Deep Video Demoiréing via Compact Invertible Dyadic Decomposition

Yuhui Quan[1,2] Haoran Huang[1] Shengfeng He[3] Ruotao Xu[1,2] *
[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[2]Pazhou Lab, Guangzhou 510335, China
[3]School of Computing and Information Systems, Singapore Management University, 188065, Singapore
csyhquan@scut.edu.cn, csherry@mail.scut.edu.cn, shengfenghe@smu.edu.sg, xrt@scut.edu.cn †

Abstract

Removing moiré patterns from videos recorded on screens or complex textures is known as video demoiréing. It is a challenging task as both structures and textures of an image usually exhibit strong periodic patterns, which thus are easily confused with moiré patterns and can be significantly erased in the removal process. By interpreting video demoiréing as a multi-frame decomposition problem, we propose a compact invertible dyadic network called CIDNet that progressively decouples latent frames and the moiré patterns from an input video sequence. Using a dyadic cross-scale coupling structure with coupling layers tailored for multi-scale processing, CIDNet aims at disentangling the features of image patterns from that of moiré patterns at different scales, while retaining all latent image features to facilitate reconstruction. In addition, a compressed form for the network's output is introduced to reduce computational complexity and alleviate overfitting. The experiments show that CIDNet outperforms existing methods and enjoys the advantages in model size and computational efficiency.

1. Introduction

Moiré patterns are visual artifacts that arise when the spatial frequency of details in a scene exceeds the density of sensors. They can occur in various scenarios, such as when photographing a screen or filming a surface with complex textures. Moiré patterns can have a negative impact on the visual aesthetics and perception of an image, as well as on the performance of downstream vision systems that require high-quality image sequences as input. Video demoiréing aims to remove moiré patterns from a video sequence while preserving all image details. This technique has applications in many fields such as photography and computer vision.

The formation model of moiré patterns is complex, as the corruption not only produces varying stripe-like patterns but also causes local and global color shifts; see *e.g.* [1, 32]. Generally, an observed image $\mathbf{I}$ corrupted by moiré patterns is composed of the underlying latent image $\mathbf{X}$ of a scene and the corresponding layer $\mathbf{M}$ of moiré patterns. The latter contains all intensity and color fluctuations caused by the moiré patterns [12]. Recovering a set of clean frames $\{\mathbf{X}^t\}_t$ from only the input moiré-corrupted frames $\{\mathbf{I}^t\}_t$ is a challenging inverse problem. It requires accurate detection and removal of all moiré patterns in each frame to recover all pixels contaminated by the moiré patterns.

There are a few methods (*e.g.* [12, 20]) proposed for image demoiréing, which used handcrafted priors for both moiré patterns and latent images to recover contaminated image pixels. However, moiré patterns exhibit diversity in density, scale, shape, and color effect across images, and even within the same image. Moreover, images of natural scenes can vary significantly in appearance. Therefore, predefined priors on moiré patterns or images are usually too simplistic to handle degraded images of complex scenes and are not adaptive to different images. Recently, deep learning has emerged as a promising approach for single image demoiréing, as demonstrated in [1, 2, 6, 7, 11, 13, 14, 22, 28, 32]. These methods train an end-to-end Deep Neural Network (DNN) that maps an image with moiré patterns to its corresponding moiré-pattern-free latent image. Although these methods have made a breakthrough over traditional methods, their frame-by-frame processing manner for video demoiréing ignores important temporal cues in video frames.

In comparison to image demoiréing, video demoiréing is still in its infant stage (see [2] for the seminal work published in 2022), with much room for improvement in terms of performance and efficiency. Therefore, there is a need

[*] Corresponding author: Ruotao Xu (xu.ruotao@outlook.com).

[†] This work is partially supported by the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012841, 2022A1515011755, 2022A1515011087); National Natural Science Foundation of China (Grant No. 62106077, 61972162); Natural Science Foundation of Guangdong Province for Distinguished Young Scholars (Grant No. 2023B1515020097); Postdoctoral Special Foundation of China (Grant No. 2022T150219); Postdoctoral Foundation of China (Grant No. 2020M682705); and Singapore Ministry of Education Academic Research Fund Tier 1 (Grant No. MSS23C002).

for developing computationally-efficient video demoiréing methods that excel in both moiré pattern removal and detail preservation, which is also the aim of this paper.

1.1. Motivation and Main Idea

Consider a single frame $\mathbf{I}^t$ with a moiré pattern layer $\mathbf{M}^t$ related by some unknown underlying model. Due to at least twice as many unknowns as equations, the problem of removing $\mathbf{M}^t$ is highly ill-posed. Suppose a scene with moiré patterns is in motion across multiple frames. As the scene's spatial frequencies change, so do the corresponding moiré patterns. Nevertheless, due to the complex nature of moiré pattern formation, the change of moiré patterns is not consistently shifted with the objects in motion or scaled with the change of scene depths, but can differ much from the change of the scene; see Fig. 1 for an illustration. This motivated us to develop a DNN that can effectively exploit such difference to decompose degraded video frames into the component of scene and the component of moiré patterns.

Efficient decomposition with perfect information fidelity Previous studies, such as [1,2,6,7,11,13,14,22,28,32], have adopted an encoding-decoding framework, which first progressively encodes an input image into latent image-related features while discarding features related to moiré patterns, and then reconstructs the latent image by decoding its corresponding features. However, typical DNNs like U-Net [6] and ResNet [22] used in these studies may result in incomplete information extraction for recovering all image details. This limitation can cause potential information loss in the encoding stage and negatively impact the reconstruction in the decoding stage. Even for the decoding stage, it may also omit important image features for reconstruction. Therefore, existing demoiréing methods may not ensure complete moiré pattern removal and detail preservation.

To overcome the challenges discussed above, we propose a different approach that interprets video demoiréing as a multi-frame decomposition process with progressive disentanglement. To ensure no information loss during feature extraction and disentanglement, we leverage invertible coupling layers [3] to form an invertible neural network (INN) acting as an invertible decomposition process. The proposed INN uses two paths, one for moiré pattern layer extraction and the other for latent image prediction, and enables rich interactions between the two paths for feature disentanglement. Compared to existing DNNs for image demoiréing, ours using coupling layers not only ensures perfect information fidelity but also improves feature extraction. Furthermore, our INN can replicate its input using its output, and thus it implicitly learns a formation model for moiré-pattern-corrupted image sequences, introducing certain implicit regularization.

Introducing multi-scale analysis into INNs via a dyadic coupling structure As both image patterns and moiré patterns vary significantly over a wide range of scales [1,22], it is important to use a multi-scale architecture for DNNs to effectively distinguish between these patterns at different scales. However, designing an effective multi-scale INN is challenging. While it is possible to introduce multi-scale analysis to an INN by defining the free-form functions in coupling layers using independent multi-scale processing blocks, this approach may not efficiently exploit and fuse cross-scale information. In addition, introducing multi-scale interactions among INN blocks may damage the invertibility.

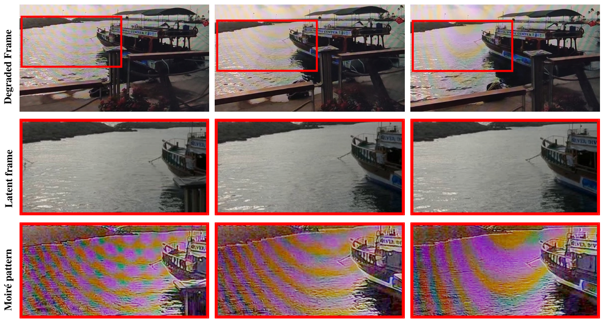


Figure 1: Three consecutive frames with moiré patterns.

We address the issues by introducing a dyadic coupling structure for multi-scale processing. It is a U-shaped structure where encoder blocks and decoder blocks for multi-scale representation are designed and linked in a coupling manner. Specifically, each encoder/decoder block consists of a series of invertible coupling layers and is also constructed with a larger coupling structure that allows partial features from the encoder blocks to be passed to the corresponding decoder blocks via a short path for cross-scale fusion. As a result, the DNN can enjoy enhanced multi-scale processing ability and invertibility simultaneously. In addition, we improve coupling blocks with spatial pyramid pooling and attention mechanisms for higher efficiency of multi-scale processing.

Overfitting reduction using a compact INN An INN requires the same dimension of input and output. One common solution for the equality is to increase the dimension of inputs by repetition. For T video frames, one can define $2T$ outputs to represent all $\mathbf{X}^t$ and $\mathbf{M}^t$, and then double the input by zero padding or frame duplication. However, this strategy not only increases the size and computational complexity of the model but also raises the risk of overfitting as the dimensionality of the problem grows. Our approach uses a compressed form for the INN's output, which reduces the solution ambiguity and mitigates the risk of overfitting.

Assuming that the frames $\{\mathbf{I}^t\}_{t=1}^T$ are aligned with respect to the scene, we can achieve $\mathbf{X}^t \approx \mathbf{X}_0$ for all t, where $\mathbf{X}_0$ denotes the aligned latent clean image. Consequently, the T frames can be decomposed into a latent image $\mathbf{X}_0$ and T layers $\{\mathbf{M}^t\}_{t=1}^T$ of moiré patterns, *i.e.*, $\mathbf{I}^t \approx \mathbf{X}_0 + \mathbf{M}^t$ for all t. Further, while the moiré pattern layers $\{\mathbf{M}^t\}_{t=1}^T$ differ from one another, they remain correlated to some degree over short time intervals after alignment operations. Therefore, we reduce the total number of layers $\mathbf{M}^t$ from T to at least $T-1$. With the above considerations, we propose inputting T aligned frames into the INN and defining the output as T layers, including a latent image and $T-1$ moiré pattern layers.

1.2. Contributions

Combining all above techniques leads to a compact invertible dyadic network (CIDNet) which provides an effective and computationally efficient solution for video demoiréing. Our main contributions are summarized below:

- We leverage deep invertible representation for video demoiréing, which ensures information fidelity in feature disentanglement and introduces implicit regularization.
- We propose a dyadic cross-scale coupling structure for invertible multi-scale processing, with a compressed form of output and improved blocks for higher efficiency.
- Our proposed CIDNet outperforms both existing simple INNs of image processing and non-invertible DNNs of video demoiréing, achieving state-of-the-art performance in visual quality, restoration accuracy, and efficiency.

2. Related Work

2.1. Prior-based Image Demoiréing

Traditional methods for demoiréing images rely on predefined assumptions to differentiate moiré patterns from clean images. Sidorov *et al.* [20] have noted that moiré patterns typically correspond to high-frequency components, and therefore they suppress these patterns by a Fourier-based band-pass filter. However, this approach can also filter out texture patterns. To address this issue, Liu *et al.* [12] proposed a variational model that incorporates a low rank prior to regulate texture components and a sparse prior in the discrete cosine transform (DCT) domain to regulate moiré components. Nonetheless, these handcrafted priors are not adaptable to real-world camera-captured screen images, where moiré patterns and natural scene content exhibit significant variations, and hence may not produce satisfactory results.

2.2. DNN-based Image Demoiréing

Most current deep learning-based methods for image demoiréing employ multi-scale processing in DNNs. Early research in this area can be traced back to [11, 22]. Liu *et al.* [11] developed a coarse-to-fine convolutional neural network (CNN) for moiré pattern removal. Sun *et al.* [22] constructed a multi-in-multi-out DNN for efficient processing and created a paired demoiréing dataset of camera-captured screen images for both training and evaluation. He *et al.* [6] employed a U-Net for multi-scale feature aggregation, introducing three additional moiré pattern-related attribute labels for improving performance. Cheng *et al.* [1] constructed a progressive multi-scale residual DNN to learn representations across multiple frequency bands. Yu *et al.* [28] proposed a semantic-aligned scale-aware module that extracts a feature pyramid and combines multi-scale features with an attention mechanism.

Given that moiré patterns exist over a broad frequency range, Liu *et al.* [13] developed a dual-branch DNN that manipulates wavelet coefficients of the input image, and Zheng *et al.* [32] proposed a DNN with a learnable band-pass filter to predict DCT coefficients. To reduce the color-shift effect in demoiréd images, Cheng *et al.* [1] introduced an adaptive instance normalization to amend the feature distribution, and Zheng *et al.* [32] integrated a learnable local/global tone mapping into the DNN. Focusing on processing full-high definition images, He *et al.* [7] released a demoiréing dataset of full-high definition images, while Yu *et al.* [28] extended the scope to ultra-high definition images. Rather than use supervised learning on clean and moiré-degraded image pairs, Liu *et al.* [14] proposed a self-adaptive scheme to learn demoiréing using only focused and defocused pairs.

2.3. Video Demoiréing

Video demoiréing requires the utilization of restoration cues from adjacent frames. Therefore, existing image demoiréing methods are not optimal for video demoiréing. However, despite the importance of video demoiréing, it has received relatively little attention from researchers. One notable exception is the work of Dai *et al.* [2], who proposed a straightforward model that aligns feature spaces and aggregates complementary information from neighboring moiré frames. They also created the first hand-held video demoiréing dataset using a dedicated data collection pipeline. There is another work [17] on arXiv, which learns filters in both frequency and spatial domains to remove moire patterns of various sizes. In contrast, our CIDNet considers information fidelity during video demoiréing, leading to performance gain. Moreover, it has the advantages of compactness and computational efficiency over previous methods.

2.4. Invertible Image Processing

In recent years, INNs have seen success in various low-level vision tasks such as image rescaling [26], real-world denoising [15], super-resolution [10, 25], steganography [5], decolorization [31], and relighting [27, 29]. However, to the best of our knowledge, no previous work has explored the utilization of INNs for demoiréing. The INNs used in most current approaches for image recovery are constructed as either a generative model in normalizing flow, such as in [25], or as a shared-weight auto-encoder with invertibility, as in [9, 15, 26]. Our approach is related to the latter method. Typically, these methods use the forward pass of the INN as an encoder to disentangle layers in a latent code space, then zero out undesired layers by zeroing their latent codes, and finally apply the reverse process as a decoder to obtain the desired image. However, if important image information is mistakenly zeroed out, it cannot be recovered in the backward process. In contrast, our CIDNet does not explicitly decompose the latent codes at intermediate outputs. Instead, it directly maps an input image to two layers, making the architecture more efficient and expressive. Further, we introduce improved INN blocks for enhancement.

3. CIDNet for Video Demoiréing

3.1. Architecture

Our proposed CIDNet is outlined in Fig. 2(a). It takes a triplet of frames as input, including a reference frame $\mathbf{I}^t \in \mathbb{R}^{H\times W\times C}$ and its adjacent frames $\mathbf{I}^{t-1}, \mathbf{I}^{t+1} \in \mathbb{R}^{H\times W\times C}$, and then decomposes them into the desired latent frame $\mathbf{L}^t \in \mathbb{R}^{H\times W\times C}$ and two moiré layers $\mathbf{M}_1^t, \mathbf{M}_2^t \in \mathbb{R}^{H\times W\times C}$:

$$\text{CIDNet}: \{\mathbf{I}^{t-1}, \mathbf{I}^t, \mathbf{I}^{t+1}\} \rightarrow \{\mathbf{L}^t, \mathbf{M}^{t-1}, \mathbf{M}^{t+1}\}. \quad (1)$$

Since image content is usually more redundant than moire patterns in adjacent video frames under certain alignment, we use two moiré components and one frame component in the CIDNet's output.

The CIDNet first aligns input frames in the feature domain by an alignment block (AB). The aligned features are then processed by a U-shaped dyadic multi-scale structure consisting of three Coupling Encoder Blocks (CEBs) and three Coupling Decoder Blocks (CDBs), separated by an Attentive Coupling Layer (ACL). Each CEB sequentially connects a pixel shuffle layer, a Spatial Pyramid Coupling Layer (SPCL), and several standard coupling layers. Each CDB sequentially connects several standard coupling layers and an unshuffle layer. The ACL in the bottleneck is used to enhance the dynamic and spatially-varying processing. Throughout the CIDNet, the information fed from previous blocks is all carried to the next blocks.

Multi-scale processing via dyadic cross-scale coupling The shuffle and unshuffle layers form a multi-scale representation in the CIDNet. Following the scheme used in [19], the shuffle layer reshapes the input feature tensor the spatial resolution is reduced to one-fourth of the ordinal one while the channel number is quadrupled. The unshuffle layer runs an inverse process of the shuffle layer, and thus the two layers can be reverted by each other.

In addition, we propose a dyadic cross-scale coupling structure to enhance multi-scale feature processing. Following the idea of coupling layers and drawing inspirations from wavelets [23], the output of each CEB is split into two parts, where one part is passed to the corresponding CDB via a path with one SPCL and several coupling layers, and the other part is passed to the next CEB. Accordingly, each CDB concatenates the features from the corresponding CEB and the features from its previous layer as the input. Indeed, each CEB and CDB can be viewed a larger coupling layer nested by some small coupling layers. Such a design empowers the CIDNet with better multi-scale analysis capability while maintaining its invertibility and perfect information fidelity.

3.2. Modules

Alignment Block The AB is implemented by the pyramid cascading deformable module (PCDM) [2, 24] trained together with other blocks. The PCDM performs alignment by a coarse-to-fine spatial pyramid structure, which transfers and fuses the feature maps aligned at different levels to the ones of higher resolution, generating an implicit motion compensation structure from rough to fine scales.

Moire patterns usually move with the image content in video frames, but their movements are not consistent with that of image contents, as illustrated in Fig. 1. We observed that the learned AB tends to align the frames based on image content instead of based on moire patterns, as moire patterns are usually much weaker than the image content. As a result, the AB further increases the redundancy of image features over the moiré patterns, resulting in a better representation for discriminating image patterns from moiré patterns for subsequent decomposition. See our supplement for details.

Standard Coupling Layers See Fig. 2(b) for the diagram of a coupling layer [3], which employs a double-branch structure so that its input can be simply reconstructed back from its output using its inverse mode. This invertibility allows perfect information fidelity during feature processing. Given input $\mathbf{X}$, the coupling layer first divides it into two parts $\mathbf{X}_1$ and $\mathbf{X}_2$ along the channel dimension. Then the two parts are transformed and interacted with each other by functions $\phi_1, \phi_2, \psi_1, \psi_2$ in a coupling manner, resulting in $\mathbf{Y}_1, \mathbf{Y}_2$ which are concatenated as $\mathbf{Y}$ for output. These steps form the forward process and can be expressed as

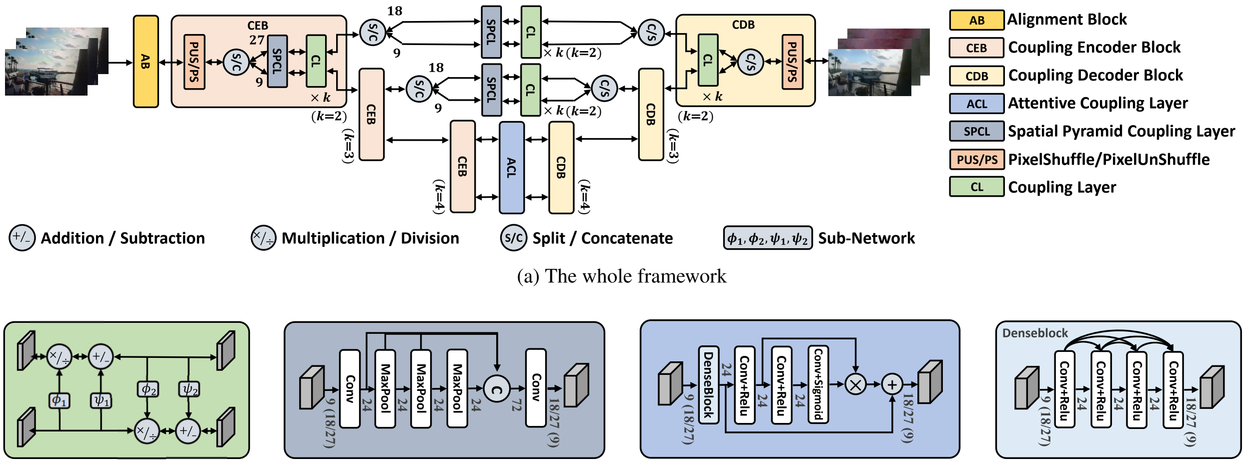
$$\mathbf{Y}_1 = \mathbf{X}_1 \odot \exp(\phi_1(\mathbf{X}_2)) + \psi_1(\mathbf{X}_2), \quad (2)$$
$$\mathbf{Y}_2 = \mathbf{X}_2 \odot \exp(\phi_2(\mathbf{Y}_1)) + \psi_2(\mathbf{Y}_1), \quad (3)$$

where $\odot$ denotes entry-wise product, $\phi_1, \phi_2, \psi_1, \psi_2$ are interactive functions that can be implemented by arbitrary DNN blocks, without damaging the invertibility of the coupling layer. The inverse process can then be simply done by

$$\mathbf{X}_2 = (\mathbf{Y}_2 - \psi_2(\mathbf{Y}_1)) \oslash \exp(\phi_2(\mathbf{Y}_1)), \quad (4)$$
$$\mathbf{X}_1 = (\mathbf{Y}_1 - \psi_1(\mathbf{X}_2)) \oslash \exp(\phi_1(\mathbf{X}_2)), \quad (5)$$

where $\oslash$ denotes entry-wise division. See Fig. 2(e) for the definitions of $\phi_1, \phi_2, \psi_1, \psi_2$ in the standard coupling layers used in our CIDNet.

Spatial Pyramid Coupling Layers To further enhance multi-scale feature extraction in CIDNet, the SPCL is constructed by defining the modules ϕ_1, ϕ_2, ψ_1 and ψ_2 of coupling layers using the spatial pyramid pooling [8], which


AB
CEB
27
9
SPCL
CL
PUS/PS
S/C
× k
(k=2)
18
9
C/S
× k (k=2)
CDB
(k=3)
ACL
(k=4)
AB Alignment Block
CEB Coupling Encoder Block
CDB Coupling Decoder Block
ACL Attentive Coupling Layer
SPCL Spatial Pyramid Coupling Layer
PUS/PS PixelShuffle/PixelUnShuffle
CL Coupling Layer
+/− Addition / Subtraction
×/÷ Multiplication / Division
S/C Split / Concatenate
$\phi_1, \phi_2, \psi_1, \psi_2$ Sub-Network
Conv
MaxPool
24
72
9 (18/27)
18/27 (9)
DenseBlock
Conv+Relu
Conv+Sigmoid
Denseblock

(a) The whole framework

(b) CL/ACL/SPCL

(c) ϕ_1, ϕ_2, ψ_1, ψ_2 in SPCL

(d) ϕ_1, ϕ_2, ψ_1, ψ_2 in ACL

(e) ϕ_1, ϕ_2, ψ_1, ψ_2 in CL

Figure 2: Architecture of proposed CIDNet. Unless specified, the spatial kernel sizes in all convolutional layers are 3×3.

captures multi-scale patterns with varying-size pooling. In detail, the ϕ_1, ϕ_2, ψ_1 and ψ_2 in SPCL shares the same SPP structure, which includes three 5×5 max pooling layers. The outputs of three max-pooling layers are upsampled and concatenated with the original feature map, generating a feature tensor that includes features extracted from receptive fields of different sizes. See Fig. 2(c) for an illustration. We apply 1×1 convolution before the first pooling and after concatenation for feature adjustment and fusion, respectively.

Attentive Coupling Layers Attention mechanism is introduced to improve the efficiency of distinguishing moiré pattern from image patterns in the latent feature space, which is implemented by inserting an ACL at the bottleneck of CIDNet. An ACL is a coupling layer with the modules ϕ_1, ϕ_2, ψ_1 and ψ_2 sharing the same structure. The structure is shown in Fig. 2(d). A dense block, with the structure illustrated in Fig. 2(e), is used to generate a feature tensor $\mathbf{F}$, then three consecutive convolutional layers, attached with two ReLUs and one Sigmoid respectively, are used to generate an attention map from $\mathbf{A}$, and finally the attentive feature tensor $\tilde{\mathbf{F}} = \mathbf{F} \odot \mathbf{A}$ is output.

3.3. Training Loss

The loss ℓ_{total} for training CIDNet consists of a content loss ℓ_{c}^t, a moiré loss ℓ_{m}^t, and a reverse loss ℓ_{r}^t:

$$\ell_{\text{total}} = \sum_t \ell_{\text{c}}^t + \alpha \ell_{\text{m}}^t + \beta \ell_{\text{r}}^t, \tag{6}$$

where the weights α, β are both set to 1 for simplicity.

The content loss ℓ_{c}^t, which measures the reconstruction accuracy of the latent clean frame $\mathbf{X}^t$, is defined by an ℓ_1 loss and a perceptual loss as follows:

$$\ell_{\text{c}}^t = \|\mathbf{L}^t - \mathbf{X}^t\|_1 + \lambda \sum_j \|\Phi_j(\mathbf{L}^t) - \Phi_j(\mathbf{X}^t)\|_2^2, \tag{7}$$

where $\Phi_j(\cdot)$ denotes the output of some VGG-16 layer for defining the perceptual loss [21] with the weight $\lambda \in \mathbb{R}^+$.

Similarly, the moiré loss $\ell_{\text{moiré}}^t$ is defined to measure the reconstruction accuracy on moiré components. Considering that the movements and changes of moiré patterns are usually slight in the aligned frames and features, we supervise $\mathbf{M}_1^t, \mathbf{M}_2^t$ by the estimated moiré layers of adjacent frames which can be represented in a residual form $\widetilde{\mathbf{M}_1}^t = \mathbf{X}^{t-1} - \mathbf{I}^{t-1}, \widetilde{\mathbf{M}_2}^t = \mathbf{X}^{t+1} - \mathbf{I}^{t+1}$. Then the moiré loss is given by

$$\ell_{\text{m}}^t = \sum_{i=1,2} \|\mathbf{M}^t - \widetilde{\mathbf{M}}_i^t\|_1 + \gamma \sum_{i=1,2} \sum_j \|\Phi_j(\mathbf{M}^t) - \Phi_j(\widetilde{\mathbf{M}}_i^t)\|_2^2, \tag{8}$$

with the weight $\gamma \in \mathbb{R}^+$ balancing the two terms.

The reverse loss is defined to utilize the invertibility of CIDNet for further regularization. By feeding $\mathbf{X}^t$ and $\mathbf{M}_1^t, \mathbf{M}_2^t$ into the reverse mode of CIDNet, we can get the estimated aligned features for the consecutive degraded frames, which should be similar to the aligned features in the forward pass. We define the reverse loss as

$$\ell_{\text{r}} = \sum_{i \in \{\pm 1, 0\}} \|\mathbf{F}^{t+i} - \bar{\mathbf{F}}^{t+i}\|_2^2, \tag{9}$$

where $\mathbf{F}^{t-1}, \mathbf{F}^t, \mathbf{F}^{t+1}$ denote the features extracted by AB in forward pass for the consecutive frames, and $\bar{\mathbf{F}}^{t-1}, \bar{\mathbf{F}}^t, \bar{\mathbf{F}}^{t+1}$ denote the corresponding estimates in the backward pass.

4. Experiments

4.1. Experimental Settings

Dataset and metrics We use the Video Demoiréing Dataset (VDD) [2] for performance evaluation, which includes 247 videos for training and 43 videos for test, with 60 frames for each captured video. The dataset is obtained from handheld cameras photographing source videos on a screen, with content covering various scenarios such as human beings, landscapes, texts, sports, and animals. The dataset adopts two device sets: Huipu v270 monitor with TCL20 Pro mobile phone and MacBook Pro with iPhone XR, and both are used for a comprehensive evaluation. For the quantitative metrics, we adopt Learned Perceptual Image Patch Similarity (LPIPS) [30], Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM).

Methods for comparison Five methods are selected for experimental comparison, including (a) three DNN-based image demoiréing methods: MBCNN [32], DMCNN [22] and ESDNet [28]; (b) an INN-based denoising method InvDN [15]; and (c) a recent video demoiréing method: VDN [2]. The DNN models of the first four methods are retrained using the same training data as ours, with their hyperparameters adjusted accordingly. The results of VDN are directly quoted from its literature [2].

Implementation details In the training of CIDNet, all input video frames are randomly cropped into 256×256 patches, with random flipping for data augmentation. Regarding the training loss, the weights γ and λ are both set to 0.5. All model weights are initialized using Xavier [4]. We utilize Adam optimizer with a cosine learning rate [16] and a batch size of 8. The CIDNet is implemented in PyTorch and executed on an NVIDIA GeForce RTX 3090 GPU.

4.2. Performance Comparison

Quantitative comparison The quantitative results are listed in Table 1. In both settings, our proposed CIDNet achieves the best results in terms of PSNR and LPIPS, with a noticeable improvement over the second-best performers. The CIDNet not only outperforms the methods designed for image demoiréing, but also performs noticeably better than VDN, a recent video demoiréing method. Specifically, the CIDNet improves the LPIPS value by 0.2 and gains a PSNR increment of approximately 0.2dB over the second-best performer ESDNet on TCL20 Pro and 0.37dB over VDN on IPhone XR. As for SSIM, it obtains the best result on IPhone XR and the second-best result on TCL20 Pro. Moreover, the CIDNet shows a significant improvement over InvDN, another INN trained for video demoiréing, which indicates the superiority of our proposed invertible architecture over the standard INNs for video demoiréing. All these results have demonstrated the effectiveness of our CIDNet.

Qualitative comparison Visual inspection on some selected demoiréd frames produced by different methods is given in Fig. 3. Apparently, the results produced by the CIDNet are more perceptually satisfactory. For instance, in the zoomed-in region of the first sample, two image-based methods MBCNN and DMCNN mistakenly removes the stripe-like texture on the pillar, while VDN generated noticeable artifacts. In comparison, our CIDNet can well preserve the texture on the pillar without introducing noticeable undesired artifacts. See also the 2nd sample, where CIDNet also preserves more spotted patterns on the clothes. In the third sample, all the other compared methods output oversmoothed results, and some also suffer from the color-shift problem. In contrast, our CIDNet outputs sharp frames without color shift. See also the 4th sample, where CIDNet is also superior in alleviating the color shift issue. To conclude, CIDNet is better than other compared methods at removing moiré patterns with varying shapes, sizes, and densities, while preserving image structures and handling the color shift problem well.

Table 1: Quantitative comparison in terms of LPIPS, SSIM and PSNR(dB). RED: best; BLUE: second-best.

Method	Source	TCL20 Pro			IPhone XR		
		LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
DMCNN	TIP18	0.321	20.32	0.703	0.218	21.63	0.701
MBCNN	CVPR20	0.260	21.53	0.740	0.201	21.93	0.713
InvDN	CVPR21	0.199	21.49	0.715	0.207	21.98	0.712
ESDNet	ECCV22	0.199	22.03	0.734	0.207	22.27	0.715
VDN-S	CVPR22	0.212	21.77	0.729	0.217	22.04	0.710
VDN	CVPR22	0.202	21.73	0.733	0.206	22.21	0.715
CIDNet	Ours	0.184	22.27	0.735	0.197	22.40	0.717

Table 2: Complexity comparison of different models.

Metric	MBCNN	DMCNN	InvDN	ESDNet	VDN	CIDNet
#Params(M)	14.21	0.66	5.80	5.96	5.82	4.57
#FLOPs(G)	66.99	11.49	86.91	18.06	28.71	28.10

Complexity comparison Table 2 compares the model size (*i.e.* number of parameters) of different methods. Our CIDNet has a smaller model size than the video demoiréing method VDN and most other image-based methods such as MBCNN and ESDNet. This implies that the CIDNet consumes less memory during inference, as well as indicates that the performance gain of CIDNet is not from enlarging the model but from the architecture design. To evaluate the efficiency, we measure and compare the number of floating points of operations (FLOPs) of different models. Table 2 shows the FLOPs required to process a 60-frame video sequence in size of 1280×720. Our CIDNet requires fewer FLOPs than VDN. Regarding the comparison to the image-based methods, CIDNet requires much fewer FLOPs than MBCNN, but a little more than DMCNN and EDSNet. Note that DMCNN and EDSNet process images while our CIDNet process multiple frames. In summary, the CIDNet exhibits relatively low model complexity and low computational cost.

4.3. Analysis and Discussion

Ablation Study To analyze the contribution of each key component of our approach, we form several baseline mod-

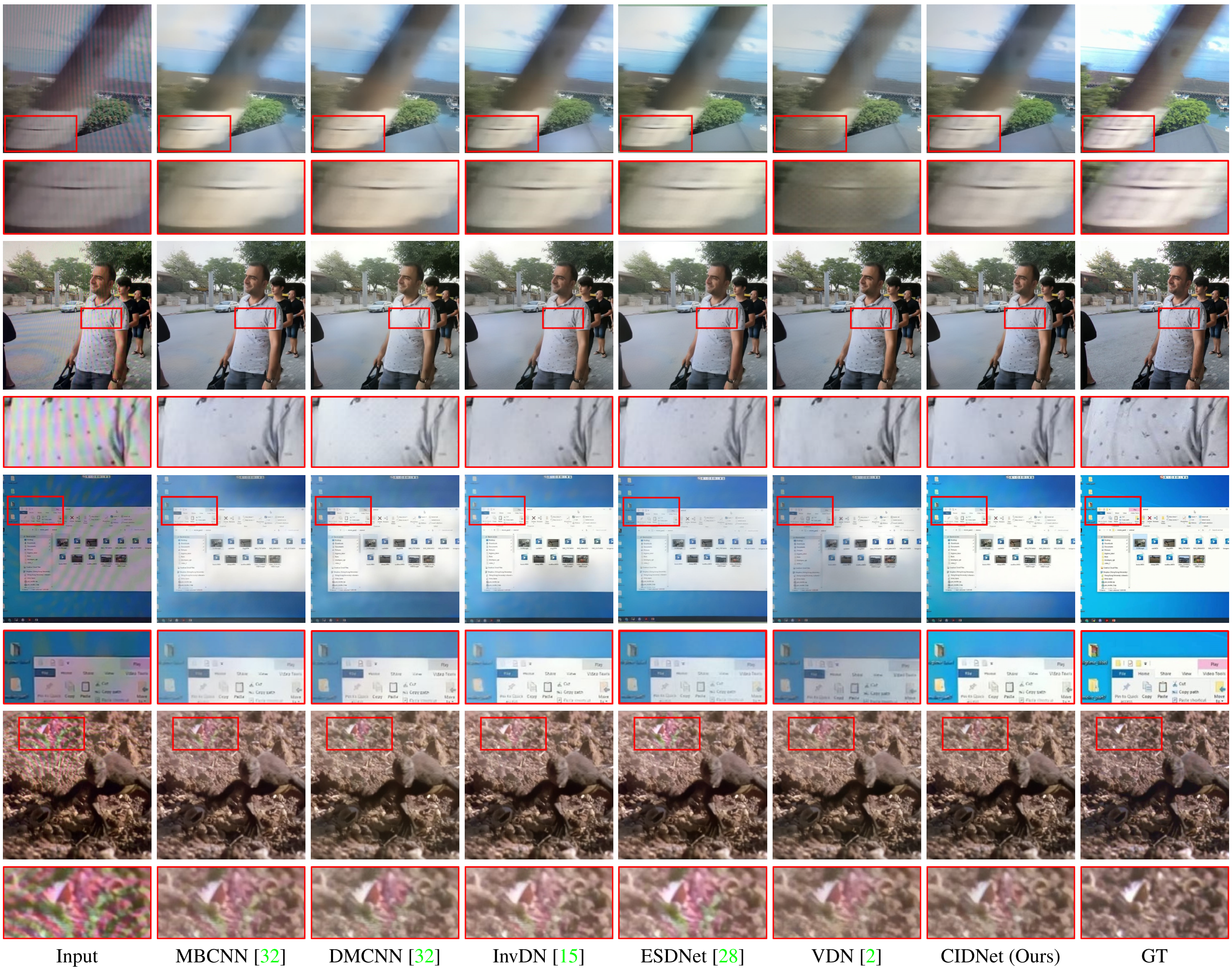


Figure 3: Visual comparison of some demoiréd frames from different methods.

Table 3: Ablation study on the key components and losses in terms of LPIPS, SSIM and PSNR(dB).

Method	w/o AB	w/o ACL	w/o SPCL	w/o MSS	w/o ℓ_{m}	w/o ℓ_{r}	Original
PSNR↑	21.73	21.96	21.86	21.43	21.96	22.20	**22.27**
SSIM↑	0.724	0.730	0.726	0.721	0.732	0.735	**0.735**
LPIPS↓	0.199	0.187	0.196	0.203	0.185	0.185	**0.184**

els. See Table 3 for the results, where "w/o X" means a simplified CIDNet model without the use of the X component or loss. Specifically, for baseline 'w/o ACL', we replace the ACL module with two coupling layers to have a similar model size as the original one for a fair comparison. For 'w/o MSS', the multi-scale structure of the coupling pipeline is replaced with several standard coupling blocks. For fairness, the number of coupling blocks is set to maintain the same number of parameters as the original CIDNet.

We can see that the baselines 'w/o AB', 'w/o ACL', 'w/o SPCL', and 'w/o MSS' all exhibit noticeable performance desegregation compared to the original CIDNet model, demonstrating the effectiveness of each key component in the CIDNet and our proposed multi-scale architecture. In addition, the noticeable improvement from using the moiré loss validates the effectiveness of using adjacent moiré patterns for supervision, and removing the reverse loss ℓ_{r} also leads to a slight performance drop. Since the reverse loss does not introduce additional parameters and has no impact on the inference cost, we keep it in CIDNet for its benefits. That the improvement from ℓ_{r} looks small is probably due to already existing regularization given by the invertible structure of CIDNet that encodes some underlying physics of the formation of moiré pattern-degraded images.

Effectiveness of dyadic multi-scale coupling design While the effectiveness of our proposed dyadic multi-scale structure in CIDNet has been demonstrated by the ablation study, we further analyze it via the comparison to two baselines: (a)

Table 4: Results using different multi-scale architectures.

Architecture	LPIPS↓	PSNR(dB)↑	SSIM↑
Wavelet + INN [15]	0.214	21.23	0.723
INN with U-Nets	0.213	21.09	0.720
CIDNet	**0.184**	**22.27**	**0.735**

"Wavelet+INN": the model proposed by [15], which introduces the wavelet transform and its inversion before and after several usual coupling layers; and (b) "INN with U-Nets": a model with similar size to CIDNet, constructed using several coupling layers with each free-form function defined by a U-Net for multi-scale analysis; see our supplement for details. The results are reported in Table 4, where the CIDNet outperforms both baselines noticeably. These results again demonstrate the effectiveness of our multi-scale design.

Effectiveness of compact decomposition Our CIDNet introduces a compressed form of its output which reduces the model size and also brings some regularization effect. To verify these benefits, we construct two baselines: (a) Repetition-4: three aligned consecutive frames together with one copied current frame are taken as input. (b) Repetition-6: three aligned consecutive frames together with three copies of the current frame are taken as input. The larger model of Repetition-4 produces almost the same result as CIDNet, but with nearly 1M additional parameters and 7G additional FLOPs. As for Repetition-6, it performs slightly better than the original model but nearly doubles the number of parameters and FLOPs. Such results demonstrate the effectiveness of the compact representation used in CIDNet.

Table 5: Results using different input/output settings.

Setting	LPIPS↓	PSNR(dB)↑	SSIM↑	#Params(M)	#FLOPs(G)
Repetition-4	0.184	22.26	0.736	5.66	34.87
Repetition-6	0.177	22.42	0.740	8.14	50.46
Original	0.184	22.27	0.735	4.57	28.10

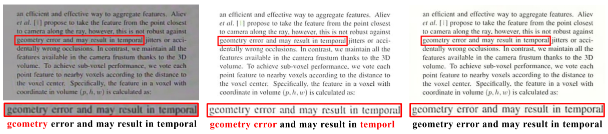
Effectiveness of one-way scheme Different from existing INNs for image recovery that use a two-way INN scheme discussed in Section 2.4, our CIDNet adopts a one-way scheme. To see the performance comparison of two-way vs. one-way schemes, we construct a baseline "two-way CIDNet" which shares the same structure with CIDNet but runs in a two-way scheme; see our supplement for more details. The quantitative results are shown in Table 6, where the two-way scheme gets a PSNR drop of 0.6dB in comparison.

Table 6: Results produced by different schemes.

Scheme	LPIPS↓	PSNR(dB)↑	SSIM↑	#FLOPs(G)
Two-way	0.201	21.54	0.724	56.20
One-way	**0.184**	**22.27**	**0.735**	**28.10**

Table 7: Recognition accuracy (%) on demoiréd frames.

Input	MBCNN	DMCNN	ESDNet	VDN	CIDNet
82.9%	83.7%	81.3%	84.6%	75.4%	**87.5%**



an efficient and effective way to aggregate features. Aliev *et al.* [1] propose to take the feature from the point closest to camera along the ray, however, this is not robust against geometry error and may result in temporal jitters or accidentally wrong occlusions. In contrast, we maintain all the features available in the camera frustum thanks to the 3D volume. To achieve sub-voxel performance, we vote each point feature to nearby voxels according to the distance to the voxel center. Specifically, the feature in a voxel with coordinate in volume (p, h, w) is calculated as:

geometry error and may result in temporal

gcometry error and may result in temporal

an efficient and effective way to aggregate features. Aliev *et al.* [1] propose to take the feature from the point closest to camera along the ray, however, this is not robust against geometry error and may result in temporal jitters or accidentally wrong occlusions. In contrast, we maintain all the features available in the camera frustum thanks to the 3D volume. To achieve sub-voxel performance, we vote each point feature to nearby voxels according to the distance to the voxel center. Specifically, the feature in a voxel with coordinate in volume (p, h, w) is calculated as:

gcometry error and may result in temporal

gcometry crror and may result in temporl

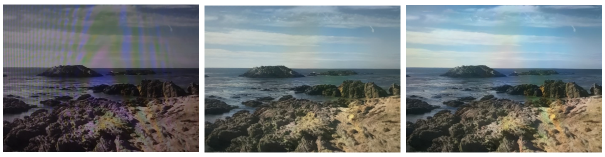
an efficient and effective way to aggregate features. Aliev *et al.* [1] propose to take the feature from the point closest to camera along the ray, however, this is not robust against geometry error and may result in temporal jitters or accidentally wrong occlusions. In contrast, we maintain all the features available in the camera frustum thanks to the 3D volume. To achieve sub-voxel performance, we vote each point feature to nearby voxels according to the distance to the voxel center. Specifically, the feature in a voxel with coordinate in volume (p, h, w) is calculated as:

geometry error and may result in temporal

geometry error and may result in temporal

Degraded frame VDN [2] CIDNet (Ours)

Figure 4: An example of text recognition. Bottom: recognized texts, with erroneous words marked in RED.



Degraded frame VDN [2] CIDNet (Ours)

Figure 5: A failure case of VDN and our CIDNet.

Evaluation on downstream tasks To further verify the practicability of our CIDNet for subsequent downstream tasks, we select some text images in the test set and use CRNN [18] for text recognition. The recognition accuracy results are listed in Table 7. The resulting demoiréd frames achieve nearly 5% improvement in accuracy over the degraded ones, which demonstrates the benefits of CIDNet in text recognition under degradation of moiré patterns. Moreover, among the demoiríng methods, CIDNet obtains the highest accuracy, which outperforms the second-best performer with nearly 3%. See also Fig. 4 for a visual example.

Challenges Our CIDNet may not well handle moiré patterns with irregular clumps, particularly when the moire patterns have similar colors with the background and the moire-degraded area is large with relatively smooth edges. See Fig. 5 and also our supplement for some examples. This is also a common challenge for other methods.

5. Conclusion

This work proposed CIDNet, a DNN with an efficient multi-scale invertible structure for video demoirêing. By incorporating a U-shaped dyadic coupling structure with improved coupling blocks, CIDNet is capable of effectively exploiting multi-scale analysis to distinguish image patterns from moirê patterns, while leveraging information fidelity for preserving image details. Our experimental results demonstrated that the advantages of CIDNet in terms of restoration quality, model complexity and computational efficiency. While performing better than existing ones, our method cannot handle some challenging cases. Improvement along this line will be our future work.

References

[1] Xi Cheng, Zhenyong Fu, and Jian Yang. Multi-scale dynamic feature encoding network for image demoiréing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3486–3493. IEEE, 2019. 1, 2, 3

[2] Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proccedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17622–17631, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[3] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2, 4

[4] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, Mar. 2010. 6

[5] Zhenyu Guan, Junpeng Jing, Xin Deng, Mai Xu, Lai Jiang, Zhou Zhang, and Yipeng Li. DeepMIH: Deep invertible network for multiple image hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[6] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Mop moire patterns using mopnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2424–2432, 2019. 1, 2, 3

[7] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. FHDe2Net: Full High Definition Demoireing Network. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of the European Conference on Computer Vision*, pages 713–729, Cham, 2020. Springer International Publishing. 1, 2, 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. 4

[9] Jiachun Li, Kunkun Qin, Ruotao Xu, and Hui Ji. Deep scale-aware image smoothing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2105–2109. IEEE, 2022. 3

[10] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4076–4085, 2021. 3

[11] Bolin Liu, Xiao Shu, and Xiaolin Wu. Demoir\'eing of Camera-Captured Screen Images Using Deep Convolutional Neural Network. *arXiv preprint arXiv.1804.03809*, Apr. 2018. 1, 2, 3

[12] Fanglei Liu, Jingyu Yang, and Huanjing Yue. Moiré pattern removal from texture images via low-rank and sparse matrix decomposition. In *Visual Communications and Image Processing*, pages 1–4. IEEE, 2015. 1, 3

[13] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Aleš Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoiréing. In *Proceedings of the European Conference on Computer Vision*, pages 86–102. Springer, 2020. 1, 2, 3

[14] Lin Liu, Shanxin Yuan, Jianzhuang Liu, Liping Bao, Gregory Slabaugh, and Qi Tian. Self-adaptively learning to demoiré from focused and defocused image pairs. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 22282–22292, 2020. 1, 2, 3

[15] Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. Invertible denoising network: A light solution for real noise removal. In *Proccedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13365–13374, 2021. 3, 6, 7, 8

[16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[17] Gyeongrok Oh, Heon Gu, Sangpil Kim, and Jinkyu Kim. Fpanet: Frequency-based video demoireing using frame-level post alignment. *arXiv preprint arXiv:2301.07330*, 2023. 3

[18] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2016. 8

[19] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proccedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 4

[20] Denis N Sidorov and Anil Christopher Kokaram. Suppression of moiré patterns via spectral analysis. In *Visual Communications and Image Processing*, volume 4671, pages 895–906. SPIE, 2002. 1, 3

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the IEEE International Conference on Learning Representations*, 2015. 5

[22] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *IEEE Transactions on Image Processing*, 27(8):4160–4172, 2018. 1, 2, 3, 6

[23] C. G. Torrence and G. P. Compo. A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998. 4

[24] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proccedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 4

[25] Jay Whang, Erik Lindgren, and Alex Dimakis. Composing normalizing flows for inverse problems. In *Proceedings of the International Conference on Machine Learning*, pages 11158–11169. PMLR, 2021. 3

[26] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *Proceedings of the European Conference on Computer Vision*, pages 126–144. Springer, 2020. 3

[27] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proccedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6287–6296, 2021. 3

[28] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards Efficient and Scale-Robust Ultra-High-Definition Image Demoiréing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Proceedings of the European Conference on Computer Vision*, pages 646–662, Cham, 2022. Springer Nature Switzerland. 1, 2, 3, 6, 7

[29] Jize Zhang, Haolin Wang, Xiaohe Wu, and Wangmeng Zuo. Invertible network for unpaired low-light image enhancement. *arXiv preprint arXiv:2112.13107*, 2021. 3

[30] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proccedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[31] Rui Zhao, Tianshan Liu, Jun Xiao, Daniel PK Lun, and Kin-Man Lam. Invertible image decolorization. *IEEE Transactions on Image Processing*, 30:6081–6095, 2021. 3

[32] Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, and Ales Leonardis. Image demoireing with learnable bandpass filters. In *Proccedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3636–3645, 2020. 1, 2, 3, 6, 7