Freqformer: Image-Demoiréing Transformer via Effective Frequency Decomposition

Xiaoyang Liu*, Bolin Qiu*, Zheng Chen, Libo Zhu,
Zihan Zhou, Kai Liu, Jiezhang Cao, and Yulun Zhang**

Shanghai Jiao Tong University

Abstract. Image demoiréing remains a challenging task due to the complex interplay between texture corruption and color distortions caused by moiré patterns. Existing methods, especially those relying on direct image-to-image restoration, often fail to disentangle these intertwined artifacts effectively. While frequency-aware approaches offer a promising direction, their potential is hindered by the discrete transform (e.g., Haar wavelet or block-based DCT), which may suffer from spatial discontinuity, channel redundancy, and further cause error accumulation during their fixed inverse processes. In this paper, we present Freqformer, a Transformer-based framework specifically designed for image demoiréing through targeted frequency separation. Our method performs an effective frequency decomposition that splits moiré patterns into high-frequency spatially-localized textures and low-frequency scale-robust color distortions, which are then handled by a dual-branch architecture and an asymmetric training scheme tailored to their distinct characteristics. We further propose a learnable Frequency Composition Transform (FCT) module to adaptively fuse the frequency-specific outputs, enabling consistent and high-fidelity reconstruction. To better aggregate the spatial dependencies and the inter-channel complementary information, we introduce a Spatial-Aware Channel Attention (SA-CA) module that refines moiré-sensitive regions without incurring high computational cost. Extensive experiments on various demoiréing benchmarks demonstrate that Freqformer achieves state-of-the-art performance with a compact model size. The code will be made publicly available at `https://github.com/xyLiu339/Freqformer`.

Keywords: Image Demoiréing · Frequency Decomposition

1 Introduction

Capturing digital screens with a camera often introduces moiré patterns—irregular, colorful artifacts caused by aliasing between the screen's subpixel layout and the camera's color filter array (CFA) [32]. These patterns significantly degrade image

* Equal contribution
** Corresponding author: Yulun Zhang, yulun100@gmail.com

quality and are difficult to remove due to their complexity and variability. Consequently, single-image demoiréing remains a challenging low-level vision task that continues to attract attention from academia and industry.

Recent advances in deep learning have led to progress in image demoiréing, with CNN-based models achieving promising results [9, 10, 31, 38, 39, 45, 46, 53]. However, most existing approaches treat moiré removal as a holistic restoration task, directly learning to map the input image to a clean output. This straightforward strategy is inherently limited due to the complex nature of moiré patterns: the **color distortions** caused by moiré patterns are mixed with the original image colors, and the **corrupted textures** often resemble natural structures. To address this, some recent works have turned to conventional frequency decomposition tools, such as the Haar wavelet [18, 24, 42] or block-based DCT [10, 53], to disentangle image components. While frequency-aware methods have shown improved performance, their potential has not been fully explored. For instance, some methods [18, 24] simply concatenate wavelet subbands for joint learning, failing to exploit the distinct semantics of different frequency components. Other methods like [53] may fail to capture and eliminate large-scale moiré patterns effectively due to the restricted receptive field.

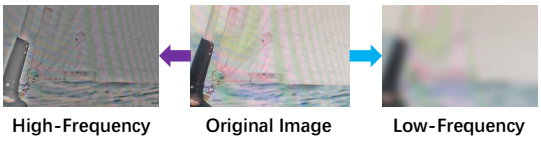

Fig. 1: Our Frequency Decomposition.

To overcome these limitations, we introduce a novel frequency decomposition to effectively **separate moiré textures and color distortions into high-frequency and low-frequency components**, respectively (Fig. 1). This enables the design of a dual-branch architecture that learns to handle each frequency band independently, providing more specialized restoration. In addition, benefiting from the favorable properties of our frequency decomposition, the high-frequency moiré textures exhibit strong **locality**, while the low-frequency components demonstrate **robustness to scale variations**. Leveraging these characteristics, we adopt an asymmetric training scheme: **crop-**based training for high-frequency features and **resize-**based training for low-frequency features. In particular, the resize-based training for the low-frequency component largely avoids detail loss [10] while concentrating on color distortions, making them easier to correct, thereby improving overall performance. Moreover, while traditional frequency-based methods rely on predefined, non-trainable inverse transforms, our decomposition is designed to support a learnable **F**requency **C**omposition **T**ransform (**FCT**) module, which adaptively fuses the two branches and ensures better consistency in reconstruction.

Additionally, most previous demoiréing methods are built on heavy U-Net-style backbones [9, 10, 45, 53] (10M∼50M), which are effective but computationally inefficient, especially for ultra-high-resolution images (*e.g.*, 4K). Recent findings [38] have shown that moiré artifacts exhibit variance across color channels, which motivates more refined attention mechanisms. To this end, we introduce a **S**patial-**A**ware **C**hannel **A**ttention (**SA-CA**) module that enhances spatial

and channel-wise feature modeling while maintaining a compact design. Unlike standard attention modules, SA-CA selectively emphasizes moiré-prone regions and channels without incurring excessive computational overhead, making our model suitable for high-resolution inference and deployment on edge devices.

By integrating the above components, we propose **Freqformer**, a hierarchical dual-branch Transformer tailored for image demoiréing. Freqformer leverages frequency feature separation, adaptive fusion, and efficient attention mechanisms to remove moiré artifacts. It achieves state-of-the-art performance on various high-resolution image demoiréing benchmarks, while maintaining a compact model size. Our main contributions are summarized as follows:

- We introduce a novel frequency decomposition strategy to effectively separate moiré patterns into high-frequency (texture) and low-frequency (color) components, enabling targeted learning for each.
- We design a dual-branch, asymmetric architecture that separately processes high- and low-frequency components using crop- and resize-based training strategies, enabling specialized feature extraction tailored to different aspects of moiré patterns. We also introduce a learnable FCT module that adaptively fuses the outputs from both branches.
- We propose Freqformer, with a lightweight SA-CA module to enhance both spatial and channel-wise feature interactions. Extensive experiments on image demoiréing benchmarks show that Freqformer achieves state-of-the-art performance with a small number of parameters.

2 Related Works

2.1 Demoiréing Methods

Image demoiréing. Early image demoiréing methods focused on specific moiré patterns. These approaches, such as smooth filtering [28], spectral analysis [29], and image decomposition [17], rely heavily on handcrafted features and mathematical models and often fail to generalize effectively to diverse moiré patterns and varying image conditions. In recent years, deep learning-based methods [9,27,31,46,52,53] have emerged as a powerful tool for image demoiréing. Yu et al. [45] proposed ESDNet, the first lightweight model for ultra-high-resolution image demoiréing. MoiréDet [41] further extracts moiré edge maps from images with moiré patterns using a triplet generation strategy. Xiao et al. [38] developed a patch bilateral compensation network, leveraging the green channel's resistance to moiré patterns. Peng et al. [27] present DMSFN, a network using dilated-dense attention and multiscale feature interaction for effective moiré pattern removal, along with a moiré data augmentation technique.

Video demoiréing. Video demoiréing poses added challenges due to the need for temporal coherence. Dai et al. [5] introduced the first hand-held video demoiréing dataset and a baseline model using relation-based temporal consistency loss. Cheng et al. [3] extended image and video demoiréing to Raw inputs with color-separated and spatial modulations. Xu et al. [40] proposed DTNet, a unified framework that performs moiré removal, alignment, color correction, and detail refinement via directional DCT modes and temporal-guided bilateral learning.

Raw domain demoiréing. The Raw domain offers high-bit-depth data with reduced ISP interference, making it well-suited for low-level tasks. This is especially true for demoiréing, where moiré patterns are simpler in Raw [47], free from ISP-induced nonlinearities. RDNet [47] first tackles Raw image demoiréing with an encoder-decoder and class-specific learning. RawVDemoiré [3] addresses video demoiréing via temporal alignment. Recent work [39] jointly leverages Raw and sRGB inputs to learn device-dependent ISP models for restoration. Despite these advantages, Raw-domain demoiréing is inherently sensor-dependent and struggles with the subsequent Raw-to-sRGB conversion, where inaccurate ISP approximations can lead to color shifts. Thus, we perform demoiréing directly in the RGB domain to better approximate the mapping to real-world scenes.

Moiré removal dataset. Sun et al. [31] introduced a large-scale benchmark dataset (TIP-2018) containing over 100,000 image pairs. Zheng et al. [53] proposed the AIM2019 image demoiréing challenge dataset (LCDMoiré) in 2020. LCDMoiré contains only text scenes, while TIP-2018 does not include text scenes but has a limited resolution of 256×256. FHDMi dataset, presented by He et al. [10], provides high-resolution images with more diverse and complex moiré patterns. More recently, Yu et al. [45] took a further step by constructing the first large-scale real-world Ultra-High-Definition demoiréing dataset (UHDM), containing real-world 4K resolution image pairs with more challenging and practical scenarios. Yue et al. [47] built the first well-aligned raw moiré image dataset (TMM-2022) by pixel-wise alignment between the recaptured images and source ones.

2.2 Frequency-based Methods

Wavelet-based methods, due to their ability to decompose an image into multiple frequency bands, have been widely applied in some computer vision tasks, including classification [6] [16] [26] [36], network compression [7] [15], face aging [22], super-resolution [13] [19], style transfer [43], etc. In recent years, wavelet-based methods have emerged as powerful tools for image demoiréing. Liu et al. [18] proposed a wavelet-based dual-branch network (WDNet) that operates in the wavelet domain to effectively separate moiré patterns from image content by leveraging dense and dilated convolution modules. Luo et al. [24] propose AWUDN, a deep network that combines wavelet-domain feature mapping with a domain adaptation mechanism. Nevertheless, these two methods simply mix different frequency features without exploiting the hierarchical frequency features of wavelet decomposition. Further, Yeh et al. [42] introduced a multibranch wavelet-based network (MBWDN) that decomposes moiré images into multiple sub-band images using wavelet transform and processes them through different branches. However, due to its multi-level structure, the method incurs high computational cost. Additionally, the Haar transform only partially separates moiré patterns, limiting its effectiveness for high-definition or complex cases.

Besides wavelets, other frequency tools like DCT have also been explored. MBCNN [53] explicitly removes moiré patterns by configuring the Inverse DCT (IDCT) matrix and learnable weights as a bandpass filter. Nevertheless, its effectiveness is bottlenecked by a limited receptive field, hindering the removal of

large-scale moiré structures. Meanwhile, FHDe²Net [10] uses DCT on the luminance channel to disentangle high-frequency details from moiré interference. However, utilizing DCT primarily to compensate for downsampling-induced detail loss significantly increases system design complexity.

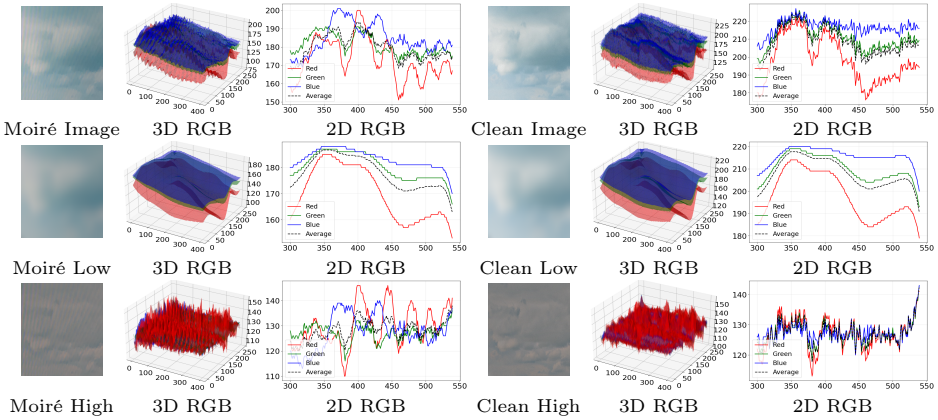

Fig. 2: Visualization of the frequency-decomposed images (zoom in for better analysis). The left three columns show the moiré images along with their 3D and 2D (a column of pixels) visualizations, while the right three columns display the clean images. The first, second, and third rows represent the original images, low- and high-frequency components, respectively. The low-frequency curves of moiré and clean images exhibit similar trends, primarily reflecting color distortions. The high-frequency curves differ significantly in shape, indicating structural differences caused by moiré patterns.

3 Method

Our whole pipeline is shown in Fig. 4. Given a moiré image $\boldsymbol{I_m}$, we first apply a frequency decomposition (Fig. 2, Sec. 3.1) to obtain a high-frequency component $\boldsymbol{I_h}$ and a low-frequency component $\boldsymbol{I_l}$, both retaining the original spatial resolution. During training, the high-frequency branch adopts a cropping strategy, while the low-frequency branch uses resizing (Sec. 3.2). During testing, the high branch directly processes the full-resolution input, while the low branch continues to operate with the resize-based strategy.

The two branches process $\boldsymbol{I_h}/\boldsymbol{I_l}$ using the encoder-decoder architecture. Each encoder and decoder block is a Spatial-Aware Channel Attention (SA-CA) module (Sec. 3.2). Moreover, a hierarchical feature fusion strategy is adopted in the high-frequency branch, where multi-scale features are concatenated with minimal overhead.

The network is trained in two phases (Sec. 3.3). First, the high- and low-frequency branches are trained independently. Then, both branches and the learnable Frequency Composition Transform (FCT) (Sec. 3.1) are jointly fine-tuned to fuse their features and reconstruct a cleaner output.

3.1 Frequency Decomposition and Learnable FCT

As shown in Fig. 2, we decompose both the moiré-contaminated image and the clean image into high- and low-frequency components and visualize the 2D/3D RGB curves. It is evident that the moiré patterns are predominantly preserved in the high-frequency part, while the low-frequency component exhibits minimal moiré textures but suffers from noticeable color distortions. Notably, although the low-frequency curve of the moiré image follows a similar overall trend to that of the clean image, there is a significant deviation in magnitude. This discrepancy can be attributed to two main factors. First, the presence of moiré patterns inherently distorts the original color information during the frequency decomposition, which complicates the accurate recovery of low-frequency components and remains a significant challenge. Second, the dataset construction process itself introduces severe color mismatches — or more precisely, color shifts (also mentioned in [32, 38, 53]) — due to varying external conditions such as lighting during the moiré image capture in the dataset construction process. Therefore, following the frequency decomposition, we design a dual-branch approach (will be detailed in Sec. 3.2): a high-frequency branch dedicated to restoring fine textures affected by moiré patterns, and a low-frequency branch focused on recovering accurate color information, addressing both moiré-induced distortions and dataset-related color shifts.

While frequency decomposition is crucial for demoiréing, conventional tools like the Discrete Wavelet Transform (DWT, e.g., multi-level Haar) [18, 24, 42] and Discrete Cosine Transform (DCT) [10, 53] are fundamentally ill-suited for our design needs due to several inherent limitations:

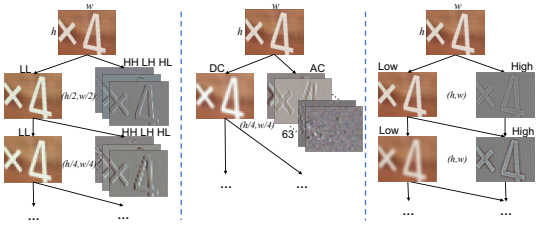

Fig. 3: Frequency decomposition methods. Left: Haar DWT; Middle: Block DCT; Right: Ours.

- **Lack of Spatial Smoothness and Error Accumulation:** Both DWT and DCT yield discrete representations lacking spatial smoothness (see the left LL and middle DC figures in Fig. 3). With rigidly fixed inverse transforms, subtle frequency-domain restoration errors inevitably accumulate during reconstruction, causing severe pixel-level and structural distortions.
- **Resolution Degradation and Shift Variance:** Multi-level recursive decomposition aggressively downsamples spatial resolution (e.g., Haar, from H×W to $H/2^i \times W/2^i$, and similarly for DCT), risking secondary aliasing. Furthermore, standard downsampled transforms (e.g., Haar) are shift-variant, yielding highly unstable representations for phase-sensitive moiré artifacts.
- **Incomplete Decoupling and Block Artifacts:** DWT's rapid resolution drop often leaves high-frequency moiré entangled within low-frequency components. Similarly, patch-based DCT assumes local signal stationarity; applying it to non-stationary, curved moiré patterns misses global structures and may introduce block artifacts.

– **Architectural Incompatibility and Channel Redundancy:** DWT scatters features across directional sub-bands (HH, HL, LH, LL), while block DCT splits signals into up to 64 isolated channels. Feeding such highly fragmented, redundant representations into neural networks drastically inflates computational overhead without proportional restoration gains.

To overcome these issues, we utilize an anti-aliasing and smoothness-preserving frequency transform with a learnable inverse process, which better separates moiré artifacts and supports a more training-friendly Transformer framework. Inspired by Wang et al. [33], our decomposition design is as follows. Given a moiré image $\boldsymbol{I_m}$, we utilize convolution with kernel $\boldsymbol{k}$, which is defined as

$$\boldsymbol{k} = \begin{bmatrix} 1/16 & 1/8 & 1/16 \\ 1/8 & 1/4 & 1/8 \\ 1/16 & 1/8 & 1/16 \end{bmatrix}. \tag{1}$$

The recursive decomposition is comprised of $\boldsymbol{L}$-level. For level $\boldsymbol{i}$'s decomposition process ($\boldsymbol{i} = 1, 2, ..., \boldsymbol{L}$), the decomposition outcome $\boldsymbol{I_l^i}$ and $\boldsymbol{I_h^i}$ is calculated as

$$\boldsymbol{I_l^i} = Conv\left(\boldsymbol{I_l^{i-1}}, \boldsymbol{k}, i\right), \tag{2}$$

$$\boldsymbol{I}_h^i = \boldsymbol{I}_h^{i-1} + \boldsymbol{I}_l^{i-1} - \boldsymbol{I}_l^i = \boldsymbol{I}_l^0 - \boldsymbol{I}_l^i, \tag{3}$$

where $\boldsymbol{I}_l^0 = \boldsymbol{I_m}$, $\boldsymbol{I}_h^0 = \boldsymbol{0}$ and $Conv$ represents the convolution operation with a dilation rate of 2^i. Ultimately, the $\boldsymbol{L}$-level decomposition results of image $\boldsymbol{I}$ are obtained as $\boldsymbol{I_l} = \boldsymbol{I}_l^L$ and $\boldsymbol{I_h} = \boldsymbol{I}_h^L$.

After the dual-branch demoiré process, we can get the clean representation $\hat{\boldsymbol{I}}_l$ and $\hat{\boldsymbol{I}}_h$. Normally, according to Eq. 3, the final output is calculated as

$$\hat{\boldsymbol{I}} = \hat{\boldsymbol{I}}_{\boldsymbol{l}} + \hat{\boldsymbol{I}}_{\boldsymbol{h}}, \tag{4}$$

But this traditional composition process cannot fuse the two branches effectively, as the reconstruction error in the same area on the low and high frequency will further lead to more severe distortions. So we instead design a learnable Frequency Composition Transform (FCT) which operates on the feature space. According to Fig. 4, we extract the feature before the convolution that projects the feature dimension to image dimension and the PixelShuffle, and denote it as $\hat{\boldsymbol{F}}_l$ and $\hat{\boldsymbol{F}}_h$. The learnable FCT module is defined as

$$\boldsymbol{F} = Conv(\hat{\boldsymbol{F}}_{\boldsymbol{l}}) + Conv(\hat{\boldsymbol{F}}_{\boldsymbol{h}}), \tag{5}$$

$$\hat{\boldsymbol{F}} = PostFusion(\boldsymbol{F}), \tag{6}$$

$$\hat{\boldsymbol{I}} = PixelShuffle(Conv(\hat{\boldsymbol{F}})). \tag{7}$$

The addition operation of the separate convolution of the output features is mathematically equivalent to $\boldsymbol{F} = Conv(Concat(\hat{\boldsymbol{F}}_{\boldsymbol{l}}, \hat{\boldsymbol{F}}_{\boldsymbol{h}}))$, only doubles the number of bias parameters. The $PostFusion$ module is an N_f layer's transformer block, which will be detailed in Sec. 3.2.

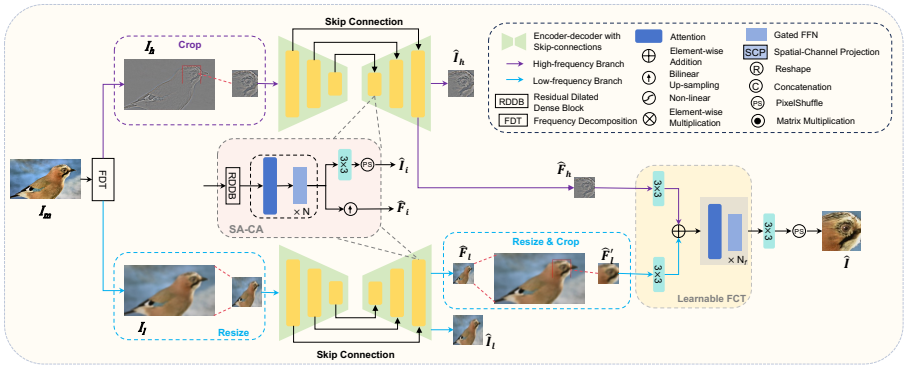

Fig. 4: The whole training pipeline. The figure shows the dual-branch learning with crop/resize strategy and a learnable FCT to aggregate the two branches' features.

3.2 Dual-Branch Design and the SA-CA Module

After decomposing the moiré image $\boldsymbol{I_m}$ into $\boldsymbol{I_l}$ and $\boldsymbol{I_h}$ different parts, some awesome benefits are shown for later processing:

1. The high-frequency component shows strong spatial locality, carrying little color and mainly encoding texture. Thus, only local context is needed for effective demoiré, making the global image structure less important. This allows robust training via random cropping and direct application to ultra-high-resolution demoiré tasks.
2. The low-frequency component is smooth and highly scale-robust. It can be downsampled to a much lower resolution (*e.g.*, 0.1×) and still be accurately reconstructed. We experimented with $\boldsymbol{I_h}$ and $\boldsymbol{I_l}$ separately resized, and then restored to their original size to assess information loss. Results show reconstruction PSNRs of ∼25 dB ($\boldsymbol{I_h}$) and ∼50 dB ($\boldsymbol{I_l}$), confirming that $\boldsymbol{I_l}$ can be safely resized with minimal degradation, avoiding the risks of detail loss caused by downsampling [10].
3. After resizing, moiré-induced color distortions become spatially concentrated. Consequently, only small, textureless patches with color deviations require correction, which still depends on local spatial features and channel-wise representations. Hence, using similar network structures for the high- and low-frequency branches is a natural and simplified choice.
4. Aggressively downsampling the low-frequency component simplifies color shift correction, as full color information is preserved compactly. This lets the low-frequency branch use the same small-scale input during both training and inference phases, avoiding potential issues with varying input sizes.

As analyzed, we adopt a two-branch learning framework and apply cropping and resizing strategies to the high- and low-frequency branches, respectively, shown in Fig. 4. During inference, the high-frequency branch operates on the full-resolution input, while the low-frequency branch processes a downsampled image that is later upsampled back to full resolution. The resulting full-resolution features from both branches are then fused by the FCT module.

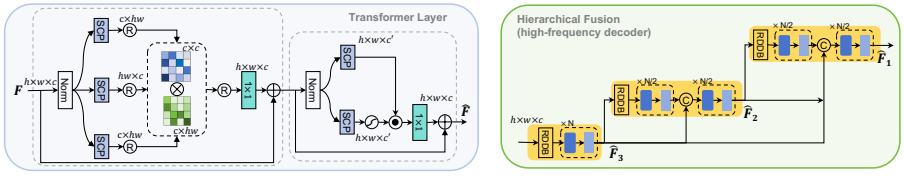

Fig. 5: Left part is the detailed transformer architecture of the SA-CA module utilized in all the encoders and decoders. The right part shows the hierarchical fusion of the high-frequency branch.

Spatial-Aware Channel Attention Module. Inspired by the work [38], which reveals that the green channel is less affected by the moiré patterns, and according to Fig. 2, we find that the RGB curves are not the same and sometimes show complementary effects, we decide to let each channel fully attend to others and adjust the feature. We adopt the channel attention mechanism here, which is also commonly used in [1, 39, 49]. However, channel information alone is insufficient for precise image demoiréing—spatial information from the adjacent domain remains crucial for texture learning and color projection. This is because the frequency characteristics or directional patterns of a wave can only be inferred from multiple spatially adjacent points, instead of a single isolated point. We adopt the Residual Dilated Dense Block (RDDB) [11, 12, 44, 51] for feature extraction with a large receptive field. Additionally, projection layers in the attention and FFN modules are crucial for aggregating spatial and channel features for local-feature and inter-channel learning.

Each block of the encoders and decoders uses the Spatial-Aware Channel Attention (SA-CA) Module. As shown in Fig. 5, our whole Spatial-Aware Channel Attention Module is defined as follows: Given an input feature $\boldsymbol{F} \in \mathcal{R}^{h\times w\times c}$, it first goes through an RDDB block to extract the spatial information with a large receptive field. Then, it goes through N layers of transformer block, each consisting of a channel attention and a gated Feed-Forward Network. For the channel Attention, given the input $\boldsymbol{F}$, we first generate the $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathcal{R}^{h\times w\times c}$ using the spatial-channel projection (SCP). Considering efficiency and functionality, we finally utilize the 1×1 convolution followed by 3×3 depth-wise convolution, expecting the former to learn the channel pattern projection and the latter to aggregate the adjacent information. Then using reshape and transpose, we get $\boldsymbol{Q'} \in \mathcal{R}^{c\times hw}$, $\boldsymbol{K'} \in \mathcal{R}^{hw\times c}$ and $\boldsymbol{V'} \in \mathcal{R}^{c\times hw}$. And then, we perform channel attention $\hat{\boldsymbol{F}} = \mathit{Attention}(\boldsymbol{Q'}, \boldsymbol{K'}, \boldsymbol{V'}) + \boldsymbol{F}$, with channels divided into heads to learn different features. As for the subsequent Feed-Forward Network, the traditional one can only conduct channel-wise learning and hardly handle the spatial-channel variant demoiréing task. As a result, we decide to let the FFN be a gated module that filters the inappropriate information. Firstly, through the spatial-channel projection (also a 1×1 convolution and a 3×3 depth-wise convolution), we get $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{R}^{h\times w\times c'}$. We utilize a non-linear function to get the filter scores and conduct the element-wise multiplication, followed by a convolution and a residual connection from the input $\boldsymbol{F}$.

Apart from the block design, we also design the high-frequency decoder's hierarchical fusion. As sometimes the moiré texture may have a wide range of variety, the demoiréed feature from the lower level of the decoder might help to produce the higher clean feature. We denote the output of each decoder block as $\hat{\boldsymbol{F}}_{\boldsymbol{3}}, \hat{\boldsymbol{F}}_{\boldsymbol{2}}, \hat{\boldsymbol{F}}_{\boldsymbol{1}}$. For the mid-block of the decoder, the $\hat{\boldsymbol{F}}_{\boldsymbol{3}}$ from the bottom block is concatenated into the medium output of the N/2 layers and then fed to the later N/2 layers for fusion. For the top block of the decoder, the $\hat{\boldsymbol{F}}_{\boldsymbol{3}}$ and $\hat{\boldsymbol{F}}_{\boldsymbol{2}}$ are concatenated into the middle output of the N/2 layers and then fed to the later N/2 layers for fusion.

3.3 Training Strategy and Loss Functions

The training is set in two stages. For stage one, the high-frequency branch and the low-frequency branch are trained separately, with the crop/resize strategy, respectively. And for stage two, we jointly trained the two branches and the learnable FCT. It is noteworthy that, due to the different crop/resize strategy, we first interpolate the $\hat{\boldsymbol{F}}_{\boldsymbol{l}}$ to the original full size. Then based on the crop location (x_1, y_1, x_2, y_2) of $\hat{\boldsymbol{I}}_{\boldsymbol{h}}$, we crop the interpolated feature $\hat{\boldsymbol{F}}_{\boldsymbol{l}}'$ and feed both $\hat{\boldsymbol{F}}_{\boldsymbol{l}}'$ and $\hat{\boldsymbol{F}}_{\boldsymbol{h}}$ into the learnable FCT to get the final $\hat{\boldsymbol{I}}$.

For the loss function of the first stage, we take the deep supervision strategy proposed in [53]. The high- and low-branch loss function is defined as:

$$L_{stage_1}^{high} = \sum_{i=1}^{3}(L_1(\hat{\boldsymbol{I}}_{\boldsymbol{h_i}}, \boldsymbol{I}_{\boldsymbol{h_i}}^{\boldsymbol{GT}}) + \lambda_1 \times L_p(\hat{\boldsymbol{I}}_{\boldsymbol{h_i}}, \boldsymbol{I}_{\boldsymbol{h_i}}^{\boldsymbol{GT}})). \tag{8}$$

$$L_{stage_1}^{low} = \sum_{i=1}^{3}(L_1(\hat{\boldsymbol{I}}_{\boldsymbol{l_i}}, \boldsymbol{I}_{\boldsymbol{l_i}}^{\boldsymbol{GT}}) + \lambda_1 \times L_p(\hat{\boldsymbol{I}}_{\boldsymbol{l_i}}, \boldsymbol{I}_{\boldsymbol{l_i}}^{\boldsymbol{GT}})). \tag{9}$$

For the second stage, the loss function is defined as:

$$L_{stage_2} = L_1(\hat{\boldsymbol{I}}, \boldsymbol{I}^{\boldsymbol{GT}}) + \lambda_2 \times L_p(\hat{\boldsymbol{I}}, \boldsymbol{I}^{\boldsymbol{GT}}). \tag{10}$$

We use the perceptual loss from the pretrained VGG16 [30] Network. λ_1 and λ_2 are utilized to balance the two losses.

4 Experiments

Datasets and metrics. In the experiments, we separately train and test our Freqformer on TIP2018 [31], LCDMoire [53], FHDMi [10] and UHDM [45] datasets. TIP2018 contains 150,000 image pairs, with 135,000 used for training and 15,000 for testing. LCDMoire provides 10,000 images for training and 100 for validation. FHDMi consists of 9,981 and 2,019 image pairs for training and testing, respectively, at a resolution of 1920×1080. UHDM includes 4,500 and 500 4K image pairs for training and testing, respectively, and we follow [45] by evaluating at 3840×2160 resolution. To measure the performance of all models, we use PSNR, SSIM [35] and LPIPS [50] as the metrics. Higher PSNR and SSIM, as well as lower LPIPS, indicate higher quality of restored images. PSNR is more sensitive to color variations, SSIM better captures structural moiré patterns, and LPIPS aligns more closely with human perceptual quality [10].

Dataset	Metrics	Input	DMCNN [31]	MDDM [2]	WDNet [18]	MopNet [9]	MBCNN [53]	FHDe^{2}Net [10]	ESDNet [45]	ESDNet-L [45]	Freqformer
TIP2018 [31]	PSNR↑	20.30	26.77	N/A	28.08	27.75	30.03	27.78	29.81	30.11	30.63
	SSIM↑	0.7380	0.8710	N/A	0.9040	0.8950	0.8930	0.8960	0.9160	0.9200	0.9269
LCDMoire [53]	PSNR↑	10.44	35.48	42.49	29.66	N/A	44.04	41.40	44.83	45.34	45.62
	SSIM↑	0.5717	0.9785	0.9940	0.9670	N/A	0.9948	N/A	0.9963	0.9966	0.9967
FHDMi [10]	PSNR↑	17.97	21.54	20.83	N/A	22.76	22.31	22.93	24.50	24.88	25.26
	SSIM↑	0.7033	0.7727	0.7343	N/A	0.7958	0.8095	0.7885	0.8351	0.8440	0.8518
	LPIPS↓	0.2837	0.2477	0.2515	N/A	0.1794	0.1980	0.1688	0.1354	0.1301	0.1253
UHDM [45]	PSNR↑	17.12	19.91	20.09	20.36	19.49	21.41	20.34	22.12	22.42	22.24
	SSIM↑	0.5089	0.7575	0.7441	0.6497	0.7572	0.7932	0.7496	0.7956	0.7985	0.8021
	LPIPS↓	0.5314	0.3764	0.3409	0.4882	0.3857	0.3318	0.3519	0.2551	0.2454	0.2424
-	Params (M)	-	1.426	7.637	3.360	58.565	14.192	13.571	5.934	10.623	6.065

Table 1: Quantitative comparisons with sRGB image demoiréing methods. Red: best and Blue: second-best.

Dataset	Metrics	OSEDiff [37]	AdaIR* [4]	AdaIR [4]	MoCE-IR* [48]	MoCE-IR [48]	RRID [39]	STD-Net [25]	DTNet [40]	Freqformer
FHDMi [10]	PSNR↑	21.81	17.08	21.87	17.40	21.80	24.39	25.05	23.01	25.26
	SSIM↑	0.7764	0.6986	0.8041	0.7031	0.7943	0.8300	0.8414	0.8128	0.8518
	LPIPS↓	0.1516	0.3057	0.1924	0.2969	0.2026	N/A	0.1293	0.1548	0.1253
UHDM [45]	PSNR↑	20.66	16.68	20.96	16.97	20.46	21.98	N/A	19.74	22.24
	SSIM↑	0.7607	0.4949	0.7807	0.4971	0.7644	0.7935	N/A	0.7637	0.8021
	LPIPS↓	0.2394	0.5440	0.3040	0.5366	0.3249	0.3032	N/A	0.2624	0.2424
-	Params (M)	950	28.785	28.785	11.478	11.478	1.448	27.960	7.360	6.065

Table 2: Quantitative comparison with image restoration models and other types of demoiréing models. Red: best and Blue: second-best.

Implementation details. We implement our algorithm using Pytorch on an NVIDIA RTX 4090 GPU. For training on the TIP2018/LCDMoire/FHDMi/UHDM, we set the crop/resize size as 256/256, 512/512, 512/512 and 768/512, respectively. Following the settings of [45], we adopt the Adam optimizer [14] and the cyclic cosine schedule [23], 150 epochs for both the training stages, and the batch size of 2. We set the $\lambda_1 = 1$ and $\lambda_2 = 0.1$. For the high-branch of the Freqformer, from the highest to the lowest resolution, the encoder's N is set to [2,2,2] and the decoder's N is set to [4,4,2]. For low-branch, both the encoder and decoder's N is [1,2,2]. N_f of the learnable FCT is set to 2.

4.1 Comparisons with State-of-the-Art Methods

Quantitative comparison. In Table 1, our proposed Freqformer achieves state-of-the-art performance compared with other image demoiréing methods, including DMCNN [31], MDDM [2], WDNet [18], MopNet [9], MBCNN [53], FHDe^{2}Net [10], and ESDNet [45]. On TIP2018/LCDMoire/FHDMi datasets, Freqformer significantly outperforms other methods by a large margin in all evaluation metrics. This indicates superior reconstruction quality, structural fidelity, and perceptual similarity. On the more challenging UHDM dataset, our Freqformer ranks first in SSIM and LPIPS, and second in PSNR, closely following ESDNet-L. Notably, Freqformer achieves these with only 6.065M parameters, outperforming or matching methods with significantly larger model sizes (*e.g.*, MopNet with 58.565M and MBCNN with 14.192M), highlighting its excellent trade-off between performance and model efficiency.

Additionally, in Table 2, we compare our method with several recent general-purpose image restoration models, including AdaIR [4] and MoCE-IR [48], as well as the generative restoration model OSEDiff [37]. The inclusion of OSEDiff is motivated by the recent success of generative priors in related image restoration

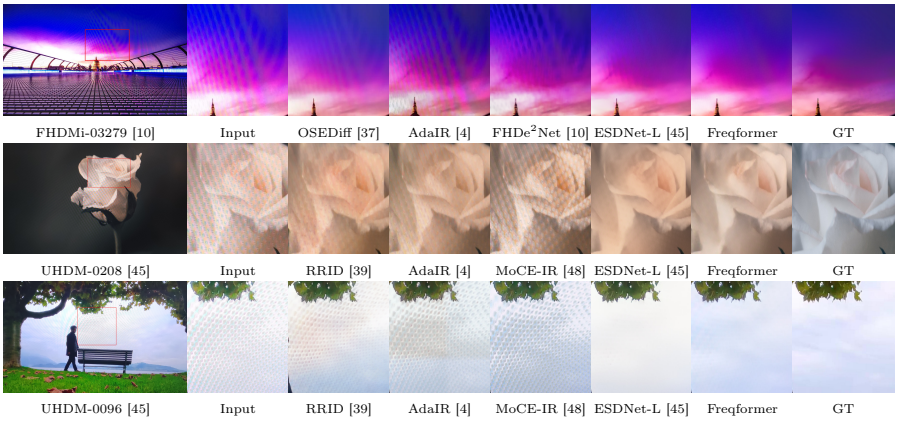

Fig. 6: Visual comparisons of different baselines on FHDMi and UHDM datasets (please zoom in for better details). While OSEDiff, AdaIR, MoCE-IR, and FHDe^{2}Net all struggle with removing moiré artifacts, ESDNet-L also leaves residual traces behind. In addition, these baselines suffer from color distortions and fail to accurately recover the original tones. In contrast, our approach effectively eliminates both moiré and color artifacts, as shown in the last row.

tasks, like [8,20,21,34]. We further compare our method with the state-of-the-art RAW-sRGB image demoiréing model RRID [39], as well as the video demoiréing methods STD-Net [25] and DTNet [40]. The AdaIR* and MoCE-IR* denote direct testing with their all-in-one model, while without * represents the retraining on the two datasets. For models RRID, STD-Net, and DTNet, we adapt their network architectures according to their papers and retrain them on the datasets. Our model surpasses all these models, only inferior to OSEDiff in LPIPS, which has nearly 1B parameters.

Qualitative comparison. Figure 6 presents visual comparisons of image restoration results on the FHDMi and UHDM datasets. Compared with existing methods, our Freqformer not only removes moiré more effectively, preserves fine details with high fidelity, but also mitigates the effects of color distortions and recovers the true color representation, resulting in more realistic and visually pleasing outputs. Competing methods such as OSEDiff, AdaIR, MoCE-IR, and RRID all struggle with moiré patterns that are embedded within the original content. Although ESDNet-L performs better than the others, it still leaves residual moiré artifacts, as observed in the top two rows of Fig. 6. Moreover, in the last row, ESDNet-L and all other baselines exhibit moiré-induced color distortions and fail to restore the original colors accurately. In contrast, our method successfully eliminates both the moiré patterns and color distortions, demonstrating Freqformer's superior ability in both artifact removal and detail restoration, consistent with its leading performance in PSNR, SSIM, and LPIPS.

Model Complexity. As compared in Tabs. 1 and 2, Freqformer not only has fewer parameters compared to MopNet, MBCNN, AdaIR and MoCE-IR, etc, but also achieves a smaller size than ESDNet-L, while superior to both ESDNet

and ESDNet-L. Moreover, Freqformer requires only 2.46 TFLOPS to process 4K images, which is on par with ESDNet (2.24T) and lower than ESDNet-L (3.68T), AdaIR (31.06T) and MoCE-IR (5.43T). In addition, for 4K image inference, Freqformer achieves an inference speed of 0.58 s/image, which is slower than that of ESDNet-L (0.25 s/image). Freqformer requires fewer FLOPS but is slower, likely due to the standard attention operations being less optimized for operator fusion, kernel scheduling, and hardware acceleration than purely convolution-based models. This will be one of our future optimization directions.

4.2 Ablation Study

In this section, we conduct ablation studies on FHDMi to validate our frequency decomposition, training strategy, and network architecture.

Our frequency decomposition v.s. Haar DWT and Block DCT. We compare our approach against two alternative transforms: the 2-level Haar DWT [18,24] (yielding a 10.921M-parameter model with band-specific demoiréing modules) and the Block DCT [10, 53] with a block size of 8 (yielding a 12.715M-parameter model). As reported in Tab. 3, Freqformer equipped with our proposed transform significantly outperforms both variants. Furthermore, visual comparisons in Fig. 7 demonstrate that images restored using either Haar DWT or Block DCT suffer from severe detail degradation. Specifically, the first row of Fig. 7 shows that text edges appear distorted, jagged, and blurred. Moreover, as illustrated in the second row of Fig. 7, neither of these alternative transforms completely eradicates the moiré patterns, leaving visible residual artifacts.

Model	PSNR	SSIM	LPIPS	Params (M)
Freqformer	25.26	0.8518	0.1253	6.065
Freqformer w/ Haar DWT	23.01	0.8036	0.1810	10.921
Freqformer w/ Block DCT	23.22	0.8073	0.1569	12.715

Table 3: Frequency Decomposition Comparison.

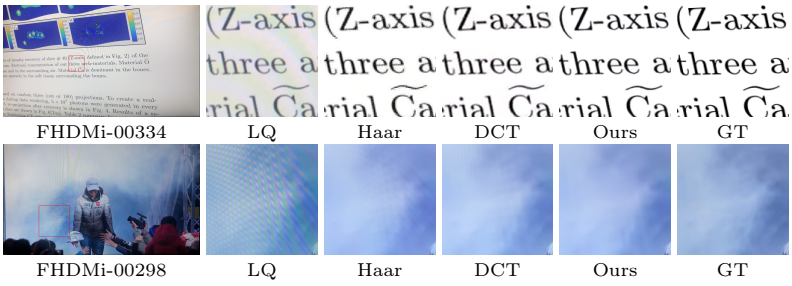

Fig. 7: Visual comparisons of the Haar DWT, Block DCT, and ours.

Low-frequency branch training via resize. As we proposed in our method, the resize operation contains many benefits during the low-branch training, which are demonstrated in Tab. 4. We also train an ESDNet-L (10M) to verify our findings in the same setting. Outcomes show that the crop-training strategy prevents the model from leveraging the low-frequency component's inherent robustness to resizing and hinders effective global information aggregation for color

Model (Low branch)	PSNR	SSIM	LPIPS	Params (M)
Freqformer w/ crop	23.13	0.9220	0.0695	2.695
Freqformer w/ crop & tile	25.90	0.9363	0.0486	2.695
Freqformer w/ resize	29.90	0.9568	0.0291	2.695
ESDNet-L w/ crop	23.01	0.9192	0.0736	10.623
ESDNet-L w/ crop & tile	25.35	0.9295	0.0575	10.623
ESDNet-L w/ resize	30.39	0.9570	0.0277	10.623

Table 4: Low-branch resizing.

correction, resulting in significant degraded performance. We further introduce a crop-tile inference strategy (crop the low-frequency part to 512×512 patches with 32 overlap and then assemble them) for the crop-training model, but it is still inferior to the resize-strategy.

Learnable Frequency Composition Transform and dual-branch training. In Tab. 5, we demonstrate the importance of our frequency decomposition and two-branch training strategy. As shown, ESDNet and Freqformer, using direct modeling without frequency decomposition (w/o F.D.), perform worse than the two-branch approach. Moreover, incorporating a learnable FCT improves fusion consistency and pattern coherence with few parameters.

Model	PSNR	SSIM	LPIPS	Params (M)
ESDNet w/ learnable FCT	25.20	0.8511	0.1244	12.567
ESDNet w/ fixed FCT	25.11	0.8469	0.1296	11.868
ESDNet-L (w/o F.D.)	24.88	0.8440	0.1301	10.623
Freqformer w/ learnable FCT	25.26	0.8518	0.1253	6.065
Freqformer w/ fixed FCT	25.13	0.8478	0.1298	5.876
Freqformer-L (w/o F.D.)	23.92	0.8409	0.1284	5.851

Table 5: Effects of learnable FCT and frequency decomposition.

Freqformer component ablation. As shown in Tab. 6, hierarchical fusion (denoted as H.F. in the Table) improves performance with minimal computational and parameter overhead. Replacing RDDB with deeper transformer layers shows that RDDB is more efficient and outperforms the pure transformer design with faster speed. We further replace SCP (1×1 convolution and a 3×3 depth-wise convolution) with a full 3×3 convolution that uses twice as many parameters, yet we observe no performance improvement and even a speed reduction. We also explored the sophisticated DAT [1] block as a baseline, but it underperforms our SA-CA, which more effectively aggregates local features.

Model (High branch)	PSNR	SSIM	Params (M)	Speed (s)
Freqformer	29.20	0.8805	3.181	0.128
Freqformer (w/o H.F.)	29.15	0.8804	2.997	0.126
Freqformer (Deeper, w/o RDDB)	29.19	0.8801	2.944	0.140
Freqformer (Full Conv. SCP)	29.18	0.8804	6.451	0.187
Freqformer (DAT Arch.)	29.10	0.8793	3.107	0.196

Table 6: The ablation of the Freqformer's architecture.

5 Conclusion

In this paper, we propose Freqformer, a novel demoiréing framework leveraging frequency separation to address complex moiré patterns. By introducing a tailored frequency decomposition strategy and a learnable FCT module, we effectively disentangle moiré artifacts into high-frequency textures and low-frequency color distortions. Furthermore, we design a dual-branch architecture with asymmetric learning strategies: crop-based training for high frequencies and resize-based training for low frequencies. Additionally, we introduce a lightweight yet effective SA-CA module to boost moiré-sensitive feature learning while maintaining efficiency, making our model suitable for deployment on high-resolution images and edge devices. This study highlights how integrating frequency decomposition, learnable fusion, and spatial-channel attention advances image demoiréing, offering promising directions for future exploration.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62501386), CCF-Tencent Rhino-Bird Open Research Fund, and CAAI-Tencent Rhino-Bird Open Research Fund. This work is also sponsored by AI Hundred Schools Program and is carried out using the Ascend AI technology stack.

References

1. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: ICCV (2023)
2. Cheng, X., Fu, Z., Yang, J.: Multi-scale dynamic feature encoding network for image demoiréing. In: ICCVW (2019)
3. Cheng, Y., Liu, X., Yang, J.: Recaptured raw screen image and video demoireing via channel and spatial modulations. NeurIPS (2023)
4. Cui, Y., Zamir, S.W., Khan, S., Knoll, A., Shah, M., Khan, F.S.: AdaIR: Adaptive all-in-one image restoration via frequency mining and modulation. In: ICLR (2025)
5. Dai, P., Yu, X., Ma, L., Zhang, B., Li, J., Li, W., Shen, J., Qi, X.: Video demoireing with relation-based temporal consistency. In: CVPR (2022)
6. Fujieda, S., Takayama, K., Hachisuka, T.: Wavelet convolutional neural networks for texture classification. arXiv preprint arXiv:1707.07394 (2017)
7. Gueguen, L., Sergeev, A., Kadlec, B., Liu, R., Yosinski, J.: Faster neural networks straight from jpeg. NeurIPS (2018)
8. Guo, J., Chen, Z., Li, W., Guo, Y., Zhang, Y.: Compression-aware one-step diffusion model for jpeg artifact removal. In: ICCV (2025)
9. He, B., Wang, C., Shi, B., Duan, L.Y.: Mop moire patterns using mopnet. In: ICCV (2019)
10. He, B., Wang, C., Shi, B., Duan, L.Y.: Fhde 2 net: Full high definition demoireing network. In: ECCV (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

12. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: CVPR (2017)
13. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: ICCV (2017)
14. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
15. Levinskis, A.: Convolutional neural network feature reduction using wavelet transform. ELEKTRON ELEKTROTECH (2013)
16. Li, Q., Shen, L., Guo, S., Lai, Z.: Wavelet integrated cnns for noise-robust image classification. In: CVPR (2020)
17. Liu, F., Yang, J., Yue, H.: Moiré pattern removal from texture images via low-rank and sparse matrix decomposition. In: VCIP (2015)
18. Liu, L., Liu, J., Yuan, S., Slabaugh, G., Leonardis, A., Zhou, W., Tian, Q.: Wavelet-based dual-branch network for image demoiréing. In: ECCV (2020)
19. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-cnn for image restoration. In: CVPRW (2018)
20. Liu, X., Wang, Y., Chen, Z., Cao, J., Zhang, H., Zhang, Y., Yang, X.: One-step diffusion model for image motion-deblurring. arXiv preprint arXiv:2503.06537 (2025)
21. Liu, X., Zhou, Z., Xu, Z., Cao, J., Chen, Z., Zhang, Y.: Fidediff: Efficient diffusion model for high-fidelity image motion deblurring. In: ICLR (2026)
22. Liu, Y., Li, Q., Sun, Z.: Attribute-aware face aging with wavelet-based generative adversarial networks. In: CVPR (2019)
23. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2017)
24. Luo, X., Zhang, J., Hong, M., Qu, Y., Xie, Y., Li, C.: Deep wavelet network with domain adaptation for single image demoireing. In: CVPRW (2020)
25. Niu, Y., Xu, R., Lin, Z., Liu, W.: Std-net: Spatio-temporal decomposition network for video demoiréing with sparse transformers. TCSVT (2024)
26. Oyallon, E., Belilovsky, E., Zagoruyko, S.: Scaling the scattering transform: Deep hybrid networks. In: ICCV (2017)
27. Peng, Y.T., Hou, C.H., Lee, Y.C., Yoon, A.J., Chen, Z., Lin, Y.T., Lien, W.C.: Image demoiréing via multi-scale fusion networks with moiré data augmentation. IEEE Sens. J. (2024)
28. Siddiqui, H., Boutin, M., Bouman, C.A.: Hardware-friendly descreening. TIP (2009)
29. Sidorov, D.N., Kokaram, A.C.: Suppression of moiré patterns via spectral analysis. In: VCIP (2002)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)
31. Sun, Y., Yu, Y., Wang, W.: Moiré photo restoration using multiresolution convolutional neural networks. TIP (2018)
32. Wang, C., He, B., Wu, S., Wan, R., Shi, B., Duan, L.Y.: Coarse-to-fine disentangling demoiréing framework for recaptured screen images. TPAMI (2023)
33. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. IJCV (2024)
34. Wang, J., Tang, Y., Gong, J., Li, J., Li, S., Liu, L., Lan, J., Liu, Y., Zhang, Y.: Spectral and trajectory regularization for diffusion transformer super-resolution. arXiv preprint 2603.06275 (2026)
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
36. Williams, T., Li, R.: Wavelet pooling for convolutional neural networks. In: ICLR (2018)

37. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. NeurIPS (2024)
38. Xiao, Z., Lu, Z., Wang, X.: P-bic: Ultra-high-definition image moiré patterns removal via patch bilateral compensation. In: ACM MM (2024)
39. Xu, S., Song, B., Chen, X., Liu, X., Zhou, J.: Image demoireing in raw and srgb domains. In: ECCV (2024)
40. Xu, S., Song, B., Chen, X., Zhou, J.: Direction-aware video demoireing with temporal-guided bilateral learning. In: AAAI (2024)
41. Yang, C., Yang, Z., Ke, Y., Chen, T., Grzegorzek, M., See, J.: Doing more with moiré pattern detection in digital photos. TIP (2023)
42. Yeh, C.H., Lo, C., He, C.H.: Multibranch wavelet-based network for image demoiréing. Sensors (2024)
43. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: ICCV (2019)
44. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. ICLR (2015)
45. Yu, X., Dai, P., Li, W., Ma, L., Shen, J., Li, J., Qi, X.: Towards efficient and scale-robust ultra-high-definition image demoiréing. In: ECCV (2022)
46. Yuan, S., Timofte, R., Slabaugh, G., Leonardis, A., Zheng, B., Ye, X., Tian, X., Chen, Y., Cheng, X., Fu, Z., et al.: Aim 2019 challenge on image demoireing: Methods and results. In: ICCVW (2019)
47. Yue, H., Cheng, Y., Mao, Y., Cao, C., Yang, J.: Recaptured screen image demoiréing in raw domain. TMM **25** (2022)
48. Zamfir, E., Wu, Z., Mehta, N., Tan, Y., Paudel, D.P., Zhang, Y., Timofte, R.: Complexity experts are task-discriminative learners for any image restoration. In: CVPR (2025)
49. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022)
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
51. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)
52. Zhang, Y., Lin, M., Li, X., Liu, H., Wang, G., Chao, F., Ren, S., Wen, Y., Chen, X., Ji, R.: Real-time image demoireing on mobile devices. ICLR (2023)
53. Zheng, B., Yuan, S., Slabaugh, G., Leonardis, A.: Image demoireing with learnable bandpass filters. In: ICCV (2020)