

Moiré Zero: An Efficient and High-Performance Neural Architecture for Moiré Removal

Seungryong Lee^{1*}, Woojeong Baek^{2*}, Younghyun Kim²
 Eunwoo Kim³, Haru Moon³, Donggon Yoo³, Eunbyung Park^{2†}
¹Sungkyunkwan University ²Yonsei University ³Samsung Display

Abstract

Moiré patterns, caused by frequency aliasing between fine repetitive structures and a camera sensor’s sampling process, have been a significant obstacle in various real-world applications, such as consumer photography and industrial defect inspection. With the advancements in deep learning algorithms, numerous studies—predominantly based on convolutional neural networks—have suggested various solutions to address this issue. Despite these efforts, existing approaches still struggle to effectively eliminate artifacts due to the diverse scales, orientations, and color shifts of moiré patterns, primarily because the constrained receptive field of CNN-based architectures limits their ability to capture the complex characteristics of moiré patterns. In this paper, we propose MZNet, a U-shaped network designed to bring images closer to a ‘Moiré-Zero’ state by effectively removing moiré patterns. It integrates three specialized components: Multi-Scale Dual Attention Block (MSDAB) for extracting and refining multi-scale features, Multi-Shape Large Kernel Convolution Block (MSLKB) for capturing diverse moiré structures, and Feature Fusion-Based Skip Connection for enhancing information flow. Together, these components enhance local texture restoration and large-scale artifact suppression. Experiments on benchmark datasets demonstrate that MZNet achieves state-of-the-art performance on high-resolution datasets and delivers competitive results on lower-resolution datasets.

1. Introduction

Moiré patterns arise from frequency aliasing when fine repetitive structures interfere with a camera sensor’s sampling process. These artifacts occur in various scenarios, such as capturing on-screen content, textiles with fine weaves, or architectural structures with dense grids. In particular, when photographing digital displays, moiré is

caused by the interaction between the camera’s color filter array (CFA) and the display’s subpixel structure, resulting in severe color banding and rippling, which significantly degrade image quality (See Fig. 1). In industrial applications, such as display manufacturing and quality control, moiré artifacts can obscure critical defects during inspection, potentially leading to product failures. Therefore, developing efficient and robust moiré removal techniques has become urgent and essential.

Conventional methods [27, 38, 40, 50, 52, 53], which rely on decomposition, filtering, or handcrafted priors, have attempted to mitigate moiré artifacts, but their effectiveness remains limited. With the advent of deep learning, various demoiré techniques [6, 16, 17, 28, 34, 42, 45, 49, 51, 54, 55, 63, 65] have emerged, exploring frequency-domain strategies [28, 63, 64] and multi-scale architectures [6, 42, 55]. For example, MBCNN [63] employs a learnable bandpass filter to target moiré frequencies, while ESDNet [55] incorporates a semantic-aligned scale-aware module for UHD demoiré, achieving promising results.

Despite these advances, moiré removal remains challenging due to its diverse scales, complex shapes, and severe color distortions as shown in Fig. 1. High-definition images further exacerbate this issue, as moiré artifacts tend to be more intricate and varied within a single image [55]. While Transformer-based models have demonstrated promising results in image restoration by capturing long-range spatial dependencies, their high computational cost and the limited availability of large-scale training pose significant challenges for their subsequent application to moiré removal. Consequently, almost all existing approaches adopt CNN-based architectures thanks to their efficiency. However, due to their limited receptive fields, CNNs often struggle to capture global structures and unique moiré patterns, resulting in suboptimal artifact suppression across diverse scenarios. To comprehensively address these challenges, we propose a novel framework, MZNet, which integrates three complementary components to form a design well-suited to the characteristics of moiré patterns.

First, we introduce the Multi-Scale Dual Attention Block

*Equal contribution.

†Corresponding author.



Figure 1. Visualization of moiré patterns from the UHDM dataset. (A) shows the full image with moiré patterns, while (B), (C), and (D) present magnified regions. They illustrate the varying directions and scales of moiré within a single image. (E) is extracted from the ground truth corresponding to the same region as (D), highlighting the color distortions caused by moiré patterns.

(MSDAB), which is designed to effectively capture different scales of the moiré patterns. MSDAB in our network extracts multi-scale features from the given input and refines them using Multi-Dilation Convolution Module (MDCM) and Dual Attention Module (DAM), respectively. MDCM consists of depth-wise dilated convolutions, which expand the receptive field of the network with minimal computational cost, while DAM employs Simplified Channel Attention (SCA) [3] and Large Kernel Attention (LKA) [14] to suppress both local textures and large-scale artifacts. Given that the moiré patterns exhibit diverse shapes and orientations, we further incorporate the Multi-Shape Large Kernel Convolution Block (MSLKB) to better capture directional information of the moiré patterns. Placed at the bottleneck of our network, MSLKB utilizes depth-wise large-kernel convolutions with square, horizontal, and vertical shapes. This design not only effectively represents the various features with diverse orientations, but also avoids excessive computational cost of the network, providing a tailored solution for subsequent demoiréing.

Finally, we incorporate a Feature Fusion-Based Skip Connection (FFSC), which augments the information flow between the encoder and decoder of our network. By aggregating features extracted from all encoder levels and injecting them into each decoder level, this strategy ensures that each decoder leverages rich, multi-scale contextual information, therefore enabling superior detail reconstruction and notable artifact suppression.

Experiments on high-resolution benchmarks (UHDM, FHDMi) demonstrate that MZNet outperforms state-of-the-art demoiréing methods, achieving PSNR gains of over 0.33 dB and 0.67 dB, respectively, while maintaining a low computational cost among competitive approaches. Our contributions are as follows:

- We propose MZNet, which incorporates three specialized components for moiré removal: Multi-Scale Dual Attention Block, Multi-Shape Large Kernel Convolution

Block, and Feature Fusion-Based Skip Connection.

- MZNet demonstrates superior demoiré performance on high-resolution benchmark datasets while maintaining a low computational cost.

2. Related Work

Image Demoiréing With the advent of deep learning, various demoiré techniques have emerged, exploring frequency-domain strategies [17, 28, 63, 64] and multi-scale architectures [6, 16, 42, 55]. For example, MBCNN [63] employs a learnable bandpass filter to target moiré frequencies, while ESDNet [55] incorporates a semantic-aligned scale-aware module for UHD demoiréing, achieving promising results. Building on the 4K UHDM benchmark [55], subsequent studies introduced a coarse-to-fine pipeline [45], the memory-efficient patch-based method [49], and RAW-level approaches that suppress moiré artifacts before demosaicing [7, 51, 56]. Despite these efforts, existing demoiréing methods generally rely on the conventional CNN-based architectures, which often struggle to capture long-range moiré structures. While many previous approaches have explored multi-scale feature extraction or frequency-based processing, effectively handling the diverse shapes and scales of moiré patterns still poses significant challenges. In this work, we propose an efficient architecture by enhancing its adaptability to diverse moiré patterns, resulting in a more accurate demoiréing network.

Large Kernel Convolution Early CNN architectures [22, 24] explored large kernels, but the success of 3×3 convolutions in deep neural networks [18, 39] established them as the standard due to their efficiency. Recently, inspired by Transformer, CNNs have revisited large kernels to model long-range dependencies using depth-wise decompositions [23, 29] and large kernel attention mechanisms [13, 14]. In image restoration, large kernels have been applied for dehazing, deblurring [33, 36], and super-resolution [46]. In addition, other approaches integrate various kernel structures to jointly capture both global and local information [8, 15].

Image Restoration As deep learning advanced, CNN-based models [1, 10, 12, 57, 58, 61, 62] achieved strong performance in image restoration. Following the success of Vision Transformer (ViT) [9], Transformer-based models have been widely explored for image restoration, leveraging self-attention to effectively capture global interactions. This approach has been applied to various restoration tasks, including deblurring [21, 44, 48], denoising [2, 11, 25], and super-resolution [4, 5, 26, 32, 41]. In partic-

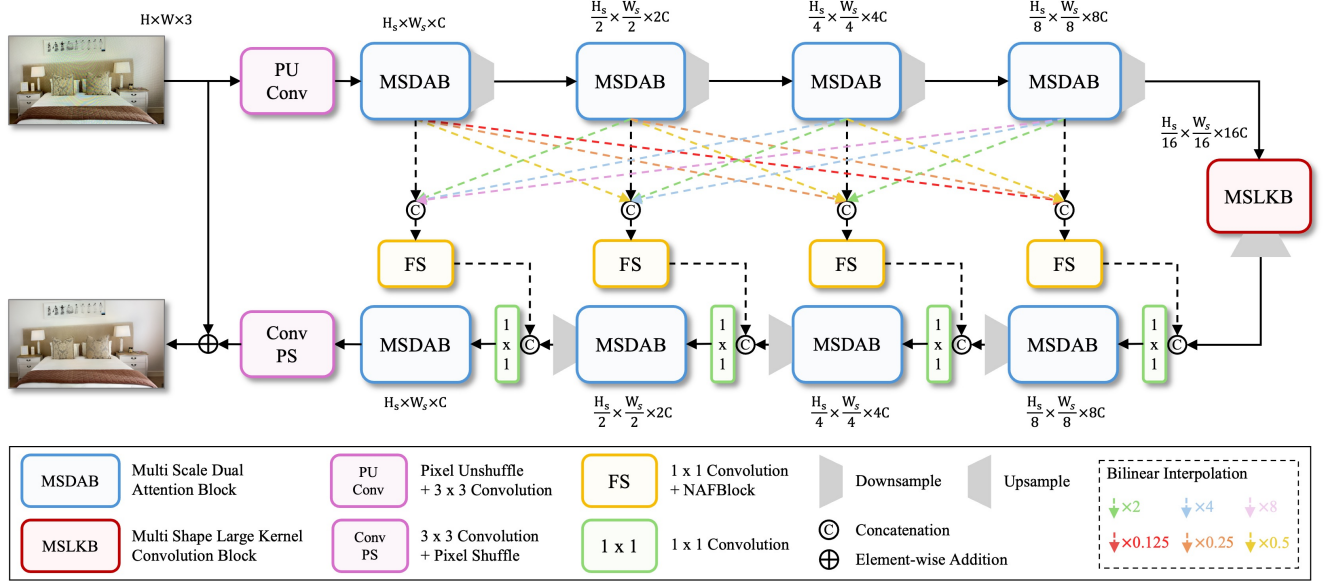


Figure 2. Overview of MZNet. The model follows a U-Net architecture with four levels of encoders and decoders. MSDAB are used in both the encoder and decoder, while the middle block incorporates a single MSLKB. FS modules in the skip connections facilitate feature fusion across scales.

ular, Restormer [59] introduced a transposed self-attention, which applies self-attention across the channel dimension, enabling efficient processing of high-resolution images. However, Transformer-based models often incur high computational costs. To address this issue, NAFNet [3] was proposed, replacing activation functions and Transformer operations with a simplified channel attention and gating mechanism, achieving both high efficiency and strong performance.

3. Method

3.1. Overall Pipeline

As illustrated in Fig. 2, our framework adopts an encoder-latent-decoder design with four levels on both the encoder and decoder sides. Given an input moiré image $I \in \mathbb{R}^{H \times W \times 3}$, we first apply a pixel unshuffle [37] operation followed by a 3×3 convolution to extract the shallow feature $F_0 \in \mathbb{R}^{H_s \times W_s \times C}$. This feature is then passed through the four-stage encoder, which progressively halves the spatial resolution while doubling the number of channels at each level. Consequently, for each level l , we obtain latent features $F_l \in \mathbb{R}^{\frac{H_s}{2^l} \times \frac{W_s}{2^l} \times 2^l C}$. Before decoding, the last latent feature F_4 is refined by our proposed Multi-Shape Large Kernel Convolution Block (MSLKB). On the decoder side, the features are upsampled by a factor of 2 at each of the four stages, while the channel dimension is reduced accordingly. At each decoder level l , the latent features from all encoders are spatially aligned to match the decoder's current resolution $\frac{H_s}{2^l} \times \frac{W_s}{2^l}$. These aligned features are then

fused through a fusion mechanism, resulting in the fused feature $\tilde{F}_l \in \mathbb{R}^{\frac{H_s}{2^l} \times \frac{W_s}{2^l} \times 2^l C}$. This feature is concatenated with the current decoder feature and passed through a 1×1 convolution to reduce the channel dimension by half before proceeding to the next level. Finally, the decoder's output features are processed through a 3×3 convolution followed by pixel shuffle upsampling to reconstruct the residual image $R \in \mathbb{R}^{H \times W \times 3}$. The clean image $\hat{I}$ is obtained by adding the residual image R to the input moiré image I .

3.2. Multi-Scale Dual Attention Block

To effectively capture and remove moiré patterns at varying scales, we propose the Multi-Scale Dual Attention Block (MSDAB), which consists of two key components: the Multi-Dilation Convolution Module (MDCM) and the Dual Attention Module (DAM). MDCM extracts multi-scale features and DAM selectively refines them through complementary attention mechanisms, ensuring effective moiré pattern removal. The input feature map is first normalized and transformed through a 1×1 convolution followed by a 3×3 depth-wise convolution, generating the initial feature representation. MDCM then processes this representation by leveraging multiple receptive fields, DAM enhances critical information, and a Feed-Forward Network (FFN) finalizes the output. The structure of MSDAB is illustrated in Fig. 3d.

3.2.1. Multi-Dilation Convolution Module

MDCM is designed to extract multi-scale features by employing multiple parallel 3×3 depth-wise convolutions with

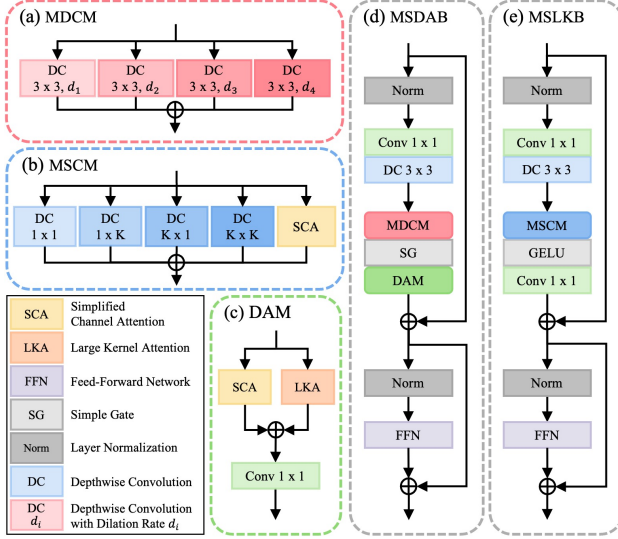


Figure 3. Illustration of the core components in MZNet. Simple Gate from NAFNet serves as an activation function.

distinct dilation rates $\{d_i\}_{i=1}^4$, which expand the receptive field to $\{2d_i+1\}_{i=1}^4$. This design allows the network to capture both fine-grained details and broader spatial structures efficiently. The module is defined as:

$$\text{MDCM}(X) = \sum_{i=1}^4 \text{DConv}_{3 \times 3, d_i}(X), \quad (1)$$

where X is an input feature and $\text{DConv}_{3 \times 3, d_i}(\cdot)$ represents a 3×3 depth-wise convolution layer with dilation factor d_i . Inspired by GoogLeNet [43], which utilizes parallel convolutional branches, our method adopts a parallel implementation of multiple convolutions with distinct dilation rates to ensure comprehensive coverage of moiré patterns while maintaining computational efficiency through depth-wise convolutions.

3.2.2. Dual Attention Module

DAM refines the multi-scale features extracted by MDCM, selectively enhancing important spatial and channel-wise information. It incorporates two complementary attention mechanisms: Large Kernel Attention (LKA) [14] and Simplified Channel Attention (SCA) [3]. LKA captures long-range spatial dependencies through an expanded receptive field, while SCA focuses on the most relevant frequency components by adaptively reweighting feature channels. The attention-enhanced features are computed as:

$$\text{DAM}(X) = \text{Conv}_{1 \times 1}(\text{SCA}(X) + \text{LKA}(X)). \quad (2)$$

First, the feature is processed by LKA and SCA modules, and these outputs are fused using a 1×1 convolution to integrate both channel-wise and spatially significant features.

By integrating LKA and SCA, the DAM module ensures that both “what is important” and “where it is important” are jointly considered, leading to more precise moiré pattern removal.

3.3. Multi-Shape Large Kernel Convolution Block

Prior works [8, 15] have shown that strip-shaped depth-wise convolutions improve the capture of elongated structures. However, despite the prevalence of such structures in moiré patterns, research on utilizing different kernel shapes for moiré removal remains underexplored. Given the varied nature of moiré artifacts, we propose the Multi-Shape Large Kernel Convolution Block (MSLKB) within the latent space of an encoder-decoder architecture to better capture the diverse spatial characteristics of moiré patterns. MSLKB integrates the Multi-Shape Convolution Module (MSCM) as its core component, applying parallel depth-wise convolutions with diverse kernel forms. The structures of MSLKB and MSCM are illustrated in Fig. 3e and Fig. 3b, respectively.

Specifically, MSCM employs kernels of size $K \times K$, $K \times 1$ and $1 \times K$ for horizontal and vertical structures, and 1×1 to refine local details. By employing a large K , MSLKB expands the receptive field, enabling the network to capture long-range moiré structures more effectively. However, since large kernels tend to incur high computational costs, we place MSLKB at the bottleneck of the network, where the feature resolution is the lowest. We also adopt SCA in MSCM to selectively enhance the most informative feature channels, complementing the extraction of diverse structural patterns. By integrating these operations in a parallel manner, MSCM is formulated as:

$$\begin{aligned} \text{MSCM}(X) &= \sum_{s \in S} \text{DConv}_{s, 1}(X) + \text{SCA}(X), \\ S &= \{K \times K, K \times 1, 1 \times K, 1 \times 1\}. \end{aligned} \quad (3)$$

By fusing the outputs of these branches, MSCM effectively integrates multi-shape representations, ensuring robust moiré pattern removal. The kernel size K of MSLKB is empirically set to the largest odd number that does not exceed the size of the bottleneck feature.

3.4. Feature Fusion-Based Skip Connection

Instead of one-to-one feature passing scheme used in U-Net [35], our method enables global feature aggregation, enriching decoder representations with multi-scale information. Let F_l be the feature map obtained from the l -th encoder level. To ensure consistency in spatial resolution across all levels, each feature map is resized using bilinear interpolation to match the resolution of the target level, which is given by $\frac{H_s}{2^l} \times \frac{W_s}{2^l}$. The resized feature maps are

then concatenated to form the aggregated feature representation for each level, denoted as $\bar{F}_l$, which can be written as follows:

$$\bar{F}_l = \text{Concat}(\{\text{Interp}(F_k, H_s/2^l, W_s/2^l)\}_{k=1}^4), \quad (4)$$

where $\text{Interp}(F, h, w)$ represents the bilinear interpolation operation that resizes the feature map F to a spatial resolution of $h \times w$ and $\text{Concat}(\cdot)$ operator takes a set of interpolated features and concatenates them along the channel dimension. The aggregated feature map $\bar{F}_l$ for each layer is then passed through a 1×1 convolution to reduce the channel dimension. To further refine the aggregated features, we employ Nonlinear Activation-Free Block (NAF-Block) [3], which improves feature extraction efficiency with a lightweight design. This results in the final output feature map $\tilde{F}_l$, expressed as:

$$\tilde{F}_l = \text{NAFBlock}(\text{Conv}_{1 \times 1}(\bar{F}_l)). \quad (5)$$

By effectively fusing features from all encoder levels, our approach ensures that each decoder level receives rich, multi-scale information. We found that this comprehensive aggregation further helps the decoder to better reconstruct fine details and suppress artifacts.

3.5. Loss Function

We adopt a combination of pixel-wise and perceptual losses to train our model. Since moiré datasets often contain spatial misalignment between input and ground-truth images (See Fig. 7), the perceptual loss [19] helps preserve structural details despite these discrepancies.

Following previous works [34, 55, 63], we predict images from the intermediate features independently and calculate the losses with ground truth image. The final decoder level produces the restored image, while the preceding two levels generate intermediate predictions using a 3×3 convolution followed by pixel shuffle upsampling. These three outputs are all supervised by the ground-truth image, ensuring progressive feature refinement. Accordingly, we compute a combination of the $\mathcal{L}_1$ loss and perceptual loss at each level, and the final loss function is formulated as their weighted sum:

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_{\text{ipps}}. \quad (6)$$

We empirically set $\lambda = 1$ for balanced optimization.

4. Experiments

In this section, we report the qualitative and quantitative results of our proposed MZNet in comparison with state-of-the-art models, and conduct ablation studies to validate the efficacy of each component. We also discuss alignment errors in high-resolution demoiré datasets.

4.1. Experimental settings

Datasets and metrics. We evaluate our model on three widely used demoiré datasets: TIP2018 [42], FHDMi [17], and UHDM [55], and compare its performance with existing state-of-the-art methods. For evaluation, we employ PSNR, SSIM [47], and LPIPS [60] which are commonly used metrics in image restoration tasks, including demoiré. However, for the TIP2018 dataset, we follow previous works and report only PSNR and SSIM, as LPIPS scores were not previously benchmarked.

Implementation Details. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. Following the training setup of ESDNet [55], we apply random cropping with a size of 768×768 for the UHDM and 512×512 for the FHDMi. For the TIP2018 dataset, images are first resized to 286×286 and then center-cropped to 256×256. During inference, full resolution images are used. MACs are measured using 4K resolution image. We optimize the model using Adam optimizer [20], setting β_1 and β_2 to 0.9, and employ a CosineAnnealingLR scheduler [30] to adjust the learning rate throughout training. For the large model in Tab. 1, we use 4, 4, 6, and 8 MSDAB blocks in the four encoder layers and 4, 4, 6, and 6 blocks in the decoder.

The kernel size K of MSCM is set to 7 for TIP2018, 15 for FHDMi, and 23 for UHDM. Remaining Detailed training settings are provided in the supplementary material.

4.2. Comparison with State-of-the-Art Methods

Quantitative comparison. Tab. 1 shows that MZNet outperforms prior state-of-the-art methods across all metrics on the high-resolution FHDMi (1920×1080) and UHDM (3840×2160) datasets. Specifically, on FHDMi, MZNet surpasses P-BiC [49] by 0.67 dB in PSNR, and on UHDM, it outperforms P-BiC by 0.33 dB. These results highlight MZNet’s strong ability to handle the complex moiré patterns in high-resolution images. On the low-resolution TIP2018 (256×256) dataset, MZNet still achieves the second-best result, demonstrating competitive performance under simpler moiré conditions. Notably, despite its high accuracy, MZNet achieves lower computational cost among the compared models, confirming its high accuracy and computational efficiency.

Qualitative comparison. Fig. 4 and Fig. 5 provide visual comparisons on the UHDM and FHDMi datasets, highlighting the qualitative advantages of our method over existing approaches. By leveraging large kernels and dilation to expand the receptive field, our model effectively addresses the limitations of existing approaches in removing large-scale moiré patterns. Additional qualitative results in the supplementary material further substantiate its effectiveness.

Dataset	Metrics	Input	DMCNN [42]	MDDM [6]	WDNet [28]	MopNet [16]	MBCNN [63]	FHDe ² Net [17]	ESDNet [55]	ESDNet-L [55]	MCFNet [34]	P-BiC [49]	MZNet (Ours)
TIP2018	PSNR $\uparrow$	20.30	26.77	-	28.08	27.75	30.03	27.78	29.81	30.11	30.13	30.56	<u>30.18</u>
	SSIM $\uparrow$	0.738	0.871	-	0.904	0.895	0.893	0.896	0.916	0.920	0.920	0.925	<u>0.921</u>
FHDMi	PSNR $\uparrow$	17.974	21.538	20.831	-	22.756	22.309	22.930	24.500	24.882	24.823	<u>25.450</u>	26.120
	SSIM $\uparrow$	0.7033	0.7727	0.7343	-	0.7958	0.8095	0.7885	0.8351	0.8440	0.8426	<u>0.8473</u>	0.8624
	LPIPS $\downarrow$	0.2837	0.2477	0.2515	-	0.1794	0.1980	0.1688	0.1354	0.1301	<u>0.1288</u>	0.1493	0.1042
UHDM	PSNR $\uparrow$	17.117	19.914	20.088	20.364	19.489	21.414	20.388	22.119	22.422	22.484	<u>23.30</u>	23.632
	SSIM $\uparrow$	0.5089	0.7575	0.7441	0.6497	0.7572	0.7932	0.7496	0.7956	0.7985	0.8001	<u>0.8007</u>	0.8096
	LPIPS $\downarrow$	0.5314	0.3764	0.3409	0.4882	0.3857	0.3318	0.3519	0.2551	0.2454	0.2536	<u>0.2324</u>	0.2237
-	Params $\downarrow$ (M)	-	1.426	7.637	<u>3.360</u>	58.565	14.192	13.571	5.934	10.623	6.181	4.922	14.824
	MACs $\downarrow$ (T)	-	2.258	3.679	1.757	-	8.522	33.23	2.247	3.689	6.903	<u>1.223</u>	1.190

Table 1. Quantitative comparisons on benchmark datasets. Bold and underlined values indicate the best and second-best results, respectively. Parameters and MACs are for our configuration on the UHDM dataset.

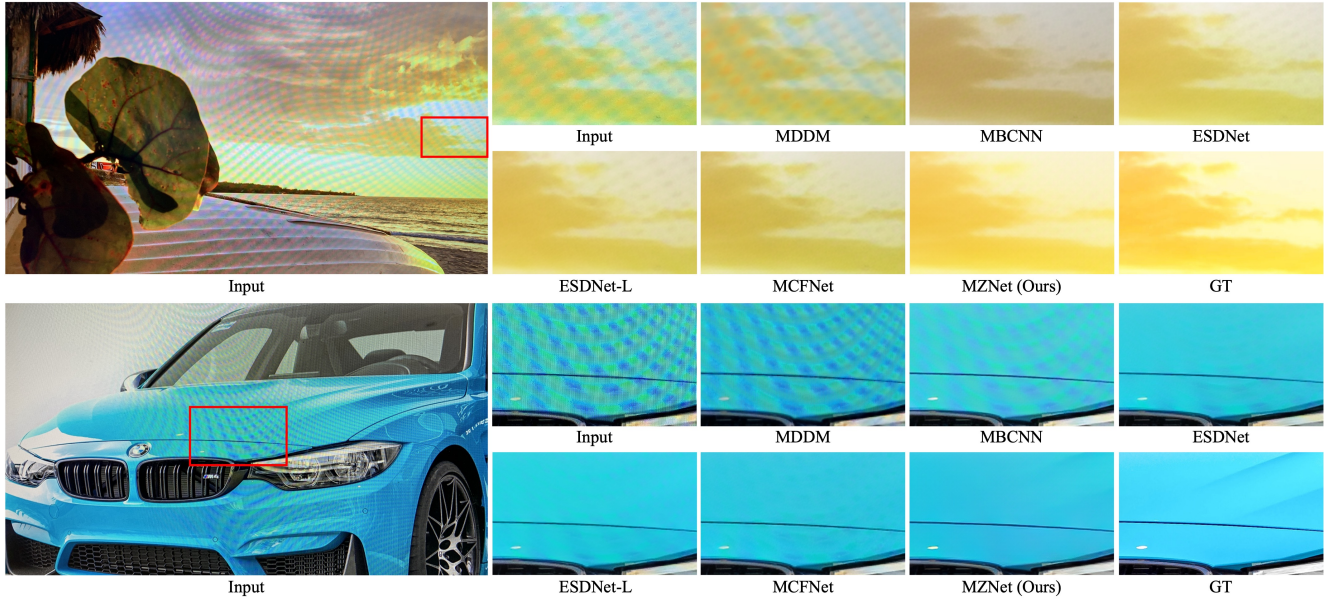


Figure 4. Qualitative comparisons with state-of-the-art methods on the UHDM dataset.

Computational Cost Analysis. While our model exhibits a relatively large number of parameters due to the use of high channel widths and subsequent 1×1 convolutions, most of the 3×3 convolutions are implemented as depth-wise convolution (as shown in Fig. 3), which helps reduce the overall MACs compared to other methods. This design balances computational efficiency and feature capacity, ensuring sufficient representation power to handle the complex moiré patterns in high-resolution images while keeping the computational cost low.

4.3. Ablation Studies

In this section, we analyze the impact of our proposed MZNet by ablating each component of the model. We train a smaller version of our model on the UHDM dataset, using four blocks at each level in both the encoder and decoder. All models are trained for 200 epochs. The initial learning rate is set to 0.0002. All other settings remain consistent

with the main experiments.

We also evaluated efficient transformer method [59] under the same protocol, but it either failed with Out-of-Memory at 4K or produced lower PSNR/SSIM than NAFNet. Therefore, we focus on efficient CNN backbones for our main comparisons. Note that this small version is different from the one provided in the Sec. 4.5.

Effect of Individual Components. Tab. 2 presents the ablation study results evaluating the contributions of each component in our model. The baseline configuration (Tab. 2a) is a simple model where all blocks are NAFBlock [3], with skip connections implemented by concatenation and a 1×1 convolution. Replacing NAFBlock with MSDAB (Tab. 2b) enhances PSNR to 23.04 dB, demonstrating its effectiveness in handling moiré patterns by capturing multi-scale features with large receptive fields. As illustrated in the first row of Fig. 6, a comparison between

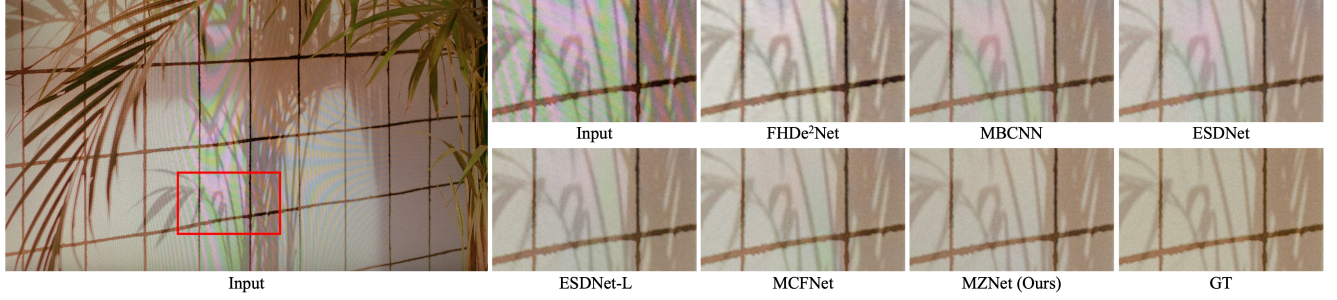


Figure 5. Qualitative comparisons with state-of-the-art methods on the FHDMi dataset.

	MSDAB	FFSC	MSLKB	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
(a)				22.70	0.792	0.242
(b)	✓			23.04	0.797	0.240
(c)		✓		22.86	0.791	0.226
(d)			✓	22.88	0.792	0.249
(e)	✓	✓		23.16	0.798	0.226
(f)	✓		✓	23.04	0.796	0.237
(g)		✓	✓	22.97	0.795	0.238
(h)	✓	✓	✓	23.23	0.801	0.224

Table 2. Ablation studies of each component in MZNet.

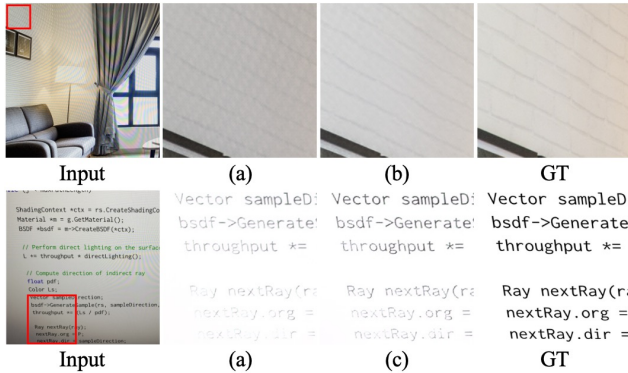


Figure 6. Qualitative comparisons of ablation study.

(a) and (b) shows that MSDAB better suppresses moiré artifacts. Incorporating FFSC (Tab. 2c) improves PSNR to 22.86 dB, while MSLKB (Tab. 2d) achieves 22.88 dB, validating their respective contributions to moiré suppression.

Notably, models equipped with FFSC (Tab. 2c, Tab. 2e) significantly improve perceptual quality, reducing LPIPS from 0.242 to 0.226. This is attributed to FFSC’s ability to aggregate multi-scale features across encoder stages, helping the decoder better reconstruct fine details, as seen in the second row of Fig. 6, where (c) better preserves subtle structures that are lost in (a). Integrating all components (Tab. 2h) yields the best performance, showing that each module contributes to moiré removal and their combination

Block	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
MSDAB	w/o LKA	22.75	0.792	0.236
	w/o SCA	23.03	0.799	0.229
	w/o MDCM	23.18	0.799	0.228
MSLKB	w/o SCA	23.15	0.799	0.229
	w/o stripe	23.17	0.800	0.227
	w/o square	23.22	0.800	0.225
Ours		23.23	0.801	0.224

Table 3. Ablation studies on the MSDAB and MSLKB.

is most effective. Despite its compact size, our model remains competitive with state-of-the-art methods.

Effect of MSDAB Components Tab. 3 presents the results of ablating each component of the MSDAB, highlighting their respective contributions.. Removing LKA results in a significant PSNR drop of 0.48 dB, confirming its critical role in the suppressing moiré patterns by effectively capturing long-range spatial dependencies. Similarly, removing SCA leads to a PSNR decrease of 0.20 dB, suggesting that channel-wise attention contributes to refining moiré-relevant features. Replacing the MDCM with a 3×3 depth-wise convolution demonstrates that MDCM is more effective at suppressing moiré distortions compared to a simple convolution.

Effect of MSLKB Components Tab. 3 presents the ablation study on the key components of the MSLKB. While multiple kernel shapes alone provide a strong baseline, the largest PSNR drop is observed when removing SCA, suggesting its complementary effect to the kernel operations. Meanwhile, removing the “stripe kernel” leads to a slightly greater performance drop than omitting the “square kernel”, suggesting that capturing elongated contextual cues is particularly beneficial for moiré removal.

Model	Params↓ (M)	MACs↓ (T)	FHDMi			UHDM		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MBCNN[63]	14.192	8.522	22.31	0.8095	0.1980	21.41	0.7932	0.3318
FHDe ² Net[17]	13.571	33.23	22.93	0.7885	0.1688	20.39	0.7496	0.3519
ESDNet[55]	5.934	2.247	24.50	0.8351	0.1354	22.12	0.7956	0.2551
ESDNet-L[55]	10.623	3.689	24.88	0.8440	0.1301	22.422	0.7985	0.2454
MCFNet[34]	6.181	6.903	24.82	0.8426	0.1288	22.48	0.8001	0.2536
P-BiC[49]	4.922	1.223	25.45	0.8473	0.1493	23.30	0.8007	0.2324
MZNet-Small	7.46	0.596	25.61	0.8522	0.1111	23.26	0.8022	0.2242
MZNet-Medium	12.61	0.823	25.91	0.8582	0.1078	23.47	0.8043	0.2230
MZNet-Large	14.82	1.190	26.12	0.8624	0.1042	23.63	0.8096	0.2237

Table 4. Performance of lightweight MZNet variants on high-resolution datasets. The “Large” model is the original version reported in Tab. 1, while “Small” and “Medium” are lightweight variants for supplementary analysis. Params and MACs are measured on 4k input images.

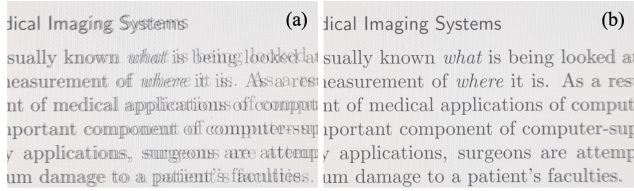


Figure 7. (a) Overlay of the moiré image and ground truth from the original UHDM dataset, showing noticeable misalignment. (b) Overlay after additional alignment.

4.4. Moiré Dataset Alignment Issues

Achieving precise alignment between moiré images and ground truth images remains challenging due to distortions introduced during image capture and inaccuracies in homography estimation, as illustrated in Fig. 7. This issue becomes particularly critical for high-resolution datasets such as UHDM, where even slight misalignments can lead to unstable training and significant performance degradation. To mitigate these alignment errors, we further refined the UHDM dataset by applying SIFT-based [31] keypoint detection and matching, leading to improved homography estimation and more precise image warping. To investigate the impact of dataset alignment issues, we trained and evaluated MZNet and ESDNet-L on the refined UHDM using identical settings. As shown in Tab. 5, both models showed substantial improvements across all metrics, underscoring the importance of precise alignment for optimal demoiré performance. These results highlight the need for further research into methods for addressing alignment issues or developing demoiré models more robust to misalignment.

4.5. Lightweight Variants on High-Res Datasets

Although our main model already demonstrates strong performance across all datasets, we additionally evaluate two lightweight variants to verify that the performance gains do not stem solely from the model’s parameter count. Specifically, we design MZNet-Medium and MZNet-Small by re-

Dataset	MZNet (Ours)			ESDNet-L		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Original	23.63	0.810	0.223	22.42	0.799	0.245
Aligned	24.50	0.841	0.205	23.09	0.822	0.235

Table 5. Impact of UHDM dataset alignment on model performance. Alignment improves all metrics for both models, with a more significant gain in our MZNet.

ducing the number of MSDAB in each encoder and decoder stage to (2, 2, 4, 6) and (2, 2, 2, 2), respectively. All Multi-Dilation Convolution Module (MDCM) use dilation rates of 1, 4, and 9, and all other experimental settings are kept identical to those of the main model. These variants are evaluated on high-resolution datasets (FHDMi and UHDM), where moiré artifacts tend to be more complex and the performance gap between different model capacities becomes more apparent. Note that this small version is different from the one provided in Sec. 4.3 As shown in Tab. 4, both lightweight variants show reasonable performance across all metrics, suggesting that the proposed architecture can remain effective even under constrained model size.

5. Conclusion

In this study, we propose MZNet, an efficient and effective solution for moiré pattern removal. Our model integrates MSDAB for multi-scale feature extraction, MSLKB for capturing diverse moiré structures, and FFSC for enhancing information flow, enabling robust moiré elimination while preserving fine details. Experiments on benchmark datasets demonstrate that MZNet outperforms existing methods in both quantitative metrics and visual quality, achieving state-of-the-art performance in challenging high-resolution scenarios. We also discuss the impact of moiré dataset alignment issues and demonstrate how precise alignment improves model performance.

References

- [1] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1192–1204, 2020. 2
- [2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2021. 2
- [3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 2, 3, 4, 5, 6
- [4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023. 2
- [5] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12312–12321, 2023. 2
- [6] Xi Cheng, Zhenyong Fu, and Jian Yang. Multi-scale dynamic feature encoding network for image demoreing, 2019. 1, 2, 6
- [7] Yijia Cheng, Xin Liu, and Jingyu Yang. Recaptured raw screen image and video demoreing via channel and spatial modulations. *Advances in Neural Information Processing Systems*, 36:40414–40425, 2023. 2
- [8] Yuning Cui, Wenqi Ren, and Alois Knoll. Omni-kernel network for image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1426–1434, 2024. 2, 4
- [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] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022. 2
- [11] Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Sunet: Swin transformer unet for image denoising. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2333–2337. IEEE, 2022. 2
- [12] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2511–2520, 2019. 2
- [13] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in neural information processing systems*, 35:1140–1156, 2022. 2
- [14] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational visual media*, 9(4):733–752, 2023. 2, 4
- [15] Xiaojiao Guo, Yihang Dong, Xuhang Chen, Weiwen Chen, Zimeng Li, FuChen Zheng, and Chi-Man Pun. Underwater image restoration via polymorphic large kernel cnns. *arXiv preprint arXiv:2412.18459*, 2024. 2, 4
- [16] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Mop moire patterns using mopnet. In *ICCV*, pages 2424–2432, 2019. 1, 2, 6
- [17] Bin He, Ce Wang, Boxin Shi, and Ling yu Duan. Fhde2net: Full high definition demoreing network. In *European Conference on Computer Vision*, 2020. 1, 2, 5, 6, 8
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2
- [19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 5
- [20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5
- [21] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 2
- [22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2
- [23] Kin Wai Lau, Lai-Man Po, and Yasar Abbas Ur Rehman. Large separable kernel attention: Rethinking the large kernel attention design in cnn. *Expert Systems with Applications*, 236:121352, 2024. 2
- [24] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2
- [25] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 2
- [26] Jingyun Liang, Jiezhong 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
- [27] Fanglei Liu, Jingyu Yang, and Huanjing Yue. Moiré pattern removal from texture images via low-rank and sparse matrix decomposition. In *2015 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2015. 1
- [28] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Ales Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoreing, 2020. 1, 2, 6

- [29] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022. 2
- [30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 5
- [31] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 8
- [32] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tiejong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022. 2
- [33] Pinjun Luo, Guoqiang Xiao, Xinbo Gao, and Song Wu. Lkd-net: Large kernel convolution network for single image dehazing. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1601–1606. IEEE, 2023. 2
- [34] Duong Hai Nguyen, Se-Ho Lee, and Chul Lee. Multiscale coarse-to-fine guided screenshot demoiréing. *IEEE Signal Processing Letters*, 30:898–902, 2023. 1, 5, 6, 8
- [35] 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. 4
- [36] Lingyan Ruan, Mojtaba Bermana, Hans-peter Seidel, Karol Myszkowski, and Bin Chen. Revisiting image deblurring with an efficient convnet. *arXiv preprint arXiv:2302.02234*, 2023. 2
- [37] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3
- [38] Hasib Siddiqui, Mireille Boutin, and Charles A Bouman. Hardware-friendly descreeing. *IEEE Transactions on Image Processing*, 19(3):746–757, 2009. 1
- [39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2
- [40] Bin Sun, Shutao Li, and Jun Sun. Scanned image descreeing with image redundancy and adaptive filtering. *IEEE transactions on image processing*, 23(8):3698–3710, 2014. 1
- [41] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13190–13199, 2023. 2
- [42] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *IEEE Transactions on Image Processing*, 27(8):4160–4172, 2018. 1, 2, 5, 6
- [43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4
- [44] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European conference on computer vision*, pages 146–162. Springer, 2022. 2
- [45] Ce Wang, Bin He, Shengsen Wu, Renjie Wan, Boxin Shi, and Ling-Yu Duan. Coarse-to-fine disentangling demoiréing framework for recaptured screen images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9439–9453, 2023. 1, 2
- [46] Yan Wang, Yusen Li, Gang Wang, and Xiaoguang Liu. Multi-scale attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5950–5960, 2024. 2
- [47] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5
- [48] 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*, pages 17683–17693, 2022. 2
- [49] Zeyu Xiao, Zhihe Lu, and Xinchao Wang. P-bic: Ultra-high-definition image demoiréing via patch bilateral compensation. In *ACM Multimedia 2024*, 2024. 1, 2, 5, 6, 8
- [50] Dejia Xu, Yihao Chu, and Qingyan Sun. Moiré pattern removal via attentive fractal network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 472–473, 2020. 1
- [51] Shuning Xu, Binbin Song, Xiangyu Chen, Xina Liu, and Jiantao Zhou. Image demoiréing in raw and srgb domains. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024. 1, 2
- [52] Jingyu Yang, Fanglei Liu, Huanjing Yue, Xiaomei Fu, Chunping Hou, and Feng Wu. Textured image demoiréing via signal decomposition and guided filtering. *IEEE Transactions on Image Processing*, 26(7):3528–3541, 2017. 1
- [53] Jingyu Yang, Xue Zhang, Changrui Cai, and Kun Li. Demoiréing for screen-shot images with multi-channel layer decomposition. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017. 1
- [54] Zemin Yang, Yujing Sun, Xidong Peng, Siu Ming Yiu, and Yuexin Ma. Unidemoiré: Towards universal image demoiréing with data generation and synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9354–9362, 2025. 1
- [55] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022. 1, 2, 5, 6, 8

- [56] Huanjing Yue, Yijia Cheng, Yan Mao, Cong Cao, and Jingyu Yang. Recaptured screen image demoiréing in raw domain. *IEEE Transactions on Multimedia*, 25:5589–5600, 2022. [2](#)
- [57] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. [2](#)
- [58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. [2](#)
- [59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration, 2022. [3](#), [6](#)
- [60] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. [5](#)
- [61] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. [2](#)
- [62] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. [2](#)
- [63] Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, and Ales Leonardis. Image demoiréing with learnable bandpass filters, 2020. [1](#), [2](#), [5](#), [6](#), [8](#)
- [64] Bolun Zheng, Shanxin Yuan, Chenggang Yan, Xiang Tian, Jiyong Zhang, Yaoqi Sun, Lin Liu, Aleš Leonardis, and Gregory Slabaugh. Learning frequency domain priors for image demoiréing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7705–7717, 2021. [1](#), [2](#)
- [65] Yunshan Zhong, Yuyao Zhou, Yuxin Zhang, Fei Chao, and Rongrong Ji. Learning image demoiréing from unpaired real data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7623–7631, 2024. [1](#)