electronics

MDPI

Article

Enhanced Deblurring for Smart Cabinets in Dynamic and Low-Light Scenarios

Yali Sun [1,†], Siyang Hu [1,†], Kai Xie [1,*], Chang Wen [2], Wei Zhang [3] and Jianbiao He [4]

[1] School of Electronic Information and Electrical Engineering, Yangtze University, Jingzhou 434023, China; 2023720752@yangtzeu.edu.cn (Y.S.); 2023710670@yangtzeu.edu.cn (S.H.)
[2] School of Computer Science, Yangtzeu University, Jingzhou 434023, China; 400100@yangtzeu.edu.cn
[3] School of Electronic Information, Central South University, Changsha 410083, China; csuzwzbn@csu.edu.cn
[4] School of Computer Science, Central South University, Changsha 410083, China; jbhe@mail.csu.edu.cn
* Correspondence: 500646@yangtzeu.edu.cn; Tel.: +86-136-9731-5482
† These authors contributed equally to this work.

Abstract: In this paper, we propose a novel method to address dynamic blur and low-light issues in smart cabinets, which is named the MIMO-IMF (Multi-input Multi-output U-Net Integrated Motion Framework). This method combines a Frequency-Domain Adaptive Fusion Module (FDAFM), built on the blind deblurring framework MIMO-UNet (Multi-input Multi-output U-Net), to improve the capture of high-frequency information and enhance the accuracy of blur region recovery. Additionally, a low-light luminance information extraction module (IFEM) is designed to complement the multi-scale features of the FDAFM by extracting valuable luminance information, significantly improving the efficiency of merchandise deblurring under low-light conditions. To further optimize the deblurring effect, we introduce an enhanced residual block structure and a novel loss function. The refined multi-scale residual block, combined with the FDAFM, better restores image details by refining frequency bands across different scales. The New Loss Function improves the model's performance in low-light and dynamic blur scenarios by effectively balancing luminance and structural information. Experiments on the GOPRO dataset and the self-developed MBSI dataset show that our method outperforms the original model, achieving a PSNR improvement of 0.21 dB on the public dataset and 0.23 dB on the MBSI dataset.

Keywords: localized dynamic blur; low-light scenes; MIMO-UNet; frequency-domain attention; luminance information

Academic Editor: Chiman Kwan

Received: 27 December 2024
Revised: 21 January 2025
Accepted: 24 January 2025
Published: 25 January 2025

Citation: Sun, Y.; Hu, S.; Xie, K.;Wen, C.; He, J.; Zhang, W. Enhanced Deblurring for Smart Cabinets in Dynamic and Low-Light Scenarios. *Electronics* **2025**, *14*, 488. https://doi.org/10.3390/electronics14030488

1. Introduction

Motion blur refers to the loss of image details and the blurring of object edges caused by rapid object motion or camera shake [1]. This phenomenon not only degrades image quality but also impacts the accuracy and efficiency of subsequent tasks, such as target detection [2,3] and image recognition [4,5]. In the field of unmanned retailing, the smart retail dynamic cabinet (referred to as the smart cabinet) is an emerging intelligent device designed for automatic goods recognition and detection. However, the rapid picking of goods by consumers and the hardware frame rate limitations of cameras make it challenging to capture fast actions in real time, resulting in blurred areas that are difficult to recover. Moreover, the use of smart cabinets in low-light conditions, such as nighttime or poorly lit environments, further complicates the accurate recognition and categorization of goods. Existing image deblurring methods [6–9], while effective in multi-scale feature extraction and global blur removal, are not well suited to address the challenges posed by smart

cabinets in dynamic and variable scenarios. Specifically, these methods struggle to handle the localized motion blur caused by rapid hand movements during goods retrieval and fail to adequately restore luminance and details in low-light environments, limiting their deblurring capabilities in such contexts.

Therefore, we propose a deblurring method, MIMO-IMF, based on an improved MIMO-UNet. This method is designed to simultaneously address irregular local dynamic blurring and deblurring challenges in low-light environments. During the feature extraction of blurred images, a luminance information extraction channel is incorporated to provide meaningful luminance features for the model. The subsequently generated light attention maps are fused with multi-scale features to enhance the deblurring process. Additionally, the deblurring effect is optimized by introducing an improved residual block structure and a novel loss function, ensuring high performance and real-time response across low light and normal light scenes of smart cabinets. Experiments on one standard and one custom dataset will be conducted to validate the method's effectiveness and applicability. The main contributions of this work are as follows:

1. **Low-light brightness information extraction module (IFEM):** This module enhances the deblurring efficiency of smart cabinets in low-light environments by extracting meaningful brightness information.
2. **Frequency-Domain Adaptive Fusion Module (FDAFM):** By incorporating Fourier Transform and frequency-domain attention mechanisms, this module strengthens the ability to capture high-frequency information, significantly improving the recovery accuracy in blurred regions.
3. **Improved Multi Residual Block (MRB) and New Loss Function:** By introducing multi-scale residual connections and an adaptive weight adjustment mechanism, combined with frequency-domain attention to refine the selected frequency bands, the restoration capability is enhanced, further optimizing the deblurring performance.

The rest of this study is organized as follows: Section 2 reviews previous studies on deblurring tasks. Section 3 provides a detailed explanation of the proposed deblurring method. Section 4 introduces the datasets used and presents the experiments and analyses conducted with the proposed method. Finally, Section 5 summarizes the findings and concludes this study.

2. Related Study

In this section, we review previous approaches to image-based deblurring tasks. Existing research on deblurring has primarily focused on image restoration, with various algorithms proposed for different scenarios. Section 2.1 discusses conventional image deblurring methods, while Section 2.2 focuses on deep learning-based approaches. Sections 2.2.1–2.2.3 detail deblurring methods based on neural networks, generative adversarial networks, and Transformers, respectively.

2.1. Traditional Image Deblurring Methods

Traditional methods generally rely on the combination of physical models and prior knowledge of images. These methods were extensively studied and applied in the early stages of image deblurring research. Early frequency-domain methods, such as the Wiener filter, used inverse filtering techniques to remove motion blur but had limited applicability due to their sensitivity to noise. With technological advancements, spatial domain optimization methods gradually became mainstream, particularly those based on image priors. Fergus et al. were the first to propose a deblurring approach based on blur kernel estimation, combining natural image statistics with a Bayesian inference method [10]. This work sparked widespread adoption of prior models, such as image sparse priors and Total

Variation (TV). For example, Levin et al. proposed a sparse gradient prior that effectively preserves the edge information of an image [11].

In the 2010s, image prior-based deblurring methods experienced further advancements [12–14]. For instance, deblurring algorithms based on L_0 norm minimization, as proposed in the literature [15], significantly enhanced image edge sharpness by promoting image sparsity. However, such traditional methods rely heavily on hand-designed prior models and often struggle to handle complex and dynamic blurred scenes. Additionally, the high computational complexity of these methods limits their applicability in real-time scenarios, making them less effective in practical, complex environments [16]. In contrast, deep learning methods demonstrate superior performance when addressing real-world non-uniform blurring.

2.2. Deep Learning-Based Deblurring Method

In recent years, deep learning techniques have achieved significant advancements in the field of image deblurring. The literature [17] provided a comprehensive summary of deep learning-based deblurring techniques, covering various methods (e.g., DAE, GAN, and multi-scale networks), commonly used loss functions, and evaluation datasets. The study demonstrates that deep learning methods substantially enhance deblurring performance through diverse network architectures and optimization strategies. However, challenges remain in addressing non-uniform blurring and achieving real-time performance. To further explore the specific implementations of deep learning deblurring techniques and their advantages, this section summarizes the innovative approaches and practical applications of deep learning in image deblurring tasks, focusing on three key categories: neural networks, generative adversarial networks, and Transformer-based networks.

2.2.1. Neural Network Based Deblurring Methods

In recent years, deep learning-based methods have taken center stage, demonstrating remarkable results in image deblurring. Compared to traditional methods, deep learning algorithms excel in capturing complex blurring features through data-driven approaches and leverage their powerful learning capabilities to restore images. Nah et al. proposed one of the earliest algorithms for deblurring using convolutional neural networks (CNNs), known as multi-scale convolutional neural networks (MSCNNs). This model significantly enhances blurred images by progressively recovering them at different scales, achieving notable deblurring effects [1]. Sun et al. introduced an end-to-end deblurring model based on convolutional networks, which demonstrated initial deblurring capabilities and laid the foundation for subsequent deep learning approaches [18].

Subsequently, Tao et al. proposed a deblurring algorithm utilizing a recursive convolutional neural network, which combines convolutional layers with a recurrent structure to progressively restore image clarity in a recursive manner [7]. Gao et al. [8] designed the PSS-NSC network, which employs parameter sharing and nested skip connections to effectively enhance deblurring efficiency and performance. Cho et al. [9] proposed MIMO-UNet, which reduces computational complexity while maintaining high-precision deblurring through a multiple-input-multiple-output structure and an asymmetric feature fusion module. Although these methods have shown promising results in multi-scale feature extraction and global blur removal, they fall short in restoring brightness and fine details in images. Furthermore, they struggle to handle local motion blur caused by fast-moving objects, making them less effective in addressing specific challenges in smart cabinet scenarios.

2.2.2. Generative Adversarial Network-Based Deblurring Methods

Generative adversarial network (GAN)-based deblurring methods have made significant progress in the field of image deblurring in recent years. The GAN model employs adversarial learning between a generator and a discriminator, where the generator aims to produce realistic deblurred images while the discriminator distinguishes between the generated and real images. This adversarial process continuously optimizes the deblurring effect.

One of the classic works applying GANs to image deblurring was proposed by Kupyn et al., known as DeblurGAN [6]. DeblurGAN introduces a conditional GAN structure that employs perceptual loss to maintain the perceptual quality of an image while generating a clear version. This approach achieved excellent results in terms of visual quality and particularly excelled in recovering fine details. Subsequently, DeblurGAN-v2 further enhanced the generative network structure by introducing a multi-scale network design, improving the ability to handle large-scale blurred scenes. Additionally, it increased model efficiency through a lightweight network architecture, making it suitable for real-time applications [19].

However, such methods have limitations. For instance, the training of GANs is often unstable and prone to mode collapse, which can lead to texture distortion or blurred details in the generated images. Yang et al. [20] proposed a method that utilizes latent spatial mapping noise and features of a neural network's shallow-to-deep feature-generating GAN, enabling the recovery of both global structures and local details of blurred images. Nevertheless, the limitations of GAN-based approaches, such as instability and generalization issues, have yet to be fully addressed. In summary, while GAN-based deblurring methods have significantly improved visual quality, there remains substantial room for improvement in terms of stability and general applicability.

2.2.3. Transformer-Based Deblurring Methods

With the development of the self-attention mechanism and Transformer architecture, the Restormer model proposed by Zamir et al. [21], based on the Transformer framework, employs a global self-attention mechanism to effectively capture long-range dependencies, demonstrating excellent results in deblurring and denoising tasks for high-resolution images. Further advancements in this area include an efficient frequency-domain-based Transformer model proposed in the literature [22], which recovers high- and low-frequency details in images through a frequency-domain feedforward network. This approach enhances the effectiveness of the feedforward network in deblurring tasks.

Another study published in the same year [23] combines local feature preservation with global representation learning. SharpFormer achieves more accurate motion blur removal by utilizing a local feature enhancement module to better capture details while leveraging the Transformer architecture to process global information.

Although these methods achieve remarkable results across various deblurring scenarios, the computational complexity inherent in Transformer architectures reduces real-time performance in smart cabinet applications. This limitation impacts the consumer shopping experience, as high computational costs can delay operations.

3. Method

In addressing the challenges of localized dynamic blurring and low-light conditions in smart cabinets, it is important to recognize that these issues primarily stem from the limitations imposed by the frame rate of the camera hardware and the rapid removal of goods by consumers. To mitigate these challenges, this paper introduces a novel method termed MIMO-IMF, the structure of which is depicted in Figure 1. Building upon the MIMO-

UNet architecture, we propose a frequency-domain attention-enhanced adaptive feature fusion module aimed at recovering dynamic blur in merchandise images. Additionally, we have designed a low-light brightness feature extraction module to enhance the performance of smart cabinets in low-light environments.

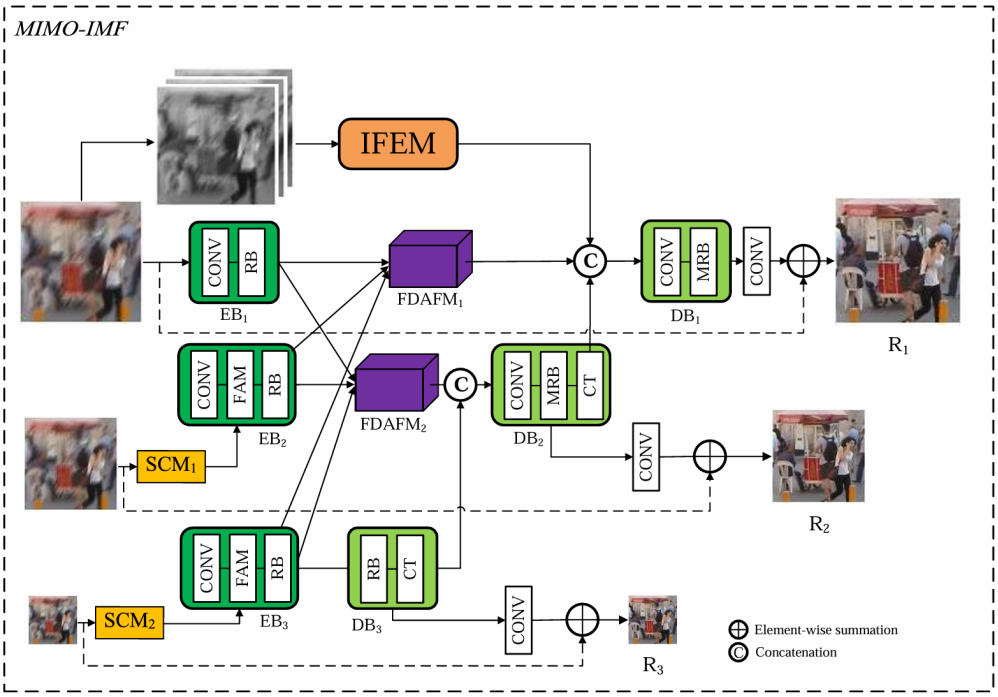

Figure 1. Algorithmic process.

3.1. Introduction to MIMO-UNet

MIMO-UNet, a multi-scale image blind deblurring model, the network architecture is divided into four main parts: shallow feature extraction module (SCM), encoder (EB), asymmetric feature fusion module (AFF), and decoder (DB). The structure is shown in Figure 2. Through the design of multi-scale feature inputs and outputs, the model is able to efficiently capture the feature information of the blurred region and recover it.

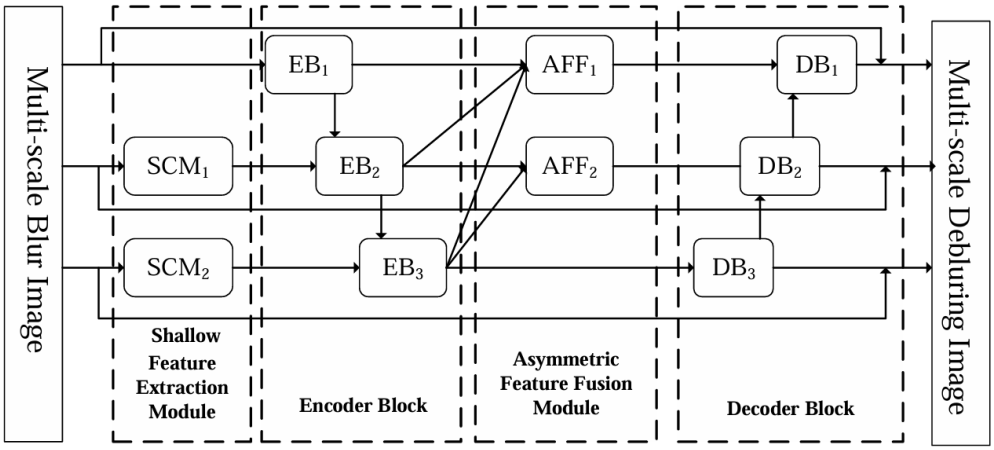

Figure 2. The Structure of MIMO-UNet.

3.2. Low-Light Brightness Information Extraction Module (IFEM)

In practical application scenarios of smart cabinets, lighting conditions often have a large variation, with most physical container placements being in low-light or dim environments. This makes it challenging to capture sufficient luminance information

using traditional image blind deblurring methods, thereby affecting the performance of deblurring and object recognition. The module architecture is shown in Figure 3; the input full resolution RGB image I_1 is separated into three channels, red (R), green (G), and blue (B), each channel contains luminance information at different wavelengths, and the purpose of the separation is to extract the specific luminance features in each channel, which may have a different impact on the luminance recovery in the image. After channel separation, the luminance information of the R, G, and B channels are weighted and fused to generate the intermediate luminance feature L_0, which is represented as the overall lighting characteristics of the image and can be expressed as

$$L_0 = \alpha R + \beta G + \gamma B \tag{1}$$

where α, β, and γ are the weighting coefficients of the channels, respectively, which are used to measure the degree of the channel's contribution to the luminance. The fused luminance features L_0 are subjected to two convolution operations and processed with the ReLu activation function to obtain a light attention map L_1 for highlighting key luminance regions in the image. Enabled by low-light information extraction, we are able to obtain more feature information in scene-constrained as well as hardware-constrained smart cabinets to help subsequent feature fusion.

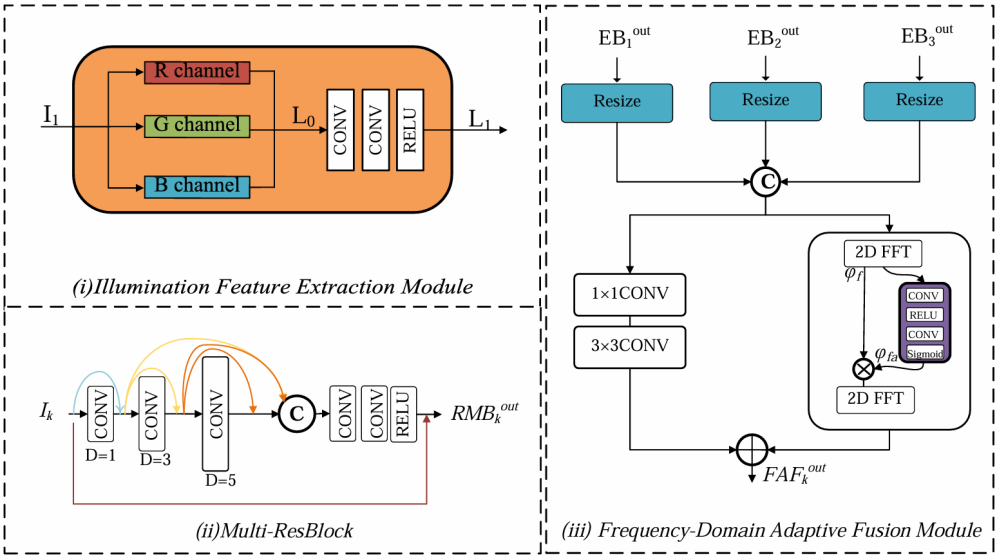

Figure 3. The component structure of MIMO-IMF.

3.3. Frequency-Domain Adaptive Fusion Module (FDAFM)

In Section 3.1, we introduced the core asymmetric feature fusion module (AFF). However, it struggles to handle the dynamic ambiguity caused by consumers' rapid picking up of goods in smart cabinet scenarios. To address this, we propose a frequency-domain attention-enhanced adaptive feature fusion module called FDAFM to replace the AFF module in the original model. The core idea of the FDAFM is to retain the spatial domain details preserved by the AFF during feature fusion while introducing a Fourier Transform and frequency-domain attention mechanism. This dual-path approach transfers spatial domain features to the frequency domain to extract high-frequency and low-frequency information. The two paths are then fused to enrich the overall feature representation, effectively addressing the loss of high-frequency information caused by dynamic blurring and further improving the deblurring performance.

Specifically, the multi-scale features extracted by the encoder (EB) are denoted as $E_1^{out}, E_2^{out}, E_3^{out}$, where E_1^{out} represents the original resolution output feature E_2^{out} for the 1/2

resolution output feature and E_3^{out} for the 1/4 resolution output feature. We resize them to unify the sizes of features with different resolutions. In the original model, the resized features are directly concatenated along the channel dimension to produce a fused feature containing rich multi-scale information, denoted as

$$F_{\text{concat}} = \text{C}(EB_1^{out}, EB_2^{out}, EB_3^{out}). \tag{2}$$

Afterward, we apply a 1×1 convolution and a 3×3 convolution to F_{concat} in the spatial domain branch to obtain the processed features F_{conv} .

Simultaneously, F_{concat} is fed into the frequency domain branch, where the fused features are transformed from the spatial domain to the frequency domain. A frequency-domain attention enhancement mechanism is introduced to strengthen the capture of high-frequency information, which is subsequently converted back to the spatial domain for fusion output. Specifically, the input features F_{concat} are transformed into the frequency domain using Fast Fourier Transform (FFT) to obtain their frequency domain representation $F_{\mathcal{F}}$.

$$F_{\mathcal{F}} = \text{FFT}(F_{\text{concat}}) \tag{3}$$

The frequency domain representation contains the spectral information of the image, where the high-frequency components correspond to the feature details. Subsequently, the frequency domain features $F_{\mathcal{F}}$ are processed by a small convolutional network $\mathcal{A}$, consisting of convolutional layers and nonlinear activation functions, to generate frequency-domain attention weight maps ϕ_{f_a}:

$$\phi_{f_a} = \mathcal{A}(F_{\mathcal{F}}) \tag{4}$$

Then, the frequency-domain attention map ϕ_{f_a} is combined with the frequency domain feature $F_{\mathcal{F}}$ by performing an element-wise multiplication to generate the frequency-domain attention-enhanced feature F_{enh}, which can be expressed as

$$F_{enh} = F_{\mathcal{F}} \odot \phi_{f_a} \tag{5}$$

Where $\odot$ denotes the element-wise product, and F_{enh} is the enhanced feature. Finally, the outputs from both branches—the enhanced frequency-domain features F_{enh} and the convolved spatial domain features F_{conv}—are combined by element-wise addition to obtain the final fused feature output:

$$FAF_k^{out} = F_{\text{conv}} + F_{enh} \tag{6}$$

This output contains not only spatial information at different scales but also compensates for missing high-frequency information, thereby providing a richer feature representation for the subsequent decoder.

After the FDAFM, the fused features I_1 and I_2 can be represented as

$$\begin{cases} I_1 = F_{\Sigma_1}\left(\text{FAF}_1^{\text{out}}, L_1\right) \\ I_2 = F_{\Sigma_2}\left(\text{FAF}_2^{\text{out}}, DB_3(x)\right) \end{cases} \tag{7}$$

Where $F_{\Sigma_1}, F_{\Sigma_2}$ represent two concatenation operations, L_1 is the light information obtained through the IFEM, and $\text{DB}_3(x)$ denotes the features acquired by the decoder $\text{DB}_3(x)$.

3.4. Multi-scale Residual Block (MRB)

Inspired by the cascaded residual blocks used in the defogging domain in the CasDyF-Net paper [24], which implements image defogging by cascading inflated convolutional

networks, we similarly design the Multi-scale Residual Block (MRB) for image deblurring. After the fusion of asymmetric feature frequency-domain features in Section 3.3, the MRB is applied in the decoder to refine the frequency band information selected by the frequency-domain attention mechanism. The MRB is defined as

$$M_i = \mathcal{MRB}(I_k^{D=1}, I_k^{D=3}, I_k^{D=5}). \tag{8}$$

It consists of three 3×3 convolutional layers with different dilation rates, D = 1, 3, 5, and introduces a 1×1 convolutional block to merge information across different scales.

As shown in Figure 1, we continue to use the ResBlock in the encoder to maintain a lightweight model structure. However, our focus in this paper is on the decoder. We keep the ResBlock in DB_3 at the low-resolution stage because, at this stage, the feature map resolution is low, and the frequency-domain advantage of the MRB is not as pronounced. In this case, we mainly rely on time-domain convolutions to enhance global information. In contrast, we use the MRB module in DB_1 and DB_2 at the high-resolution stage, where its multi-scale convolutional capabilities are leveraged to filter the bands selected by the previous frequency-domain attention mechanism, thereby enhancing detail restoration and optimizing the deblurring effect.

3.5. Loss Function

In the loss function design, we add the loss terms of the LFEM and FDAFM to the original model's multi-scale content loss function. This aims to evaluate the model's performance using more precise loss metrics, thereby further improving the deblurring effect. The content loss function L_c is defined as follows:

$$L_c = \sum_{k=1}^{K} \frac{1}{t_k} \|\hat{S}_k - S_k\|_1, \tag{9}$$

where k is the number of layers, and t_k represents the total number of elements. For the frequency domain loss, we use a multi-scale frequency domain reconstruction loss to preserve the high-frequency details in the image:

$$L_{freq} = \sum_{k=1}^{K} \frac{1}{t_k} \|\mathcal{F}(\hat{S}_k) - \mathcal{F}(S_k)\|_1, \tag{10}$$

where $\mathcal{F}$ denotes the Fast Fourier Transform (FFT), which converts the image from the spatial domain to the frequency domain. In the low-light feature extraction module (LFEM), we introduce the low-light feature loss to ensure the network effectively extracts luminance information in low-light environments. This can be expressed as

$$L_{LFEM} = \sum_{k=1}^{K} \frac{1}{t_k} \|L(\hat{S}_k) - L(S_k)\|_1, \tag{11}$$

where L represents the luminance feature extraction operation, which measures the difference between the luminance information extracted by the network and the ground truth luminance information. Based on the above loss terms, the final loss function for the proposed network is defined as

$$L_{total} = L_c + \lambda_1 L_{freq} + \lambda_2 L_{LFEM}, \tag{12}$$

where λ_1 and λ_2 are the weighting factors for each loss term.

4. Experiments

In this section, we evaluate the proposed methodology and compare it with state-of-the-art methods using both publicly available and custom datasets.

4.1. Experimental Dataset and Parameter Settings

4.1.1. Commodity Fuzzy Image Dataset

Currently, there is a lack of publicly available datasets dedicated to the motion blurring of goods in smart cabinets. Additionally, training deep learning networks typically requires a large amount of paired data to capture more detailed features. Moreover, external factors make it difficult to consistently pick up goods at the same speed and angle under identical conditions, such as position and illumination. To address these challenges, we synthesize a commodity motion blur dataset using the method proposed in [6], which simulates random motion trajectories. The synthesized images effectively mimic motion blur caused by object movement, providing a practical solution when real motion blur images are difficult to obtain. The process can be modeled as follows:

$$B(x, y) = I(x, y) * K, \tag{13}$$

$$K = S(x). \tag{14}$$

Here, x is a random trajectory vector generated by a Markov randomization process; $B(x, y)$ denotes the image after motion blurring; $I(x, y)$ represents the clear image; K is the blurring kernel; $S(x)$ is the result of the computation of x using a sub-pixel interpolation algorithm. To adapt the dataset specifically for smart cabinet scenarios, we created a custom dataset consisting of 12 merchandise categories, each with 100 clear images acquired using an RGB camera mounted on the smart cabinet. We then applied the aforementioned algorithm to generate motion blur images, resulting in 1200 clear-blurred image pairs. These pairs were divided into 80% (960) for the training set and 20% (240) for the test set, ensuring no overlap between the two subsets. This dataset, named **MBSI (Motion Blur Shop Images)**, is tailored for smart cabinet scenarios and is used to evaluate the applicability of our method in real-world environments. Examples of the dataset are shown in Figure 4.

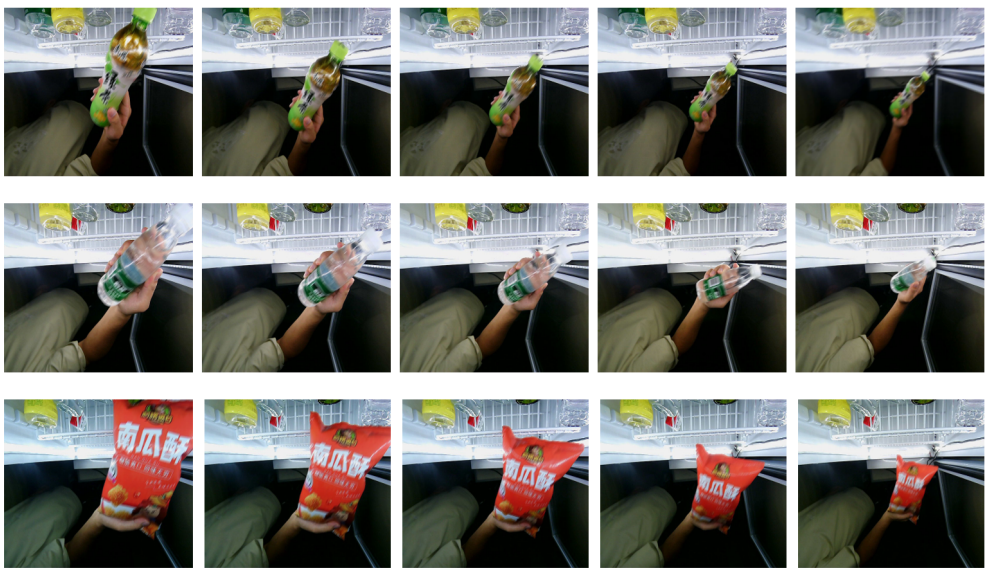

Figure 4. MBSI dataset examples.

4.1.2. Parameter Setting

To ensure the validity and generalizability of our method, we first use the GoPro dataset for experiments. The GoPro dataset is a widely used benchmark in the deblurring domain, containing 3214 pairs of high-quality sharp-blurred images, with 2103 pairs used for training and 1111 pairs for testing. Its extensive coverage of various motion blur scenarios makes it suitable for validating the performance and robustness of deblurring methods.

In each training iteration, we randomly sample four images from the GoPro dataset and crop them to 256 $\times$ 256. Data augmentation is performed with a 50% chance of horizontal flipping for each patch. To ensure sufficient convergence, the network is trained for 3000 epochs with an initial learning rate of 1×10^{-4}, which decays by a factor of 0.5 every 200 epochs.

The choice of the GoPro dataset allows us to evaluate our method's performance in a standardized and comparable setting, as it is commonly used in the field. Despite its differences from the smart cabinet application scenario, the physical properties of motion blur (e.g., caused by object movement and camera frame rate limitations) are similar. Therefore, the GoPro dataset provides a valuable benchmark for demonstrating the general effectiveness of our approach.

To further validate the applicability of our method in specific real-world settings, we use the MBSI dataset as a complementary evaluation. The MBSI dataset simulates smart cabinet scenarios, providing a tailored assessment of our method's ability to handle practical challenges such as diverse lighting conditions, object categories, and motion patterns. The experiments on the MBSI dataset follow the same training settings as the GoPro dataset, with 3000 epochs, the same initial learning rate, and a similar decay schedule.

In terms of evaluation, we use standard metrics such as the PSNR and SSIM. PSNR (Peak Signal-to-Noise Ratio) is a common metric for measuring image restoration quality, where higher PSNR values indicate better recovery. The SSIM (Structural Similarity Index) measures the structural similarity between the restored image and the ground truth, with higher SSIM values indicating better preservation of image structures and textures. Additionally, we perform ablation studies (Section 4.2) and module validation experiments (Sections 4.3 and 4.4) to assess the contribution of each module and validate the effectiveness of the IFEM and FDAFM in different conditions.

The combination of experiments on the GoPro and MBSI datasets enables a comprehensive performance evaluation, demonstrating both the generalizability and specificity of our proposed method.

4.2. Ablation Study

We conducted a series of ablation experiments to assess the impact of the IFEM, FDAFM, and MRB on model performance under low-light and normal lighting conditions. The results, shown in Table 1, highlight the performance improvements achieved by adding different modules.

The baseline model, without any additional modules, yields a PSNR of 31.68 on the GoPro dataset and 23.37 on the MBSI dataset. When the FDAFM is added, the PSNR increases to 31.93 on GoPro and 23.46 on MBSI, demonstrating its ability to enhance high-frequency detail recovery. The FDAFM is particularly effective in capturing sharp edges, leading to significant improvements in the PSNR.

When combined with the MRB module, the FDAFM further improves the PSNR to 31.94 (GoPro) and 23.48 (MBSI). This combination results in an even better recovery of image details, especially in sharp edges.

Adding the IFEM module, which focuses on extracting luminance features, improves the model's performance under low-light conditions. In this case, the PSNR reaches 32.02 on GoPro and 23.83 on MBSI, showing that the IFEM module contributes effectively to enhancing deblurring in low-light scenarios. When the IFEM and FDAFM are combined, the model achieves a PSNR of 31.96 on GoPro and 23.51 on MBSI, further improving performance.

Finally, the full model, incorporating all three modules (IFEM + FDAFM + MRB), achieves the best performance, with a PSNR of 32.03 on GoPro and 23.99 on MBSI. This combination effectively captures high-frequency information and extracts luminance features, improving deblurring performance in both low-light and normal lighting conditions.

In summary, the FDAFM provides the most substantial improvement in the PSNR by enhancing high-frequency detail recovery, while the MRB module further refines edge sharpness. The IFEM contributes by extracting critical luminance information, especially under low-light conditions. The combination of all three modules results in the most significant performance gains, demonstrating the complementary nature of each module.

Table 1. Performance comparison on GoPro and MBSI datasets.

	IFEM	**FDAFM**	**MRB**	**GoPro Dataset**		**MBSI Dataset**	
				PSNR	**SSIM**	**PSNR**	**SSIM**
Baseline				31.68	0.903	23.37	0.824
w/only IFEM	✓			31.80	0.913	23.44	0.849
w/only FDAFM		✓		31.93	0.916	23.46	0.851
IFEM + MRB	✓		✓	31.94	0.914	23.48	0.850
FDAFM + MRB		✓	✓	32.02	0.921	23.83	0.859
IFEM + FDAFM	✓	✓		31.96	0.915	23.51	0.852
Ours	✓	✓	✓	32.03	0.923	23.99	0.861

4.3. Supplementary Validation of IFEM Effectiveness

To rigorously assess the effectiveness of the proposed IFEM in handling low-light deblurring scenarios, we conducted extensive experiments using the MBSI dataset's extended subset (low-light simulation). This extended subset was generated by simulating real-world low-light environments through targeted manipulations, including brightness reduction and noise injection, to emulate challenging lighting conditions encountered in smart cabinet applications.The experiments conducted in this study focus on addressing the challenges of image deblurring in low-light and normal lighting conditions encountered in smart cabinet applications. Comparative evaluations were carried out across three representative images using five different methods in Figure 5: MIMO-IMF (our proposed method), MIMO-Plus [9], Restormer [21], MPRNet [25], and DeblurGANv2 [19]. The experimental results reveal that the MIMO-IMF consistently outperforms competing methods, delivering superior image clarity under low-light conditions. The enhanced performance of our model is attributed to the unique design of the IFEM, which effectively extracts channel-specific luminance features and fuses them to generate an adaptive luminance attention map. This mechanism emphasizes critical luminance regions in the image, enabling better feature enhancement and integration in downstream processes. Consequently, the IFEM equips the MIMO-IMF with a distinctive advantage in recovering structural and visual details, ensuring higher accuracy in identifying merchandise under adverse lighting conditions, a challenge inadequately addressed by existing deblurring models.

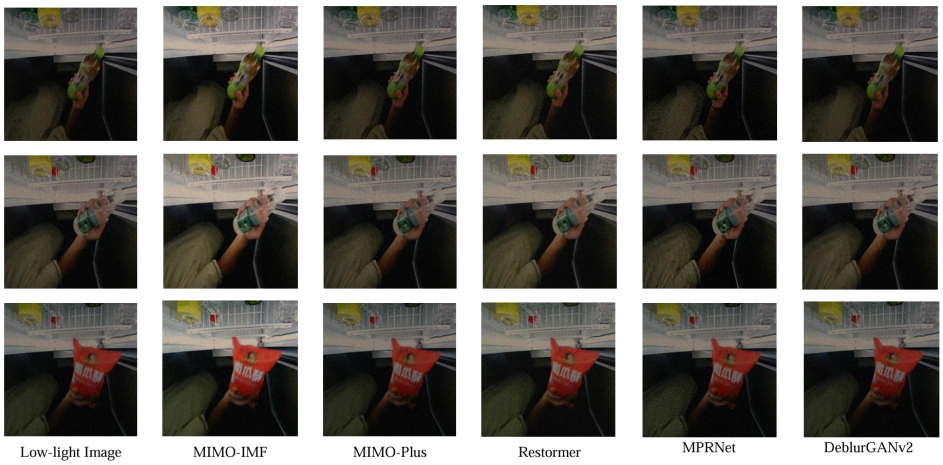

Figure 5. Comparison of deblurring results in MBSI (**low-light**).

4.4. Supplementary Validation of FDAFM Effectiveness

In this section of the experiments, we validate the effectiveness of the FDAFM through image comparisons. Figures 6 and 7 illustrate the visualization results of heatmaps on the GoPro and MBSI datasets, with and without the FDAFM, respectively.

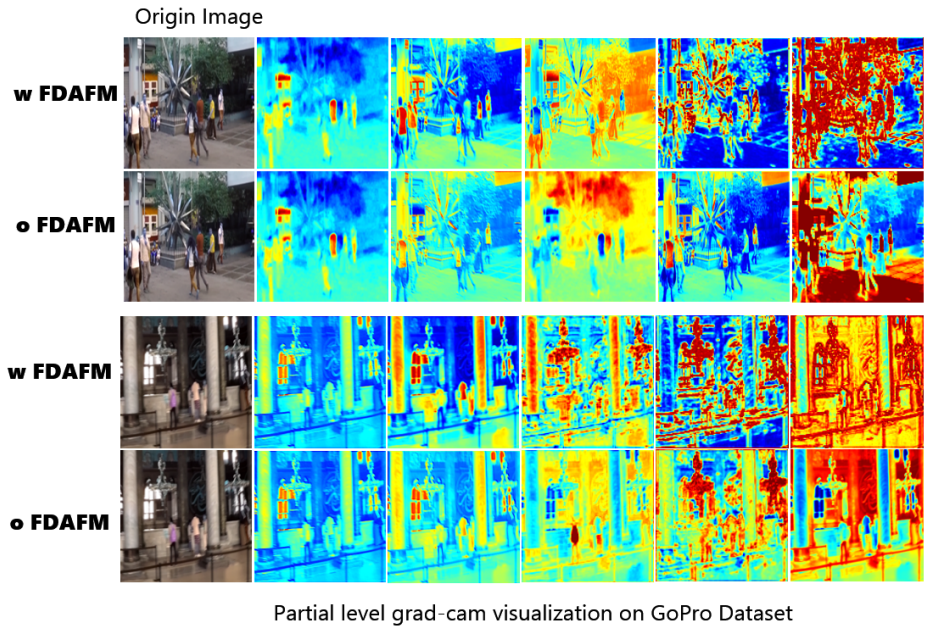

Figure 6. Visualization on GoPro dataset.

From the visualization results on the GoPro dataset (Figure 6), the addition of the FDAFM leads to a more concentrated focus area for the network. This allows the network to effectively capture key details of the image and gradually focus its attention on the high-frequency regions, thereby improving the deblurring effect. This improvement is particularly evident in the edge regions of the image. In contrast, the network without the FDAFM demonstrates more dispersed attention regions and struggles to efficiently capture the critical features required for deblurring in complex scenes.

Similarly, the experimental results on the MBSI dataset (Figure 7) further corroborate the effectiveness and rationality of the FDAFM. Without the FDAFM, the network exhibits scattered attention and fails to accurately capture object details and edges. After integrating the FDAFM, the network clearly focuses on the edges and details of the image subjects, facilitating a better deblurring process.

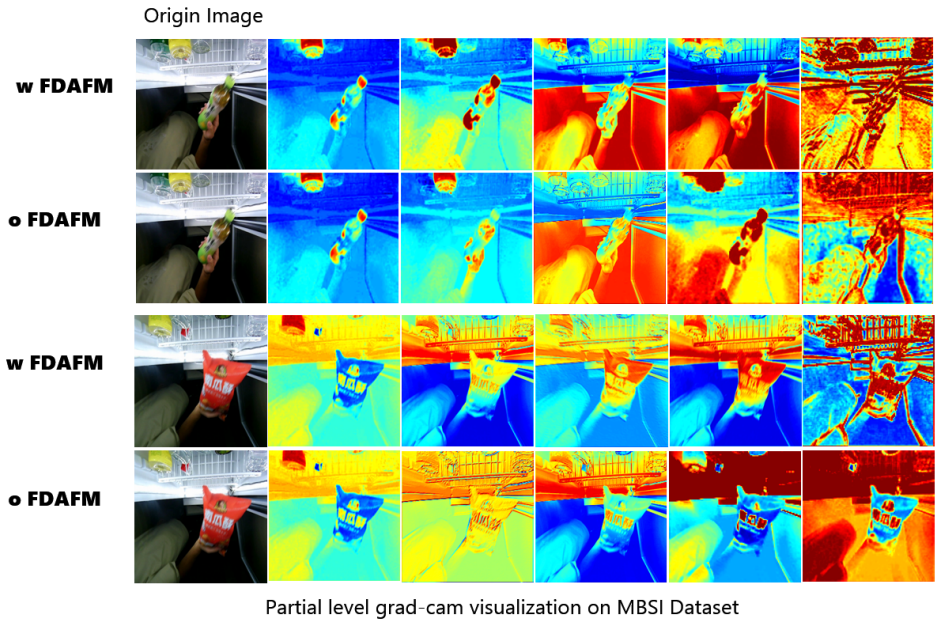

Figure 7. Visualization on MBSI dataset.

Two images from each dataset were selected for heatmap visualization. From the results, it is evident that the FDAFM enhances the network's attention to high-frequency features by implementing a frequency-domain attention mechanism. This enables the network to more precisely identify and process the locally blurred regions within the image. This experiment clearly demonstrates the superiority of the FDAFM in addressing image blurring, providing robust support for its application in deblurring tasks across various scenarios.

4.5. Comparison Experiment

To validate the performance of the algorithmic models proposed in this paper for blind deblurring tasks, a series of comparative experiments are designed, covering the current mainstream deblurring algorithms. We have replicated several deblurring models to ensure the fairness and accuracy of the performance tests. Table 2 presents the quantitative comparison of different models on the GoPro and homemade MBSI datasets, using the PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) as the primary evaluation metrics.

Table 2. Quantitative comparison on GoPro and MBSI datasets.

Method	GoPro		MBSI	
	PSNR	**SSIM**	**PSNR**	**SSIM**
DeblurGAN [6]	28.70	0.858	21.50	0.604
DeblurGAN-v2 [19]	29.55	0.934	22.36	0.618
Zhang et al. [26]	29.19	0.931	22.10	0.610
SRN [7]	30.26	0.934	23.50	0.705
DBGAN [27]	31.10	0.942	23.90	0.720
DMPHN [28]	31.20	0.940	24.00	0.728
SPAIR [29]	32.06	0.953	24.20	0.740
MIMOUNetPlus [9]	32.45	0.957	24.34	0.734
MPRNet [25]	32.66	0.959	24.42	0.743
Restormer [21]	32.92	0.961	24.55	0.757
Ours	**32.98**	**0.962**	**24.87**	**0.770**

4.5.1. Quantitative Assessment

As shown in Table 2, our method outperforms other compared methods on both datasets in terms of the PSNR and SSIM. On the GoPro dataset, our model achieves a PSNR of 32.98 and an SSIM of 0.9623, outperforming other classical models such as MPRNet and Restormer. On the MBSI dataset, the algorithm also performs well, achieving a PSNR of 24.87 and an SSIM of 0.7697, significantly improving both detail recovery and global information retention.

In contrast, the DeblurGAN family of models, while showing some deblurring effects, performs poorly on the MBSI dataset, especially in recovering merchandise details. Newer algorithms such as Restormer and MPRNet, although improved in deblurring ability, still fall short in handling some complex scenes, such as distant buildings and merchandise text detail recovery.

4.5.2. Comparison of Subjective Visual Effects

Figure 8 shows the visual comparison of the algorithm proposed in this paper with other methods in a real-world scene. Five models are compared and analyzed using two images selected from the GoPro and MBSI datasets, respectively.

For the GoPro dataset, the selected images are from a traffic scene. In the comparison, the MPRNet and Restormer algorithms successfully remove background blur but at the cost of some detail loss, especially in the clarity of distant buildings and road signs. In contrast, the algorithm proposed in this paper not only effectively removes motion blur but also preserves finer details, particularly around vehicle edges and billboard text, resulting in a clearer restoration.

For the MBSI dataset, which contains various trading goods scenarios, the MPRNet and Restormer methods show slight underperformance in areas with higher brightness, where text on bottled goods and bags remains partially blurred. In contrast, the algorithms in this paper excel at recovering fine textual details, especially under low and normal light conditions, demonstrating a stronger ability to process fine details.

Figure 8. Deblurred results on the GoPro dataset and MBSI dataset.

4.5.3. Synthesize and Analyze

From the above quantitative and qualitative comparisons, it is clear that the MIMO-IMF method proposed in this paper offers significant advantages in the blind deblurring task. Especially in scenes with complex backgrounds or small objects, the proposed method not only improves deblurring accuracy but also effectively preserves image detail. Furthermore, the low-light brightness information extraction module (IFEM) performs particularly well in deblurring tasks under low-light conditions, effectively enhancing the clarity of

product images in smart cabinet scenarios. The results demonstrate the applicability and robustness of the algorithms in real-world applications such as smart cabinets, showcasing their potential for high performance and real-time response.

5. Conclusions

In this study, we proposed the MIMO-IMF method. First, we designed a low-light brightness information extraction module (IFEM) to effectively capture luminance information in low and normal light conditions. Second, addressing hardware constraints and the local motion blur of goods caused by consumers' rapid picking behavior, we introduced a Frequency-Domain Adaptive Fusion Module (FDAFM), which enhances the model's ability to capture high-frequency information, improving the deblurring performance. Finally, we redesigned a multi-scale residual block tailored for deblurring tasks and introduced a novel loss function that incorporates both luminance and high-frequency information, further optimizing the deblurring effect. Overall, compared to other state-of-the-art models, the method proposed in this paper is better suited for real-world applications, such as smart cabinet systems.

Author Contributions: Conceptualization, Y.S., S.H., and K.X.; methodology, Y.S. and S.H.; software, Y.S. and S.H.; writing—original draft preparation, Y.S. and S.H.; writing—review and editing, Y.S. and S.H.; visualization, Y.S., S.H., and J.H.; supervision, K.X.; investigation, W.Z. and C.W.; funding acquisition, J.H.;All authors have read and agreed to the published version of this manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62272485 and 62373372

Data Availability Statement: The data are unavailable due to privacy.

Acknowledgments: We gratefully acknowledge all the members who participated in this work.

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

References

1. Nah, S.; Kim, T.H.; Lee, K.M. Deep multi-scale convolutional neural network for dynamic scene deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3883–3891.
2. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276.
3. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232.
4. Shafiq, M.; Gu, Z. Deep residual learning for image recognition: A survey. *Appl. Sci.* **2022**, *12*, 8972.
5. Yang, X.; Zhang, Y.; Lv, W.; Wang, D. Image recognition of wind turbine blade damage based on a deep learning model with transfer learning and an ensemble learning classifier. *Renew. Energy* **2021**, *163*, 386–397.
6. Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; Matas, J. Deblurgan: Blind motion deblurring using conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8183–8192.
7. Tao, X.; Gao, H.; Shen, X.; Wang, J.; Jia, J. Scale-recurrent network for deep image deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8174–8182.
8. Gao, H.; Tao, X.; Shen, X.; Jia, J. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3848–3856.
9. Cho, S.J.; Ji, S.W.; Hong, J.P.; Jung, S.W.; Ko, S.J. Rethinking coarse-to-fine approach in single image deblurring. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4641–4650.
10. Fergus, R.; Singh, B.; Hertzmann, A.; Roweis, S.T.; Freeman, W.T. Removing camera shake from a single photograph. In *Acm Siggraph 2006 Papers*, Association for Computing Machinery: New York, NY, USA, 1 July 2006, pp. 787–794.
11. Levin, A.; Weiss, Y.; Durand, F.; Freeman, W.T. Understanding and evaluating blind deconvolution algorithms. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1964–1971.

12. Xu, L.; Zheng, S.; Jia, J. Unnatural l0 sparse representation for natural image deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–528 June 2013; pp. 1107–1114.
13. Guo, Y.; Ma, H. Image blind deblurring using an adaptive patch prior. *Tsinghua Sci. Technol.* **2018**, *24*, 238–248.
14. Pan, J.; Sun, D.; Pfister, H.; Yang, M.H. Deblurring images via dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2315–2328.
15. Pan, J.; Hu, Z.; Su, Z.; Yang, M.H. *l*_0-regularized intensity and gradient prior for deblurring text images and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 342–355.
16. Sheng, B.; Li, P.; Fang, X.; Tan, P.; Wu, E. Depth-aware motion deblurring using loopy belief propagation. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 955–969.
17. Zhang, K.; Ren, W.; Luo, W.; Lai, W.S.; Stenger, B.; Yang, M.H.; Li, H. Deep image deblurring: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 2103–2130.
18. Sun, J.; Cao, W.; Xu, Z.; Ponce, J. Learning a convolutional neural network for non-uniform motion blur removal. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 769–777.
19. Kupyn, O.; Martyniuk, T.; Wu, J.; Wang, Z. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8878–8887.
20. Yang, T.; Ren, P.; Xie, X.; Zhang, L. Gan prior embedded network for blind face restoration in the wild. In Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 672–681.
21. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
22. Kong, L.; Dong, J.; Ge, J.; Li, M.; Pan, J. Efficient frequency domain-based transformers for high-quality image deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5886–5895.
23. Yan, Q.; Gong, D.; Wang, P.; Zhang, Z.; Zhang, Y.; Shi, J.Q. SharpFormer: Learning local feature preserving global representations for image deblurring. *IEEE Trans. Image Process.* **2023**, *32*, 2857–2866.
24. Yinglong, W.; Bin, H. CasDyF-Net: Image Dehazing via Cascaded Dynamic Filters. *arXiv* **2024**, arXiv:2409.08510.
25. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Multi-stage progressive image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14821–14831.
26. Zhang, J.; Pan, J.; Ren, J.; Song, Y.; Bao, L.; Lau, R.W.; Yang, M.H. Dynamic scene deblurring using spatially variant recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2521–2529.
27. Zhang, K.; Luo, W.; Zhong, Y.; Ma, L.; Stenger, B.; Liu, W.; Li, H. Deblurring by realistic blurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2737–2746.
28. Zhang, H.; Dai, Y.; Li, H.; Koniusz, P. Deep stacked hierarchical multi-patch network for image deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5978–5986.
29. Purohit, K.; Suin, M.; Rajagopalan, A.; Boddeti, V.N. Spatially-adaptive image restoration using distortion-guided networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2309–2319.