameters. Our method ranks among the top solutions in numerous tracks, including shadow removal in the NTIRE'24 challenge [36], without applying score-boosting techniques like ensembling.

2. Background

This section details the related works in the field of image reconstruction.

2.1. Image Restoration

Image restoration methods focus on restoring original high-quality images from poor-quality versions. The SOTA CNN [12, 42, 51] models, like U-Net[24], SRCNN [8], and AR-CNN [47], utilize various processing algorithms to achieve comprehensive image restoration and reconstruction. With the continual development of computer vision, researchers have proposed vast models and techniques for image reconstruction, including GAN [10, 14, 19], diffusion models [4, 11, 30, 43], and denoising [20]. The development of vision transformers have increased the popularity of attention mechanisms for the domain of image reconstruction. For instance, UFormer[40] uses an attention mechanism to learn multi-scale features, while Restormer[50] uses an attention mechanism to capture long-range pixel interactions to restore high-quality images. Comparatively, transformers outperform CNNs [34, 48] among various tasks, being able to due to their ability to capture ranged dependencies in data efficiently through self-attention mechanisms.

2.2. Vision Transformer

Transformer architecture [37] was initially introduced for the Natural Language Processing domain, showing a significant performance in perceiving and working upon extended connections in sequential data. It found its application in the vision domain, particularly due to transformer's ability to comprehend global features dynamically, which are crucial for image recognition, segmentation, and object detection [7, 9, 32]. The main difference from well-known SOTA CNN-based models is that Vision Transformer [9] methods decompose an image into patches and use it as input to learn the relationships between them. Transformer architecture allows the capture of long-run dependencies between sequences of patches, and due to that, researchers found it reasonable to apply Vision Transformer methods for low-level vision tasks, too, like image restoration and super-resolution [22, 48]. The self-attention mechanism allows capture of relevant image regions, which can be effectively used to improve image quality and image restoration. However, computation complexity is the main difficulty of generating high-quality images, which increases quadratically with the number of image patches used.

2.3. Retinex Theory on Image Enhancement

The Retinex theory [16, 18] proposes ideas on how the human visual system perceives color and brightness of images. Rather than directly capturing the colors and brightness, the theory suggests that the brain compares different

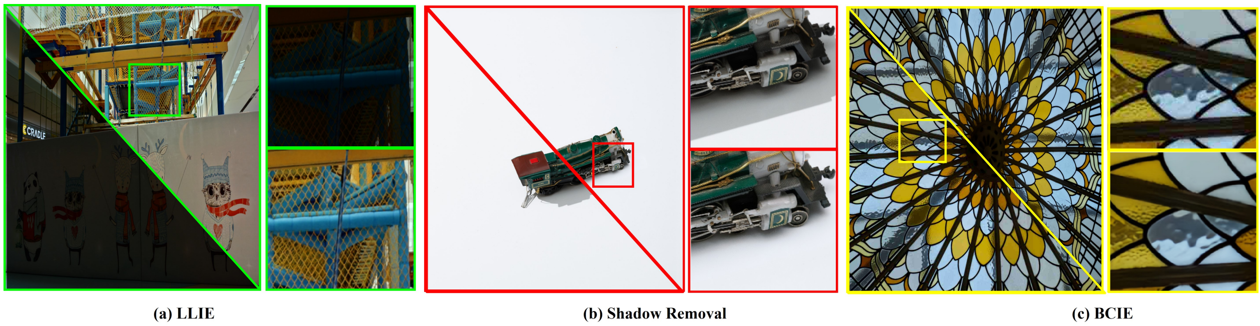

Figure 2. Examples of image restoration tasks perceived by the proposed Dformer. Our proposed method can handle numerous image restoration tasks without illustrating visually disturbing artifacts. (a) LLIE. (b) Shadow Removal. (c) BCIE

parts of the image to determine their true colors and brightness levels. In line with this theory, adjusting the brightness and color can considerably enhance the quality of images [41]. Consequently, the Retinex theory can be leveraged to develop algorithms for various low-vision tasks, including image restoration, low-light enhancement, and shadow removal. Low-light image enhancement models [41, 42] use a Retinex-based deep unfolding network to learn a network that decomposes the low-light image into reflectance and illumination layers. Retinex-based transformer architectures [5, 25, 46], are also widely used for image enhancement. RetinexFormer [5] utilizes a stage transformer to capture illumination information from images, light up the low-light regions further, and restore the corrupted areas. Additionally, diffusion models are combined with Retinex theory to produce effective results on image enhancement [46].

3. Methodology

This section provides insights into the proposed Dformer, as well as the details of perceptual optimization and training.

3.1. Dformer

Figure 3 provides an overview of the proposed Dformer model, which is designed to directly enhance low-quality images ($\mathbf{I_L}$) through a mapping function $\mathrm{D} : \mathbf{I_L} \rightarrow \mathbf{I_R}, \mathbf{I_F}$. Here, $\mathbf{I_R}$ represents the intermediate reconstruction image, while $\mathbf{I_F}$ denotes the final refined output. Our approach begins with a first encoder-decoder stage that incorporates a luminance-boosted image ($\mathbf{I_B}$) alongside the input ($\mathbf{I_L}$), drawing inspiration from the Retinex theory [16–18]. This stage aims to generate an intermediate reconstructed image ($\mathbf{I_R}$) with enhanced visual quality. Moreover, to improve feature encoding and decoding efficiency, our encoder-decoders utilize a residual illumination-guided attention block (RIGAB), inspired by the success of the illumination-guided attention block (IGAB) [5].

Furthermore, we exploit the reconstructed image $\mathbf{I_R}$ along with the input image ($\mathbf{I_L}$) to establish correlations between reconstructed images and their corresponding inputs, leveraging a multi-head self-attention mechanism (MCA block) for accelerated learning. In the subsequent refinement stage, we employ a second set of encoder-decoder blocks to enhance the quality of the reconstructed images further, ultimately yielding the final refined image $\mathbf{I_F}$. By integrating principles from the Retinex theory, attention mechanisms, and recent advancements in encoder-decoder architectures, Dformer emerges as a promising approach for effectively enhancing the visual quality of low-quality images.

Residual IGAB. Residual blocks are well-known for addressing the problem of vanishing gradients in deeper models. It also helps the deep model to extract salient features from a given tensor. On the other hand, IGAB has proven to be one of the most efficient blocks for image enhancement. In this study, we proposed incorporating IGAB in a residual manner to accelerate our restoration performance. The proposed RIGAB can be represented as follows:

$$\mathbf{F} = \mathbf{W} + G(\mathbf{W}) \tag{1}$$

Here, $G(\cdot)$ represents the consecutive IGAB and W input weights for the RIGAB block.

Multi-head cross attention. We adopted MCA from a recent study [28]. Notably, MCA has leveraged in the previous study to correlate luminance-chrominance correlation to reduce visual artifacts. However, in this study, we utilize the MCA to correlate the refined and input images. Here, we leverage a multi-head attention block to refine illumination mapping of refinement and input images along with their corresponding RGB counterparts, as shown in Fig. 4. We perceive multi-head attention as follows:

$$\text{attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

Here, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices, respectively, and d_k represents the dimensionality of the key vectors.

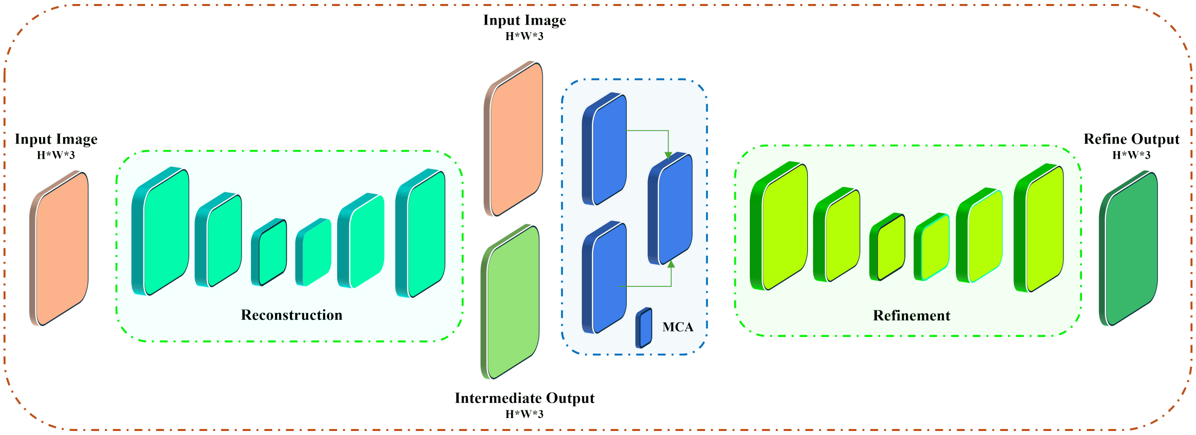

Figure 3. Overview of the proposed Dformer. The proposed model leverages illumination guidance on a dual encoder-decoder architecture to learn generic image restoration. To perceive visually plausible images, the proposed method has been guided with a multi-term perceptual loss, including reconstruction loss, regularized feature loss, and luminance-chrominance loss.

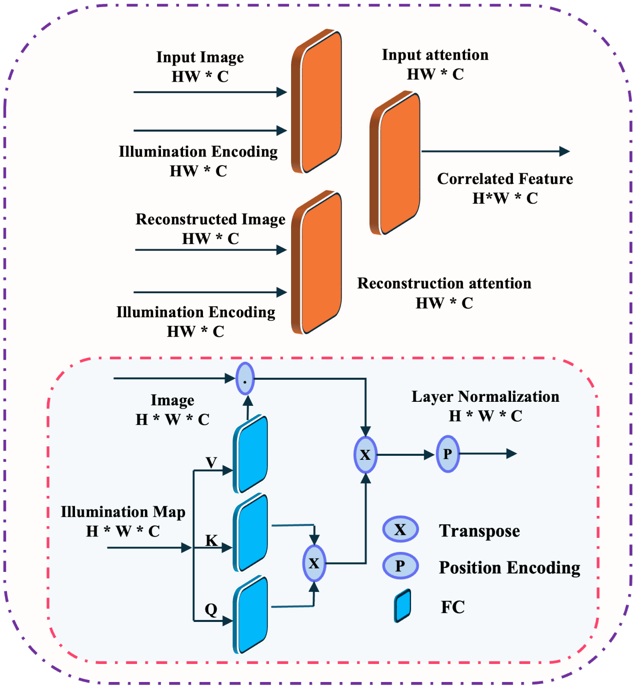

Figure 4. Overview of the multi-head cross attention block. The MCA correlates reconstructed and input images to feed the second encoder-decoder for effective image restoration.

3.2. Perceptual Optimization

In this context, $\mathcal{L}_T$ represents the proposed multi-term objective function. Its purpose is to enhance the reconstructed image's perceptual quality by considering various factors such as details, texture, color, etc.

Reconstruction loss. The L1-norm is recognized for its effectiveness in generating sharper images [26, 27, 52]. Consequently, it has been adopted to compute the pixel-wise reconstruction error as follows:

$$\mathcal{L}_R = \parallel \mathbf{I_G} - \mathbf{I_F} \parallel_1 \tag{3}$$

Here, $\mathbf{I_G}$ and $\mathbf{I_F}$ present the ground truth image and refine output of $\mathrm{D}(\mathbf{I_L})$ respectively.

Regularized feature loss (RFL).: VGG-19 feature-based loss functions aim to improve a reconstructed image's perceptual quality by encouraging it to have identical feature representations like the reference images. Typically, such activation-map loss functions are represented as follows:

$$\mathcal{L}_{FL} = \lambda_P \times \mathcal{L}_{VGG} \tag{4}$$

Where $\mathcal{L}_{VGG}$ can be extended as follows:

$$\mathcal{L}_{VGG} = \frac{1}{H_j W_j C_j} \parallel \psi_t(\mathbf{I_G}) - \psi_t(\mathbf{I_F}) \parallel_1 \tag{5}$$

Here, ψ and j denote the pre-trained VGG network and its j^{th} layer.

In Equation 4, λ_P represents the regulator controlling the feature loss. Traditionally, setting the value of this regulator is crucial, and improper tuning can harm the reconstruction process [39]. To overcome this challenge, we implement a total variation regularization (λ_R) for λ_P, as recommended by research [1, 31]. This adjustment allows for more flexibility in determining the value of λ_R, enhancing the reconstruction process. The expression for λ_R is as follows:

$$\lambda_R = \frac{1}{H_j W_j C_j} \parallel \Delta H \parallel + \parallel \Delta W \parallel \tag{6}$$

Here,$\parallel \Delta W \parallel$ and $\parallel \Delta H \parallel$ presents the summation of the gradients in the vertical and horizontal directions.

Overall, we perceive RFL loss as follows:

$$\mathcal{L}_{RFL} = \lambda_R \times \mathcal{L}_{VGG} \tag{7}$$

Here, λ_R represents the total variation regularization.

Luminance-Chrominance Loss. The proposed Dformer utilizes illumination mapping to focus on luminance information for efficient reconstruction. However, the illumination-boosted images are prone to produce color inconsistency in the reconstructed images. To address this limitation, we adopted luminance-chrominance loss as a part of our perceptual optimization [28]. We calculated the per-pixel distance between the reconstructed and reference luminance-chrominance as follows:

$$\mathcal{L}_{LC} = \| \mathbf{I_G} - \mathbf{I_F} \|_1 \tag{8}$$

Total Perceptual loss. The final perceptual objective function ($\mathcal{L}_T$) has calculated as follows:

$$\mathcal{L}_T = \mathcal{L}_R + \mathcal{L}_{RFL} + \lambda_{LC}.\mathcal{L}_{LC} \tag{9}$$

Here, λ_{LC} represents regularization coefficient for $\mathcal{L}_{LC}$. Throughout this study, we tuned λ_R = 0.3.

3.3. Training Details

We trained our model across three tasks, all linked to NTIRE 2024 Challenges: Image Shadow Removal, Blind Compressed Image Enhancement, and Low-Light Enhancement. We further utilize three datasets $D = \{D_{shadow}, D_{blind}, D_{low}\}$ sampled from NTIRE 2024 Challenges training datasets [2, 23, 33, 36, 44] and their subsequent ground-truth pairs $G = \{G_{shadow}, G_{blind}, G_{low}\}$. The D_{shadow} and G_{shadow} both consist of 1000, 1920 $\times$ 1440 images. For each image in the G_{blind} dataset, there are 10 images of different compression levels, ranging from 10 to 90, available in D_{blind}. There are 800 images in G_{blind} of various sizes. The D_{blind} dataset is then converted into D_{blindC} by introducing a random cropping mechanism, with crop size 128 $\times$ 128. The D_{low} and G_{low} datasets consist of 230 images.

Apart from the data preparation, we leverage Adam optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.99$ for training our Dformer. Also, we tuned the learning rate of 1×10^{-4} while initiating the training process. We developed a custom dataloader that selects a pool of 4 images. The samples from D_{shadow} were cropped into 512×512 patches and resized to 320×320 squares. The training for D_{shadow} dataset reached 95,000 steps and lasted approximately 48 hours. The training for D_{blindC} dataset reached 110,000 steps and lasted approximately 64 hours. The training for D_{low} consisted of 100,000 steps. The hardware used featured an NVIDIA A100 GPU, 32GB RAM, and a 16-core CPU.

All models were trained using the same D.

4. Experiments

This section discusses the experimental setup for evaluating the proposed method's performance compared to other state-of-the-art methods. Next, we assess performance based on 3 NTIRE'24 Challenge datasets [23, 36, 44] for image restoration: BCIE, Shadow removal, and low-light enhancement. Finally, an ablation study is done to see each component's importance in the proposed method.

4.1. Comparison with SOTA Methods

We compared our proposed Dformer and SOTA image restoration methods, including MIRNet[49], UFormer[40], and Restormer[50]. Additionally, we explored the feasibility of applying Retinexformer[5] to image restoration tasks. Notably, Dformer was inspired by RetinexFormer[5], and we assessed its effectiveness in generic restoration tasks beyond just LLIE. To ensure fairness in comparison, we utilized a three-stage Retinexformer[5] instead of its one-stage variants. All compared methods were retrained for restoration tasks using recommended hyperparameters and the same dataset.

The performance of the deep learning methods was summarized using standard evaluation metrics such as PSNR and SSIM. Additionally, we evaluated the models based on perceptual metrics like LPIPS [26], which calculates the distance between image patches; a lower LPIPS value indicates more similar image patches. This comprehensive evaluation approach provides insights into the comparative effectiveness of the different restoration methods across various evaluation criteria.

4.1.1 Shadow removal

Model	SSIM	PSNR	LPIPS
RetinexFormer [5]	21.42	0.8711	0.3297
UFormer [40]	21.83	0.8762	0.3011
Restormer [50]	22.30	0.8929	0.2266
MIRNet [49]	21.30	0.8788	0.2977
Ours	**23.85**	**0.8973**	**0.2219**

Table 1. Comparison between existing restoration models and proposed Dformer for Shadow removal dataset . Our Dformer outperforms the existing methods in all evaluation metrics by a notable margin.

Next, we observe our model's performance compared to RetinexFormer [5], UFormer [40], Restormer [50], and MIRNet [49] on the Shadow removal dataset. Table 1 clearly shows that based on SSIM, PSNR, and LPIPS [26] scores, our model outperforms all other state-of-the-art models, achieving 23.84, 0.8972, and 0.2219, respectively.

According to Fig. 5, our proposed model demonstrates a unique capability to accurately distinguish between texts

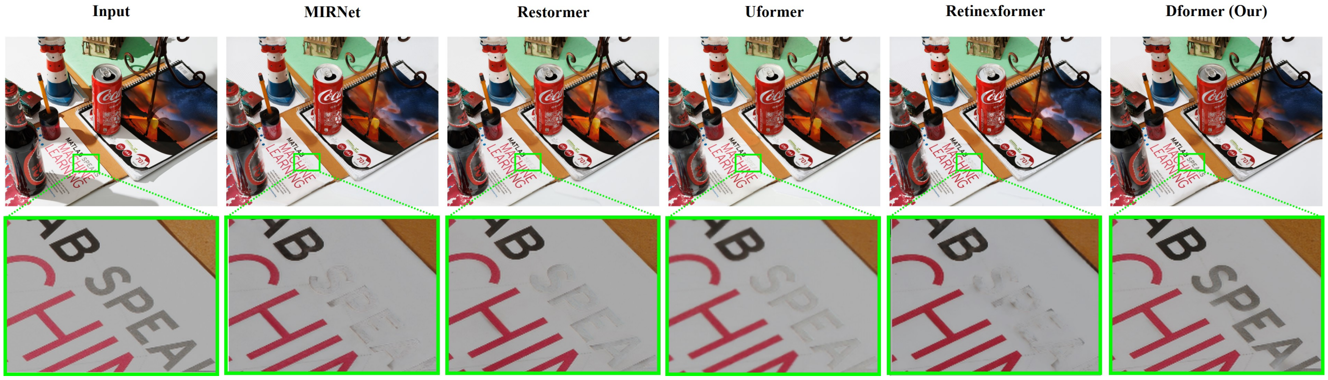

Figure 5. Example of shadow removal from image scene with multiple objects. The existing restoration method failed to recover shadow-free images with complex object structures. Also, these methods cannot distinguish objects with similar color consistency, like shadows. From left to right, input image, MIRNet[49] output, Restormer [50], UFormer [40], Retinexformer[5], Dformer (ours).

and shadows. Since the writings are in black, most models confuse them for being a part of shadow. We found the existing methods illustrate significant limitations in distinguishing shadows and objects with gray color tones. Consequently, these methods are prone to mitigate such shadow-free regions despite understanding the scene's context. In contrast, our proposed Dformer with MCA and illumination strategy helps our model to understand such critical and complex scenarios. Notably, our feature correlation with input images helps our second encoder-decoder to produce refined output for better visual representation.

4.1.2 Low-light enhancement

Our model is compared with 4 state-of-the-art models: RetinexFormer [5], UFormer [40], Restormer [50], MIRNet [49] on Low light enhancement dataset. The results were reported based on 2 different real-world low-light benchmark datasets, LOL-v1 [41], and LOL-v2 [45]. Table 2 demonstrates the quantitative evaluation of image restoration and the LLIE method on real-world LLIE. Models were evaluated on PSNR, SSIM, and LPIPS [26], and our method outperforms existing methods for both datasets. Compared to existing models, It performs significantly better on the SSIM score, outperforming by at least 0.2. Same as for the Blind dataset, average results were also reported, showing that our method performs better based on all 3 metrics for Low Light Enhancement datasets.

Apart from the quantitative evaluation, we also perform a subjective assessment of the NTIRE24 challenge testing set. Fig. 6 illustrates the LLIE performance proposed and existing restoration methods. The proposed method can produce visually plausible images with better denoising performance. Also, our perceptual optimization helped our Dformer to maintain color consistency while performing challenging LLIE.

4.1.3 Blind Compressed Image Enhancement

Table 3 shows results on NTIRE Challenge 2024 BCIE dataset based on PSNR, SSIM and LPIPS. We compare our method with RetinexFormer[5], UFormer[40], Restormer[50] and MIRNet[49] for 3 levels of compression: 20%, 40%, 60%.

At 20% compression level, our model outperforms all other models by at least 0.8 dB on the PSNR score. It also outperforms them based on the SSIM score, showing the highest value among all models - 0.9226. Meanwhile, the LPIPS [26] score shows the lowest value of 0.2683, indicating better performance than RetinexFormer[5], UFormer[40], Restormer[50], and MIRNet[49] on the Blind dataset.

The same trend is observed with 40% and 60% image compression, as our model outperforms other state-of-the-art models based on PSNR, SSIM, and LPIPS [26] scores. The best performance is observed with 60% image compression, where our method reaches 35.30 dB scores on PSNR and 0.9701 on SSIM. Overall, average results for 3 levels of image compression were shown, which indicate that at any % of image compression, the proposed method performs better than previously existing methods.

Fig. 7 illustrates the visual comparison between comparing methods. Please note that BCIE was relatively easier than shadow removal and LLIE. The SOTA method can be part of the proposed method while evaluating subjectively. However, the proposed method is lighter than its counterpart and has proven efficient in making a trade-off between fidelity and computation complexity.

4.2. Ablation Study

We conducted an extensive ablation study to rigorously assess the efficacy of integrating novel components into the Dformer architecture to enhance its capabilities in im-

	LOL-v1			LOL-v2			Average		
Model	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
RetinexFormer	22.53	0.7846	0.3072	25.66	0.8261	0.2862	24.09	0.8053	0.2967
UFormer	17.08	0.7260	0.3267	18.13	0.7439	0.3153	17.60	0.7350	0.3210
Restormer	18.40	0.7220	0.3388	20.50	0.7447	0.3231	19.45	0.7334	0.3309
MIRNet	20.00	0.7305	0.3402	23.31	0.7777	0.3159	21.66	0.7541	0.3280
Ours	**22.57**	**0.8035**	**0.2917**	**27.10**	**0.8494**	**0.2636**	**24.84**	**0.8264**	**0.2777**

Table 2. Comparison between SOTA methods for LLIE image enhancement on real-world low-light benchmark dataset. The proposed method can outperform the existing methods in evaluation metrics.

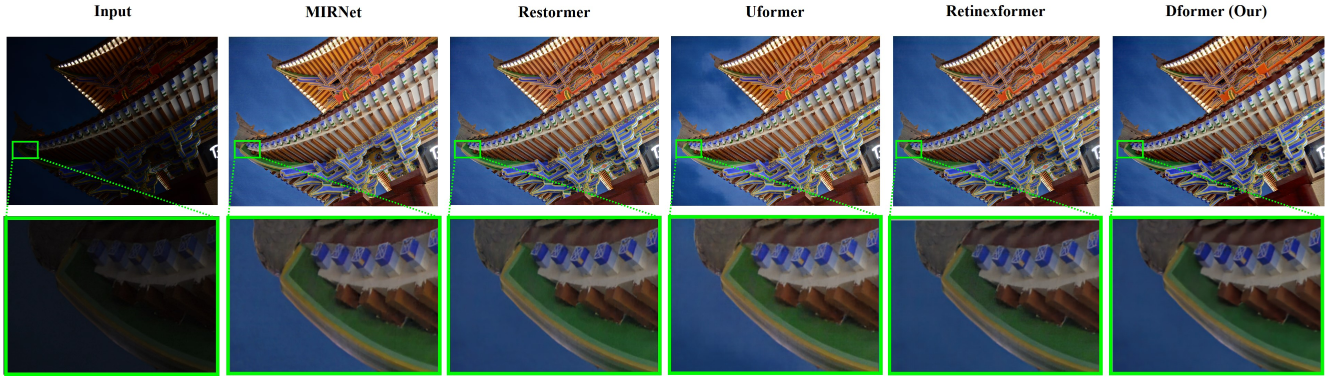

Figure 6. Comparison between SOTA restoration methods for LLIE. The proposed method can reconstruct visually plausible highlight images. Our double encoder-decoder strategy helps Dformer perform efficient denoising, while perceptual optimization helps our model to reproduce color-consistent images. From left to right, input image, MIRNet[49] output, Restormer [50], UFormer [40], Retinexformer[5], Dformer (ours).

Comp. Level	20			40			60			Average		
Model	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
RetinexFormer	31.48	0.9217	0.2914	33.92	0.9571	0.2270	35.24	0.9690	0.1855	33.55	0.9493	0.2346
UFormer	29.21	0.8910	0.3259	30.27	0.9176	0.2847	30.53	0.9238	0.2686	30.00	0.9108	0.2931
Restormer	30.71	0.9129	0.2938	32.49	0.9448	0.2367	33.23	0.9554	0.2046	32.14	0.9377	0.2450
MIRNet	31.43	0.9209	0.2966	33.93	0.9569	0.2313	35.26	0.9698	0.1891	33.54	0.9492	0.2390
Ours	**31.50**	**0.9227**	**0.2683**	**33.96**	**0.9576**	**0.2086**	**35.30**	**0.9701**	**0.1692**	**33.59**	**0.9501**	**0.2154**

Table 3. Comparison for BCIE among existing restoration models and proposed Dformer. Our proposed method outperforms the existing methods by a notable margin in quantitative evaluation. Notably, the performance of the proposed Dformer remains constant for different compression levels.

					Shadow			Blind			LowLight			Average		
Model	DED	MCA	PO	Parm. (M)	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Base	y	x	x	4.81	21.42	0.8711	0.3297	32.24	0.9341	0.2392	23.87	0.7910	0.3038	25.84	0.8654	0.2909
DDH	y	y	x	5.8	22.44	0.8671	0.2334	33.37	0.9465	0.2389	23.95	0.8020	0.2909	26.59	0.8719	0.2544
Proposed	y	y	y	5.8	**23.85**	**0.8973**	**0.2219**	**33.59**	**0.9502**	**0.2154**	**24.84**	**0.8265**	**0.2777**	**27.42**	**0.8913**	**0.2383**

Table 4. Ablation study on the proposed Dformer. Each of our proposed components has its own contribution to achieving efficient image registration.

age restoration tasks. The study systematically removed and reintroduced these components individually to evaluate their contributions. Abbreviations such as DED (double encoder-decoder), MCA (multi-head cross-attention), and PO (perceptual optimization) represent crucial elements within the proposed Dformer. By systematically reintroducing each component, we discerned their unique roles in facilitating comprehensive feature extraction, selective attention mechanisms, and perceptual alignment, culminating in enhanced image restoration capabilities. Table 4 illustrates the impact of the proposed component on Dformer architecture.

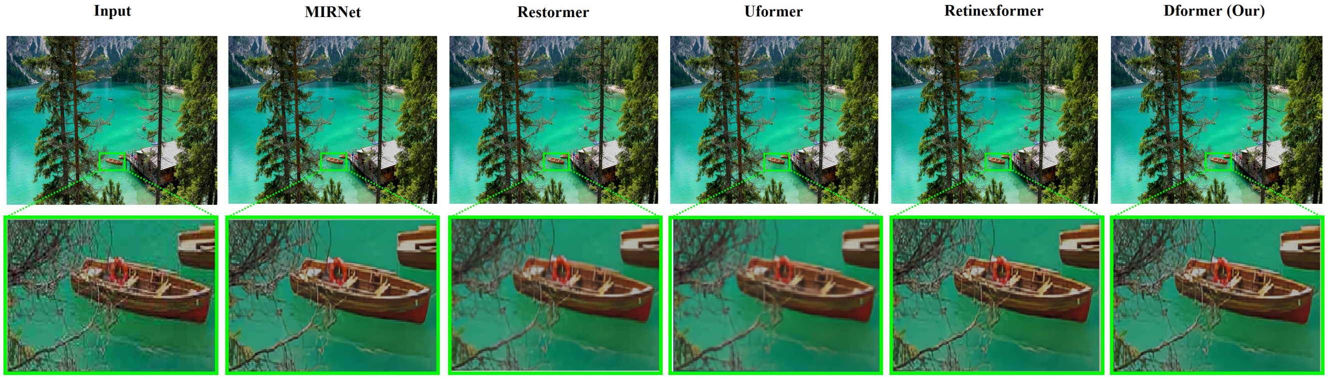

Figure 7. Real-world BCIE with image restoration methods. The proposed method can outperform the SOTA method in perceptual evaluation with fewer trainable parameters. It is best viewed in zoom and color. From left to right, input image, MIRNet[49] output, Restormer [50], UFormer [40], Retinexformer[5], Dformer (ours)

Our proposed DED architecture ensures salient feature extraction and reconstruction by incorporating encoding and decoding stages within the Dformer architecture. This design enables the model to capture intricate image details effectively. Additionally, integrating MCA mechanisms enhances the model's ability to focus on spatial image regions, prioritizing information across multiple heads for optimal reconstruction. Furthermore, incorporating PO techniques refines the restored output images by considering perceptual characteristics. Overall, the findings underscore the critical roles of DED, MCA, and PO components in our proposed Dformer architecture for efficient and effective image restoration tasks.

4.3. Discussion

The Dformer model we propose shows significant improvement over existing image restoration methods. Our approach uses only 5.8 million trainable parameters, six times less than the current SOTA MIRNet [49] and four times fewer than the Restormer [50]. It is worth noting that the well-known RetinexFormer [5] inspires our encoder-decoder blocks. RetinexFormer has proven to be one of the most effective models for trainable parameter efficiency, with less than 2 million parameters for a single-stage model. In addition to such an efficient encoder-decoder design, our illumination boosting technique for generic image restoration helps the Dformer achieve top-tier restoration performance without requiring large-sized window attention. Additionally, our dual encoder-decoder strategy, combined with MCA, allows the proposed method to delve more deeply into salient feature extraction. Moreover, incorporating perceptual optimization enhances the model's effectiveness by boosting performance without adding any noticeable increase in inference complexity.

The Dformer demonstrates notable performance in desktop environments. However, its potential for application in edge computing still needs to be discovered. It is worth noting that image restoration tasks have a widespread use case on edge devices. Particularly, generic vision tasks like object detection, stereo matching, segmentation, etc., on edge devices substantially suffer due to image degradation [29]. An efficient and optimized method for such hardware can help improve the performance of generic vision tasks on edge platforms. Thus, evaluating and optimizing Dformer for low-power devices could be an interesting future direction. Additionally, the practicability of the proposed Dformer in combined enhancement tasks like joint demosaicing and denoising is yet to be explored. We planned to study the scope of generic image restoration on joint enhancement tasks and its practicability in edge devices through a future study.

5. Conclusion

This study proposed an illumination-guided double encoder-decoder deep network for addressing challenging image restoration tasks. Our Dformer comprises a multi-head attention block to perceive cross-attention between intermediate reconstructed and input images. We also proposed a novel perceptual optimization strategy to learn image restoration efficiently. We have demonstrated the practicability of our proposed method through extensive experimentation across various challenging tasks, including shadow removal, LLIE, and BCIE. Our method outperformed the state-of-the-art image restoration methods with fewer trainable parameters. Despite promising results in desktop environments, the proposed method's performance and deployment challenges still need to be discovered. We planned to study the practicability of the proposed Dformer in an edge environment in the foreseeable future.

References

[1] SM A Sharif, Rizwan Ali Naqvi, and Mithun Biswas. Beyond joint demosaicking and denoising: An image processing pipeline for a pixel-bin image sensor. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 233–242, 2021. 4

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5

[3] Stuti N Ahuja and Seema Biday. A survey of satellite image enhancement techniques. *Int. J. Adv. Innov. Res.(IJAIR)*, 2 (8), 2013. 1

[4] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019. 2

[5] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023. 3, 5, 6, 7, 8

[6] Gang Cao, Yao Zhao, Rongrong Ni, and Xuelong Li. Contrast enhancement-based forensics in digital images. *IEEE transactions on information forensics and security*, 9(3): 515–525, 2014. 1

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[12] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 2

[13] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 536–537, 2020. 1

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[15] Furkan Kınlı, Sami Menteş, Barış Özcan, Furkan Kıraç, Radu Timofte, Yi Zuo, Zitao Wang, Xiaowen Zhang, Yu Zhu, Chenghua Li, et al. Aim 2022 challenge on instagram filter removal: methods and results. In *European Conference on Computer Vision*, pages 27–43. Springer, 2022. 1

[16] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 2, 3

[17] Edwin H Land. Recent advances in retinex theory. In *Central and Peripheral Mechanisms of Colour Vision: Proceedings of an International Symposium Held at The Wenner-Gren Center Stockholm, June 14–15, 1984*, pages 5–17. Springer, 1985.

[18] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971. 2, 3

[19] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[20] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 2

[21] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3867–3876, 2019. 1

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2

[23] Xiaoning Liu, Zongwei Wu, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, and Radu Timofte. NTIRE 2024 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 1, 5

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[25] Junxiang Ruan, Xiangtao Kong, Wenqi Huang, and Wenming Yang. Retiformer: Retinex-based enhancement in transformer for low-light image. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[26] Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2):912–923, 2018. 4, 5, 6

[27] SMA Sharif, Rizwan Ali Naqvi, and Woong-Kee Loh. Two-stage deep denoising with self-guided noise attention for multimodal medical images. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2024. 1, 4

[28] SMA Sharif, Prokash Sarkar, Abidin Zain Ul, and Rizwan Ali Naqvi. Illuminating darkness: Enhancing real-world low-light scenes with smartphone images. Under review, 2024. 2, 3, 5

[29] S M A Sharif, Azamat Myrzabekov, Nodirkhuja Khudjaev, Roman Tsoy, Seongwan Kim, and Jaeho Lee. Learning optimized low-light image enhancement for edge vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 1, 8

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[31] David Strong and Tony Chan. Edge-preserving and scale-dependent properties of total variation regularization. *Inverse problems*, 19(6):S165, 2003. 4

[32] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[33] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 1

[36] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei WU, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 1, 2, 5

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[38] An Gia Vien, Hyunkook Park, and Chul Lee. Dual-domain deep convolutional neural networks for image demoireing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 470–471, 2020. 1

[39] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 4

[40] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, 2022. 1, 2, 5, 6, 7, 8

[41] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 3, 6

[42] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2022. 2, 3

[43] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023. 2

[44] Ren Yang, Radu Timofte, et al. NTIRE 2024 challenge on blind enhancement of compressed image: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 1, 5

[45] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 6

[46] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 3

[47] Ke Yu, Chao Dong, Chen Change Loy, and Xiaoou Tang. Deep convolution networks for compression artifacts reduction. *arXiv preprint arXiv:1608.02778*, 2016. 2

[48] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 2

[49] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020. 5, 6, 7, 8

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 2, 5, 6, 7, 8

[51] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1, 2

[52] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 4