Deep Video Demoiréing via Compact Invertible Dyadic Decomposition

Yuhui Quan\textsuperscript{1,2}, Haoran Huang\textsuperscript{1}, Shengfeng He\textsuperscript{3}, Ruotao Xu\textsuperscript{1,2}*

\textsuperscript{1}School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
\textsuperscript{2}Pazhou Lab, Guangzhou 510335, China
\textsuperscript{3}School of Computing and Information Systems, Singapore Management University, 188065, Singapore
cyhquan@scut.edu.cn, caharry@mail.scut.edu.cn, shengfenghe@amul.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 I corrupted by moiré patterns is composed of the underlying latent image X of a scene and the corresponding layer M of moiré patterns. The latter contains all intensity and color fluctuations caused by the moiré patterns [12]. Recovering a set of clean frames \( \{ X' \} \), from only the input moiré-corrupted frames \( \{ I' \} \), 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

\textsuperscript{*}Corresponding author: Ruotao Xu (ruotao.xu@outlook.com).
\textsuperscript{1}This work is partially supported by the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012341, 2022A1515011355, 2022A1515011087), National Natural Science Foundation of China (Grant No. 62106077, 61971262), 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. 2022M662705); and Singapore Ministry of Education Academic Research Fund Tier 1 (Grant No. MSS23C002).
Introducing multiscale analysis into DNs via a dyadic coupling structure. As small image patterns and image patterns very significantly over a wide range of scales [1, 23].

Efficient decomposition with perfect information fidelity. Previous studies, such as [12, 13, 14, 15, 16, 17] have shown that using an encoding-decoding network (e.g., a DNN) can effectively exploit the spatial frequency characteristics of the images. However, typical DNNs like T-Net [6] can only achieve limited improvements in reconstruction quality.

We address this issue by introducing a dyadic coupling strategy. Specifically, each encoder layer and decoder layer are designed to operate in a series of invertible blocks. This approach enables parallel processing of multi-scale features, resulting in higher performance and efficiency.

Overlapping reduction using a compact DNN. An DNN reduces overlapping by adapting to the specific characteristics of the image. This allows for more efficient use of resources and improved performance in terms of reconstruction quality.

Figure 1: Three consecutive frames with multi-patterns. 1.1 Motivation and Main Idea. For dealing computationally efficient video denoising methods that can provide multi-pattern removal and detail preservation, which is also the aim of the paper. This covers a multi-pattern removal and detail preservation.
over short time intervals after alignment operations. Therefore, we reduce the total number of layers $M'$ 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 pre-defined 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


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 demoiréed 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 moiré 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 resizing [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 $I' \in \mathbb{R}^{H \times W \times C}$ and its adjacent frames $I'^{-1}, I'^{+1} \in \mathbb{R}^{H \times W \times C}$, and then decomposes them into the desired latent frame $L' \in \mathbb{R}^{H \times W \times C}$ and two moiré layers $M'_1, M'_2 \in \mathbb{R}^{H \times W \times C}$:

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

Since image content is usually more redundant than moiré 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 [9], the shuffle layer reshapes the input feature tensor so that 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.

Moiré 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 moiré patterns, as moiré 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 $X$, the coupling layer first divides it into two parts $X_1$ and $X_2$ 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 $Y_1, Y_2$ which are concatenated as $Y$ for output. These steps form the forward process and can be expressed as

$$Y_1 = X_1 \odot \exp(\phi_1(X_1)) + \psi_1(X_1), \quad (2)$$
$$Y_2 = X_2 \odot \exp(\phi_2(Y_1)) + \psi_2(Y_1), \quad (3)$$

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

$$X_2 = (Y_2 - \psi_2(Y_1)) \odot \exp(\phi_2(Y_1)),$$
$$X_1 = (Y_1 - \psi_1(X_2)) \odot \exp(\phi_1(X_2)),$$

where $\odot$ 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 $\phi_1, \phi_2, \psi_1$ and $\psi_2$ of coupling layers using the spatial pyramid pooling [8], which
captures multi-scale patterns with varying-size pooling. In detail, the $\phi_1$, $\phi_2$, $\psi_1$ and $\psi_2$ in SPCL shares the same SPP structure, which includes three $5 \times 5$ 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 $1 \times 1$ 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 $\phi_1$, $\phi_2$, $\psi_1$ and $\psi_2$ 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 $F$, then three consecutive convolutional layers, attached with two ReLUs and one Sigmoid respectively, are used to generate an attention map from $A$, and finally the attentive feature tensor $F = F \odot A$ is output.

### 3.3. Training Loss

The loss $\ell_{\text{total}}$ for training CIDNet consists of a content loss $\ell_c$, a moiré loss $\ell_m$, and a reverse loss $\ell_r$:

$$\ell_{\text{total}} = \sum_t \ell_c^t + \alpha \ell_m^t + \beta \ell_r,$$

where the weights $\alpha$, $\beta$ are both set to 1 for simplicity.

The content loss $\ell_c$, which measures the reconstruction accuracy of the latent clean frame $X^t$, is defined by an $\ell_1$ loss and a perceptual loss as follows:

$$\ell_c = ||L^t - X^t||_1 + \lambda \sum_j ||\Phi_j(L^t) - \Phi_j(X^t)||^2_2,$$

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 $\ell_m$ 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 $M_1^t, M_2^t$ by the estimated moiré layers of adjacent frames which can be represented in a residual form $\tilde{M}_1^t = X^{t-1} - I^{-1}, \tilde{M}_2^t = X^{t+1} - I^{t+1}$. Then the moiré loss is given by

$$\ell_m = \sum_{i=1,2} ||M_i^t - \tilde{M}_i^t||_1 + \gamma \sum_{i=1,2} \sum_j ||\Phi_j(M_i^t) - \Phi_j(\tilde{M}_i^t)||^2_2,$$

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 $X^t$ and $M_1^t, M_2^t$, 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

$$\ell_r = \sum_{i \in \{+1, -1\}} ||F^{t+i} - F^{t+i}||^2_2,$$

where $F^{t-1}, F^t, F^{t+1}$ denote the features extracted by AB in forward pass for the consecutive frames, and $F^{t-1}, F^t, F^{t+1}$ 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 45 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 demoiréng methods: MBCNN [32], DMCNN [22] and ESDNet [28]; (b) an INN-based denoising method InVDN [15]; and (c) a recent video demoiréng 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 demoiréng, but also performs noticeably better than VDN, a recent video demoiréng method. Specifically, the CIDNet improves the LPIPS value by 0.2 and gains a PSNR increment of approximately 0.2 dB over the second-best performer ESDNet on TCL20 Pro and 0.37 dB 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 demoiréng, which indicates the superiority of our proposed invertible architecture over the standard INNs for video demoiréng. All these results have demonstrated the effectiveness of our CIDNet.

Qualitative comparison Visual inspection on some selected demoiréed 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 over-smoothed 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.

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 demoiréng 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 point 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 MBCNN 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 component of our approach, we form several baseline mod-
Table 3: Ablation study on the key components and losses in terms of LPIPS, SSIM and PSNR (dB).

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o AB</th>
<th>w/o ACL</th>
<th>w/o SPCL</th>
<th>w/o MSS</th>
<th>w/o $\ell_w$</th>
<th>w/o $\ell_t$</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR↑</td>
<td>21.73</td>
<td>21.96</td>
<td>21.86</td>
<td>21.93</td>
<td>22.20</td>
<td>22.27</td>
<td></td>
</tr>
<tr>
<td>SSIM↑</td>
<td>0.724</td>
<td>0.730</td>
<td>0.726</td>
<td>0.721</td>
<td>0.732</td>
<td>0.735</td>
<td>0.735</td>
</tr>
<tr>
<td>LPIPS↓</td>
<td>0.199</td>
<td>0.187</td>
<td>0.196</td>
<td>0.205</td>
<td>0.185</td>
<td>0.185</td>
<td>0.184</td>
</tr>
</tbody>
</table>

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

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 $\ell_t$ 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 $\ell_t$ 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)
'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.

<table>
<thead>
<tr>
<th>Setting</th>
<th>LPIPS↓</th>
<th>PSNR (dB)↑</th>
<th>SSIM↑</th>
<th>#Params(M)</th>
<th>#FLOPs(G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repetition-4</td>
<td>0.184</td>
<td>22.26</td>
<td>0.736</td>
<td>5.66</td>
<td>34.87</td>
</tr>
<tr>
<td>Repetition-6</td>
<td>0.177</td>
<td>22.42</td>
<td>0.740</td>
<td>8.14</td>
<td>50.46</td>
</tr>
<tr>
<td>Original</td>
<td>0.184</td>
<td>22.27</td>
<td>0.735</td>
<td>4.57</td>
<td>28.10</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>LPIPS↓</th>
<th>PSNR (dB)↑</th>
<th>SSIM↑</th>
<th>#FLOPs(G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-way</td>
<td>0.201</td>
<td>21.54</td>
<td>0.724</td>
<td>56.20</td>
</tr>
<tr>
<td>One-way</td>
<td>0.184</td>
<td>22.27</td>
<td>0.735</td>
<td>28.10</td>
</tr>
</tbody>
</table>

**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 demoiré 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 demoiré 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 moiré patterns have similar colors with the background and the moiré-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 demoiré. 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 moiré 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


