FPANet: Frequency-based Video Demoiréing using Frame-level Post Alignment

Gyeongrok Oh[a], Heon Gu[b], Jinkyu Kim*[c], Sangpil Kim*[a]

[a]*Department of Artificial Intelligence, Korea University, Seoul, 02841, South Korea*
[b]*LG Display Research Center, Seoul, 07795, South Korea*
[c]*Department of Computer Science and Engineering, Korea University, Seoul, 02841, South Korea*

Abstract

Interference between overlapping gird patterns creates moiré patterns, degrading the visual quality of an image that captures a screen of a digital display device by an ordinary digital camera. Removing such moiré patterns is challenging due to their complex patterns of diverse sizes and color distortions. Existing approaches mainly focus on filtering out in the spatial domain, failing to remove a large-scale moiré pattern. In this paper, we propose a novel model called FPANet that learns filters in both frequency and spatial domains, improving the restoration quality by removing various sizes of moiré patterns. To further enhance, our model takes multiple consecutive frames, learning to extract frame-invariant content features and outputting better quality temporally consistent images. We demonstrate the effectiveness of our proposed method with a publicly available large-scale dataset, observing that ours outperforms the state-of-the-art approaches, including ESDNet, VDmoire, MBCNN, WDNet, UNet, and DMCNN, in terms of the image and video quality metrics, such as PSNR, SSIM, LPIPS, FVD, and FSIM.

Keywords: Moiré Removal, Video Restoration

1. Introduction

Moiré patterns are commonly observed in images, which are taken by ordinary digital cameras, capturing a screen of a digital display device. This

*Corresponding authors: Sangpil Kim (spk7@korea.ac.kr) and Jinkyu Kim (jinkyukim@korea.ac.kr)

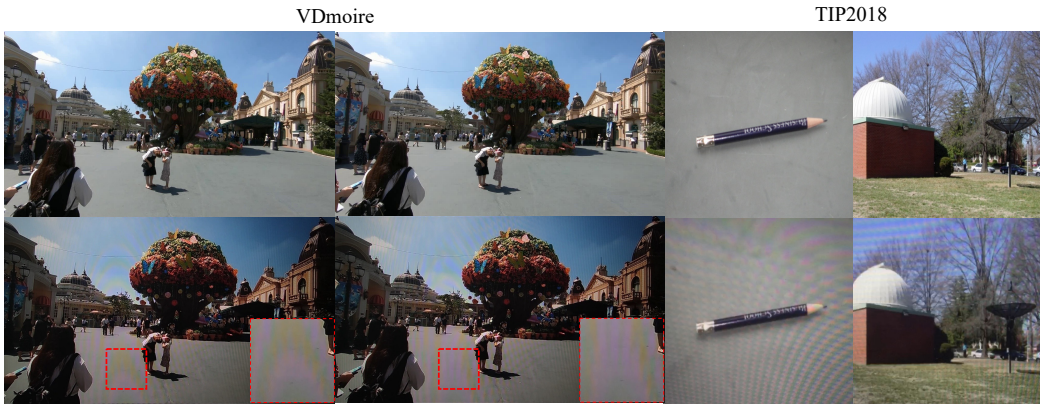

Figure 1: Examples of video frames or images that have moiré patterns as visual artifacts. Note that (a) and (b) are consecutive frames, which are extracted from the publicly available VDmoire dataset, while (c) is from the TIP2018 dataset. Target images are shown in the first row, and images with moiré patterns are shown at the bottom. For better visualization, we also provide magnified patches.

is mainly due to frequency aliasing – an interference between overlapping grid patterns, such as camera sensor grid and display pixel grid or textures on clothes. Such interference depends on the degree of overlap, resulting in diverse and complex patterns, including stripes, curves, and ripples, which are analytically infeasible (see Figure 1). They also significantly degrade the visual quality of images, often causing severe color distortions of the original content. This makes it challenging to remove such moiré patterns and restore the original image.

Learning-based approaches have been introduced to train a model to filter out visual artifacts and restore the original contents automatically. Recently, ConvNet-based hierarchical architectures have been explored to remove various sizes of moiré patterns [1, 2, 3, 4]. However, their performance depends on their receptive fields, often failing to remove large-scale artifacts.

To address these issues, recent works [5, 2, 3] suggest that co-learning in the frequency domain is useful to deal with various sizes of moiré patterns. However, a frequency spectrum of the moiré pattern is often intermingled with those of the original contents. Moreover, the camera's Bayer filter mosaic is imbalanced over the RGB channel, producing different intensities of moiré patterns. To address this issue, we advocate for leveraging amplitude and phase components separately. In our experiment (see Figure 2), color

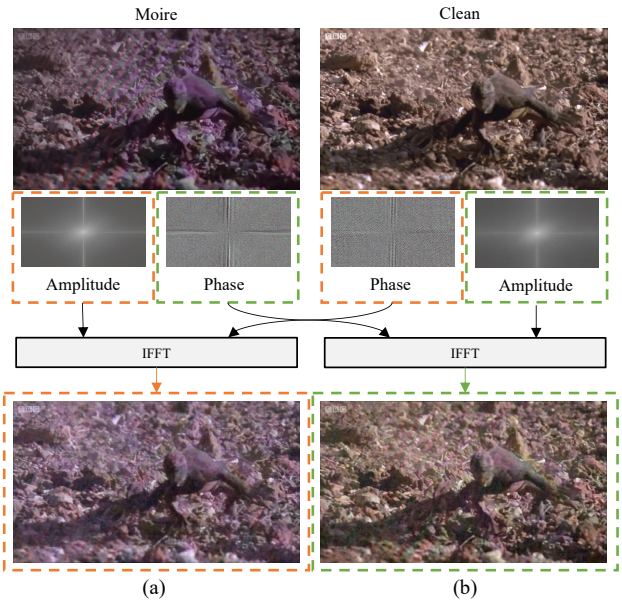

Figure 2: Visualization on effect of amplitude and phase component over moiré patterns. The orange box generates synthetic image combining with moiré image amplitude and clean image phase. The green box generates synthetic image combining with moiré image phase and clean image amplitude.

distortion or degradation is often caused by the signal's amplitude component, while moiré patterns remain in the phase component. Thus, instead of directly utilizing frequency components as done in conventional approaches, encoding these components in a separate branch will accelerate the learning procedure.

Thus, we propose a novel module called Frequency Selection Fusion (FSF), which first transforms the spatial information into the frequency domain spectrum using Fast Fourier Transform (FFT). Its amplitude and phase components are extracted and encoded separately. We further apply a selective fusion strategy to merge both components. Moreover, to maximize the representation power in the spatial domain, we adopt multi-scale architecture to improve restoring fine-grained details in the spatial domain.

This is only a part of the story. We further expand our work for video demoireing tasks where we now take multiple consecutive frames, outputting temporally consistently restored images. We argue that leveraging multiple frames with similar contents but slightly different moiré patterns is helpful to

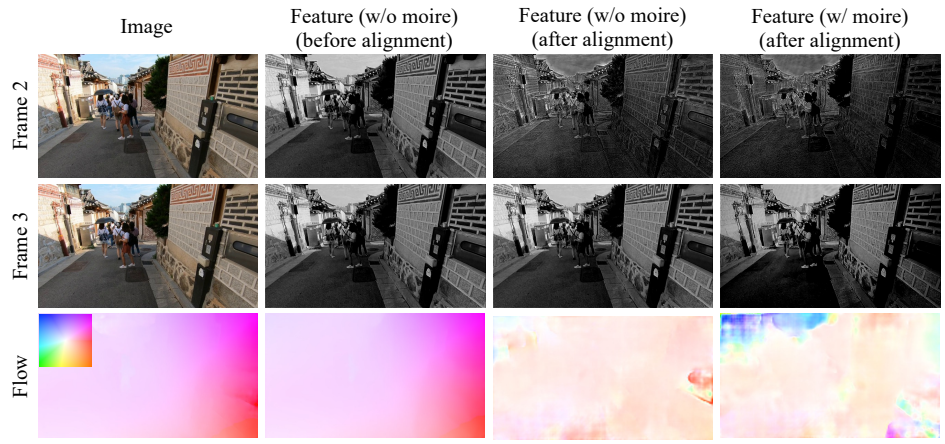

Figure 3: Example for misalignment caused by moiré patterns. Aligned features between reference frame (Frame 2) and target frame (Frame 3) with or without moiré patterns are listed in first row. To measure the accuracy for alignment, we calculate optical flow using PWC-Net [12], illustrating the relative motion by the color coding that indicates the motion vectors (direction and magnitude) with color intensity.

filter out such distortions, while augmented views of original contents are used to restore main content features. Thus, a key module for video demoiréing models is aligning multiple consecutive frame inputs. Conventional aligning approaches [6, 7] used recurrent architectures (e.g. use bidirectional or unidirectional recurrent units followed by using an image warping technique), but they are prone to accumulate misalignment errors in a long sequence input, resulting in poor restoration performance. Enlarging receptive fields by applying deformable convolution [8, 9, 10, 11] is an alternative way to align features of multiple consecutive frames. However, aligning features without removing moiré patterns often yields a large misalignment error, as shown in Figure 3. Thus, to address this issue, we propose an improved alignment module called Post Align Module (PAM). Unlike existing approaches that use an alignment module in the early stage (and are separated from the main network), we apply such an alignment module in the multiple intermediate stages where more distortions are getting removed. This allows the alignment module to be robust against moiré distortions.

To demonstrate the effectiveness of our methods, we conducted numerous ablation studies and show competitive results in various evaluation metrics, peak signal-to-ratio (PSNR), structural similarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Frechet Video Distance (FVD), and

Feature-SIMilarity (FSIM). To quantitatively evaluate FPANet, we compare the difference between ground truth clean images and estimated images. Furthermore, we observed that our proposed methods are effective in removing moiré patterns and reconstructing a fine-detailed image compared with previous state-of-the-art methods. Our main contributions are summarized as follows:

- We propose a novel building module called Frequency Selection Fusion (FSF), which consists of two modules: (i) Frequency Selection Module (FSM) and (ii) Cross Scale Fusion Module (CSFM). Our FSM transforms the input into the frequency domain spectrum and operates on their amplitude and phase components to remove large-scale moiré patterns without creating undesired color artifacts. CSFM extracts multi-scale features to help restore fine-grained details in the spatial domain. We demonstrate that these modules are effective in removing moiré patterns without creating noticeable visual artifacts.

- To deal with multiple consecutive frames for video demoiréing tasks, we introduce an improved temporal feature alignment module (PAM) deployed in the multiple intermediate stages to remove various types of moiré distortions.

- We compare our model with current state-of-the-art approaches on a publicly available video demoiréing dataset called VDmoire, and ours outperforms existing approaches in various image and video quality metrics, such as PSNR, SSIM, LPIPS, FVD, and FSIM.

2. Related Work

2.1. Image Demoiréing

The interference between two similar signals creates moiré pattern. In particular, when taking a picture of a display, the moiré pattern is caused by the misalignment between the grid of the display and the camera filter. Moiré patterns that are not present in the original image can seriously degrade the image quality because of the shape that looks like a ripple, ribbon, or stripe and the color change. To restore the original image without moiré pattern, previous literature [1, 13, 14, 5, 15, 16, 2, 17, 11] delicately designed neural networks. DMCNN [1] is the first convolution neural network that uses hierarchical stages in order to remove various size moiré patterns and releases

a benchmark dataset for image demoiré. MopNet [13] utilizes the properties of moiré patterns such as scale, color, and shape. FDNet [15] predicts a clean image given image pairs, degraded by moiré pattern and defocused moiré free image. However, these two methods require additional input except for noisy image and clean image pairs. WDNet [5] and MBCNN [2] utilize frequency prior with wavelet transform and implicit Discrete Cosine Transform (DCT) [18]. However, these methods are still difficult to remove moiré patterns clearly because they do not sufficiently leverage property of moiré patterns in fourier domain. MRGAN [17] is an unsupervised method using Generative Adversarial Network [19] for removing this type of pattern. But, their model may not be able to restore the color of the original image or remove the artifacts because they did not use moiré and clean image pairs during training. FHDe^2Net [14], ESDNet [4] are responsible for high-resolution image demoiré. These two methods use multi-scale features for handling large-scale images and release high-resolution datasets. Another explored research problem, video demoiré [11], is the first work for restoring video using relation-based consistency loss. Since the moiré patterns desaturate not only the color of the entire image but also make the shape of objects difficult to recognize, this method fails to completely recover the original color and fine-grained details.

In this paper, we concentrate on the video demoiré that is less explored in the former literature. Our proposed methods are the first works in demoiré task to remove moiré texture and recover the original color in the frequency domain using amplitude and phase. In addition, we design Post Align Module (PAM) to leverage the neighboring frames as auxiliary information without disruption.

2.2. Learning in the Frequency Domain

Recently, learning in the frequency domain has been widely studied in various fields [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. Some of these methods [21, 22, 23] applied the spectral block correspondence to the vanilla convolution block where they utilize each grid value, which is associated with the frequency components. This is advantageous to leverage a much larger receptive field than the conventional convolution operations. Also, FcaNet [27] proposed frequency channel attention using the discrete cosine transform (DCT) to compress each feature over pre-defined DCT bases. Cai *et al.* [26] suggests the usefulness of amplitude and phase information for generating photo-realistic images in the generative models. Given this, Focal

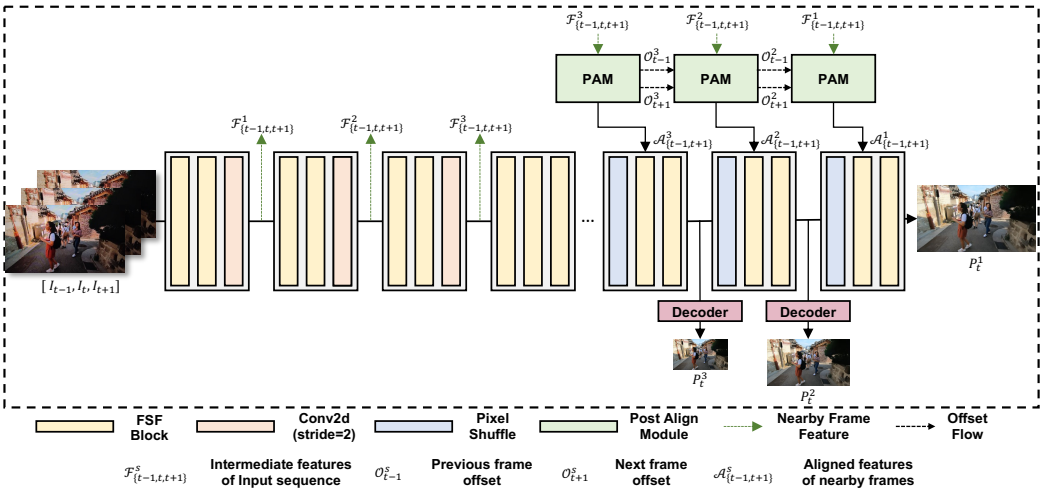

Figure 4: An overview of our proposed FPANet. FPANet is based on an encoder-decoder architecture design. Frequency Selection Fusion (FSF) block is the core part of FPANet that is responsible for removing moiré patterns using frequency domain information. Also, Post Align Module (PAM) is used for leveraging temporal information between nearby frames.

Frequency Loss [32] is utilized as a regularizer for the generative models. Also, FSDGN [25] uses amplitude information to guide restoring the original image with haze. However, they separate spatial and frequency branches with the unidirectional flow (i.e. frequency to spatial) that can not fully use both information. LaMa [29] shows remarkable performance in the inpainting task with Fast Fourier Convolution [22], which has the advantage of recovering repetitive structures such as fences. We also follow this line of work where we exploit amplitude and phase components separately via our proposed Frequency Selection Module (FSM) for the video demoiréing task.

2.3. Multi-frame Encoding and Alignment

Techniques for encoding multiple consecutive frames can mainly be categorized into two ways: (i) window-based and (ii) recurrent-based methods. Window-based approaches [10, 33, 34, 35, 9] often use deformable convolution to have enlarged receptive field followed by encoding concatenated multiple frames. For example, TDAN [9] predicts offsets of the convolution kernel to align multi-frame features, EDVR [10] used Pyramid Cascade Deformable (PCD), which is the hierarchical architecture to facilitate precise

offset prediction, which is widely adopted in the field of video restoration task [10, 33, 35, 11, 9].

Recurrent-based methods often utilize recurrent units for encoding a sequence of frames: e.g. BasicVSR [6] and BasicVSR++ [7] leverage bidirectional flow to align multiple frames over time. However, dealing with long-term dependency issues remains challenging with the recurrent architectures, and their aligning performance is sub-optimal. Thus, in this paper, we pursue to align features between neighboring frames without interference of moiré patterns and computational burden. Thus, we present the Post Align Module (PAM) that aligns nearby feature more accurately and it can be easily plugged in conventional convolution block.

3. Method

In this paper, we propose a novel video demoiréing method called FPANet (Frequency-based video demoiré using frame-level Post Alignment). Following existing work, our model also adopts encoder-decoder architecture, which encodes an input image with moiré to a high-level feature representation, retaining geometric structural information for restoring high-quality images. Instead of taking a single image I_t at a certain timestep t, our model takes the previous I_{t-1} and the next frame I_{t+1} as well. Thus, it takes three consecutive frames as input: $\mathbf{I} = \{I_{t-1}, I_t, I_{t+1}\}$. Given this, our model is trained to produce a restored image P_t.

As shown in Figure 4, our model consists of the following three main components: (i) Frequency Selection Module (FSM), (ii) Cross Scale Fusion Module (CSFM), and (iii) Post Align Module (PAM). FSM first converts the incoming spatial-domain features into frequency-domain features, then its amplitude and phase components are separately processed to remove moiré patterns, retaining geometric structural information (Section 3.2). CSFM processes the spatial information of a given image, effectively capturing multi-scale features for restoring both global and fine-grained features (Section 3.3). Lastly, PAM effectively aligns multiple consecutive frames with reducing visual artifacts (Section 3.4).

3.1. Preliminary

One key component behind our model is encoding an input image in the frequency domain. Thus, we first explain how we convert a spatial domain feature $F(x, y) \in \mathbb{R}^{H \times W \times 3}$ into frequency components $\mathcal{F}(u, v)$. One common

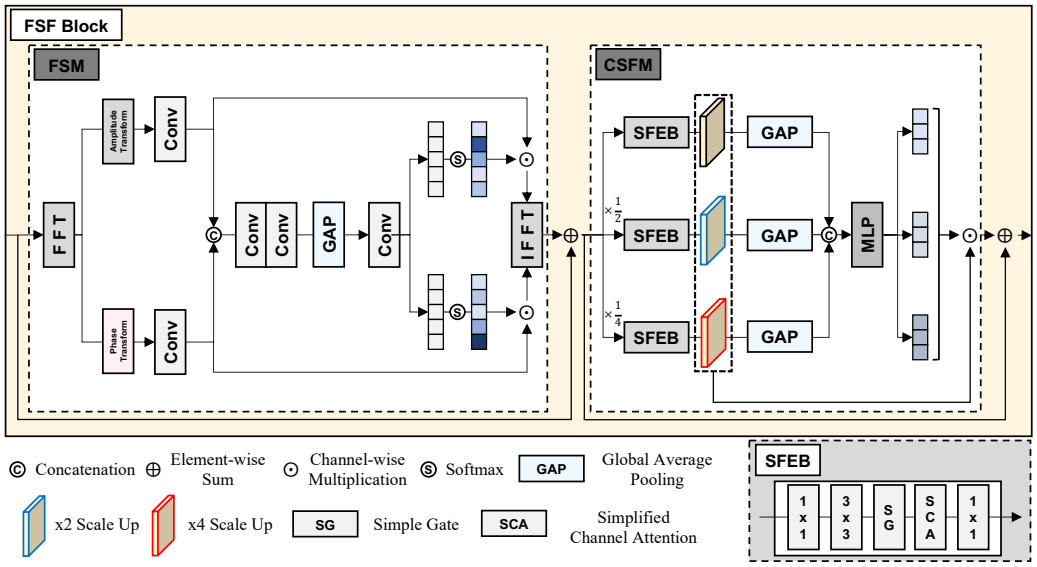

Figure 5: An overview of Frequency Selection Fusion Module, which consists of two main building blocks: (i) Frequency Selection Module (FSM) and (ii) Cross Scale Fusion Module (CSFM). FSM converts spatial information into frequency domain components, then encodes such frequency information to remove various types of moiré patterns effectively. CSFM processes a given spatial domain feature in a multi-scale processing manner.

way to convert an image into frequency components is via Discrete Fourier Transform (DFT), which decomposes spatial information (i.e. images) into frequency components using the following equation:

$$\mathcal{F}(u,v) = \sum_{x=0}^{W-1}\sum_{y=0}^{H-1} F(x,y)\cdot e^{-i2\pi(\frac{x}{W}u+\frac{y}{H}v)}\,, \tag{1}$$

where H and W indicate the height and width of the image, respectively. This frequency information $\mathcal{F}(u,v)$ can further be decomposed into amplitude $\mathcal{F}^{\mathcal{A}}(u,v)$ and phase $\mathcal{F}^{\mathcal{P}}(u,v)$ components as follows:

$$\begin{aligned}\mathcal{F}^{\mathcal{A}}(u,v) &= \sqrt{\mathrm{Re}^2(u,v)+\mathrm{Im}^2(u,v)}\\ \mathcal{F}^{\mathcal{P}}(u,v) &= \arctan(\mathrm{Im}(u,v)/\mathrm{Re}(u,v))\end{aligned} \tag{2}$$

where $\mathrm{Re}(u,v)$ and $\mathrm{Im}(u,v)$ denote real and imaginary value of complex Fourier coefficient, respectively (i.e. $\mathcal{F}(u,v) = \mathrm{Re}(u,v) + \mathrm{Im}(u,v)$).

3.2. Frequency Selection Module (FSM)

As shown in Figure 4, our model consists of multiple Frequency Spatial Fusion (FSF) blocks in series, each of which contains two components: (i) Frequency Selection Module (FSM) and (ii) Cross Scale Fusion Module (CSFM). In this section, we explain the details of the FSM, and we will explain the CSFM in the next section. Note that H_i, W_i, and C_i represent the height, width, and channel dimensions at i-th stage, respectively.

As shown in Figure 5 (left), the i-th FSM block takes as an input a (spatial domain) feature $F_i(x, y) \in \mathbb{R}^{H_i \times W_i \times C_i}$, outputting an encoded feature of the same dimension with a skip connection, i.e. $\texttt{FSM}(F_i(x, y)) + F_i(x, y)$. The first step in the FSM block transforms the spatial features into frequency features $\mathcal{F}_i(u, v) \in \mathbb{R}^{H_i \times (\lfloor \frac{W_i}{2} \rfloor + 1) \times C_i}$ using 2D Fast Fourier Transform (FFT) [36]. We only take half of the matrix to improve computational efficiency because 2D FFT produces a conjugate symmetric Hermitian matrix.

The resulting frequency features are then decomposed into amplitude $\mathcal{F}_i^{\mathcal{A}}(u, v) \in \mathbb{R}^{H_i \times (\lfloor \frac{W_i}{2} \rfloor + 1) \times C_i}$ and phase $\mathcal{F}_i^{\mathcal{P}}(u, v) \in \mathbb{R}^{H_i \times (\lfloor \frac{W_i}{2} \rfloor + 1) \times C_i}$ components using Eq. 2. Given these features, we apply a selective fusion strategy to filter out moiré patterns effectively, which generally depends on both amplitude and phase components. Inspired by prior works [37, 38, 39], we (channel-wise) concatenate both features, followed by 1×1 and 3×3 convolution layers in series to encode combined information. Then, we generate confidence maps ($\alpha_i \in \mathbb{R}^{C_i}$ and $\beta_i \in \mathbb{R}^{C_i}$) using a global average pooling (GAP) layer, another 1×1 convolution layer, and a softmax layer. These (channel-wise) confidence maps are then multiplied by $\mathcal{F}_i^{\mathcal{A}}(u, v)$ and $\mathcal{F}_i^{\mathcal{P}}(u, v)$ accordingly. Lastly, we apply 2-D Inverse Fast Fourier Transform (IFFT) to transform frequency features into spatial features as follows:

$$\texttt{FSM}(F_i(x, y)) = \texttt{IFFT}(\texttt{Conv}(\alpha_i \odot \mathcal{F}_i^{\mathcal{A}}(u, v)), \texttt{Conv}(\beta_i \odot \mathcal{F}_i^{\mathcal{P}}(u, v)))) \tag{3}$$

where $\odot$ is channel-wise multiplication.

3.3. Cross Scale Fusion Module (CSFM)

As shown in Figure 5 (right), we further use Cross Scale Fusion Module (CSFM) to encode multi-scale features. As suggested by recent works [1, 2, 3, 4], such a multi-scale encoding is advantageous to restore fine-grained visual details. Following Yu *et al.* [4], we use the pyramid feature architecture to encode multi-scale features (i.e. original, $\times 1/2$ down-scaled, and $\times 1/4$ down-scaled features). However, we use a Simple Feature Extraction Block

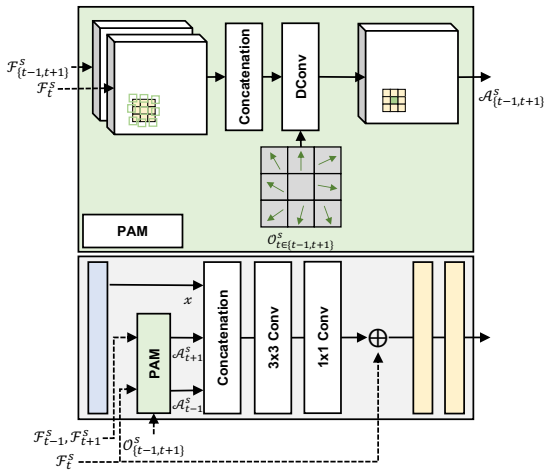

Figure 6: Schematic for Post Align Module (PAM). It is plugged into the decoder of whole architecture to fuse multiple features. The dotted line indicates the skip connection from encoder and solid line represents the straightforward flow in the module.

(SFEB) instead of dilated residual dense block, which is due to improve the computational efficiency. In Figure 5 (bottom right), we illustrate SFEB, which consists of five layers including 1×1 and 3×3 convolution layers, Simple Gate (SG), and Simplified Channel Attention (SCA) [40]. After extracting multi-scale features from SFEB, we apply the Dynamic Fusion method [4] in the same way to calibrate each scale feature into a unified feature map maintaining important information.

3.4. Post Align Module (PAM)

Recent feature alignment methods [10] using Pyramid Cascade Deformable (PCD) or its deformation need to utilize a separated module and align each feature before reconstructing high-quality frames. However, moiré patterns are different in both size and properties, compared with the other artifacts such as blur, rain streak, and gaussian noise. Specifically, moiré patterns lead to severe color degradation and confused texture, which is an undifferentiated original texture.

Therefore, we propose the Post Align Module (PAM) for aligning features between nearby frames without the interference of moiré patterns. As shown in Figure 6, PAM is placed in the decoder per each stage. PAM uses features that are the output of each stage encoders over the sequence of frames

$\mathbf{I} = \{I_{t-1}, I_t, I_{t+1}\}$ as an input. We define each input as $\mathcal{F}^s_{\{t-1,t,t+1\}}$, where s denotes each stage (e.g. $s \in \{1, 2, 3\}$). At the first stage of the PAM, we calculate learnable offset $\mathcal{O}_{\{t-1,t+1\}}$ using the output of the last stage of encoder $\mathcal{F}_{\{t-1,t,t+1\}}$:

$$\mathcal{O}_{\{t-1,t+1\}} = \text{Conv}([\mathcal{F}_{\{t-1,t+1\}}, \mathcal{F}_t]), \tag{4}$$

where $\text{Conv}(\cdot)$ represents the general convolution operation and $[\cdot, \cdot]$ indicates the concatenation operation. Because the direction of the offset is identical and only changes magnitude regardless of the scales, we utilize the previous predicted learnable offset using bilinear interpolation. We formulate following process as:

$$\mathcal{O}^s_{\{t-1,t+1\}} = \text{Conv}([\text{Conv}([\mathcal{F}^s_{\{t-1,t+1\}}, \mathcal{F}^s_t]), \text{Up}(\mathcal{O}^{s-1}_{\{t-1,t+1\}})]), \tag{5}$$

where $\text{Up}(\cdot)$ indicates the bilinear upscaling operation.

Formally, given learnable offset $\mathcal{O}_{\{t-1,t+1\}}$, we apply deformable convolution in order to align features according to the motion of objects. The overall PAM process is represented as:

$$\mathcal{A}^s_{\{t-1,t+1\}} = \text{Dconv}(\mathcal{F}^s_t, \mathcal{O}^s_{\{t-1,t+1\}}), \tag{6}$$

where $\mathcal{A}$ represents the aligned feature and $\text{Dconv}(\cdot)$ denotes the deformable convolution [8]. Last, obtained aligned features $\mathcal{A}_{\{t-1,t+1\}}$ and upsampled features x from the previous block are processed using a conventional fusion method that consists of concatenation and convolution operation for leveraging each information. Note that the PAM can be easily plugged into the decoder and can obtain the aligned features over neighboring frames.

3.5. Loss Function

We train our model end-to-end using the following three loss functions: (i) Multi-scale L_1-based spatial domain loss, (ii) Multi-scale perceptual loss, and (iii) L_1-based frequency domain loss. Our Multi-scale L_1-based pixel-wise loss $\mathcal{L}_s$ quantifies the pixel-wise differences between the target image and the restored image, which is defined as follows:

$$\mathcal{L}_s = \sum_t ||P_t - \hat{I}_t||_1 \tag{7}$$

where $\hat{I}_t$ is the ground-truth (target) image without moiré patterns, and P_t is the restored image by our proposed method.

Following [4], we also use multi-scale perceptual loss by utilizing ImageNet pre-trained VGG19 [41] network. We extract high-level feature representations from the intermediate layer and train a model to minimize the feature-level difference between the target and the predicted images. We define multi-scale perceptual loss as follows:

$$\mathcal{L}_p = \sum_t ||\texttt{VGG}_{19}(P_t) - \texttt{VGG}_{19}(\hat{I}_t)||_1 \tag{8}$$

where $\texttt{VGG}_{19}(\cdot)$ denotes the pre-trained network.

Lastly, we also use a L_1-based frequency domain loss, which quantifies differences between frequency components (i.e. amplitude and phase) of the target and restored images. Our loss function is defined as follows:

$$\mathcal{L}_f = \sum_t ||\mathcal{F}^{\mathcal{A}}(P_t) - \mathcal{F}^{\mathcal{A}}(\hat{I}_t)||_1 + ||\mathcal{F}^{\mathcal{P}}(P_t) - \mathcal{F}^{\mathcal{P}}(\hat{I}_t)||_1 \tag{9}$$

where $\mathcal{F}^{\mathcal{A}}$ and $\mathcal{F}^{\mathcal{P}}$ represent amplitude and phase components. Concretely, we use the following loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_s + \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f \tag{10}$$

where λ_p and λ_f are the hyper-parameters to control the strength of each loss terms.

4. Experiments

4.1. Implementation and Evaluation Details

We train our model end-to-end with AdamW [42] optimizer with the initial learning rate set to 10^{-3}. We use cyclic cosine annealing learning rate schedule [43] that enables partial warm restart optimization, generally improving the convergence rate in gradient-based optimization. Note that we use a grid search to find a better hyperparameter set: we set λ_{vgg} and λ_{freq} to 0.1. Moreover, we set the number of FSF blocks at the encoding and decoding stage to [2, 2, 4] for each scale. The entire architecture also contains 12 FSF blocks between the encoder and decoder.

We use the following three evaluation metrics: (1) Peak Signal to Noise Ratio (PSNR), (2) Structural Similarity index (SSIM) [44], and (3) Learned Perceptual Image Patch Similarity (LPIPS) [45]. These metrics are widely

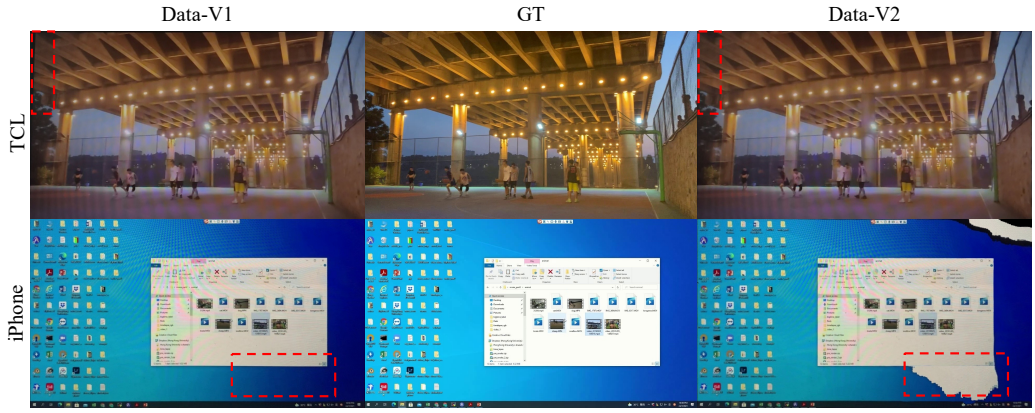

Figure 7: Example of sampled frames in VDmoire [11] that is divided into two types of dataset conforming to different characteristics. The frame captured by two different types of device pairs (TCL and iPhone) is listed in each row. In addition, each column represents an example of two versions of the dataset (V1 and V2) before and after refinement respectively except fot the second column that indicates the ground truth.

used in the image demoiréing task (as well as other image restoration tasks) by quantifying the quality of generated outputs. PSNR quantifies pixel-level similarity, while SSIM [44] uses structural information (from pixel intensities, luminance, and contrasts), providing a more human-like perception metric. Lastly, LPIPS [45] measures perceptual similarity by comparing high-level visual representations from pre-trained networks (e.g., ImageNet [46] pre-trained). To measure the performance of our model for the video demoiré task, we conduct single-frame and multi-frame experiments with the following settings. Specifically, at the multi-frame experiments, we randomly sample three consecutive frames with batch size 8 and crop a 384×384 patch for all experiments over VDmoire dataset [11]. In the single-frame experiments, we randomly sample a single frame and leverage it repetitively instead of consecutive frames.

4.2. Dataset

To evaluate the effectiveness of our proposed methods, we use the following publicly available dataset, i.e. VDmoire [11] dataset. This provides a video demoiréing dataset, which provides 290 source videos and corresponding videos with moiré patterns. To obtain such pairs, the 720p (1080×720) source videos are displayed on the MacBook Pro display (or Huipu v270 display). At the same time, a hand-held camera (iPhoneXR or TCL20 pro

Table 1: Quantitative comparison with the state-of-the-art image (or video) demoiréing approaches: UNet [47], DMCNN [1], ESDNet [4], WDNet [5], MBCNN [3], and VDmoire [11]. We use the publicly available VDmoire dataset (contains TCL-V1 and iPhone-V2 splits) for this experiment. Note that we use **bold** to highlight the best scores among different models. ↑ represents a higher score is better, while ↓ indicates a lower score is better. *Abbr.* Freq: use the frequency components, Multi: use the multiple consecutive frame inputs

Method	Freq	Multi	TCL-V1			iPhone-V2		
			PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DMCNN [1]	-	-	20.321	0.703	0.321	21.816	0.749	0.496
UNet [47]	-	-	20.348	0.720	0.225	21.678	0.790	0.338
WDNet [5]	✓	-	20.576	0.697	0.234	23.971	0.834	0.205
MBCNN [3]	✓	-	21.534	0.740	0.260	24.060	0.849	0.211
VDmoire [11]	-	✓	21.725	0.733	0.202	25.230	0.860	0.157
ESDNet [4]	-	-	**22.026**	0.734	0.199	25.064	0.853	0.165
Ours (*single frame only*)	✓	-	21.577	0.772	0.189	25.215	0.875	0.157
Ours	✓	✓	21.953	**0.784**	**0.173**	**25.446**	**0.883**	**0.146**

camera) captures the screen to create moiré patterns in the recorded frames. Further, to reduce the effect of the misaligned frame correspondences, they estimate the homography matrix using the RANSAC algorithm to align two frames. Despite of these efforts, errors of spatial alignment between the captured frame and ground truth are still remained (see first row in Figure 7). To handle this problem, they present a refined new dataset using optical flow. But, this dataset also has the problem showing distortion caused by inaccurate optical flow (see second row in Figure 7). For comparing ours with the former state-of-the-art methods in diverse settings, we utilize two different datasets (TCL-V1 and iPhone-V2). Note that we call TCL-V1 and iPhone-V2 based on the cameras used (i.e. TCL20 pro camera and iPhone XR camera).

4.3. Quantitative Evaluation

As we summarized in Table 1, we start by quantitatively comparing the quality of generated outputs with state-of-the-art approaches, including UNet [47], DMCNN [1], ESDNet [4], WDNet [5], MBCNN [3], and VDmoire [11]. Note that VDmoire [11] and ours utilize multi-frame inputs, while others are based only on single-frame image input. Also, we use the VDmoire dataset [11] for this evaluation, and the same hyper-parameters commonly used for image demoiréing, such as patch size, are used in this experiment

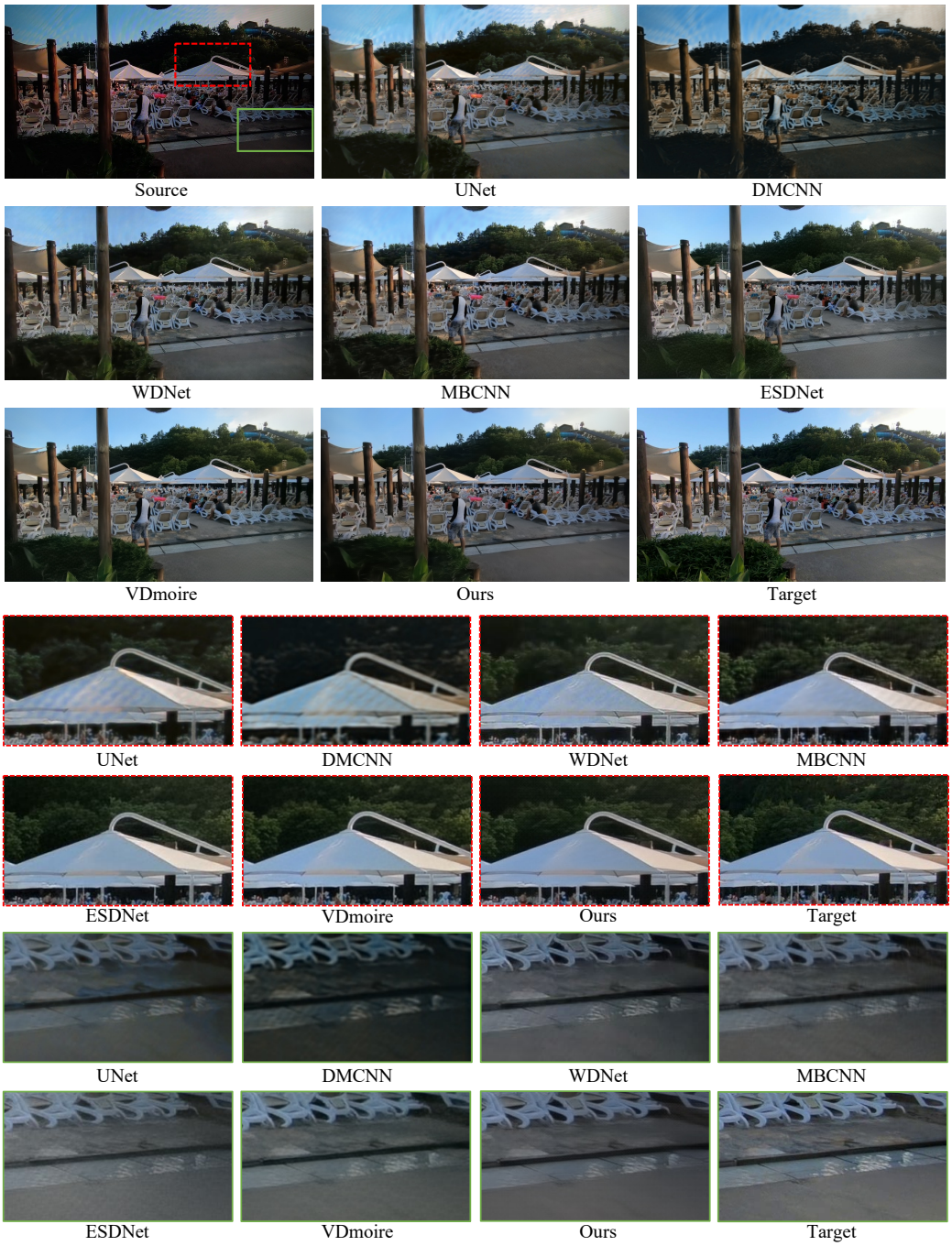

Figure 8: Qualitative comparison on the TCL-V1. The red and green boxes zoom in on frames to obviously compare the result.

Figure 9: Qualitative comparison on the iPhone-V2. The red and green boxes zoom in on frames to obviously compare the result.

for a pair comparison, while we use the default values for the other hyper-parameters.

We observe in Table 1 that our proposed method generally outperforms the other approaches in all metrics (except PSNR on TCL-V1 data), showing a significant gain over other state-of-the-art approaches. In specific, compared with VDMoire, a state-of-the-art video demoiréing method, ours shows 0.38 dB PSNR gain on TCL-V1 data and 0.23 dB PSNR gain on iPhone-V2 data. Such a gain is also apparent in other perceptual evaluation metrics, i.e. SSIM and LPIPS. This indicates that ours produces more realistic and high perceptual quality.

4.3.1. Temporal Consistency

Consistent with the recent work [11], we observe that using multiple consecutive frame inputs is helpful to improve the overall quality of image demoiréing (compare bottom two rows vs. others). Note that even without utilizing multi-frame image inputs, our proposed method still outperforms the other existing approaches, which justifies the effectiveness of using amplitude and phase components (in our Frequency Selection Module (FSM) and Post Align Module (PAM)) for removing moiré patterns.

Table 2: Quantitative comparison in terms of FVD [48] and FSIM [49] metrics to further analyze the quality of generated videos (or image frames). Note that lower FVD and higher FSIM scores are better.

Method	FVD↓	FSIM↑
DMCNN [1]	992.86	0.857
UNet [47]	928.81	0.878
WDNet [5]	697.28	0.895
MBCNN [3]	694.28	0.857
VDmoire [11]	697.09	0.911
ESDNet [4]	633.93	0.906
Ours	**633.87**	**0.967**

Following existing works [50, 51], we further use the following two metrics to analyze the quality of video outputs: FVD [48] and FSIM [49]. FVD adapts Frechet Inception Distance (FID) to capture the temporal coherence of a video, while FSIM emphasizes low-level features in IQA metric inspired by the human visual system (HVS). As shown in Table 2, our model outperforms the other existing approaches in terms of both metrics, demonstrating that our outputs are more similar to the target distribution of the entire video sequences, maintaining per-pixel and structural visual information.

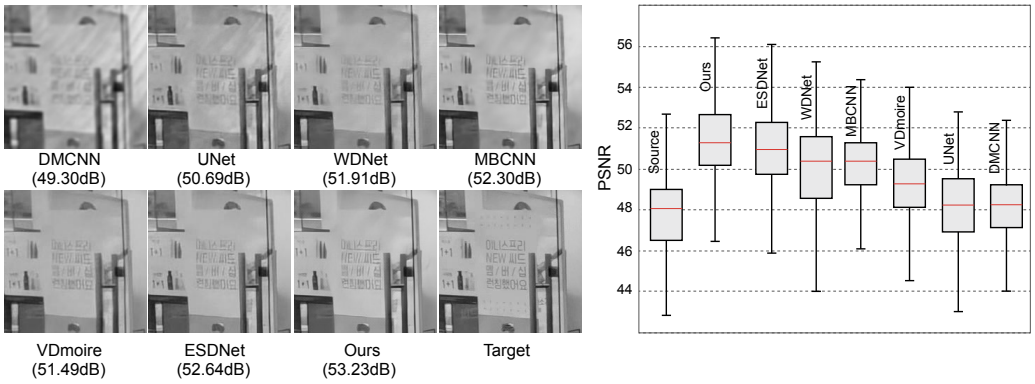

Figure 10: (left) Comparison of the luminance (Y channel in YCbCr color space) with state-of-the-art approaches (right) and their box-plots of PSNR values.

4.4. Qualitative Analysis

Further, we qualitatively compare the quality of the restored images with the state-of-the-art approaches: UNet [47], DMCNN [1], ESDNet [4], WDNet [5], MBCNN [3], and VDmoire [11].

4.4.1. Effect on Restoring Fine-grained Details

As shown in Figure 8 and 9, we provide (randomly sampled) restored output images as well as a source image (see top left, which clearly has moiré patterns) and a target image (see bottom right). At the bottom of the figure, we provide two magnified image regions (see red and green boxes in the source image) for an effective comparison. Our proposed FPANet shows better image restoration quality, preserving fine-grained details (e.g., sharper edges). Compared with conventional approaches, such as UNet [47], DMCNN [1], WDNet [5], and MBCNN [3], which often fail to filter out large-scale moiré patterns, our method effectively removes various sizes of moiré patterns without showing visually obvious artifacts across the image. Importantly, WDNet [5] and MBCNN [3] also rely on the frequency domain components using wavelet transform and implicit Discrete Cosine Transform, respectively. This may confirm that our method clearly outperforms the other frequency-based approaches, effectively dealing with various artifacts. We provide more diverse examples in Appendix A.

Further, following [52, 53, 54], we analyze the image's luminance channel (Y channel in YCbCr color space) to compare the restoration quality, comparing the ability to restore fine-grained features. We also measure PSNR to

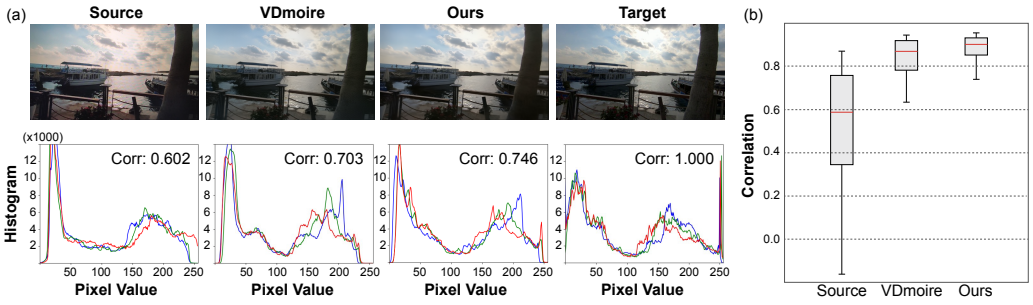

Figure 11: (a) We provide a sample of images (top) and their color histograms (bottom). Red, blue, and green lines indicate histograms for each RGB component: the x-axis represents the pixel values in [0, 255], while the y-axis denotes the number of pixels. (b) Box-plots for correlation between the color distributions of the input and the target images.

evaluate the performance on the Y channel. As shown in Figure 10, we visualize a source/target image and restored image patches from existing methods and ours. For better visualization, we only provide an enlarged image patch (see full-size images in the supplemental material). Our proposed method performs the best restoration without showing obvious visual artifacts across the image, which is confirmed by box-plots of PSNR values (see right).

4.4.2. Robustness against Color Degradation

ESDNet [4] and VDmoire [11] also provide a compelling quality demoiréing, but we observe VDmoire [11] often suffers from color shifts, probably due to a lack of the model's color restoration power. VDmoire [11] depends on the pixel's statistical information (i.e. mean and variance) across the multiple consecutive frames for temporal consistency, which makes it difficult to restore accurate pixel value per frame, resulting in making toned-down images. However, ours, which uses Frequency Spatial Fusion (FSF) module, shows fewer artifacts in color degradation. We also observe that ESDNet [4] often generates images with low visual acuity, failing to restore the original visual contents. This is more apparent in examples shown in Figure 9 (see some blurry characters, which is not the case for ours). These confirm that our proposed method effectively deals with moiré restoration with fewer artifacts, such as color degradation, lack of sharpness, and remaining large-scale moiré patterns. We provide more diverse examples in Appendix A.

Further, we also compare RGB color histograms with VDmoire to compare the amount of color degradation as shown in Figure 11 (a). To quantify

Table 3: We provide our ablation study to analyze the effect of individual building blocks: (i) Amplitude and Phase, (ii) FSM, (iii) CSFM, and (iv) PAM. We measure PSNR and SSIM for each combination of our four components. Note that ✓ represents that the module is deployed.

Components	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6 (Ours)
Amplitude & Phase (Section 3.2)	-	-	✓	✓	✓	✓
FSM (Section 3.2)	-	-	-	✓	✓	✓
CSFM (Section 3.3)	-	✓	✓	-	✓	✓
PAM (Section 3.4)	-	✓	✓	✓	-	✓
PSNR↑/SSIM↑	20.51/0.75	21.00/0.76	21.42/0.77	21.73/0.77	21.13/0.77	**21.95/0.78**

the color degradation, we measure the average correlation (for each channel) with those of the target (clear) image. A source image with moiré patterns shows the smallest correlation value, 0.602, while our proposed method shows the highest correlation value, 0.746, which is better than VDmoire. In Figure 11 (b), we provide box plots of such correlations for all test images, further confirming that our proposed method is more robust to color degradation.

4.5. Ablation Study

4.5.1. Effect of Individual Modules

Our proposed model consists of four main components: (i) Amplitude and Phase, (ii) FSM (Frequency Selection Module), (iii) CSFM (Cross Scale Fusion Module), and (iv) PAM (Post Align Module). To analyze their contributions, we conduct an ablation experiment with various combinations of these building blocks. Note that we use the TCL-V1 dataset for this experiment. We observe in Table 3 that each building component equally improves the overall image demoiréing performance in terms of PSNR and SSIM. For example, removing each building block decreases the overall performance (compare Model 6 vs. Model 2, 3, 4, and 5). Note that our FSM module depends on the Amplitude and Phase modules; thus, we remove both modules in Model 2.

4.5.2. Effect of Loss Functions

Recall from Section 3.5; we train our model with the following three loss terms: (i) Multi-scale spatial domain loss $\mathcal{L}_s$, (ii) Multi-scale perceptual loss $\mathcal{L}_p$, and (iii) Frequency domain loss $\mathcal{L}_f$. We also conduct an ablation study with different combinations of the loss terms to see their individual impacts. As shown in Table 4 (left), we observe all loss terms have noticeable

Table 4: We provide our ablation study to demonstrate (left) the effect of each loss terms ($\mathcal{L}_s$, $\mathcal{L}_p$, and $\mathcal{L}_f$) and (right) the effect of our Frequency Selection Fusion (FSF) module by replacing it with existing frequency-based building blocks: (i) FFC [22] and (ii) DeepRFT [21].

Loss function	PSNR↑	SSIM↑	LPIPS↓
$\mathcal{L}_s$	21.653	0.770	0.240
$\mathcal{L}_s + \mathcal{L}_p$	21.896	0.780	0.185
$\mathcal{L}_s + \mathcal{L}_f$	21.883	0.781	0.207
$\mathcal{L}_s + \mathcal{L}_p + \mathcal{L}_f$ (Ours)	**21.953**	**0.784**	**0.173**

Methods	PSNR↑	SSIM↑	LPIPS↓
Ours w/ FFC [22]	20.959	0.766	0.193
Ours w/ DeepRFT [21]	20.340	0.749	0.216
Ours w/ FSF	**21.572**	**0.772**	**0.189**

effects on the overall performance, and using all loss terms provides the best performance in all three metrics. Especially the multi-scale perceptual loss function helps generate more realistic images.

4.5.3. Effect of Frequency Spatial Fusion (FSF) block

We propose a novel encoding module called the Frequency Spatial Fusion module, which utilizes frequency domain components. To further demonstrate the effectiveness of our FSF module, we replace it with existing frequency-based encoding modules: FFC [22] and DeepRFT [21], which directly use the frequency components instead of dealing with amplitude and phase components separately. As shown in Table 4 (right), our proposed module outperforms the other approaches in all metrics. Note that we only use a single frame for all models.

4.5.4. Effect of Temporal Feature Alignment by PAM

We further compare our temporal feature alignment called Post Align Module (PAM) with existing Pyramid Cascade Deformable technique [10], which is utilized in VDmoire [11] to extract implicitly aligned features between consecutive frames. As we summarized in Table 5, our proposed method outperforms the alternative in all metrics, including PSNR, SSIM, and LPIPS: 0.45 dB, 0.004, and 0.009 gains in PSNR, SSIM, and LPIPS, respectively.

Table 5: We provide our ablation study to demonstrate the effect of our Post Align Module (PAM) module.

Methods	PSNR↑	SSIM↑	LPIPS↓
Ours w/ PCD [10]	21.507	0.780	0.184
Ours w/ PAM	**21.953**	**0.784**	**0.173**

5. Conclusion

We introduced a novel video demoiréing method called FPANet by proposing the following three components: (i) Frequency Selection Module (FSM), (ii) Cross Scale Fusion Module (CSFM), and (iii) Post Align Module (PAN). FSM utilizes amplitude and phase components in the frequency domain to address undesired color changes and large-scale moiré patterns, while CSFM is used to capture multi-scale features to recover both global and fine-grained features. Lastly, PAM is utilized to align features from multiple consecutive frames with reduced visual artifacts. We demonstrated the effectiveness of using our proposed method with a public video demoiréing dataset called VDmoire, and ours generally outperforms existing state-of-the-art approaches in terms of various image and video quality metrics, such as PSNR, SSIM, LPIPS, FVD, and FSIM.

Acknowledgment

This work was supported by LG Display Research Center and the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2023-2020-0-01819) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). S. Kim and G. Oh are partially supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2022R1F1A1074334) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University)).

Appendix A. More Qualitative Results

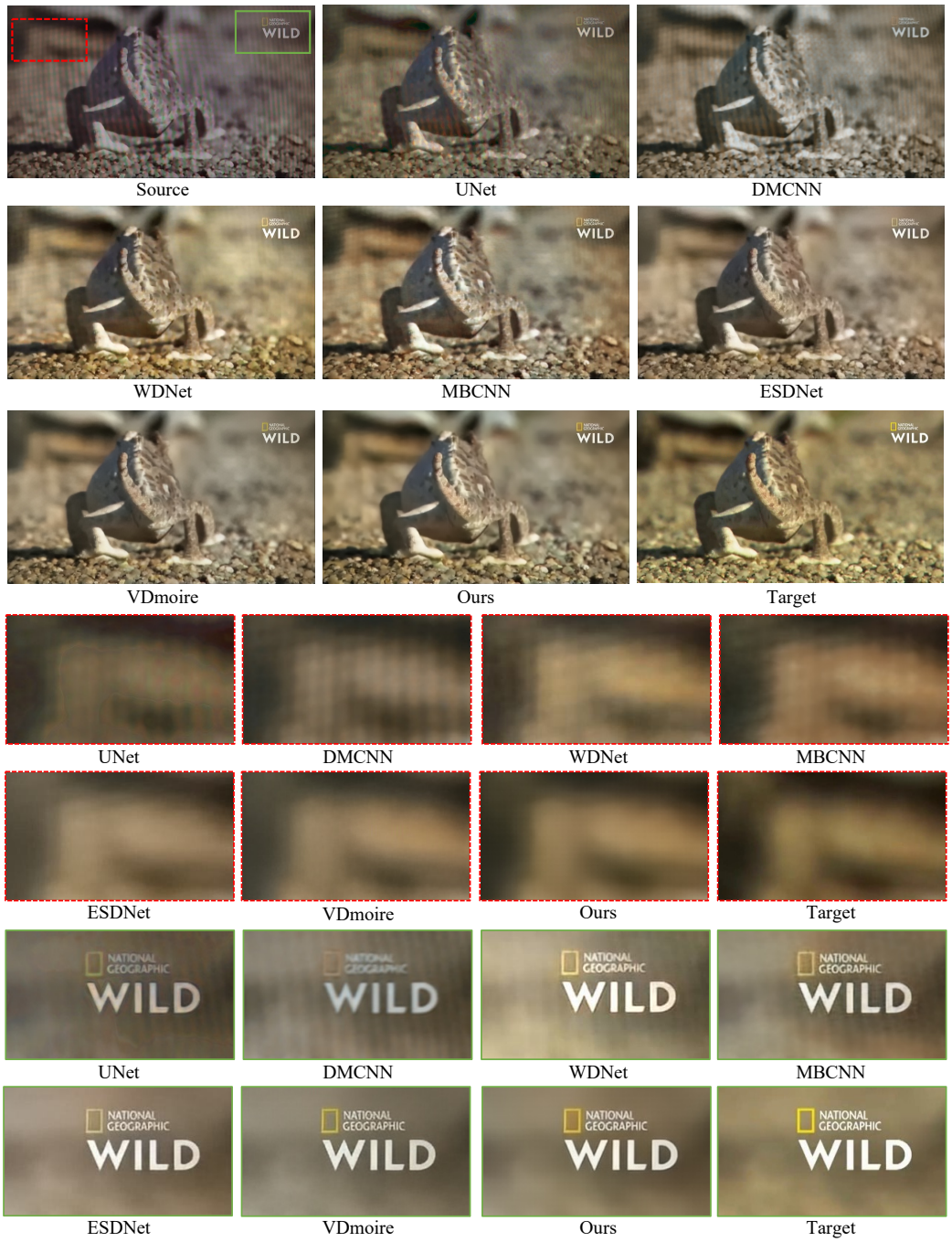

Figure A.12: Qualitative comparison on the TCL-V1. The red and green boxes zoom in on frames to obviously compare the result.

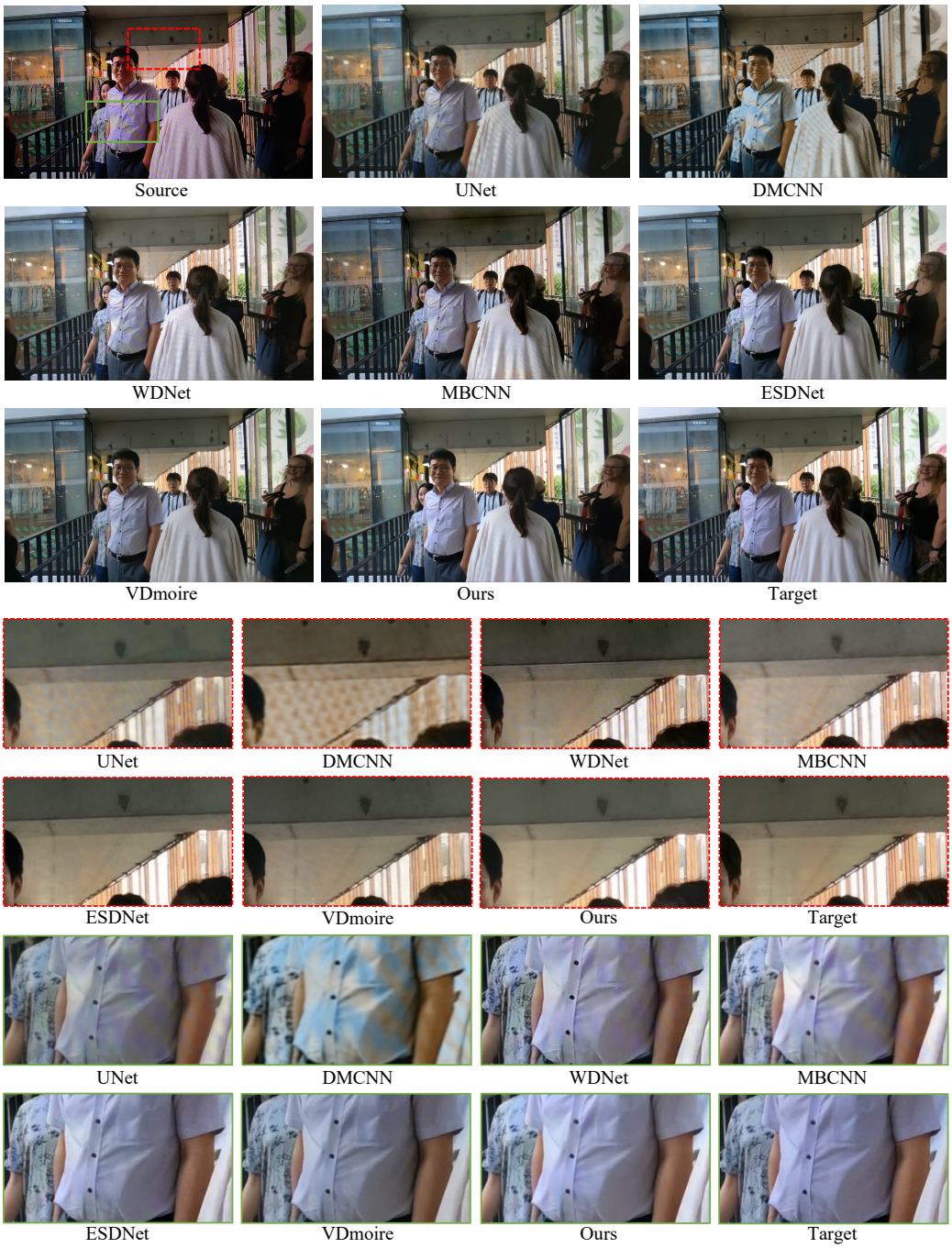

Figure A.13: Qualitative comparison on the TCL-V1. The red and green boxes zoom in on frames to obviously compare the result.

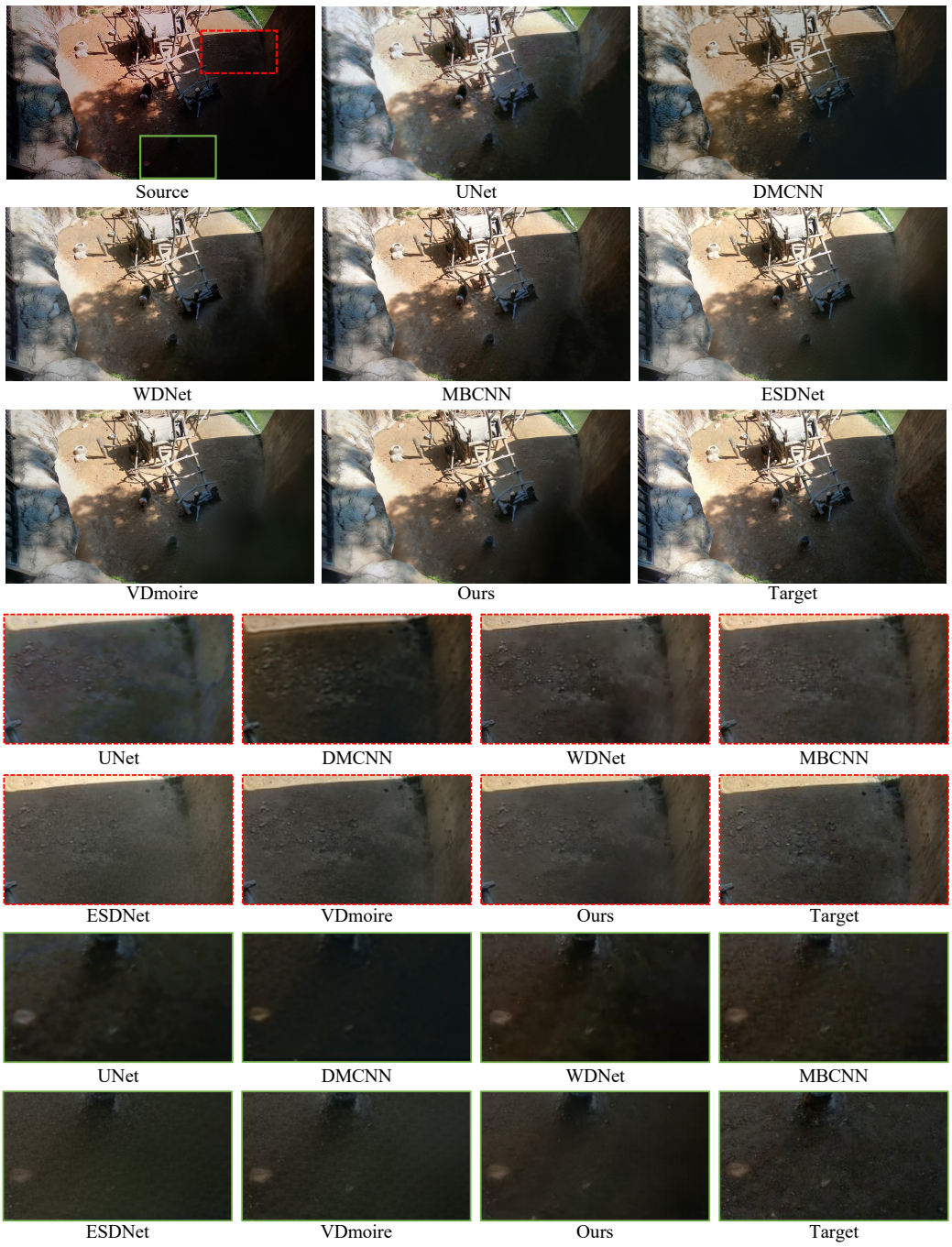

Figure A.14: Qualitative comparison on the TCL-V1. The red and green boxes zoom in on frames to obviously compare the result.

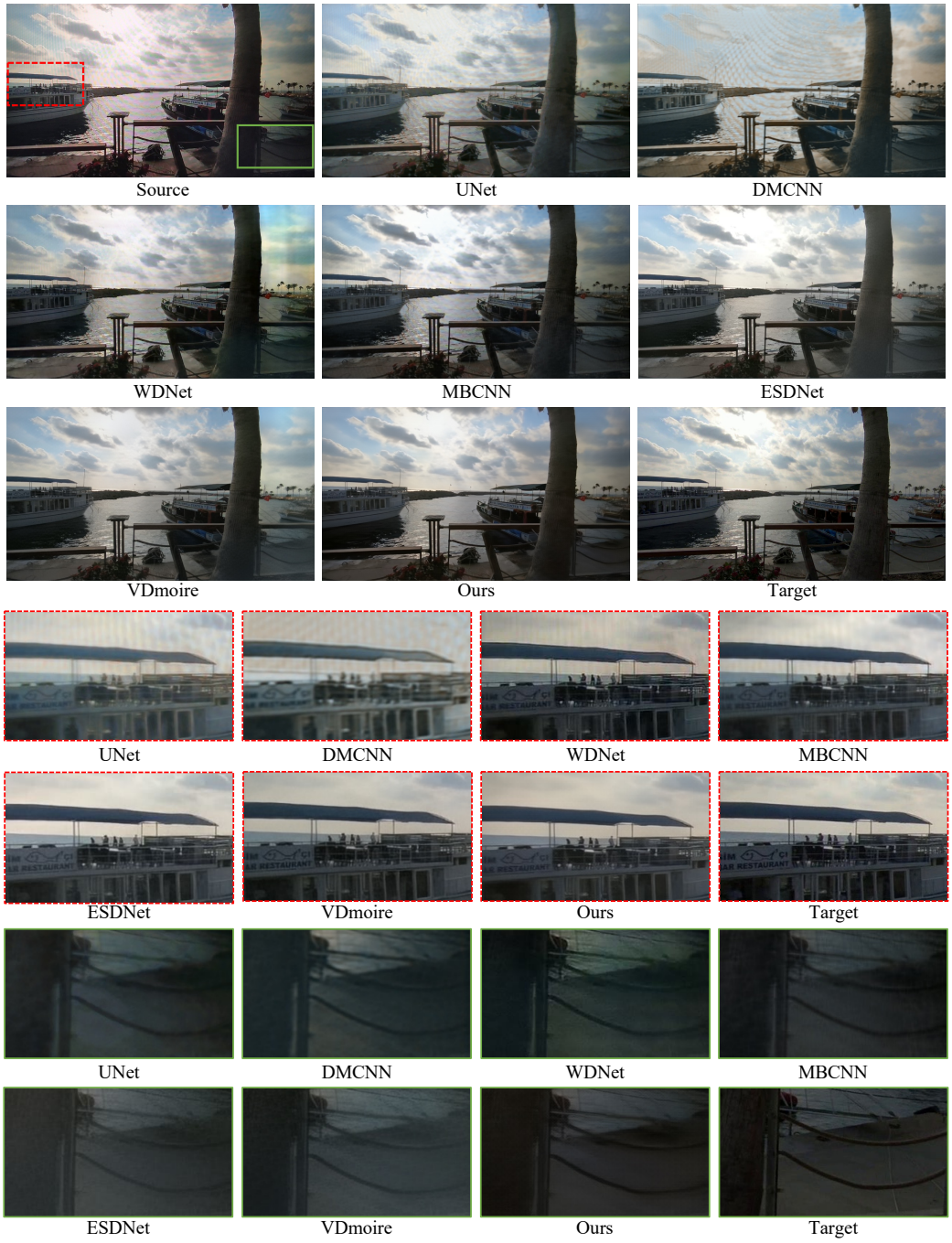

Figure A.15: Qualitative comparison on the TCL-V1. The red and green boxes zoom in on frames to obviously compare the result.

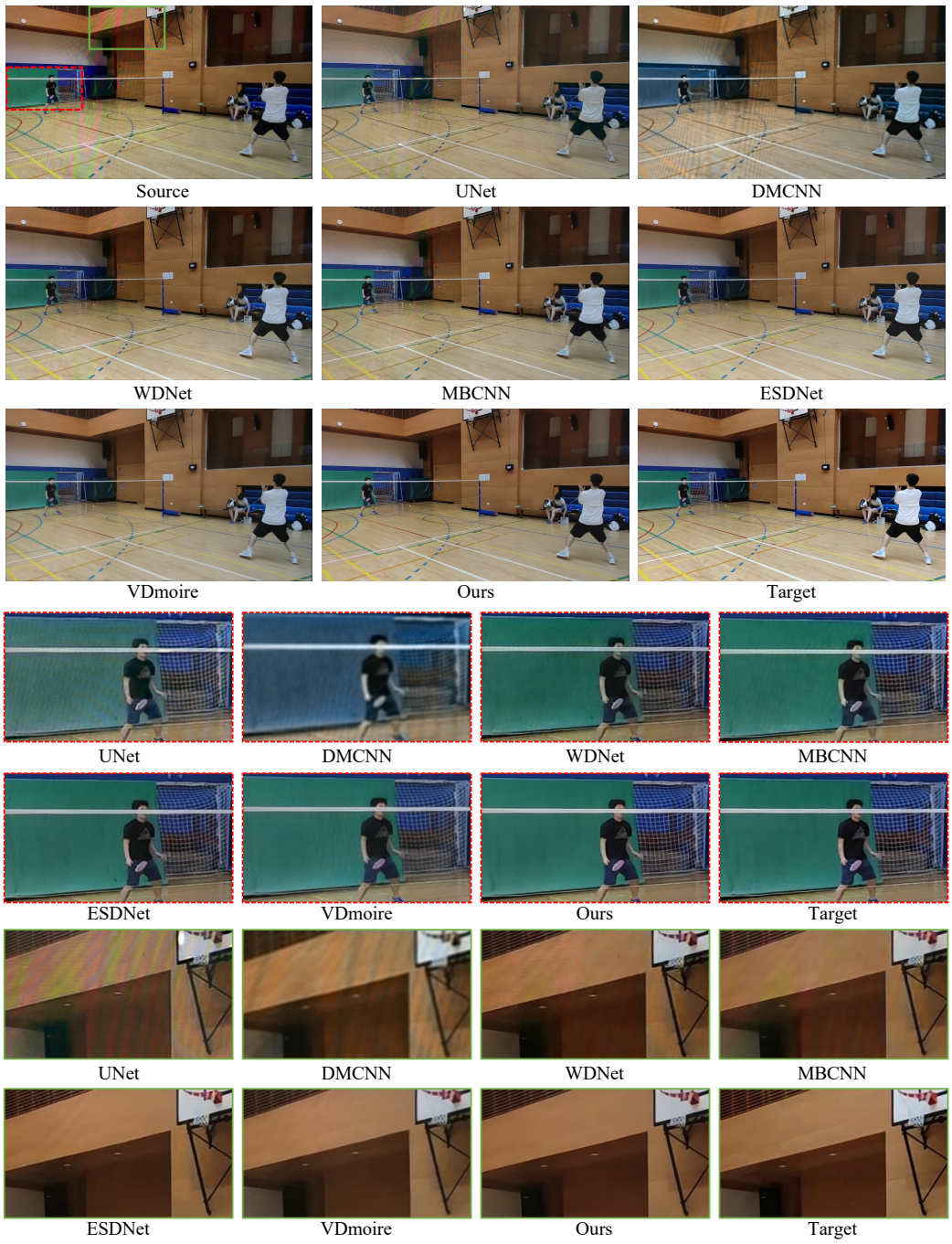

Figure A.16: Qualitative comparison on the iPhone-V2. The red and green boxes zoom in on frames to obviously compare the result.

References

[1] Y. Sun, Y. Yu, W. Wang, Moiré photo restoration using multiresolution convolutional neural networks, IEEE Transactions on Image Processing 27 (8) (2018) 4160–4172.

[2] Z. Bolun, Y. Shanxin, Y. Chenggang, T. Xiang, Z. Jiyong, S. Yaoqi, L. Lin, L. Ales, S. Gregory, Learning frequency domain priors for image demoireing, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[3] B. Zheng, S. Yuan, G. Slabaugh, A. Leonardis, Image demoireing with learnable bandpass filters, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3636–3645.

[4] X. Yu, P. Dai, W. Li, L. Ma, J. Shen, J. Li, X. Qi, Towards efficient and scale-robust ultra-high-definition image demoiréing, in: European Conference on Computer Vision, Springer, 2022, pp. 646–662.

[5] L. Liu, J. Liu, S. Yuan, G. Slabaugh, A. Leonardis, W. Zhou, Q. Tian, Wavelet-based dual-branch network for image demoiréing, in: European Conference on Computer Vision, Springer, 2020, pp. 86–102.

[6] K. C. Chan, X. Wang, K. Yu, C. Dong, C. C. Loy, Basicvsr: The search for essential components in video super-resolution and beyond, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4947–4956.

[7] K. C. Chan, S. Zhou, X. Xu, C. C. Loy, Basicvsr++: Improving video super-resolution with enhanced propagation and alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5972–5981.

[8] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.

[9] Y. Tian, Y. Zhang, Y. Fu, C. Xu, Tdan: Temporally-deformable alignment network for video super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3360–3369.

[10] X. Wang, K. C. Chan, K. Yu, C. Dong, C. Change Loy, Edvr: Video restoration with enhanced deformable convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[11] P. Dai, X. Yu, L. Ma, B. Zhang, J. Li, W. Li, J. Shen, X. Qi, Video demoireing with relation-based temporal consistency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17622–17631.

[12] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8934–8943.

[13] B. He, C. Wang, B. Shi, L.-Y. Duan, Mop moire patterns using mopnet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[14] B. He, C. Wang, B. Shi, L.-Y. Duan, Fhde2net: Full high definition demoireing network, in: ECCV (22), 2020, pp. 713–729.
URL https://doi.org/10.1007/978-3-030-58542-6-43

[15] L. Liu, S. Yuan, J. Liu, L. Bao, G. Slabaugh, Q. Tian, Self-adaptively learning to demoiré from focused and defocused image pairs, Advances in Neural Information Processing Systems 33 (2020) 22282–22292.

[16] S. Kim, H. Nam, J. Kim, J. Jeong, C3net: Demoiréing network attentive in channel, color and concatenation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 426–427.

[17] H. Yue, Y. Cheng, F. Liu, J. Yang, Unsupervised moiré pattern removal for recaptured screen images, Neurocomputing 456 (2021) 352–363.

[18] B. Zheng, Y. Chen, X. Tian, F. Zhou, X. Liu, Implicit dual-domain convolutional network for robust color image compression artifact reduction, IEEE Transactions on Circuits and Systems for Video Technology 30 (11) (2019) 3982–3994.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[20] S. Yuan, R. Timofte, G. Slabaugh, A. Leonardis, Aim 2019 challenge on image demoireing: Dataset and study, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3526–3533. doi:10.1109/ICCVW.2019.00437.

[21] X. Mao, Y. Liu, W. Shen, Q. Li, Y. Wang, Deep residual fourier transformation for single image deblurring, arXiv preprint arXiv:2111.11745 (2021).

[22] L. Chi, B. Jiang, Y. Mu, Fast fourier convolution, Advances in Neural Information Processing Systems 33 (2020) 4479–4488.

[23] L. Chi, G. Tian, Y. Mu, L. Xie, Q. Tian, Fast non-local neural networks with spectral residual learning, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2142–2151.

[24] Y. Yang, S. Soatto, Fda: Fourier domain adaptation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4085–4095.

[25] H. Yu, N. Zheng, M. Zhou, J. Huang, Z. Xiao, F. Zhao, Frequency and spatial dual guidance for image dehazing, in: European Conference on Computer Vision, Springer, 2022, pp. 181–198.

[26] M. Cai, H. Zhang, H. Huang, Q. Geng, Y. Li, G. Huang, Frequency domain image translation: More photo-realistic, better identity-preserving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13930–13940.

[27] Z. Qin, P. Zhang, F. Wu, X. Li, Fcanet: Frequency channel attention networks, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 783–792.

[28] Y. Rao, W. Zhao, Z. Zhu, J. Lu, J. Zhou, Global filter networks for image classification, Advances in Neural Information Processing Systems 34 (2021) 980–993.

[29] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust large mask inpainting with fourier convolutions, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2149–2159.

[30] W. Zou, M. Jiang, Y. Zhang, L. Chen, Z. Lu, Y. Wu, Sdwnet: A straight dilated network with wavelet transformation for image deblurring, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1895–1904.

[31] M. Zhou, J. Huang, K. Yan, H. Yu, X. Fu, A. Liu, X. Wei, F. Zhao, Spatial-frequency domain information integration for pan-sharpening, in: European Conference on Computer Vision, Springer, 2022, pp. 274–291.

[32] L. Jiang, B. Dai, W. Wu, C. C. Loy, Focal frequency loss for image reconstruction and synthesis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13919–13929.

[33] Z. Liu, W. Lin, X. Li, Q. Rao, T. Jiang, M. Han, H. Fan, J. Sun, S. Liu, Adnet: Attention-guided deformable convolutional network for high dynamic range imaging, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 463–470.

[34] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, J. Jia, Mucan: Multi-correspondence aggregation network for video super-resolution, in: European conference on computer vision, Springer, 2020, pp. 335–351.

[35] A. Dudhane, S. W. Zamir, S. Khan, F. S. Khan, M.-H. Yang, Burst image restoration and enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5759–5768.

[36] E. O. Brigham, R. E. Morrow, The fast fourier transform, IEEE Spectrum 4 (12) (1967) 63–70. doi:10.1109/MSPEC.1967.5217220.

[37] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510–519.

[38] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, L. Shao, Learning enriched features for fast image restoration and enhancement, arXiv preprint arXiv:2205.01649 (2022).

[39] S. Fan, W. Liang, D. Ding, H. Yu, Lacn: A lightweight attention-guided convnext network for low-light image enhancement, Engineering Applications of Artificial Intelligence 117 (2023) 105632.

[40] L. Chen, X. Chu, X. Zhang, J. Sun, Simple baselines for image restoration, arXiv preprint arXiv:2204.04676 (2022).

[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[42] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[43] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983 (2016).

[44] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[47] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[48] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, S. Gelly, Towards accurate generative models of video: A new metric & challenges, arXiv preprint arXiv:1812.01717 (2018).

[49] L. Zhang, L. Zhang, X. Mou, D. Zhang, Fsim: A feature similarity index for image quality assessment, IEEE transactions on Image Processing 20 (8) (2011) 2378–2386.

[50] J. Cho, S. Kim, K. Sohn, Memory-guided image de-raining using time-lapse data, IEEE Transactions on Image Processing (2022).

[51] I. Skorokhodov, S. Tulyakov, M. Elhoseiny, Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3626–3636.

[52] J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[53] C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European conference on computer vision, Springer, 2016, pp. 391–407.

[54] X. Mao, C. Shen, Y.-B. Yang, Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections, Advances in neural information processing systems 29 (2016).