Temporal-Spatial Fusion Transformer for Video Demoiréing

Ji-Wei Wang, Li-Yong Shen
School of Mathematical Sciences
University of Chinese Academy of Sciences,100049, Beijing, China
wangjiwei21@mails.ucas.ac.cn, lyshen@ucas.ac.cn

Abstract

When utilizing digital cameras to capture the videos on screen, the occurrence of moiré patterns can lead to color distortions, significantly degrading the quality of both images and videos. Considering the escalating demand for video acquisition, it becomes necessary to design algorithms for video demoiréing. In this paper, we introduce a novel attention-based network for this task, named Temporal-Spatial Fusion Transformer (TSFT). By introducing Temporal and Spatial Attention Encoder and multi-scale feature fusion method, TSFT can learn the dynamic variations of moiré patterns in both temporal and spatial dimensions. In the decoding phase, the self-attention mechanism is employed to learn the temporal dependencies at both image-level and video-level, enhancing model performance in moiré removal. Experimental results demonstrate a significant improvement in the performance of the proposed model compared to existing methods on public datasets. Furthermore, TSFT can output visual attention maps for analyzing the distribution of moiré and the focus of model learning. The outstanding performance in the video deraining task also proves the robustness of our model, highlighting its enormous potential for application in other restoration tasks.

***Keywords:** Video demoiréing, deep learning, transformer, attention mechanism.*

1. Introduction

With the advancement of electronic devices such as cameras, computers, and LCD, there is a growing demand for capturing videos. In our daily lives and work, videos are frequently required for documenting and disseminating information. However, there will be colored stripes when capturing videos on screens, known as moiré, this is caused by the overlapping of color filter arrays (CFA) and subpixels on the screen. These patterns can significantly degrade video quality and result in information loss. For example, when display devices are tested for quality control in industrial settings, eliminating moiré effects is essential to avoid obscuring defects; during the video rendering process, moiré may arise, thereby impacting the quality of the rendered video; for some spatial computing devices, such as Apple Visual Pro, moiré can adversely affect the device's imaging capabilities. There are a lot of studies on single-image demoiréing, including traditional and deep learning methods, but only a few studies attempt to address video demoiréing problem. Considering that video is an essential carrier of information in real life, the research on video demoiréing is of great significance. But only a few studies attempt to address video demoiréing problem.

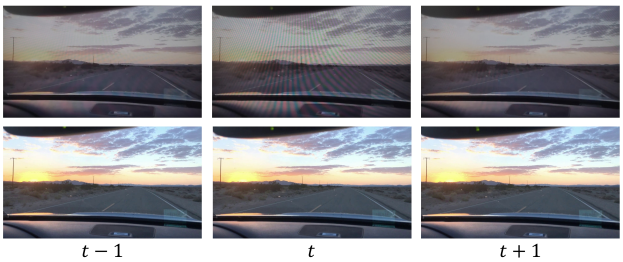


Figure 1. Moiré example on Video Demoiréing Dataset, the moiré intensity of different frames changes significantly.

From the perspective of the moiré pattern, in comparison to other video restoration tasks such as video deraining[53, 55] and denoising[30, 40], moiré exhibits different characteristics (shape, color, and distribution variations) when there are changes in shooting angles and devices. This variability increases the difficulty of moiré removal. From the perspective of videos, compared to single-image demoiréing, the model can leverage information from adjacent frames for restoration in video demoiréing task. However, moiré shows dynamic changes over time, causing difficulties in feature alignment and information extraction. Furthermore, the variations in moiré patterns do not align with the illumination of video scenes, they show different intensities that are correlated with the filming environment, as shown in Fig. 1. Factors such as changes in focal length can also influence the intensity of moiré, causing information fluctuations across different frames of moiré videos. Effectively leveraging video information to remove moiré is still a challenging problem.

Existing methods for single image demoiréing mainly include multi-scale[39, 7, 54], frequency domain processing[60, 15, 28], and moiré classification methods[14]. However, these methods are not suitable for video demoiréing task as they fail to utilize video information to enhance restoration efficiency. In Video Restoration, the key point is the registration of multiple frames, so as to effectively use the information of auxiliary frames. Existing methods[2, 26, 41] mainly focus on using optical flow[1] or deformable convolution[8] for feature alignment, but they do not effectively use spatial information of noise images. Due to the significant spatial variations of moiré over time, which are more complex compared to other noise, it is necessary to integrate spatial and temporal information to model the moiré distribution. Existing video demoiréing methods[9, 34, 27, 32] primarily employ feature alignment techniques and adjacent frame information learning for restoration. However, the irregularities of stripes and color variations in moiré video still make it hard for performance improvement.

In this paper, we propose an attention-based deep neural network for Video Demoiréing, named Temporal-Spatial Fusion Transformer (TSFT). TSFT primarily utilizes Temporal and Spatial Attention Encoders (TAE and SAE) to learn information from multiple frames and spatial domain, employing Feature Fusion Network (FFN) and multi-head self-attention modules to implicitly learn variations in moiré intensity. TSFT is built upon self-attention mechanisms and demonstrates high efficiency in video demoiréing task. Our model addresses this task by leveraging the information of temporal and spatial dimensions. In the encoder stage, the network transforms moiré video frames into a series of deep features that encapsulate the intrinsic content and structural information. Our model utilizes two types of encoders to extract features at different scales. Notably, unlike other types of noise, moiré patterns often exhibit irregular color variations across different spatial regions. The spatial encoder is designed to fit the diverse moiré textures across regions and apply weighting to scales with dense moiré patterns. Besides, the temporal attention encoder extracts features from adjacent frames while mitigating the interference of redundant information, thereby improving the model's learning efficiency. Moreover, TSFT employs an efficient single-stage end-to-end training method during the training phase without the need for extra adjustments while also demonstrating outstanding performance in this task.

We analyze the characteristics of moiré patterns and provide an intuitive modeling method. Based on the analysis, we constructed TSFT to remove moiré stripes and reconstruct high-frequency details by considering three aspects: information from adjacent frames, spatial distribution of noise, and variations in moiré intensity.

The role of Temporal Attention Encoder (TAE) is to extract features of different frames and integrate them, thereby learning temporal attention and weighting the original features. TAE utilizes the additional information provided by temporal consistency and reduces the contribution of unnecessary features. Spatial Attention Encoder (SAE) shares a similar structure as TAE but focuses on learning the spatial features of multi-scale. Given the complex distribution of moiré across different scales, SAE aims to learn spatial attention of various scales to effectively model the spatial distribution of moiré patterns.

After the encoder phase, the features are input into the Temporal-Spatial Fusion Decoder. Firstly, the Feature Fusion Network (FFN) is utilized to merge the encoded features. Subsequently, multi-head self-attention is employed to learn the correlations between different temporal and spatial positions in the features, extracting video-level and image-level information from multiple frames. Finally, Feed-Forward and Detail Reconstruction are conducted in the last stage to obtain the moiré-free image for each frame. Self-attention mechanism is particularly suitable for modeling the variations of moiré textures in both temporal and spatial dimensions. Existing demoiréing algorithms have not considered the application of self-attention mechanisms in this field. Therefore, we introduce the self-attention methods and construct a deep network for Video Demoiréing task.

Extensive experiments on public video demoiréing datasets validate the superiority of TSFT. Compared to existing state-of-the-art methods, our algorithm achieves 0.9dB improvement in PSNR while also leading in other metrics. Our model can also generate interpretable attention weight maps, facilitating the analysis of the approximate distribution of moiré patterns. Moreover, TSFT can be effectively extended to other low-level tasks. Further experiments demonstrate the model's excellent performance in Video Deraining as well.

In summary, the contributions of our paper can be summarized as follows:

- Based on the Transformer architecture, we introduce a video demoiréing model named Temporal-Spatial Fusion Transformer. Our model utilizes multi-scale information combined with TAE and SAE to learn the dynamic variations of moiré, then employs modules based on self-attention mechanisms to learn correlations between different frames and image areas. Our model achieves superior performance and introduces self-attention in Video Demoiréing, which is rarely seen in previous demoiréing methods.

- We propose TAE and SAE to model the distribution of moiré across different scales. These two modules integrate multi-scale methods and attention mechanisms, weighting features from two dimensions to enhance

learning efficiency. Our model can also generate visual attention maps to analyze moiré variations.

- The results of quantitative and qualitative evaluations on the video demoiréing dataset show that TSFT outperforms existing methods in this task. Additionally, our model demonstrates excellent performance in other low-level tasks, such as Video Deraining, showcasing its applicability.

2. Related Work

2.1. Image Demoiréing

Moiré effect often appears when filming videos on the screen and degrades video quality. This is due to the aliasing of the camera's color filter array (CFA) and the screen's subpixel. Early research primarily focused on image filtering[49, 38] and image decomposition[25, 52] algorithms to remove moiré on single-image. These methods aim to construct optimized algorithms based on the spatial and frequency characteristics of moiré patterns. However, the restored images may lose texture details during the process, leading to issues such as excessive smoothness.

In recent years, many learning-based demoiréing networks have been proposed. Sun et al.[39] introduced a multiresolution convolutional neural network to remove moiré at different scales and constructed a large demoiréing dataset (TIP2018) for subsequent demoiréing research. Cheng et al.[7] proposed a deep model (MDDM) using a multi-scale feature encoding module to remove moiré patterns. MopNet[14] used edge detectors and convolution modules of different color channels to detect and classify moiré. MBCNN[60] divides the demoiréing process into two parts, texture removal and color restoration, and proposes a learnable bandpass filter to learn the distribution of moiré texture. WDNet[28] uses wavelet transform to separate the spectrum distribution of moiré from the original image.

Most recently, He et al.[15] proposed a demoiréing network (FHDe2Net), containing multi-branch to remove moiré and reconstruct details. Yu et al.[54] proposed a baseline model ESDNet, which used the semantic-aligned scale-aware module to deal with the scale changes of moiré, and created a 4K resolution demoiréing dataset. AMSDM[10] employed the Adaptive Multispectral Encoding Module to encode and remove moiré of different scales. Niu et al.[31] proposed PMTNet and introduced a progressive texture complementation block to correct color shifts gradually. There is also algorithm[58] that considers lightening existing demoiréing models and enabling models to inference on mobile devices.

The existing methods primarily aim to remove moiré patterns on single images. However, for video data, these approaches fail to leverage information from adjacent frames and do not consider the dynamic changes of moiré over time. The advantage of our approach is that the TAE and self-attention modules can utilize the sequential information of videos to assist in restoration and adaptively reduce the interference of redundant information.

2.2. Video Restoration

Video Restoration aims to utilize information from nearby frames, as they can provide additional information. The key point of multi-frame feature utilization is to align different frames to reduce information interference caused by pixel deviations. Early methods[37, 43, 22] primarily used optical flow[1] for multi-frame alignment, while in recent years, some methods[47, 9, 41, 42, 6] employed deformable convolutions[8] to align features from various frames implicitly. Existing demoiréing method proposed a baseline video demoiréing network (VDN) with implicit feature space alignment and selective feature aggregation to utilize complementary information from adjacent frames. VDN employs a two-stage training approach, where the first stage focuses on moiré removal, and the second stage adjusts the loss function to enhance restoration effects. CIDNet[34] proposed a compact invertible dyadic network that progressively decouples the frame and moiré. DTCENet[27] presented an invertible network to align distortion color patches in the embedding framework. However, the performance of these methods can be affected by the dynamic changes of moiré videos and leave room for improvement. STDNet[32] leverages temporal correlation and attention mechanisms to learn the spatial structure of moiré patterns, but it introduces a level of computational cost. In this paper, we drew inspiration from existing methods[47, 9] and used deformable convolutions for multi-frame alignment. Our approach adopts an encoder-decoder structure. In the decoder stage, TSFT reconstructs details and employs self-attention mechanisms to learn correlations between different frames and image regions, then reconstructs high-frequency details. Our model achieves excellent performance with single-stage training, thus enhancing training efficiency.

2.3. Attention Mechanism

Attention mechanism plays an important role in human perception. Some methods[7, 60, 54] use attention for Image and Video Demoiréing tasks. However, these methods do not consider the integration of temporal and spatial information, thereby failing to enhance the model's learning efficiency. We analyzed the characteristics of moiré patterns that the temporal and spatial variations of moiré differ from the video content. The dynamic changes of moiré are influenced by shooting angles and devices. We construct a novel attention-based model that integrates temporal and spatial attention methods to learn the distribution of moiré.

In recent years, Vision Transformer (ViT)[13] has demonstrated tremendous potential in computer vision tasks, surpassing CNN-based models in some tasks. ViT utilizes patch-wise linear embedding to project image patches into token sequences before inputting them into the transformer architecture. However, the shortages of ViT are its lack of inductive bias and relatively high computational cost. Consequently, some studies have introduced inductive bias by incorporating convolutional[50, 11, 21] or pyramid structures[45, 46]. Among these works, Swin Transformer[29] divides image patches into different windows and computes self-attention within each window. This method (shifted windows) enhances the network's modeling and representation capabilities while reducing computational costs. Many studies[44, 4, 12] improved the shifted windows operation and achieved impressive results in various visual tasks.

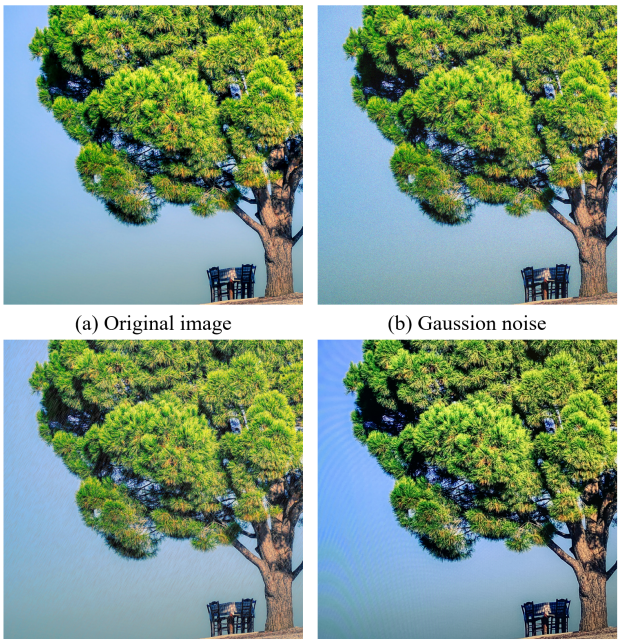
There are also some works using the Swin Transformer architecture in the restoration field. SwinIR[23] applied Swin transformer blocks to low-level tasks, which consists of three parts: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. Wang et al.[48] built a hierarchical encoder-decoder network using the Transformer block for image restoration tasks, named Uformer. Xu et al.[51] present a bidirectional transformer network to exploit the long-range informative dependency for video deblurring. Transformer architecture enjoys a high capability for capturing both local and global dependencies for this task, and it is also very suitable for learning complex variations of moiré in temporal and spatial dimensions. Based on Swin Transformer block, we make modifications and innovations of transformer architecture according to the characteristics of moiré, which significantly improves the performance of TSFT in Video Demoiréing.

3. Method

3.1. Characteristics of moiré video

Firstly, we analyze of the moiré video sequences. The dynamic variations of moiré patterns are rather complex, influenced by the device and shooting angle. Conversely, the changes displayed in the video on the screen are solely related to the content of the original video. The variation of moiré between adjacent frames also provides an opportunity for multi-frame restoration. This is because the changes in the positions of moiré stripes between nearby frames can be utilized to extract valuable information and to differentiate between various moiré regions. So it's crucial to learn the spatial and temporal changes of moiré. Our model, particularly in the encoder section, also focuses on the feature extraction of these two aspects.

Moreover, it is noted that most moiré videos are generated by filming screens, and the use of video-capturing devices inevitably leads to automatic changes in focus length. This results in variations in the moiré intensity in specific frames, which may also cause distortion, blurring, and other issues. Such intensity changes can affect the restoration effectiveness of the model.



(a) Original image (b) Gaussion noise

(c) Synthetic rain (d) Moiré

Figure 2. Noise diagram.

For a n-frames moiré video sequence $\{I_t^m\}_{t=1}^n$, the objective of video demoiréing task is to get a moiré-free video sequence $\{I_t^c\}_{t=1}^n$. The moiré video sequence can be modeled as follows:

$$\begin{aligned} \vec{I}^m &= (I_1^m, I_2^m, \cdots, I_n^m) \\ \vec{I}^c &= (I_1^c, I_2^c, \cdots, I_n^c) \\ \vec{I}^m &= \mathcal{G}(\vec{I}^c) + \vec{N}_t \end{aligned} \tag{1}$$

where $\vec{N}_t$ represents the moiré noise, and $\mathcal{G}$ represents complex, non-linear variations beyond moiré artifacts, including blurriness and tonal shifts. These variations are primarily caused by changes in focal length and shooting environment during video recording. Therefore, to obtain a moiré-free video, the restoration process can be expressed as:

$$\vec{I}^c = \mathcal{G}^{-1}(\vec{I}^m - \vec{N}_t) \tag{2}$$

We make a intuitionistic decomposition of the moiré noise sequence as shown in Eq.3:

$$\vec{I}^c = \mathcal{G}^{-1}(\vec{I}^m - (\mathbf{N} + \vec{n}_t)) \tag{3}$$

where $\mathbf{N} = (N, N, \cdots, N)$ represents the base noise of a video sequence, and $\vec{n}_t = (n_1, n_2, \cdots, n_n)$ is the variations

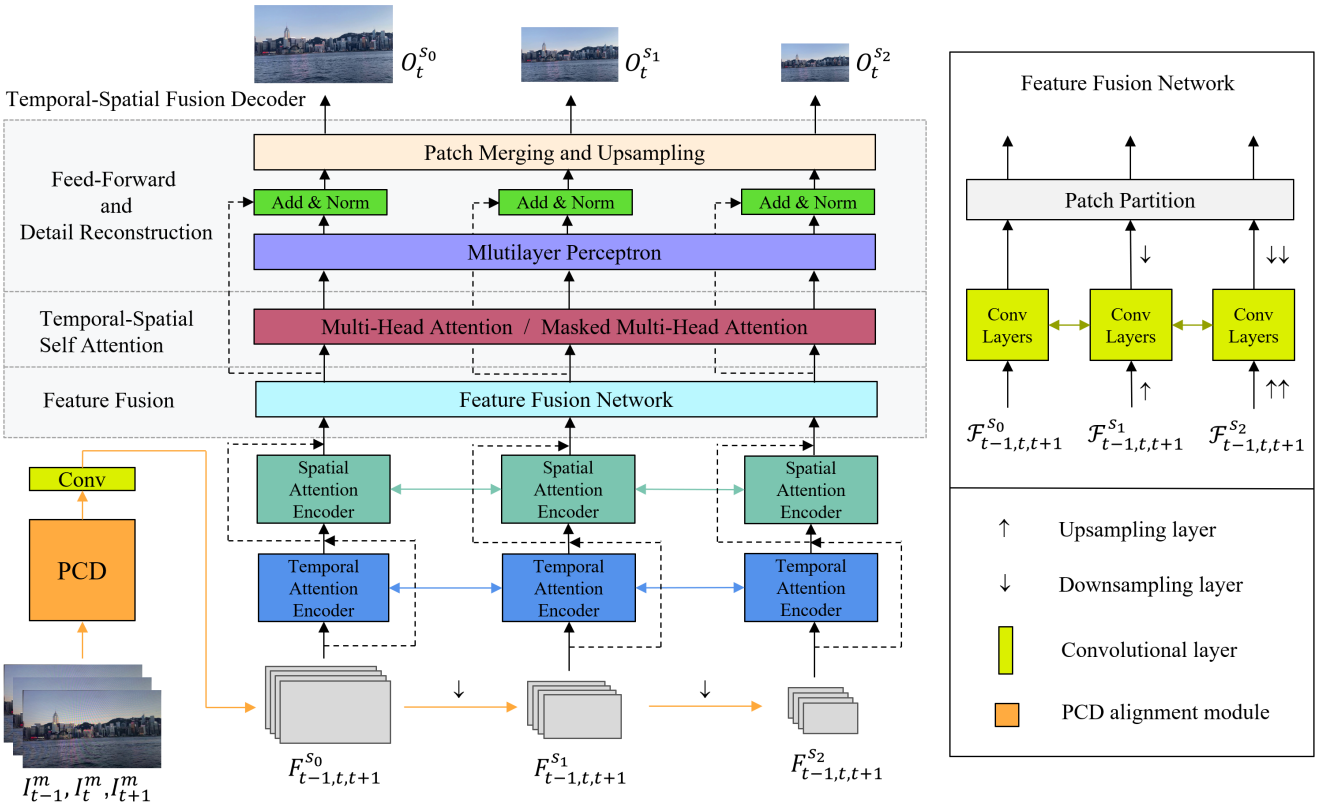


Figure 3. The architecture of Temporal-Spatial Fusion Transformer (TSFT). Our network comprises a CNN-based encoder and a Temporal-Spatial Fusion Decoder. The encoder uses consecutive moiré frames as input, extracting deep features along both temporal and spatial dimensions. The decoder integrates a Feature Fusion Network to remove moiré artifacts and reconstruct textures, ultimately producing moiré-free video frames.

of noise sequence. For simple noises such as rain, Gaussian noise, etc., their variations over t are typically constant, periodic, or linear. However, the variation of moiré $\vec{n}_t$ is more complex and irregular, as shown in Fig.2. Therefore, accurately estimating the moiré noise $\mathbf{N}$ and the variation $\vec{n}_t$ is important. This complexity underscores the challenge in video demoiréing, as it requires not only identifying the underlying noise accurately but also adapting to its unpredictable changes over time.

Based on existing research in multi-frame restoration, the remaining frames can be utilized as model inputs to assist in restoring the current frame. For the output I_t^c, its structure can be illustrated as follows:

$$\begin{aligned} I_t^c &= \mathcal{G}^{-1}(\vec{\alpha_t} \cdot \vec{I}^m - (N + n_t)) \\ \vec{I}^c &= \mathcal{G}^{-1}(\mathbf{A} \cdot \vec{I}^m - (\mathbf{N} + \vec{n}_t)) \end{aligned} \tag{4}$$

where $\mathbf{A}$ represents a time-correlated weight tensor matrix. This matrix encapsulates the varying impact of different frames on the restoration. According to intuitive human visual perception, it is evident that frames more temporally distant from the current frame exhibit lower relevance. Therefore, leveraging frames that are closer to the current frame for assistance in restoration emerges as a preferable strategy, offering the dual benefits of reduced computational demand and minimized information redundancy. Without loss of generality, our model employs three consecutive frames as inputs for the restoration.

Modeled in this way, we conducted a qualitative analysis for the construction of our model. For Video Demoiréing, the model needs to approximate the unknown transformations represented in the formula, encompassing several components: the utilization of temporal information ($\mathbf{A}$), the approximation of spatial noise distribution ($\mathbf{N}$), the variation of noise over time ($\vec{n}_t$), and non-linear variations ($\mathcal{G}^{-1}$). So, our model is designed with multiple modules that extract information across temporal and spatial dimensions to fit $\mathbf{A}$ and $\mathbf{N}$ during the encoding phase. Then, our model utilizes Feature Fusion Network to integrate information from two dimensions and implicitly learns the variation of noise $\vec{n}_t$. Finally, self-attention module is employed to learn the non-linear variations $\mathcal{G}^{-1}$. This comprehensive approach enables the model to effectively address the complex interplay of factors contributing to the presence of moiré patterns, ensuring an efficient restoration of video quality. The overall structure of Temporal-Spatial Fusion

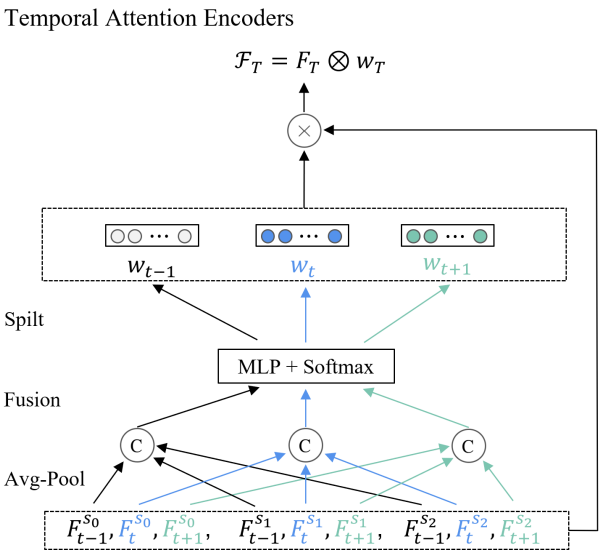


Figure 4. The structure of Temporal Attention Encoder (TAE).

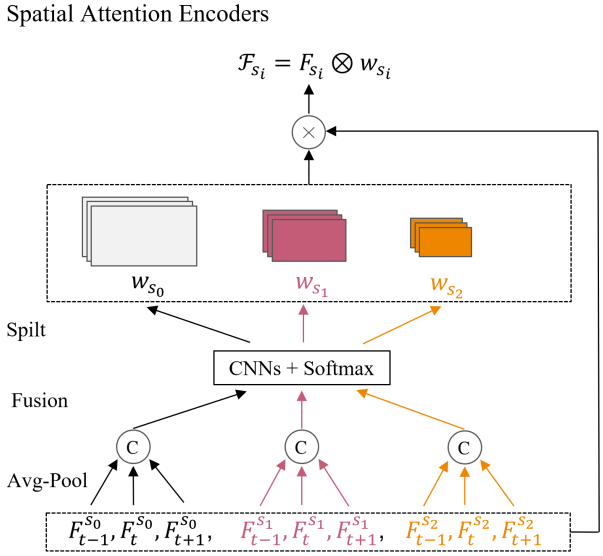


Figure 5. The structure of Spatial Attention Encoder (SAE).

Transformer (TSFT) is shown in Fig.3.

3.2. Network Architecture

3.2.1 Feature alignment and extraction

Our model utilizes three frames $(I_{t-1}^m, I_t^m, I_{t+1}^m)$ as input and outputs the moiré-free prediction I_t^c. Initially, the multi-frame images are input into the Pyramid Cascading Deformable (PCD) alignment module[47], which is used for extract deep features. Subsequently, a convolutional layer is employed to expand the feature channels. Finally, a convolution-based downsampling layer is applied to extract multi-scale features, as shown in Eq.5:

$$\begin{aligned}&(F_{t-1,t,t+1}^{s_0}, F_{t-1,t,t+1}^{s_1}, F_{t-1,t,t+1}^{s_2})\\&=\text{Down}(\text{Conv}(\text{PCD}(I_{t-1}^m, I_t^m, I_{t+1}^m)))\\&F_T^{s_i} \in \mathbb{R}^{n\times\frac{h}{2^i}\times\frac{w}{2^i}},\ T=t-1,t,t+1,\ i=0,1,2\end{aligned} \tag{5}$$

where h and w represent the height and width of the feature maps produced by the PCD. $\text{Down}(\cdot)$ and $\text{Conv}(\cdot)$ represent the downsampling and convolutional layers, respectively. Each frame is expanded through the convolutional network to n channel features. The PCD alignment module, which is composed of a pyramid structure and deformable convolutions[8], is designed to implicitly align multiple frames in the feature space without estimating optical flow. The purpose of this stage is to extract and generate multi-scale deep features.

3.2.2 Temporal and Spatial Attention Encoder

The Temporal Attention Encoder (TAE), as illustrated in Fig. 4, focuses on learning temporal attention weights from aligned features to minimize the interference of redundant information. Additionally, temporal weighting helps mitigate the effects of video illumination changes on the restoration process. For multi-scale input features, they are divided along the temporal dimension t into three groups, as shown in Eq.6:

$$F_T = (F_T^{s_0}, F_T^{s_1}, F_T^{s_2}), \quad T = t-1, t, t+1 \tag{6}$$

Subsequently, global average pooling (Avg-Pool) and concatenation (Fusion) are used to obtain three different 1D global features based on the temporal dimension. Then, the features are input into a Multilayer Perceptron (MLP) and split into three groups of temporal weights. Each temporal weight is applied to the corresponding scale of input features to enhance them:

$$\begin{aligned}&v_{t-1}, v_t, v_{t+1} = Avg(F_{t-1}), Avg(F_t), Avg(F_{t+1})\\&w_{t-1}, w_t, w_{t+1} = \sigma(\text{MLP}(v_{t-1}, v_t, v_{t+1}))\\&\mathcal{F}_T = w_T \otimes F_T, \quad w_T \in \mathbb{R}^n,\ T=t-1,t,t+1\end{aligned} \tag{7}$$

The Spatial Attention Encoder (SAE), as shown in Fig. 5, has a similar structure to the Temporal Attention Encoder (TAE). However, SAE focuses on learning the spatial distribution of moiré patterns. For multi-scale features, the inputs are divided along the spatial dimension, as shown in Eq.8:

$$F_{s_i} = (F_{t-1}^{s_i}, F_t^{s_i}, F_{t+1}^{s_i}), \quad i = 0, 1, 2 \tag{8}$$

Then, global average pooling (Avg-Pool) and concatenation (Fusion) are used to obtain three different 2D global features based on the spatial dimension. These are input into Convolutional Neural Networks (CNNs) to obtain three

spatial weights at different scales. Finally, the multi-scale spatial attention weights are respectively multiplied with the corresponding feature maps:

$$\begin{aligned} &v_{s_0}, v_{s_1}, v_{s_2} = Avg(F_{s_0}), Avg(F_{s_1}), Avg(F_{s_2}) \\ &w_{s_0}, w_{s_1}, w_{s_2} = \sigma(\mathrm{CNNs}(v_{s_0}, v_{s_1}, v_{s_2})) \\ &\mathcal{F}_{s_i} = w_{s_i} \otimes F_{s_i}, \quad w_{s_i} \in \mathbb{R}^{\frac{h}{2^i} \times \frac{w}{2^i}},\ i = 0, 1, 2 \end{aligned} \tag{9}$$

In summary of the encoding phase, the multiple frames are input to extract the aligned features $M_F^{s_i}$. Then, TAE and SAE are utilized for encoding and weighting to get output features $M_T^{s_i}$ and $\mathcal{F}_T^{s_i}$. The encoding component integrates skip-connection, which is crucial for maintaining gradient stability and enhancing the model's generalization ability, as shown in Eq.10:

$$\begin{aligned} &M_T^{s_i} = F_T^{s_i} + w_T \otimes F_T^{s_i} \\ &\mathcal{F}_T^{s_i} = M_T^{s_i} + w_{s_i} \otimes M_T^{s_i} \\ &T = t-1, t, t+1, \quad i = 0, 1, 2 \end{aligned} \tag{10}$$

3.2.3 Temporal-Spatial Fusion Decoder

Rethinking the structure of Swin Transformer [29] and SwinIR [23], we introduce patch partition and patch merging operations in the decoding phase (Fig. 3). The shifted window approach is also employed to introduce cross-window connections and reduce computational cost.

Firstly, the Feature Fusion Network (FFN) is used to extract the features and integrate the outputs from both temporal and spatial attention encoders. Composed of convolutional layers, FFN employs multiple cross-connections to merge features across different scales and temporal sequences. During the encoding phase, residual connections are employed to retain the original features from various stages, which are inputs to FNN. Unlike the fusion of features at different scales in TAE and SAE modules, FFN concentrates on the integration of weighted features from distinct stages. The convolutional layers at multi-scale capture different sizes of receptive fields, enabling a better fit to the distribution of moiré. This module is beneficial for identifying and eliminating moiré patterns of various shapes and sizes.

Secondly, the features are input into a multi-head self-attention module, which has two leading roles: 1. The self-attention mechanism is capable of learning the correlations between different image patches and frames, capturing the impacts on the video caused by variations in moiré intensity and other factors. Window-based self-attention can introduce local inductive bias, enhancing training efficiency. Specifically, self-attention allows the model to autonomously determine the shape and type of its receptive field through the attention weight map in computer vision tasks, thereby offering superior generalization and learning capabilities. 2. MLP and convolutional layers within the module are utilized for learning image textures and reconstructing high-frequency details. The process of the decoding phase is shown in Eq. 11:

$$\begin{aligned} &F_{\text{ffn}} = \mathrm{FFN}(\mathcal{F}_{t-1,t,t+1}^{s_0}, \mathcal{F}_{t-1,t,t+1}^{s_1}, \mathcal{F}_{t-1,t,t+1}^{s_2}) \\ &F_{\text{msa}} = \mathrm{Attention}(F_{\text{ffn}}) \\ &F_{\text{out}} = \mathrm{MLP}(\mathrm{LN}(F_{\text{msa}})) + F_{\text{msa}} \\ &O_t^{s_0}, O_t^{s_1}, O_t^{s_2} = \mathrm{Up}(\mathrm{PM}(F_{\text{out}})) \end{aligned} \tag{11}$$

where $\mathrm{FFN}(\cdot)$, $\mathrm{Attention}(\cdot)$, $\mathrm{Up}(\cdot)$ and $\mathrm{PM}(\cdot)$ represents feature fusion network, multi-head attention mechanism, upsampling layer and patch merging, respectively. The enhanced features are input into the FFN to fuse the attention information of two dimensions. In this step, the image is divided into non-overlapping patches to facilitate the subsequent learning of self-attention weights.

The model can efficiently reconstruct high-frequency details in combination with the Feed-Forward and Detail Reconstruction layer. Specifically, this framework is composed of Multi-Layer Perceptron (MLP), Layer Normalization (LN) layers, Patch Merging (PM), and Upsampling (Up) layers. $\mathrm{MLP}(\cdot)$ employs two fully connected layers with GELU nonlinearity in between for advanced feature transformation. $\mathrm{LN}(\cdot)$ are incorporated into both the Multi-Head Self-Attention (MSA) and MLP to ensure gradient stability. $\mathrm{PM}(\cdot)$ is designed to interact with the preceding patch partition layer, enabling information exchange and integration of features across different frames. This module utilizes linear layers and concatenation operations while controlling the dimensionality of the output features. Finally, $\mathrm{Up}(\cdot)$ is used to obtain the restored image $O_t^{s_0}$. TSFT also outputs images of different sizes ($O_t^{s_1}, O_t^{s_2}$) for multi-scale supervised strategy.

3.2.4 Analysis

Our model is specifically designed to remove moiré patterns by learning temporal information, spatial noise distribution, and other factors. TSFT effectively addresses the limitation of existing demoiréing methods that fail to fully leverage temporal information. After using PCD to generate deep features, the introduction of TAE primarily aims to mitigate the impact of disruptive information in nearby frames. SAE utilizes the multi-scale method and takes account of various receptive fields, enabling it to learn about the overall shape and detailed patterns of moiré. Spatial attention allows the model to focus more intensely on areas with higher moiré intensity, thereby enhancing the efficiency of model restoration.

The decoding phase comprises the Feature Fusion Network (FFN) and self-attention module. FFN learns the information extracted by TAE and SAE and interacts with

features across different scales to model the spatial distribution of moiré over time. The self-attention mechanism takes into account the correlation of the features, further utilizing information from adjacent frames and various patches of the image. Compared to traditional convolutional methods, the self-attention mechanism can effectively fit the complex distribution of moiré due to its superior adaptability and learning capabilities.

The architecture design of our model comes from our research on video demoiréing task. Through the analysis of video moiré patterns, we identified that the inherent challenges of these patterns make them more difficult to remove than other classic types of noise. Based on these characteristics, we develop an intuitive approach to explicitly model moiré patterns in videos. We decompose the video demoiréing task into two main components: temporal and spatial feature learning and the restoration of degraded content. Experimental results validate the advantages of our model. However, considering the complexity of moiré patterns—such as their varying texture shapes, local color shifts, and color tone deviations—efficiently modeling moiré is still a challenging problem. One of our future research directions is to improve and optimize the modeling approach, enabling the designed model to better balance computational complexity and feature learning efficiency.

3.3. Loss Function

Existing researches[24, 59] has confirmed the efficacy of L1 loss for restoration tasks. However, L1 loss focuses on pixel-level differences and overlooks the role of deeper structural information. The texture and structural information of moiré patterns on a larger scale also indicate the necessity of learning deep structural information. Moreover, we use the deep supervision strategy to supervise the output images of different sizes ($O_t^{s_0}, O_t^{s_1}, O_t^{s_2}$) and assist the model training. Building on existing algorithms [33, 9], we incorporate perceptual loss[19] into the training loss function to supervise deep features and the structural information, as shown in Eq. 12:

$$\mathcal{L}(O_t^{s_i}, G_t^{s_i}) = \mathcal{L}(O_t^{s_i}, G_t^{s_i}) + \lambda \cdot \mathcal{L}_p(O_t^{s_i}, G_t^{s_i}), \quad i = 0, 1, 2 \tag{12}$$

$O_t^{s_0}$ represents the restored image, and $O_t^{s_1}, O_t^{s_2}$ represent the multi-scale output of TSFT. $G_t^{s_0}, G_t^{s_1}, G_t^{s_2}$ represent the corresponding ground truth that downsampled to the same size. λ is the hyper-parameter used to balance two losses. The final training loss $\mathcal{L}_{train}$ is the sum of multi-scale losses:

$$\mathcal{L}_{train} = \sum_{i=0}^{2} \mathcal{L}(O_t^{s_i}, G_t^{s_i}) \tag{13}$$

4. Experiments

4.1. Experimental Setup

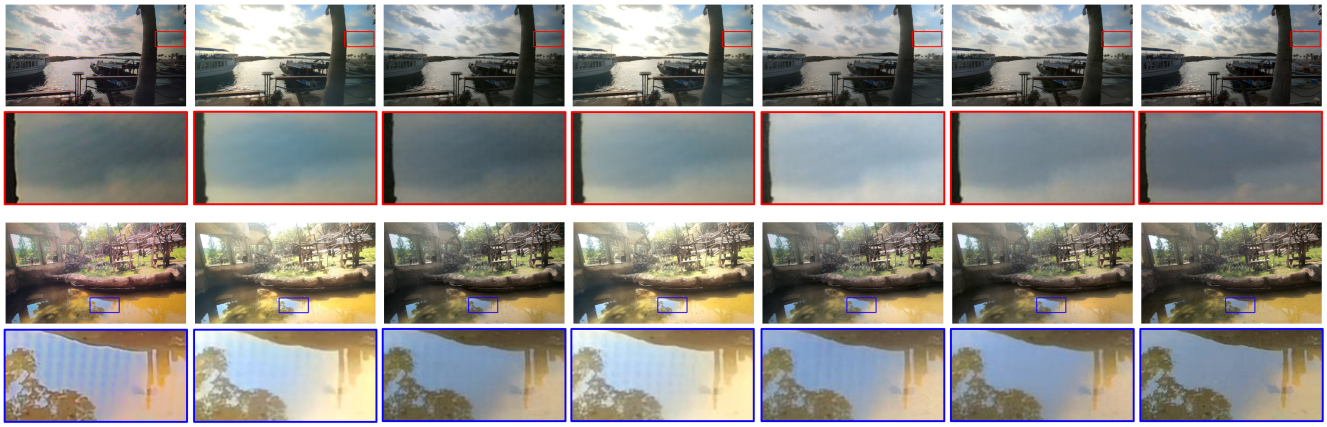
In this paper, we conduct experiments on the Video Demoiréing Dataset (VDD)[9], which comprises 290 moiré videos and the corresponding clean videos. Each video contains 60 frames, with the resolution of 720p (1280×720). The source videos were displayed on a monitor and recorded using a handheld camera to capture the moiré videos. Following the existing methods, VDD is divided into 247 training videos and 43 testing videos. The TSFT model uses sequences of three consecutive frames as input to predict a single restored frame.

We set the weight of the perceptual loss (λ) to 1, the number of heads in the self-attention component to 4, and the size of the shifted windows to 10 after extensive experimental research. The inputs are randomly cropped to 400×400 for patch embedding during the training phase. We used the Adam[20] optimizer with an initial learning rate set to 0.0002 and a batch size of 1. We train 60 epochs in the Pytorch framework on the Nvidia RTX 3090. Similar to other Transformer-based models, the self-attention module in Temporal-Spatial Fusion Decoder can be stacked to enhance model performance. We achieved significant results in Video Demoiréing task with a single-layer decoder. Experimenting with a multi-layer decoder for input images with higher resolutions could be beneficial.

4.2. Quantitative Comparison

We compare our approach with state-of-the-art demoiréing methods (DMCNN[39], MBCNN[60], VDN[9], ESDNet[54], CIDNet[34], DTCENet[27] and STDNet[32]) and the foundational backbone network in the restoration field (U-Net[36]), to demonstrate the effectiveness of TSFT. Other image demoiréing networks are also utilized adjacent frames as inputs and employ the PCD module to extract deep features for horizontal comparison.

In Quantitative Comparison, we evaluate the restoration performance of the models from three perspectives: frame-level quality, video-level quality, and human perceptual quality. Frame-level quality primarily assesses the pixel-level differences between the output images and ground truth, as well as the effectiveness of moiré texture removal. We adopt widely used metrics PSNR and SSIM for comparison. For video-level quality, we adopt FID (Fréchet Inception Distance) metric to measure the overall quality of the output video sequences, where FID utilizes a pretrained I3D[3] to extract features and calculate differences. For human perceptual quality, we utilize LPIPS[57] as the evaluation metric to assist in analyzing the quality of the output images, as it aligns more closely with human visual perception.



(a) Input (b) U-net (c) DMCNN (d) MBCNN (e) VDN (f) Ours (g) GT

Figure 6. Qualitative comparisons on VDD.

Table 1. Quantitative comparisons with state-of-the-art methods on VDD, the best results are highlighted with bold.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓
U-Net	20.348	0.720	0.225	0.204
DMCNN	20.321	0.703	0.321	0.265
MBCNN	21.534	0.740	0.260	0.212
VDN_S	21.772	0.729	0.212	0.094
VDN	21.725	0.733	0.202	0.084
ESDNet	21.812	0.731	0.206	0.101
ESDNet-L	22.003	0.739	0.203	0.097
CIDNet	22.270	0.735	0.184	-
DTCENet	21.881	0.744	**0.181**	0.078
STDNet	22.075	0.740	0.196	-
Ours	**22.653**	**0.752**	0.189	**0.063**

Table 1 presents the restoration metrics of different models on VDD, where VDN_S is a variant of VDN that inputs three repeated frames. By comparing the performance with existing methods, the results show that TSFT significantly outperforms state-of-the-art methods across most of the metrics. This demonstrates that our method not only possesses superior demoiréing capabilities but also effectively restores image details, enhancing the frame-level quality, video-level quality, and human perceptual quality. Furthermore, our research provided a baseline framework suitable for moiré analysis and removal. This framework holds potential for future expansion in constructing large-scale pre-trained models.

Further analysis of existing methods, U-Net and DMCNN utilize stacked pooling and convolution layers to progressively reduce the size of feature maps, thereby constructing multi-scale features. However, they overlook the interaction between features at different scales, which reduces the efficiency of the models. TAE, SAE and FFN in our model all incorporate interactions across various scales. Our model can fit the distribution of moiré more effectively and improve its performance.

MBCNN, VDN and ESDNet extract multi-scale information, but neither has an effective architecture for learning the temporal trends of moiré patterns in video sequences. In contrast, TAE allows the model to focus on the temporal distribution of moiré. As the intensity of moiré varies across different frames, areas with less intense patterns can provide more information of source videos, aiding the restoration of the current frame. The decoding stage focuses on learning the correlations in image-level and video-level, which proves more effective in demoiréing.

Due to the dynamic changes in moiré stripes and colors, cross-frame correlation plays an essential role in modeling moiré patterns and preserving temporal consistency. CIDNet and DTCENet, which are based on convolutional networks, leverage complementary information from nearby frames. However, they still lack the capability to learn temporal correlations, resulting in sub-optimal performance. STDNet also uses the self-attention method and decoupled spatio-temporal manner for video demoiréing. The key difference between STDNet and TSFT is that STDNet emphasizes learning moiré textures in the decoder stage by improving multi-head attention. While this architecture enhances performance, it also increases computational cost, and the feature extraction network of STDNet is relatively simple, limiting its ability to deeply learn moiré features. Our method performs multi-level information extraction and interaction during the encoder phase, greatly enhancing feature extraction capabilities. This structure enables the model to to improve performance while maintaining computational efficiency (TSFT having about one-third of the parameters of STDNet).

We also show the parameters and running time of our model and state-of-the-art methods in Table 2. TSFT exhibits a significantly lower number of parameters compared to U-Net, MBCNN, and STDNet. This indicates that our

Table 2. Computational efficiency of different methods.

Metric	U-Net	DMCNN	MBCNN	VDN	ESDNet-L	CIDNet	DTCENet	STDNet	Ours
Params(M)	17.26	1.40	14.19	5.98	10.62	4.57	5.17	27.96	10.46
Time(s)	0.089	0.110	0.264	0.261	0.252	-	-	0.300	0.197

model also possesses certain advantages in computational efficiency. Although multi-head self-attention in TSFT increases computational complexity and network size, the novel architectural design allows our model to achieve a significantly lower parameter count than other attention-based methods. Besides, our model demonstrates superior performance and generalization capabilities while maintaining comparable parameters and inference times to most state-of-the-art CNN-based approaches. Another advantage of attention-based methods is that they can achieve superior performance for large-scale datasets, and the stacking of self-attention modules can enhance the model's restoration applicability.

4.3. Qualitative Comparison

Fig. 6 shows the images restored by different methods, and TSFT demonstrates superior performance. It is evident that TSFT is more effective in removing moiré and is more accurate in restoring colors and high-frequency details, such as the sky and reflections in the water. For the picture above, U-Net, MBCNN and VDN show less efficiency in color restoration, with the overall image color noticeably brighter, while DMCNN still has some remaining moiré textures. For the picture below, the other comparison models exhibit the same issues and don't effectively remove moiré artifacts.

In contrast, our method excels in color restoration, texture removal, and detail reconstruction. This success is attributed to the architectural design of our model, which considers the characteristics of moiré patterns. TSFT utilizes auxiliary information from temporal and spatial dimensions to construct attention weights, thus fitting the distribution of moiré across different scales effectively. Moreover, the decoding phase uses feed-forward network to reconstruct the image's high-frequency details, making the restored images visually superior to those produced by existing methods.

5. Ablation Study

The ablation study consists of three parts: First, we investigate the roles of different components of TSFT to analyze their contributions. Second, we study the effect of the loss function by adjusting the hyper-parameter λ to find an appropriate balance. Third, we utilize visualization of feature weight maps to understand the role of the fused attention method in our model.

5.1. Components of TSFT

The comparative experiment included three variants of TSFT, each lacking a key component: the Temporal Attention Encoder (w/o TAE), the Spatial Attention Encoder (w/o SAE), and the Multi-head Self-Attention (w/o MSA). We also add the experimental results of the baseline(without core modules) and full model for comparison. To ensure that the reduction in parameter count didn't affect the overall model's performance, the removed modules were replaced with residual convolutional networks[16] of approximately the same parameter count. As shown in Table 3, the removal of these components resulted in varying degrees of performance degradation. Fig. 8 shows that our network lacking certain modules are inferior in color restoration and texture removal compared to the original model.

Table 3. Quantitative comparison among different variants of TSFT.

Network	PSNR↑	SSIM↑	LPIPS↓	Loss↓
Baseline	21.670	0.735	0.205	1.374
w/o TAE	22.016	0.741	0.200	1.328
w/o SAE	22.413	0.745	0.196	1.342
w/o MSA	22.425	0.745	0.190	1.359
Ours	**22.653**	**0.752**	**0.189**	**1.306**

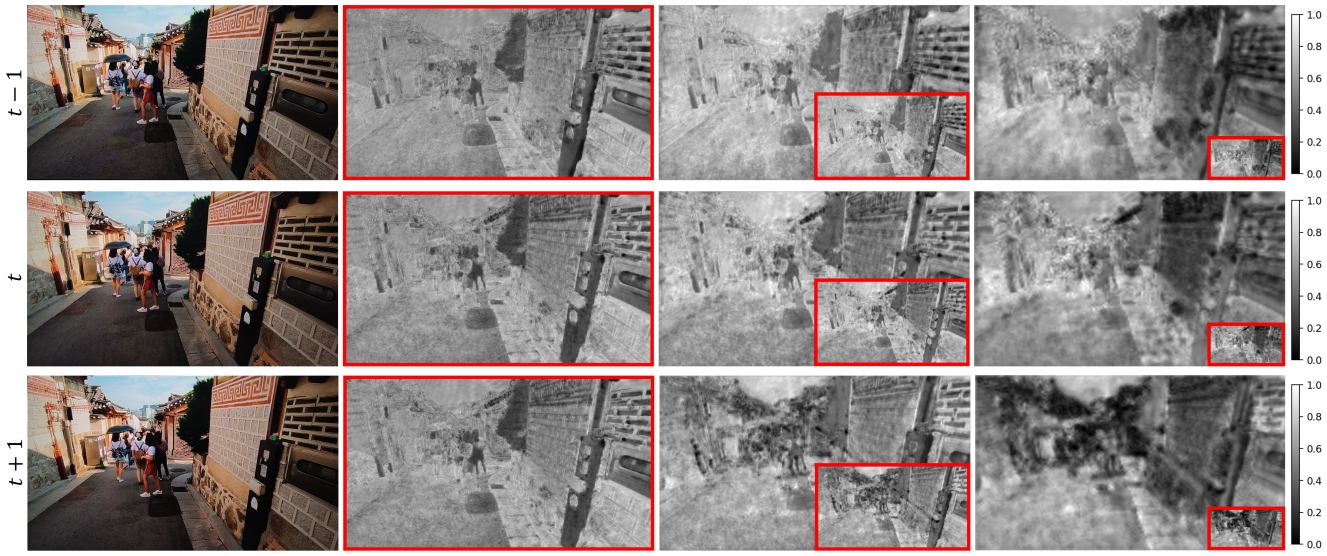
TAE efficiently utilizes information provided by adjacent frames, SAE is primarily used to fit the spatial distribution of moiré patterns, focusing the model learning on areas with higher intensity of moiré, and the decoder's role is to effectively fuse the multi-scale features learned by TAE and SAE and to learn the temporal variation of moiré based on correlations. The results from the figure and table clearly demonstrate that each component of TSFT plays a crucial role in Video Demoiréing.

5.2. Effect of perceptual loss

Table 4. Quantitative comparison of TSFT trained with different loss functions, λ represents the weight of perceptual loss.

Metrics	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 1.5$
PSNR↑	22.001	22.345	22.343	**22.653**	22.587
SSIM↑	0.743	0.747	0.746	**0.752**	0.749
LPIPS↓	0.251	0.197	0.195	**0.189**	0.190

The loss function plays an important role in guiding the model training. L1 loss is commonly used in image and video restoration tasks because it helps learn the differences between pixels. Unlike traditional noise, moiré patterns vary in shape and color across different regions. Perceptual loss can help to align the deep global information of

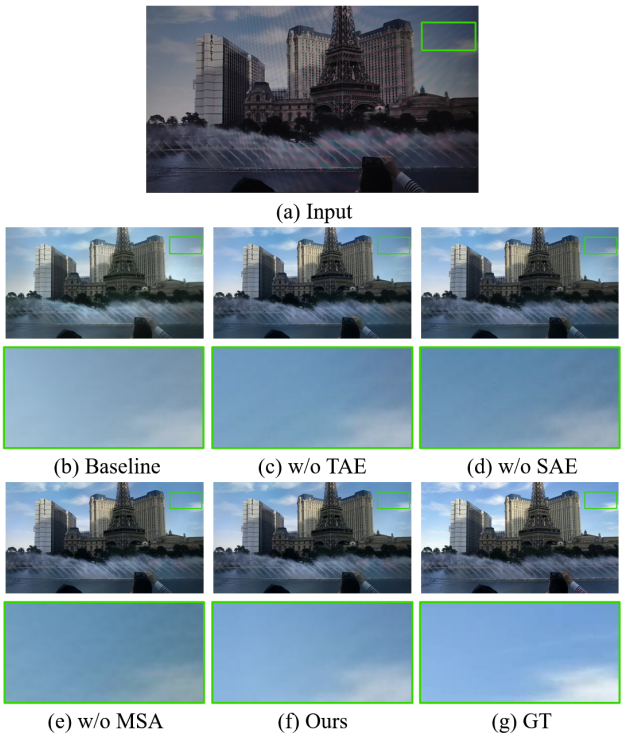


(a) Input (b) Scale 0 (c) Scale 1 (d) Scale 2

Figure 7. The attention maps on VDD, the attention maps are scaled up to the same size to better compare the results of different scales, red areas show the original size of the attention maps.

Table 5. Quantitative comparisons with state-of-the-art methods on NTURain dataset.

Metrics	Rainy	FastDeRain	PreNet	SpacCNN	SLDNet	MPRNet	S2VD	ESTINet	**Ours**
PSNR↑	30.41	30.54	32.99	33.11	34.89	36.11	37.37	37.48	**37.88**
SSIM↑	0.902	0.925	0.952	0.947	0.954	0.963	0.968	0.970	**0.972**



(a) Input

(b) Baseline (c) w/o TAE (d) w/o SAE

(e) w/o MSA (f) Ours (g) GT

Figure 8. Visual comparison among different variants of TSFT.

the restored frame and the ground truth by computing distances in the feature space, thus increasing the efficiency of video restoration.

We employ both L1 loss and perceptual loss during the training phase and use a coefficient λ to balance these two loss functions (Sec. 3.3). The value of λ being either too high or too low will affect the effectiveness of model training. So, we conduct the experiment with different values of λ. According to the results presented in Table 4, the best training outcomes were achieved when $\lambda = 1$. Moreover, the application of perceptual loss ($\lambda > 0$) significantly reduced the value of LPIPS, enhancing the human visual quality of the restored videos.

5.3. Attention visualization

In this chapter, we visualize the attention weight maps that are used for feature enhancement to demonstrate the role of the attention mechanism in TSFT. As shown in Fig. 7, the attention mechanism of TSFT effectively captures the spatial distribution of moiré artifacts, such as moire stripes in the sky. For image regions where moiré patterns are less prominent, such as the darker-colored trees, the weight of the corresponding regions is small, thereby enhancing the efficiency of parameter utilization. Large-scale weight maps focus on the overall distribution of moiré across the entire frame, and small-scale weight maps focus on the detailed textures in local areas. By integrating multi-scale features, the modeling capability of our network is

(a) Input (b) Ours (c) GT (d) Input (e) Ours (f) GT

Figure 9. Visual results on NTURain dataset.

enhanced. The figure illustrates that our model learns the differences between various frames through the attention weights and increases the efficiency of utilizing information from adjacent frames.

6. Discussion

To validate the potential of our model in other video restoration tasks, we conducted experiments on Video Deraining task to assess the robustness of TSFT in low-level tasks. Rain is a common weather condition that reduces the visibility in videos. The primary causes of degradation are rain streaks and rain accumulation (or rain veiling effect). Rain streaks obscure parts of the background scene, altering the appearance of the video and making the scene appear blurry. Rain accumulation dilutes the scene's colors, lowers overall contrast, and produces a masking effect. Both rain streaks and rain accumulation decrease the visibility of the scene. Particularly in videos, rain accumulation becomes more pronounced over time. Therefore, to obtain better visual information of the background scene, it is necessary to remove rain from videos. Similar to moiré, Rain and moiré noise both affect the source video in terms of color and texture. However, the difference between the two noises is that rain doesn't have complex streaks and colors, and the variation of rain over time is more monotonous compared to moiré.

The experiments validate the restoration capabilities of TSFT on a public video deraining dataset. We choose the NTURain dataset[5] for this study, which consists of videos taken by a camera with slow and fast movements. The training set includes 24 rainy sequences and clean sequence videos, while the testing set comprises 8 pairs. We compare the restoration performance of the state-of-the-art video deraining methods, including FastDeRain[18], PreNet[35], SpacCNN[5], SLDNet[53], MPRNet[17], S2VD[55] and ESTINet[56]. We compare the average metrics in testing set, as shown in Table 5. The results indicate that our model surpasses existing methods in video deraining and achieves excellent performance.

Fig. 9 shows the comparison between the frames restored by TSFT and the source frames. TSFT effectively removes rain noise from the videos while reconstructing the details and areas obscured by rain. We analyze that other video restoration process also utilizes information in nearby frames like Video Demoiréing. The texture and dynamics of rain noise are relatively simple and can be considered as a particular form of noise (based on the study of Sec.3.1). Therefore, our model is also applicable in Video Deraining task. The results demonstrate the generalization ability of TSFT and prove the model's capability to handle other video restoration tasks.

7. Conclusion

In this paper, we proposed an attention-based DNN architecture to solve the problem of video demoiréing through better qualitative and quantitative results. Based on the characteristics of moiré patterns, we intuitively model the moiré videos and decompose the task into several steps to construct targeted modules. Considering the efficiency of Vision Transformer and the gaps in existing demoiréing research, we introduce the transformer architecture into this field and construct an efficient model named TSFT. First, multiple frames are input into TSFT for multi-scale feature alignment and extraction. In the encoding phase, the model learns the variations of moiré in both spatial and temporal dimensions across consecutive frames while minimizing

the interference of redundant information. In the decoding phase, our model captures temporal correlations to learn the impact of moiré intensity changes and reconstructs the texture details. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods on video demoiréing dataset. Our method also achieves outstanding results in video deraining, proving that TSFT can be effectively generalized to another low-level vision task. The experiments show the potential of our model in the video restoration field.

References

[1] L. Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 7–12. IEEE, 2000. 2, 3

[2] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4778–4787, 2017. 2

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 8

[4] C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 4

[5] J. Chen, C.-H. Tan, J. Hou, L.-P. Chau, and H. Li. Robust video content alignment and compensation for rain removal in a cnn framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6286–6295, 2018. 12

[6] J. Chen, X. Tan, C. Shan, S. Liu, and Z. Chen. Vesr-net: The winning solution to youku video enhancement and super-resolution challenge. *arXiv preprint arXiv:2003.02115*, 2020. 3

[7] X. Cheng, Z. Fu, and J. Yang. Multi-scale dynamic feature encoding network for image demoiréing. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3486–3493, 2019. 2, 3

[8] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2, 3, 6

[9] P. Dai, X. Yu, L. Ma, B. Zhang, J. Li, W. Li, J. Shen, and X. Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17622–17631, 2022. 2, 3, 8

[10] Q. Dai, X. Cheng, L. Zhang, and L. Sun. Adaptive multi-spectral encoding network for image demoiréing. *IEEE Transactions on Instrumentation and Measurement*, 2023. 3

[11] Z. Dai, H. Liu, Q. V. Le, and M. Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021. 4

[12] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12124–12134, 2022. 4

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] B. He, C. Wang, B. Shi, and L.-Y. Duan. Mop moire patterns using mopnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2424–2432, 2019. 2, 3

[15] B. He, C. Wang, B. Shi, and L.-Y. Duan. Fhde2net: Full high definition demoireing network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII*, pages 713–729, 2020. 2, 3

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 10

[17] T.-W. Hui and C. C. Loy. Liteflownet3: Resolving correspondence ambiguity for more accurate optical flow estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 169–184. Springer, 2020. 12

[18] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, L.-J. Deng, and Y. Wang. Fastderain: A novel video rain streak removal method using directional gradient priors. *IEEE Transactions on Image Processing*, 28(4):2089–2102, 2018. 12

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 8

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[21] K. Li, Y. Wang, G. Peng, G. Song, Y. Liu, H. Li, and Y. Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2021. 4

[22] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 335–351. Springer, 2020. 3

[23] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 4, 7

[24] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 8

[25] F. Liu, J. Yang, and H. Yue. Moiré pattern removal from texture images via low-rank and sparse matrix decomposition. In *2015 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2015. 3

[26] J. Liu, W. Yang, S. Yang, and Z. Guo. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3233–3242, 2018. 2

[27] L. Liu, J. An, S. Yuan, W. Zhou, H. Li, Y. Wang, and Q. Tian. Video demoiréing with deep temporal color embedding and video-image invertible consistency. *IEEE Transactions on Multimedia*, 2024. 2, 3, 8

[28] L. Liu, J. Liu, S. Yuan, G. Slabaugh, A. Leonardis, W. Zhou, and Q. Tian. Wavelet-based dual-branch network for image demoiréing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 86–102. Springer, 2020. 2, 3

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 7

[30] M. Maggioni, Y. Huang, C. Li, S. Xiao, Z. Fu, and F. Song. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3466–3475, 2021. 1

[31] Y. Niu, Z. Lin, W. Liu, and W. Guo. Progressive moire removal and texture complementation for image demoireing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[32] Y. Niu, R. Xu, Z. Lin, and W. Liu. Std-net: Spatio-temporal decomposition network for video demoiréing with sparse transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 3, 8

[33] K. Oksuz, B. C. Cam, E. Akbas, and S. Kalkan. Rank & sort loss for object detection and instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3009–3018, 2021. 8

[34] Y. Quan, H. Huang, S. He, and R. Xu. Deep video demoiréing via compact invertible dyadic decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12677–12686, 2023. 2, 3, 8

[35] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2019. 12

[36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241, 2015. 8

[37] M. S. Sajjadi, R. Vemulapalli, and M. Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018. 3

[38] D. N. Sidorov and A. C. Kokaram. Removing moire from degraded video archives. In *2002 11th European Signal Processing Conference*, pages 1–4. IEEE, 2002. 3

[39] Y. Sun, Y. Yu, and W. Wang. Moiré photo restoration using multiresolution convolutional neural networks. *IEEE Transactions on Image Processing*, 27(8):4160–4172, 2018. 2, 3, 8

[40] M. Tassano, J. Delon, and T. Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1354–1363, 2020. 1

[41] Y. Tian, Y. Zhang, Y. Fu, and C. Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3360–3369, 2020. 2, 3

[42] H. Wang, D. Su, C. Liu, L. Jin, X. Sun, and X. Peng. Deformable non-local network for video super-resolution. *IEEE Access*, 7:177734–177744, 2019. 3

[43] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An. Learning for video super-resolution through hr optical flow estimation. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14*, pages 514–529. Springer, 2019. 3

[44] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu. Crossformer++: A versatile vision transformer hinging on cross-scale attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[45] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 4

[46] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 4

[47] X. Wang, K. C. Chan, K. Yu, C. Dong, and 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*, pages 0–0, 2019. 3, 6

[48] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 4

[49] Z. Wei, J. Wang, H. Nichol, S. Wiebe, and D. Chapman. A median-gaussian filtering framework for moiré pattern noise removal from x-ray microscopy image. *Micron*, 43(2-3):170–176, 2012. 3

[50] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 4

[51] Q. Xu and Y. Qian. Bidirectional transformer for video deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8450–8461, 2022. 4

[52] J. Yang, F. Liu, H. Yue, X. Fu, C. Hou, and F. Wu. Textured image demoiréing via signal decomposition and guided filtering. *IEEE Transactions on Image Processing*, 26(7):3528–3541, 2017. 3

[53] W. Yang, R. T. Tan, S. Wang, and J. Liu. Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1720–1729, 2020. 1, 12

[54] X. Yu, P. Dai, W. Li, L. Ma, J. Shen, J. Li, and X. Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 646–662, 2022. 2, 3, 8

[55] Z. Yue, J. Xie, Q. Zhao, and D. Meng. Semi-supervised video deraining with dynamical rain generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 642–652, 2021. 1, 12

[56] K. Zhang, D. Li, W. Luo, W. Ren, and W. Liu. Enhanced spatio-temporal interaction learning for video deraining: faster and better. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1287–1293, 2022. 12

[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 8

[58] Y. Zhang, M. Lin, X. Li, H. Liu, G. Wang, F. Chao, S. Ren, Y. Wen, X. Chen, and R. Ji. Real-time image demoiréing on mobile devices. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, pages 1–13. OpenReview.net, 2023. 3

[59] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 8

[60] B. Zheng, S. Yuan, G. Slabaugh, and A. Leonardis. Image demoireing with learnable bandpass filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3636–3645, 2020. 2, 3, 8