Video Demoiréing with Relation-Based Temporal Consistency

Peng Dai[1]   Xin Yu[1]   Lan Ma[2*]   Baoheng Zhang[1]   Jia Li[3]   Wenbo Li[4]   Jiajun Shen[2]   Xiaojuan Qi[1*]

[1]The University of Hong Kong   [2]TCL AI Lab

[3]Sun Yat-sen University   [4]The Chinese University of Hong Kong

Abstract

Moiré patterns, appearing as color distortions, severely degrade image and video qualities when filming a screen with digital cameras. Considering the increasing demands for capturing videos, we study how to remove such undesirable moiré patterns in videos, namely video demoiréing. To this end, we introduce the first hand-held video demoiréing dataset with a dedicated data collection pipeline to ensure spatial and temporal alignments of captured data. Further, a baseline video demoiréing model with implicit feature space alignment and selective feature aggregation is developed to leverage complementary information from nearby frames to improve frame-level video demoiréing. More importantly, we propose a relation-based temporal consistency loss to encourage the model to learn temporal consistency priors directly from ground-truth reference videos, which facilitates producing temporally consistent predictions and effectively maintains frame-level qualities. Extensive experiments manifest the superiority of our model. Code is available at `https://daipengwa.github.io/VDmoire_ProjectPage/`.

1. Introduction

Video is an important source of entertainment, information recording and dissemination through social media. When photographing a video on a screen, frequency aliasing leads to moiré patterns (Fig. 1) which appear as colored stripes, severely degrading the visual quality and fidelity of captured contents. Although many research efforts have been made to remove such moiré patterns in a single image [14, 15, 25, 31, 40, 54] and attained notable progress with deep learning [14, 15, 25, 40, 54], video demoiréing is still an unexplored research problem as far as we know, which is yet of great significance due to the ubiquity and importance of video data in our daily life.

This paper investigates the problem of video demoiréing. Compared to image demoiréing, this task offers more opportunities for high-quality frame-level restoration through

---

*Corresponding Author

Figure 1. The first row shows moiré frames at different times, and the second row shows our demoiréd results. Please see our videos, which are clean and temporally consistent.

leveraging auxiliary information from nearby video frames but is yet more challenging as it requires not only frame-level visual quality but also temporal consistency.

The state-of-the-art image demoiréing method [54] fails to recover temporally consistent videos due to its inability to access temporal information/supervision. Using existing post-processing methods such as [18, 22]; in doing so, however, the chance is lost to utilize video information for enhancing frame-level quality. Besides, these post-processing methods are susceptible to artifacts in demoiréd results, and complicate the system design, leading to increased computational costs. Another widely adopted strategy is to incorporate a flow-based consistency regularization [21, 37, 51, 52] on the predicted videos during training, which encourages aligned pixels from nearby frames to have the same pixel intensity values. While simple, such regularization ignores natural intensity changes of pixels in videos (Fig. 3 (a)), is prone to errors in estimated optical flows (Fig. 3 (b) and (c)), and has the potential to propagate artifacts of one frame to nearby frames. Consequently, the improved temporal consistency tends to sacrifice frame-level quality and fidelity, leading to blurry and low-contrast results (Fig. 7 (a): blurry textures).

In this work, we present a simple video demoiréing model to leverage multiple video frames and a new relation-based consistency loss to improve video-level temporal consistency without sacrificing frame-level qualities. Besides, we construct the first hand-held video demoiréing dataset to facilitate further studies on learning-based approaches.

We analyze the characteristics of moiré patterns in

videos and develop a video demoiréing baseline model following [40, 49, 50] with a selective aggregation scheme to adaptively combine aligned features and a pyramid architecture to enlarge the receptive field. The baseline model can effectively leverage nearby frames for a better frame-level demoiréing. Deep supervision at different scales is adopted during training to facilitate model optimization.

Moreover, inspired by the observation that human beings can perceive video flickering [11] directly from consecutive frames without using explicitly aligned videos, we propose a simple relation-based temporal consistency loss that encourages the direct relations (e.g., pixel intensity differences) of predicted video frames to follow those of ground-truth frames. In particular, we exploit such relations at multiple levels, including pixel level using pixel intensity differences and patch level using intensity statistics (e.g., mean) changes considering different patch sizes. Instead of constraining intensities of aligned pixels to be identical, our relation-based regularization directly matches the natural relations and changes of nearby video frames with those of ground-truth videos. This simple design bypasses the aforementioned drawbacks of flow-based consistency regularization and avoids sacrificing frame-level qualities while still being able to enforce the model to learn temporal consistency priors from ground-truth videos.

Further, as there are no available datasets for developing and evaluating video demoiréing methods, we collect a new video demoiréing dataset with a dedicated pipeline to ensure spatial and temporal alignments between moiré videos and corresponding ground-truth ones.

Finally, extensive experiments on our video demoiréing dataset demonstrate the superior performance of our method. In particular, our method obtains 22% improvements in terms of LIPIS in comparison with MBCNN [54] and more than 75% of users preferred our results when compared with results without using the multi-scale relation-based consistency loss.

2. Related Work

Image Demoiréing. Moiré patterns appear when two similar repetitive patterns interact with each other, and it is frequently observed while capturing images on the screen, which severely degrades image qualities. To remove it, early works have studied spectral models [38] and the sparse matrix decomposition method [23]. However, these methods can only remove certain types of moiré patterns. With the rising of deep learning, various convolution neural networks [14, 15, 25, 26, 40, 54] have been designed for image demoiréing. Sun et al. [40] built the first large-scale image demoiréing dataset and designed a multi-scale architecture to remove moiré patterns. Further, MopNet [14] integrates the characteristics of the moiré pattern into the network and achieves a better result. For high-resolution im-
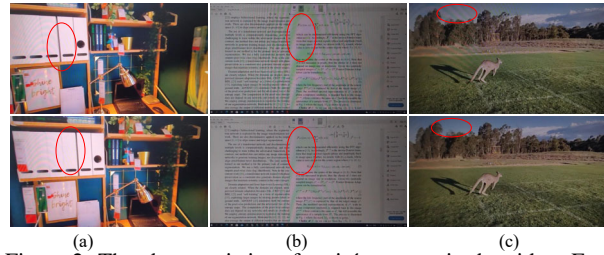

Figure 2. The characteristics of moiré patterns in the video. Each row represents frames with different time stamps, and the differences between two rows are highlighted by red circles.

age demoiréing, He et al. [15] designed a two-stage method to simultaneously remove large moiré patterns and preserve image details. In addition to the above methods which design networks in the image domain, some approaches attempt to address this problem from the perspective of frequency domain [25, 54]. Most recently, Liu et al. [26] designed a self-supervised learning method to restore the image only from a pair consisting of one focused moiré-degraded image and one defocused moiré-free image. What differentiates our work from the above research efforts is that we study the new task of video demoiréing with a collected dataset, which provides new opportunities to improve demoiréing qualities by leveraging temporal information.

Multi-Frame Restoration. Multi-frame restoration [3, 24, 39, 41, 44] aims to improve restoration performance by leveraging information from auxiliary frames and typically performs better than image-based counterparts. A key component in multi-frame restorations is the registration of multiple frames, and previous methods usually achieve this using optical flow [1, 3]. Recently, Tian et al. [43] introduced the deformable convolution [10] into video super-resolution to implicitly align multiple frames and obtain superior results. This module has been further developed and adopted by several follow-up works [5, 6, 28, 49]. In this work, we follow the method in [49] to align multiple frames in feature space and develop a module to automatically select valuable information from nearby moiré frames.

Video Temporal Consistency. To obtain temporally consistent videos, previous methods have adopted consistency regularization during network training [21, 33, 37, 47, 51] or have used it to post-process [2, 18, 22] flickering videos. The most widely adopted consistency regularization is based on dense correspondences (e.g., optical flow), which enforces the intensity of aligned pixels in different frames to be the same [21, 37, 51]. However, such a flow-based approach is sensitive to the quality of the estimated dense correspondences [12, 42] and ignores the natural changes in videos. Without optical flows, Lei et al. [22] obtained temporally consistent videos by developing a video prior method which needs time-consuming test-time training. Besides, the effectiveness of the approach relies on a temporally consistent video input which is different from our case. Some
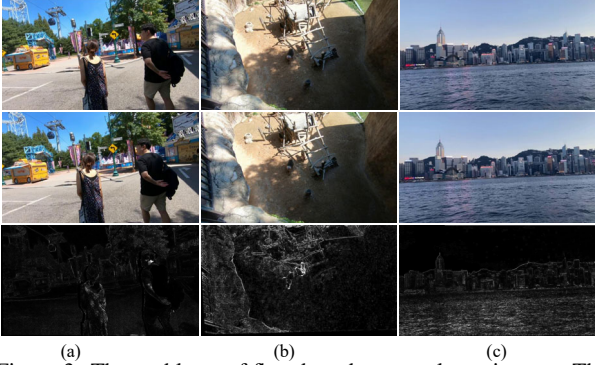
Figure 3. The problems of flow-based temporal consistency. The first two rows are two consecutive frames, and the last row visualizes the warping error using RAFT [42]. (a) Intensity changes when the person walks from shadow to sunlight. (b), (c) show misalignment between two frames.

approaches [13, 32, 48] improve temporal consistency of CNN predictions by augmenting a single frame to multiple frames and enforcing their consistency. Unfortunately, the moiré pattern in videos is difficult to simulate which makes augmentation-based methods ineffective. Compared to previous works, our relation-based regularization is simple and can take the natural changes of videos into account. Without using optical flows, our method also avoids suffering from the issues caused by inaccurate optical flow estimation.

3. Method

We first present the characteristics of video moiré patterns in Sec. 3.1, which inspires the design of our baseline video demoiréing model. Then, we elaborate on the key components of our baseline model (Fig. 4) including frame alignment, feature aggregation, and demoiré reconstruction in Sec. 3.2. Further, we analyze the weakness of flow-based temporal consistency and detail our newly proposed relation-based consistency regularization in 3.3. Finally, we show our training objectives in Sec. 3.4.

3.1. Characteristics of Moiré Patterns in Video

The color, shape and location of moiré patterns are generally influenced by camera viewpoints, as shown in Fig. 2 (a) and (b). Under a mild video-capturing setting using hand-held cameras, we observe the following characteristics of moiré patterns in captured videos. First, as a video plays, the degraded areas have a chance to be clean due to their change of appearing locations (Fig. 2 (a): the white box at different positions), which can provide valuable information to recover distorted regions in nearby frames. Second, the unavoidable hand shaking while shooting videos will slightly change camera viewpoints and induce different moiré patterns in nearby video frames (Fig. 2 (b): the different text color), which can be leveraged to better distinguish moiré regions by comparing such appear-
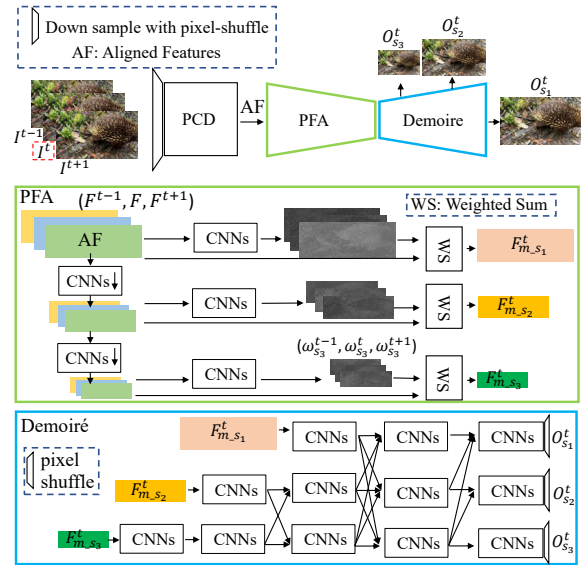

Figure 4. The overview of our method. Our video demoiréing network mainly consists of three parts: First, the *PCD* [49] takes consecutive frames as inputs to implicitly align frames in the feature space. Second, the feature aggregation module merges aligned frame features at different scales by predicting blending weights. Third, the merged features are sent to the demoiré model with dense connections to realize moiré artifacts removal.

ance changes. Third, the strength of moiré patterns varies in different video frames due to the auto-change of focal length [26], offering a chance to leverage less influenced "lucky" frames to restore severely degraded ones (Fig. 2 (c): the sky with and without moiré patterns).

Based on the above analysis, our baseline video demoiréing network (Sec. 3.2) aligns multiple frames for the purpose of appearance comparisons, effectively aggregates features from nearby frames, and incorporates a blending mechanism to select valuable information from nearby frames in a learnable manner.

3.2. Baseline Video Demoiréing Network

Our baseline video demoiréing network shown in Fig. 4 takes as inputs multiple consecutive video frames (I^{t-1}, I^t, I^{t+1}) and outputs restored prediction O^t (equal to $O_{s_1}^t$), leveraging multiple nearby video frames for restoring I^t. Note that we take three adjacent frames to illustrate our model without loss of generality.

Given the inputs (I^{t-1}, I^t, I^{t+1}), we first incorporate a pyramid cascading deformable (PCD) model in [28] to extract and generate implicitly aligned features (F^{t-1}, F^t, F^{t+1}). To deal with large moiré patterns in high-resolution videos, we apply pixel shuffle to down-sample the inputs before feeding them into the PCD module which can effectively enlarge the receptive field of the model without sacrificing original information.

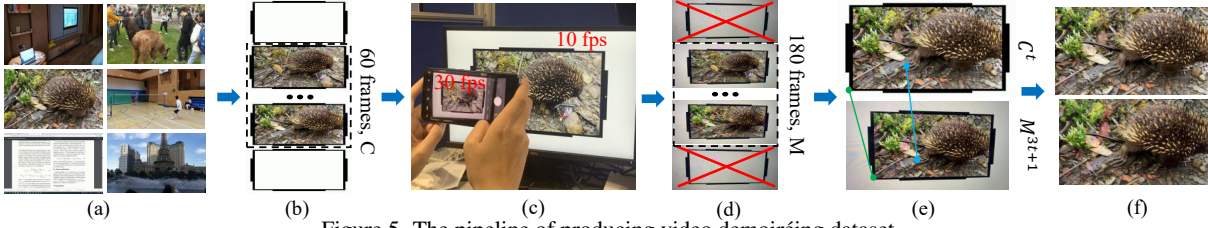Then, a pyramid feature aggregation (PFA) module

Figure 5. The pipeline of producing video demoiréing dataset.

(Fig. 4: green box) is developed to selectively aggregate aligned features at multiple scales (s_1, s_2, s_3). Specifically, the aligned features are down-sampled using convolution layers with a stride of 2 to produce a feature pyramid that allows feature aggregation to be performed at different resolutions to handle multi-scale moiré patterns. At each scale s_i, the aligned features are concatenated together and used to predict normalized blending weights ($\omega_{s_i}^{t-1}, \omega_{s_i}^t, \omega_{s_i}^{t+1} \in (0,1)$). The aggregated features $F_{m\_s_i}^t$ are further generated through a pixel-wise weighted summation of aligned features, which enables selective feature aggregation.

Finally, the demoiré reconstruction module produces the demoiréd image O^t. We densely connect features at different scales to allow them to communicate with each other following [46, 50] (Fig. 4: blue box). We apply more convolutional blocks at lower resolution branches to capture a large field of view, benefiting from identifying and removing large moiré patterns and using less convolutional blocks at higher resolution branches to preserve image details.

3.3. Temporal Consistency

Although our baseline video demoiréing network can generate high-quality frame-level results, it cannot ensure video-level consistency. Here, we study the problem of how to generate temporally consistent video demoiréing results. In the following, we start by analyzing classic flow-based temporal consistency regularization which tends to degrade frame-level qualities, and then elaborate on our simple relation-based temporal consistency loss.

Flow-Based Temporal Consistency Regularization. Classic methods achieve temporal consistency by estimating the pixel correspondences in nearby video frames with mostly optical flow methods and building a loss as Eq. (1) to enforce the intensity of matched pixels to be the same [18, 51, 52].

$$L_f = ||M \cdot (\mathcal{W}_{t+1 \to t}(O^{t+1}, \mathcal{F}_{t+1 \to t}) - O^t)||_1, \quad (1)$$

where M represents the occlusion map to rule out the influence of occluded pixels, $\mathcal{W}_{t+1 \to t}$ means the flow-based image warp [16] to align pixels based on optical flow $\mathcal{F}_{t+1 \to t}$, and O^t, O^{t+1} are nearby output frames.

Key Observations. We carried out a systematic study on flow-based temporal consistency loss and have the following key observations. First, a video often undergoes natural

changes as time passes due to environmental factors such as lighting and view directions [34], and thus a temporally satisfactory video does not necessarily mean that the intensity of the same region never changes (Fig. 3 (a): a person from shadow to sunlight). However, such natural changes will incur a large loss (Fig. 3 (a) third row: the warping error) in flow-based temporal consistency regularization, violating the natural phenomenon. Second, the effectiveness of flow-based temporal consistency is adversely affected by the inaccurate estimation of optical flows. Even the existing state-of-the-art flow estimation method, RAFT [42], suffers from many failure modes (Fig. 3 (b) and (c): warping errors due to inaccurate flow estimations), especially in objects' boundaries and repetitive textures. These mistakenly matched pixels will incur a penalty that does not exist. Finally, the above inaccurate penalties will force the model to trade off frame-level quality for temporal consistency, *e.g.*, averaging matched pixels, leading to blurry and low-contrast results (please see videos and experiments).

Relation-Based Temporal Consistency. Human beings can assess whether a video is temporally consistent or not by directly observing consecutive video frames without using explicitly aligned frames, which motivates us to rethink whether pre-aligned correspondences are needed to learn temporally consistent results and study how to learn temporally consistent results directly from ground-truth reference videos, as they are naturally consistent. Here, in order to learn temporal consistency patterns from reference videos, we propose matching the direct temporal relations of predicted video frames (O^t, O^{t+1}) to those of the reference ones (G^t, G^{t+1}), where G indicates the ground-truth video. The simplest temporal relation can be built by comparing the pixel intensity between video frames; we also investigate other options for temporal relations below.

Basic Relation Loss. The most basic relation we consider is the difference between two frames, as Eq. (2):

$$L_r = ||(O^{t+1} - O^t) - (G^{t+1} - G^t)||_1. \quad (2)$$

As opposed to the flow-based temporal consistency loss in Eq. (1), which constrains aligned predictions to have the same intensity values, the basic relation loss requires that the difference of outputs and reference frames should be similar, *i.e.*, the predicted results should follow the temporal change of the reference videos.
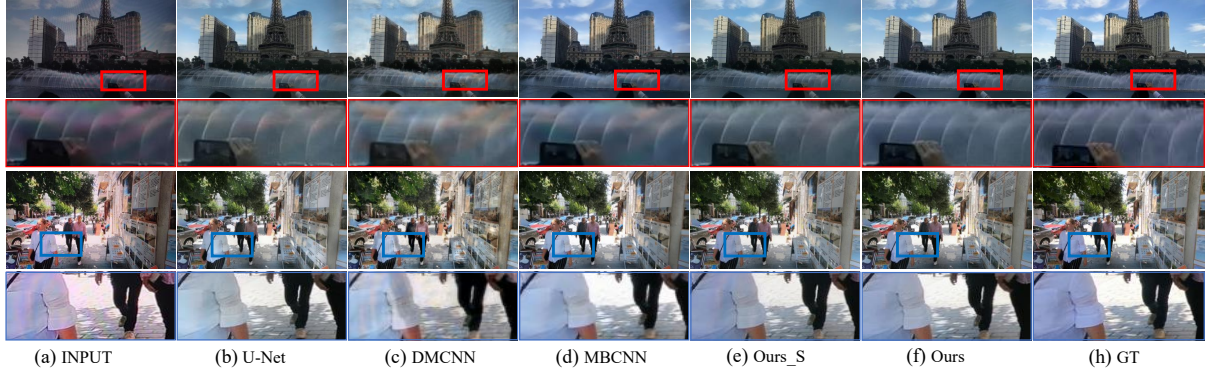
| (a) INPUT | (b) U-Net | (c) DMCNN | (d) MBCNN | (e) Ours_S | (f) Ours | (h) GT |

Figure 6. Qualitative Comparisons. We compare with other baselines and obtain better results on the moiré artifacts removal.

Multi-Scale Region-Level Relation Loss. Besides pixel-level relations, we also consider region-level relations that follow human habits [8,30]. Biologically, the retinal cell receives light from a region instead of a point, and the region size is determined by the distance between retinal cells and observed objects. For region-level relations, we use pixel statistics, such as the mean value of pixel intensities, to build the relation loss. We empirically find the mean value works very well in practice. The reason might be that the mean of a patch reflects the brightness of that area, which is closely related to flickers [9]. Specifically, we use patches with different sizes $k \in C$ to take account of various receptive fields, extract the statistics from these patches, and construct a multi-scale region-level relation loss as in Eq. (3). Moreover, we only penalize the scale that incurred the minimum difference to protect temporally consistent predictions from nearby potential flickering regions.

$$L_{mbr} = \frac{1}{N} \sum_{n=1}^{N} L_n^{k^*} |_{k^* = \arg\min_k \{|(\mathcal{T}_k(O^{t+1}) - \mathcal{T}_k(O^t))_n|\}, k \in C},$$

$$L_n^k = |((\mathcal{T}_k(O^{t+1}) - \mathcal{T}_k(O^t))_n - (\mathcal{T}_k(G^{t+1}) - \mathcal{T}_k(G^t))_n|, \quad (3)$$

where $\mathcal{T}_k$ indicates the operation of calculating the statistics of a patch with size $k \in C$ ($C = \{1\}$ is the basic relation-based loss), and n is the pixel position index.

Analysis. The relation-based loss is simple without needing to estimate dense correspondences and thus avoids the problem of misalignment caused by optical flow estimation, and the natural changes in ground-truth videos can be transferred to output frames. Meanwhile, the model can learn to produce temporally consistent results by mimicking the temporal relations of the reference video, which naturally encompasses temporal consistency priors.

3.4. Training Objectives

Our overall training objective L_{train}, in Eq. (4), is the combination of the frame-level demoiréing loss L_d^t, L_d^{t+1}, which regresses outputs at different scales to the ground truths, and the relation loss L_{mbr} of temporal consistency.

$$L_{train} = L_d^t + L_d^{t+1} + \lambda_t L_{mbr}, \quad (4)$$

λ_t is used to control the degree of temporal consistency.

To construct L_d, we adopt L_1 and perceptual loss [17], which guide the regression process. Apart from the loss on the original resolution, deep supervisions [20] are applied at different scales to assist the network training. The frame-level demoiréing loss L_d^t is formulated as Eq. (5):

$$L_d^t = \sum_{i,l} ||O_{s_i}^t - G_{s_i}^t||_1 + \lambda ||\Phi_l(O_{s_i}^t) - \Phi_l(G_{s_i}^t)||_1, \quad (5)$$

where $O_{s_i}^t$ and $G_{s_i}^t$ are output and corresponding ground truth at the s_i scale, respectively. Φ_l is a set of VGG-16 layers, and λ is the weight used to balance different parts.

4. Video Demoiréing Dataset

We collect the first video demoiréing dataset captured by hand-held cameras, *e.g.*, a smartphone camera. The capturing pipeline to ensure spatial and temporal alignments between camera-recorded and original videos is shown in Fig. 5 and elaborated below.

First, the 720p high-quality source videos displayed on the screen consist of videos from REDS [29], MOCA [19], and videos taken by ourselves. To ensure the diversity of collected videos, we manually choose videos covering various scenarios, including human beings, landscapes, texts, sports, and animals (examples in Fig. 5 (a)). We collect 290 videos, and each video has 60 frames.

Second, it is difficult to align videos recorded by cameras and source videos played on the screen considering different frame rates and asynchronous start timestamps. For example, if the camera frame rate is not divisible by the video frame rate, the recorded frame will contain multi-frame information (occurs when switching frames) from the source video, which results in blurry images. Even though the frame rate meets the requirement, different start timestamps (*i.e.*, start to play and record the video) also cause the problem of multi-frame confusion. For these obstacles, we adjust the frame rates and insert start/end flags into videos. Specifically, we set camera and source video frame rates to 30 fps and 10 fps, respectively, and extend source videos
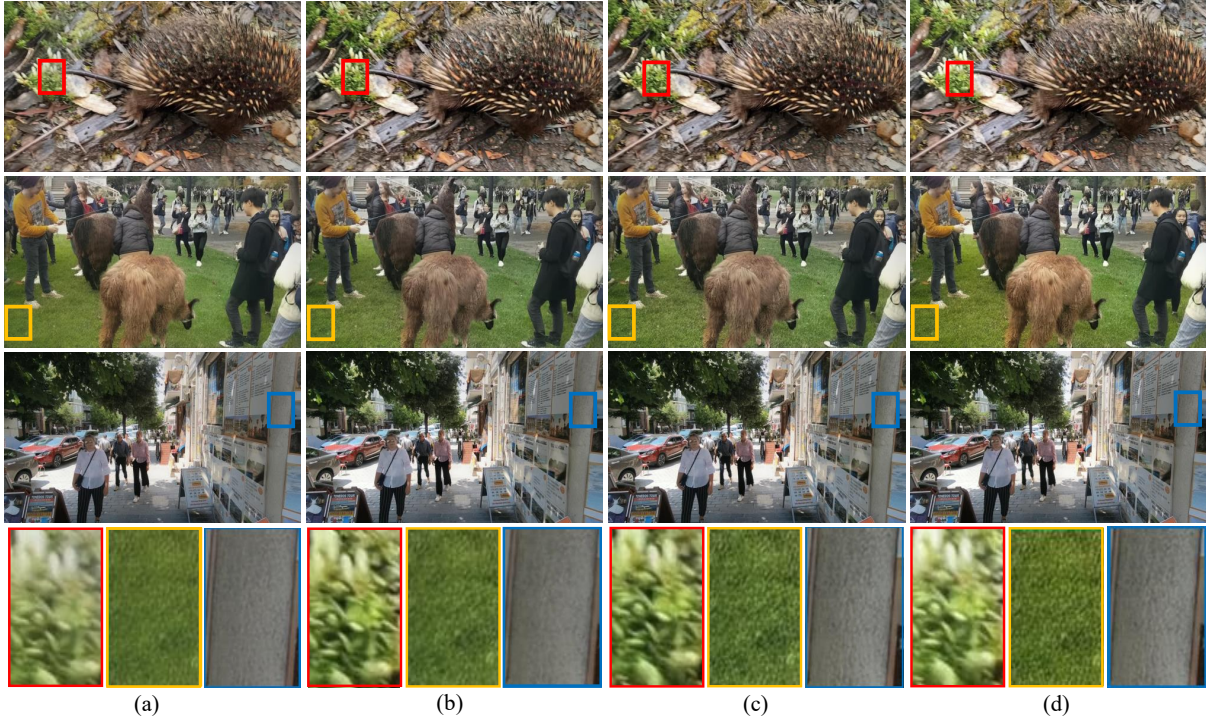
Figure 7. Different types of temporal consistency. (a) Flow-based temporal consistency. (b) Ours with basic relation loss. (c) The full version of our method. (d) Results without temporal constraints (reference). We can observe that (c) preserves details best.

with a few white frames at the beginning and the end of each video. What's more, we follow the data collection process in [40] to add some black blocks surrounding the frame to provide more robust keypoints (Fig. 5 (b) and (c)).

Third, given the source video, mobile phone, and monitor, the moiré pattern can be produced by adjusting the camera view points. While capturing, the mobile phone is hand-held by a person to simulate practical video recording senarios, and different shooting angles and distances are adopted to increase the diversity of moiré patterns (Fig. 5 (c)). After recording, we can obtain 180 frames (three times the source video) from each video after removing the pre-inserted white frames (Fig. 5 (d)), and the final moiré frame is sampled among three consecutive frames. Here, we sample the intermediate one since it is not sensitive to frame transitions (Fig. 5 (e)).

Finally, to obtain training pairs (Fig. 5 (f)), source and captured frames should be aligned through frame correspondences, such as optical flow and homography matrix. In this work, we adopt the homography matrix to align two frames (Fig. 5 (e)). Instead of using only keypoints (ORB [36]) detected on image regions [15] or auxiliary black regions [40], we utilize both of them to estimate the homography matrix using the RANSAC [45] algorithm.

5. Experiments

In this section, we first introduce training details (Sec. 5.1), then qualitatively and quantitatively compare our

method with other baselines at the frame level (Sec. 5.2) and the video level (Sec. 5.3). Finally, we validate our video demoiréing model and the relation-based consistency regularization (Sec. 5.4).

5.1. Training Details

The video demoiréing network takes three consecutive frames as inputs to predict one restored image. To train the model, we automatically divide the video demoiréing dataset into 247 train videos and 43 test videos, and the hyperparameters λ and λ_t are set to 0.5 and 50, respectively. Furthermore, we adopt four region sizes $C = \{1, 3, 5, 7\}$ to simulate different receptive fields. The optimizer in our implementation is Adam with a cosine learning rate [27]. In total, we train 60 epochs with batch size 1 on one NVIDIA 2080Ti GPU, and the temporal consistency loss is invoked in the last 10 epochs for training stability.

5.2. Frame-Level Comparisons

We compare our approach with image demoiréing methods (*i.e.*, MBCNN [54] and DMCNN [40]) and other widely used backbones, such as U-Net [35]. In order to verify the effectiveness of video demoiréing without being affected by other factors (*e.g.*, number of parameters and the choice of loss function), we adopt our video demoiréing model but change the input to repetitions of a single frame (Ours_S, see Fig. 8 (b)). To quantitatively measure the performance of demoiréing, we adopt PSNR, SSIM, and LPIPS [53] that

is more aligned with human perception as our metrics. ('↑': larger value is better, '↓': smaller value is better.)

Methods	LPIPS ↓	PSNR ↑	SSIM ↑
MBCNN [54]	0.260	21.534	0.740
DMCNN [40]	0.321	20.321	0.703
U-Net [35]	0.225	20.348	0.720
Ours_S	0.212	21.772	0.729
Ours	0.202	21.725	0.733

Table 1. Demoiréing performance of different methods. (Red: best, Blue: second best)

Methods	FID ↓	warping error ↓	user study ↑	LPIPS↓
Ours_S	0.094	5.98	14%	0.212
Ours	**0.084**	5.65	25%	0.202
Ours+F	0.109	**2.70**	9%	0.339
Ours+R	0.088	4.79	42%	0.211
Ours+M	0.085	5.03	-	0.201
GT	0.000	4.56	-	0.000

Table 2. Temporal consistency measurements when λ_t is 50. Ours_S: video demoiréing model with three repetitive frames, Ours: video demoiréing model with multiple frames, Ours +F: add flow-based consistency loss, Ours+R: add basic relation-based consistency loss, Ours+M: add multi-scale relation-based consistency loss. In user study, all other baselines are compared with Ours+M, and this table reports the percentage of each baseline being selected (Ours+M outperforms all baselines).

Qualitative Comparison. In Fig. 6, we show images restored by different methods. It clearly shows that our approach has advantages over other methods for removing moiré artifacts, such as the moiré patterns on the fountain, white T-shirt and floor. We attribute the superiority of our method to its ability to utilize auxiliary information from the nearby video frames.

Quantitative Comparison. Frame-level quantitative results are reported in Table 1. Under the circumstance of single image demoiréing, our method (Ours_S) outperforms previous methods (above the dotted line). Moreover, the performance is further improved using multiple frames (Ours), especially LPIPS, which manifests the effectiveness in leveraging multiple frames to improve perception results.

5.3. Video-Level Comparisons

Following previous works [7, 47], we adopt FID and warping error to measure video-level performance. Here, FID measures the distance between output and ground-truth videos in the feature domain using I3D [4], and the warping error calculates differences between two frames aligned by optical flows [42]. Note that the warping error cannot accurately reflect the video temporal consistency due to inaccurate optical flow and natural changes in videos. To illustrate it, we calculate the warping error of ground-truth videos (Table 2: last row), which is still very large. Besides, we also conduct user studies to assist video-level comparisons. For the user study, participants are asked to choose one out
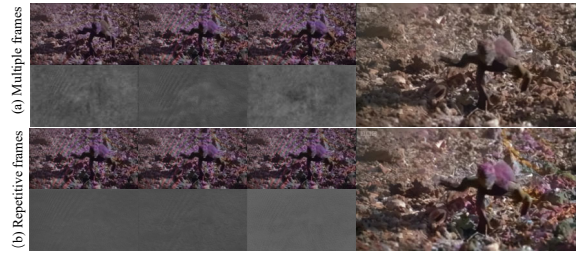

Figure 8. Visualization of weight maps. (a) Three consecutive frames and the weight maps. (b) Replace consecutive frames with repetitions of a single frame and the weight maps.
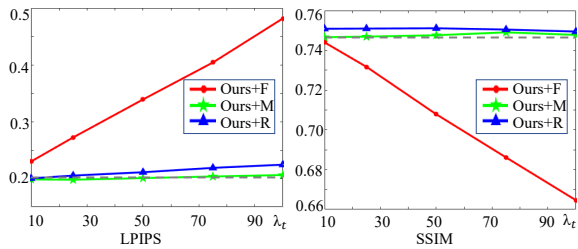

Figure 9. Demoiréing performance when increasing λ_t.

of two videos based on video quality or mark them as indistinguishable; they are given sufficient time to make the decision. In the process of our user study, two videos produced by different methods are displayed in random order, and participants can replay videos with various frame rates. In total, 14 individuals participated in our experiments.

As our baseline video demoiréing model (Ours) obtains better results than other compared methods, we take it as the baseline model for video-level evaluation. Specifically, we compare the video temporal consistency and quality with the results of single image demoiréing (Ours_S), classic flow-based consistency regularization (Ours+F, replace L_{mbr} loss with L_f loss in Eq. (1)) and multi-scale relation-based consistency regularization (Ours+M, L_{mbr} loss).

As shown in Table 2, the multi-frame demoiréing (Ours) is more consistent than the single-frame demoiréing (Ours_S). Also, the FID indicates that videos restored by multiple frames are closer to ground-truth videos with higher quality. By incorporating temporal constraints, the video temporal consistency is improved. Specifically, the flow-based method (Ours+F) has the best warping error, but the LPIPS shows that the frame-level quality may drop significantly. Furthermore, only 9% of users preferred this type of videos when compared with the full version of our method (Ours+M). In contrast, our multi-scale relation-based loss (Ours+M) can improve the video temporal consistency while maintaining the frame-level quality (LPIPS is similar to the method without using temporal consistency regularization, 0.201 v.s. 0.202). More users preferred these results in comparison with all over baselines.

More Analysis on Temporal Consistency. In the following, we perform more analysis to demonstrate the robustness of our relation-based loss. We plot the curve of
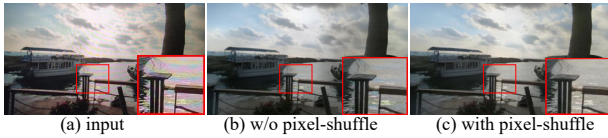
| (a) input | (b) w/o pixel-shuffle | (c) with pixel-shuffle |

Figure 10. Different receptive fields. A large receptive field (with pixel-shuffle) benefits the moiré artifacts removal.

Methods	LPIPS ↓	PSNR ↑	SSIM ↑
no pixel-shuffle	0.205	21.372	0.733
no deep supervision loss	0.216	21.153	0.728
Ours	0.202	21.725	0.733

Table 3. Ablation study on the network and loss.

demoiréing performance at different weights λ_t of the temporal consistency loss. The results are shown in Fig. 9, where the dotted line represents the performance without temporal constraints (Ours). With the increase of λ_t, the flow-based (Ours+F) consistency regularization leads to worse LPIPS and SSIM. On the contrary, our multi-scale relation-based approach (Ours+M) learns consistency priors directly from ground-truth videos without sacrificing video quality (please refer to our videos).

We show visual comparisons in Fig. 7. When compared with reference images (Fig. 7 (d)) without temporal constraints (Ours), the flow-based method (Ours+F) heavily blurs image details, such as repetitive textures of the grass and cracks on the stone. By contrast, the multi-scale relation-based method (Ours+M) preserves image details well (Fig. 7 (c)), which is comparable to reference images with improved temporal consistency.

5.4. Ablation Studies

Components of Networks. We validate our network designs from the following two aspects. 1) *Receptive field enlargement due to the pixel shuffle operation*: we remove the pixel shuffle operation to reduce the network's receptive field and evaluate the performance. From results in Table 3, we observe that the performance degrades without using pixel shuffle. Besides, a large receptive field benefits high-resolution images and large moiré patterns. This can be seen in Fig. 10, where moiré artifacts on the lake are removed under the large receptive field. 2) *Analysis of blending weights*: to better understand the role of blending weights in our model, we visualize the weight maps (see Fig. 8) that are used to merge multi-frame features. The weight maps can reflect moiré patterns and choose valuable information from nearby frames for fusion, as shown in Fig. 8 (a). Moreover, we compare with a special scenario where the inputs are repetitions of a single frame. Under this circumstance, it is difficult to infer moiré patterns without clues from auxiliary frames, as shown in weight maps (Fig. 8 (b)). Consequently, the final demoiréing results (Fig. 8 last column) become worse.

Deep Supervision Loss. To illustrate this, we build the loss function only on the original image scale. From Table 3, we observe that the deep supervision loss boosts the performance regarding all three metrics. A possible explanation is that deep supervision loss forces each branch to learn more reasonable demoiréing representations and fa-

cilitate the optimization process.

Relation-Based Temporal Consistency. We validate two variants of relation-based losses: the multi-scale relation-based loss (Ours+M) and the basic relation-based loss (Ours+R). From Fig. 7 (b), the textures are a bit blurry with the basic relation-based loss and are worse than results (Fig. 7 (c)) from our multi-scale design. The reason might be that region-level statistics (*i.e.*, mean) help reduce negative impacts of temporal-consistency regularization, which tends to average and erase image details. In comparison with the multi-scale design in Table 2, less users (42%) selected the basic single-scale design. More importantly, the multi-scale based regularization can well maintain the frame-level qualitative performance (see LPIPS in Fig. 9).

6. Limitations and Broader Impacts

Although we have designed a pipeline to ensure the alignment of captured data pairs, it is difficult to perfectly align them under different camera views. Currently, our model also suffers from generalization issues if evaluated on data captured using new devices (*e.g.*, different ISP and Bayer filters) and screens (*e.g.*, different resolution). Expanding the scale of the dataset is one potential solution that will be our future work. In addition, the relation-based loss is generic and can potentially be applied to other video tasks, such as video stabilization. In practice, we have found that the video instability caused by frame misalignments has been reduced. One possible explanation is that stabilization priors are learned from ground-truth videos.

7. Conclusion

In this work, we construct the first video demoiréing benchmark, including a hand-held video demoiréing dataset, and develop a baseline video demoiréing model, effectively leveraging multiple frames. More importantly, we design an effective relation-based consistency regularization, which simultaneously boosts video temporal consistency and maintains visual quality. Detailed analyses are carried out to assist the understanding of video moiré patterns and the weaknesses of flow-based consistency regularization. Finally, extensive experiments demonstrate the superiority of our method.

References

[1] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *ICPR*, 2000. 2

[2] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *TOG*, 2015. 2

[3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 7

[5] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, 2021. 2

[6] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021. 2

[7] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *ICCV*, 2019. 7

[8] Yang Cheng, Jie Cao, Yangkun Zhang, and Qun Hao. Review of state-of-the-art artificial compound eye imaging systems. *Bioinspiration & biomimetics*, 2019. 5

[9] Lark Kwon Choi and Alan Conrad Bovik. Video quality assessment accounting for temporal visual masking of local flicker. *Signal Processing: image communication*, 2018. 5

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2

[11] Jodi L Davenport. Consistency effects between objects in scenes. *Memory & Cognition*, 2007. 2

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 2

[13] Gabriel Eilertsen, Rafal K Mantiuk, and Jonas Unger. Single-frame regularization for temporally stable cnns. In *CVPR*, 2019. 3

[14] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Mop moire patterns using mopnet. In *ICCV*, 2019. 1, 2

[15] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Fhde 2 net: Full high definition demoireing network. In *ECCV*, 2020. 1, 2, 6

[16] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 4

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[18] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 1, 2, 4

[19] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *ACCV*, 2020. 5

[20] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, 2015. 5

[21] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *CVPR*, 2019. 1, 2

[22] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *NeurIPS*, 2020. 1, 2

[23] Fanglei Liu, Jingyu Yang, and Huanjing Yue. Moiré pattern removal from texture images via low-rank and sparse matrix decomposition. In *VCIP*, 2015. 2

[24] Jiaying Liu, Wenhan Yang, Shuai Yang, and Zongming Guo. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *CVPR*, 2018. 2

[25] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Aleš Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoiréing. In *ECCV*, 2020. 1, 2

[26] Lin Liu, Shanxin Yuan, Jianzhuang Liu, Liping Bao, Gregory Slabaugh, and Qi Tian. Self-adaptively learning to demoiré from focused and defocused image pairs. *arXiv preprint arXiv:2011.02055*, 2020. 2, 3

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

[28] Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *CVPR*, 2021. 2, 3

[29] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, 2019. 5

[30] EYK Ng, Jen Hong Tan, U Rajendra Acharya, and Jasjit S Suri. Human eye imaging and modeling. 2012. 5

[31] Kimihiko Nishioka, Naoki Hasegawa, Katsuya Ono, and Yutaka Tatsuno. Endoscope system provided with low-pass filter for moire removal, Feb. 15 2000. US Patent 6,025,873. 1

[32] Hao Ouyang, Tengfei Wang, and Qifeng Chen. Internal video inpainting by implicit long-range propagation. In *ICCV*, 2021. 3

[33] Kwanyong Park, Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Preserving semantic and temporal consistency for unpaired video-to-video translation. In *ACM MM*, 2019. 2

[34] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 4

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6, 7

[36] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011. 6

[37] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 1, 2

[38] Denis N Sidorov and Anil Christopher Kokaram. Suppression of moiré patterns via spectral analysis. In *VCIP*, 2002. 2

[39] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 2

[40] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *TIP*, 2018. 1, 2, 6, 7

[41] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *CVPR*, 2020. 2

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2, 3, 4, 7

[43] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 2

[44] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1984. 2

[45] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *ACM MM*, 2010. 6

[46] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 4

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 2, 7

[48] Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. Consistent video style transfer via compound regularization. In *AAAI*, 2020. 3

[49] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019. 2, 3

[50] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 2, 4

[51] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, 2021. 1, 2, 4

[52] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *ICCV*, 2019. 1, 4

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[54] Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, and Ales Leonardis. Image demoireing with learnable bandpass filters. In *CVPR*, 2020. 1, 2, 6, 7