Rethinking Video Deblurring with Wavelet-Aware Dynamic Transformer and Diffusion Model

Chen Rao[1†], Guangyuan Li[1†], Zehua Lan[1], Jiakai Sun[1], Junsheng Luan[1], Wei Xing[1*], Lei Zhao[1*], Huaizhong Lin[1*], Jianfeng Dong[2], and Dalong Zhang[1]

[1] College of Computer Science and Technology, Zhejiang University
[2] Zhejiang Gongshang University

Abstract. Current video deblurring methods have limitations in recovering high-frequency information since the regression losses are conservative with high-frequency details. Since Diffusion Models (DMs) have strong capabilities in generating high-frequency details, we consider introducing DMs into the video deblurring task. However, we found that directly applying DMs to the video deblurring task has the following problems: (1) DMs require many iteration steps to generate videos from Gaussian noise, which consumes many computational resources. (2) DMs are easily misled by the blurry artifacts in the video, resulting in irrational content and distortion of the deblurred video. To address the above issues, we propose a novel video deblurring framework VD-Diff that integrates the diffusion model into the Wavelet-Aware Dynamic Transformer (WADT). Specifically, we perform the diffusion model in a highly compact latent space to generate prior features containing high-frequency information that conforms to the ground truth distribution. We design the WADT to preserve and recover the low-frequency information in the video while utilizing the high-frequency information generated by the diffusion model. Extensive experiments show that our proposed VD-Diff outperforms SOTA methods on GoPro, DVD, BSD, and Real-World Video datasets. The codes will be available at `https://github.com/Chen-Rao/VD-Diff`.

Keywords: Diffusion model · Prior feature · Video deblurring · Wavelet-based transformer

1 Introduction

Portable devices are widely used to capture videos. However, camera shake and fast movement of objects lead to undesirable blur in video. To reduce the impact of video blur, researchers have done a lot of work on video deblurring. Traditional methods are mainly based on hand-crafted assumptions and priors, which are difficult to design. Besides, inappropriate priors will reduce the quality of video deblurring.

[†] co-first authors, equal contribution; [*] Corresponding author

In the past decade, video deblurring has witnessed significant progresses with the development of deep learning. Early deep learning methods mainly use CNN-based models [13, 14, 26, 28, 36, 42, 46] to extract video features for capturing the spatio-temporal information between frames. Nevertheless, these approaches have limitations in modelling long-range spatial dependencies and capturing non-local self-similarities.

Recently, Transformer has been employed in video deblurring tasks [4, 21, 22, 37] to address limitations in CNN-based methods, which can establish long-range spatial dependencies and capture non-local spatial information. Although Transformer-based methods exhibit impressive performance, they still have the following issues. On the one hand, the methods using standard global Transformer [8] tend to deblur videos based on low-frequency details and have limited perception of high-frequency information. On the other hand, when the local window-based Transformer [37] is used to reduce the computational cost, self-attention is calculated within a location-specific window, resulting in a limited receptive field. Consequently, it may ignore some key elements of similar and clearer scene patches in the spatiotemporal neighborhood. Therefore, the local window-based self-attention fails to capture valuable high-frequency information from the spatiotemporal neighborhood, resulting in the loss of high-frequency information.

Very recently, diffusion models (DMs) have shown their outstanding capabilities in video synthesis tasks [2, 9, 23, 24, 31, 32, 40], which can denoise pure Gaussian noise into high-quality video with abundant high-frequency details. Therefore, we consider employing DMs in video deblurring to restore high-frequency information. However, directly applying DMs to video deblurring tasks has the following problems. First, DMs demand a large number of iteration steps to generate videos from Gaussian noise, which costs a lot of computational resources. Second, DMs are sensitive to motion blur in videos, resulting in irrational content and distortion of the deblurred video.

To solve the above problems, we propose a novel method called VD-Diff, which combines our proposed Wavelet-Aware Dynamic Transformer (WADT) and diffusion model for video deblurring. Specifically, we employ the diffusion model in a highly compact latent space to generate prior features containing high-frequency information that conforms to the ground truth distribution, which facilitates video deblurring. Applying DM in a highly compressed latent space can significantly reduce the number of denoising iteration steps.

In addition, we design WADT to retain and restore the low-frequency information in the video while exacting the high-frequency information from the prior features generated by the diffusion model, which ensures the deblurred video is undistorted. WADT first decomposes the blurred video frames into the approximation coefficient and detail coefficients by Wavelet Transform (WT) and extracts low-frequency information from the approximation coefficient by Wavelet-Aware Dynamic Transformer Layers (WADTL). Meanwhile, WADT employs WADTL to fuse the prior features generated by DM containing high-frequency information to generate artifact-free and distortion-free videos. Furthermore, we

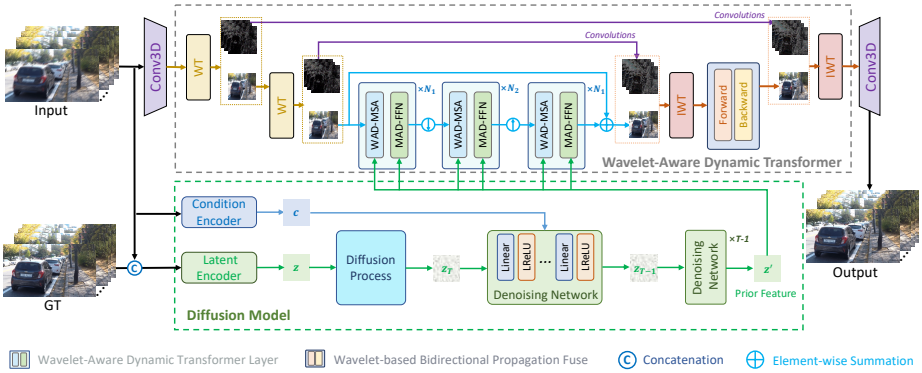


Fig. 1: The overall architecture of the proposed VD-Diff, which consists of Wavelet-Aware Dynamic Transformer (WADT) and Diffusion Model (DM). Specifically, WADT adopts Wavelet Transform (WT) for feature separation, WADT Layer (WADTL) for deep feature extraction, and Wavelet-based Bidirectional Propagation Fuse (WBPF) for spatio-temporal information propagation between frames. The DM generates prior features to supplement high-frequency information for the deblurring process in WADTL.WT: Wavelet Transform. IWT: Inverse Wavelet Transform.

introduce a Wavelet-based Bidirectional Propagation Fuse (WBPF) module in WADT to fully use the spatio-temporal information between frames, which ensure that the video is undistorted in the spatio-temporal dimension. Our main contributions are as follows:

(1) We propose a novel method combining the Wavelet-Aware Dynamic Transformer (WADT) and the diffusion model for video deblurring. To the best of our knowledge, our study is the first to employ DM for video deblurring, which requires only a few iteration steps to restore artifact-free and distortion-free videos.

(2) Our diffusion model can generate prior features containing high-frequency information that conforms to the GT distribution, which facilitates video deblurring.

(3) Our proposed WADT can capture low-frequency information while utilizing the prior features generated by the DM to supplement high-frequency information, which further enhance the quality of deblurred video.

(4) Extensive experiments show that our proposed VD-Diff outperforms SOTA methods on GoPro, DVD, BSD and Real-World Video datasets.

2 Related Work

2.1 Video Deblurring

With the development of deep learning [20], CNN is widely used in video deblurring. For instance, Zhong *et al.* [46] propose residual dense blocks and a global spatio-temporal attention module for video deblurring; Pan *et al.* [28] develope a

deep CNN model to estimate optical flow from intermediate latent frames. CNN-based methods have driven the development of video deblurring, but CNN has limitations in capturing long-range dependencies. To address these issues, some researchers applied the Transformer to video deblurring and achieved impressive results. For example, Lin *et al.* [22] propose a method to calculate self-attention by capturing key elements of adjacent frames through optical flow. Liang *et al.* [21] propose a recurrent video restoration transformer to process local neighboring frames in parallel within a globally recurrent framework. However, the above Transformer-based methods have limited ability to capture high-frequency information, leading to the deblurred videos lacking details.

2.2 Diffusion Models

DMs have shown excellent performance in the field of synthesis. LDM [31] reaches a near-optimal point between complexity reduction and detail preservation by applying DM in the latent space. Blattmann *et al.* [2] turns the image generator into a video generator by introducing a temporal dimension to the latent space diffusion model. Artbank [43] utilize textual embedding to learn the visual priors from the images, synthesizing highly realistic artistic stylized images. DMs have also shown strong capabilities in the field of image restoration. For instance, DiffIR [38], Hi-diff [5] and DiffMSR [18] use DM to generate prior representations for image restoration. Although some methods, such as DDIM [45] and LDM [31], have reduced the computational complexity of DM by changing the denoising process or compressing the image, using existing DM for video deblurring still requires unaffordable computational resources. To cope with the above issue, we apply DM in a highly compact latent space to reduce the iteration steps. We were the first to apply diffusion models to video deblurring, and our models achieved outstanding performance.

2.3 Wavelet-based Restoration Methods

Wavelet is widely used in low-visual restoration tasks [7, 15, 19, 44, 48] as it can separate low-frequency and high-frequency information. For example, Huang *et al.* [12] propose a wavelet-based CNN for multi-scale face super resolution. WavTrans [17] combines wavelet transform with Transformer for feature extraction and fusion; DSTNet [29] develops a wavelet-based feature propagation method to effectively propagate main structures from long-range frames for better video deblurring. Inspired by the above methods, we proposed Wavelet-Aware Dynamic Transformer, which can extract the global structure of the video and utilize the prior features generated by DM to restore high-frequency details.

3 Methodology

3.1 Overall Architecture

The overall framework of our propsed VD-Diff is shown in Fig. 1, which consists of two parts: (1) Wavelet-Aware Dynamic Transformer (WADT); (2) Diffusion

Model. Specifically, we first employ Diffusion Model to obtain a highly compact latent prior feature $z \in \mathbb{R}^{T\times C'}$, which contains supplementary high-frequency information. Note that C' is the number of channels of latent features. Then, we employ WADT to deblur the blurry video frames $V_{blur} \in R^{T\times 3\times H\times W}$ into $V_{HQ} \in R^{T\times 3\times H\times W}$ with the help of the prior feature z.

3.2 Wavelet-Aware Dynamic Transformer

As shown in the WADT part of Fig. 1, we first utilize Conv3D to perform shallow feature extraction on the blurry video frames $V_{blur} \in R^{T\times 3\times H\times W}$ to obtain a global feature $F_{in} \in \mathbb{R}^{T\times C\times H\times W}$. Then, we use Wavelet Transform (WT) to decompose it into approximate coefficients $F_a \in \mathbb{R}^{T\times C/4\times H\times W}$ and detail coefficients $F_d \in \mathbb{R}^{T\times 3C/4\times H\times W}$ and again use WT on F_a to obtain $F_{aa} \in \mathbb{R}^{T\times \hat{C}\times H\times W}$ and $F_{ad} \in \mathbb{R}^{T\times 3\hat{C}\times H\times W}$, where $\hat{C} = C/16$. Next, we employ WADTL to obtain the deblurred features F_{out} from the approximate coefficients F_{aa} and the prior feature z'. Then, we use WBPF to explore spatio-temporal information further. Finally, we use two IWTs to fuse features to obtain artifact-free and distortion-free videos V_{HQ}. We will elaborate on the structure of WADTL and WBPF as follows.

Wavelet-Aware Dynamic Transformer Layer. As shown in Fig. 1, we employ a stack of multiple WADT layer (WADTL) modules hierarchically to fuse approximate coefficients F_{aa} and prior features z'. As shown in Fig. 2, WADTL consists of a Wavelet-Aware Dynamic Multi-head Self-Attention (WAD-MSA) and a Wavelet-Aware Dynamic Feed-Forward Network (WAD-FFN).

In the WAD-MSA part, given a approximate coefficients F_{aa}, we use the highly compact latent prior feature z' to guide video restoration:

$$\hat{F} = W_l^1 z' \odot LN(F_{aa}) + W_l^2 z', \tag{1}$$

where $\odot$ indicates element-wise multiplication, LN denotes layer normalization [1], W_l represents linear layer. Then, we utilize a 3D convolution layer and a 2D convolution layer to project $\hat{F}$ into query $Q = W_2^Q W_3^Q \hat{F}$, key $K = W_2^K W_3^K \hat{F}$, value $V = W_2^V W_3^V \hat{F}$ where W_3 is the channel-wise 3D convolution and W_2 is the depth-wise 2D convolution. Note that we use 3D convolution in the T, H, and W dimensions and the same 2D convolution for the T features in the temporal dimension T. Next, we perform the attention calculation:

$$\mathcal{F}'_{out} = W_a \hat{V} \cdot Softmax(\hat{K} \cdot \hat{Q}/\gamma) + F_{aa}, \tag{2}$$

where γ is a learnable scaling parameter.

In the WDA-FFN part, we first integrate the prior z' into $\mathcal{F}'_{out}$ to obtain $\hat{\mathcal{F}}'_{out}$. And then we use Conv3D to aggregate temporal information and use Conv2D to aggregate spatial information from neighboring pixels. Besides, we employ a gating mechanism to obtain more valuable information. The overall process of WDA-FFN can be described as:

$$F_{out} = G(W_2^1 W_3^1 \hat{\mathcal{F}}'_{out}) \odot W_2^2 W_3^2 \hat{\mathcal{F}}'_{out} + \hat{\mathcal{F}}'_{out}, \tag{3}$$

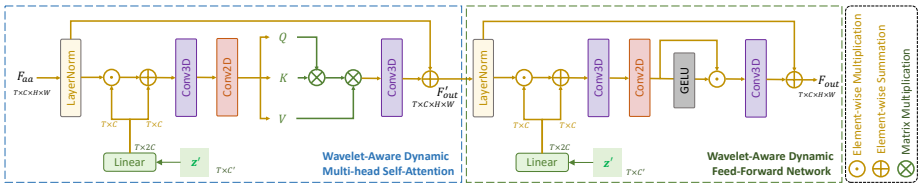


Fig. 2: The illustration of Wavelet-Aware Dynamic Transformer Layer (WADTL), which consists of Wavelet-Aware Dynamic Multi-head Self-Attention (WAD-MSA) and Wavelet-Aware Dynamic Feed-Forward Network (WAD-FFN).

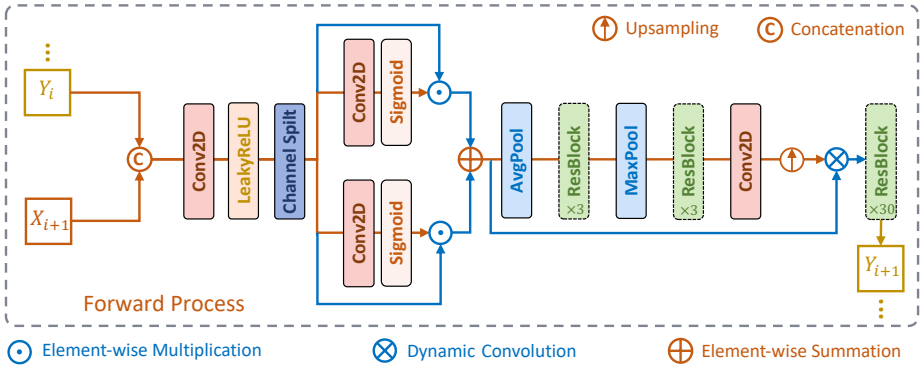


Fig. 3: The structure of the Forward Process in the WBPF. The Backward Process has the same network structure as the Forward Process, but the direction of information propagation between frames is opposite.

where G denotes Gaussian Error Linear Units (GELU).

Wavelet-based Bidirectional Propagation Fuse. Given the generated features $\{X_i\}_{i=1}^{N}$ by the first Inverse Wavelet Transform (IWT), to explore useful contents from the features of other frames, exiting methods usually simply stack $\{X_i\}_{i=1}^{N}$ or the alignment results of $\{X_i\}_{i=1}^{N}$. However, directly stacking frames without accurately estimating the features of $\{X_i\}_{i=1}^{N}$ or the alignment results of $\{X_i\}_{i=1}^{N}$ can only partially exploit the valuable information between frames and may also cause the propagation of erroneous information between frames. To further capture more useful long-range spatio-temporal information and reduce the influence of inaccurate information, we propose Wavelet-based Bidirectional Propagation Fuse (WBPF). As shown in Fig. 1, WBPF consists of a forward process and a backward process.

We take the forward process as an example to introduce the WBPF module in detail, as shown in Fig. 3. First, we connect Y_i and X_{i+1} to get $\bar{X}$, where Y_i is the output of the previous forward process and X_{i+1} is the input of the current forward process. We then use a convolution layer with LeakyReLU to project $\bar{X}$ into $\hat{X}$. Then we split $\hat{X}$ into $\hat{X}_1, \hat{X}_2$ along the channel dimension. Next, we

obtain the fused feature from $\hat{X}_1, \hat{X}_2$ by:

$$F_{pro} = \hat{X}_1 \odot S(W_1\hat{X}_1) + \hat{X}_2 \odot S(W_2\hat{X}_2), \tag{4}$$

where S denotes the Sigmoid function and W is the 3×3 2D convolution. We utilize $AvgPool$, $ResBlock$, $MaxPool$, and $Conv2D$ to better explore spatial information. The overall process can be described as follows:

$$\bar{Y}_{i+1} = F_{pro} \otimes upsample(P(F_{pro})), \tag{5}$$

where P represents the stack of $AvgPool$, $ResBlock$, $MaxPool$, $ResBlock$ and $Conv2D$. In the last, we employ 30 ResBlocks on $\bar{Y}_{i+1}$ to obtain Y_{i+1}.

3.3 Diffusion Model

Based on conditional denoising diffusion probabilistic models [11, 27, 31, 35], the DM involves a forward diffusion process and a reverse denoising process. We first adopt the Latent Encoder (LE) on ground truth video and the Condition Encoder (CE) on blurry video to generate the ground truth feature $z \in R^{T\times 4C'}$ and the condition feature $c \in R^{T\times 4C'}$. Then, we use the diffusion model to generate the prior feature $z' \in R^{T\times 4C'}$ using $c \in R^{T\times 4C'}$ as the condition and $z \in R^{T\times 4C'}$ as the target. We detail the diffusion process as follows.

In the forward diffusion process, we transform z into Gaussian noise by T iterations, which can be described as follows:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I), \tag{6}$$

where $t = 1, 2, ..., T$; $z_0 = z$; $\beta_t \in (0, 1)$ are hyperparameters that control the variance of the noise. Using the notation $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and through iterative derivation with reparameterization, Eq. (6) can be written as:

$$q(z_T|z_0) = \mathcal{N}(z_T; \sqrt{\bar{\alpha}_T}z_0, (1-\bar{\alpha}_T)I), \tag{7}$$

that is

$$z_T = \sqrt{\bar{\alpha}_T}z_0 + \sqrt{1-\bar{\alpha}_T}\epsilon. \tag{8}$$

In the reverse denoising process, DM sample a Gaussian random noise z_T and then gradually denoise z_T to z_0:

$$q(z_{t-1}|z_t, z_0) = \mathcal{N}(z_{t-1}; \mu_t(z_t, z_0), \tilde{\beta}_t I), \tag{9}$$

$$\mu_t(z_t, z_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}z_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}z_t. \tag{10}$$

We can use Eq. (8) to represent z_0 in Eq. (10) and obtain:

$$\mu_t(z_t, z_0) = \frac{1}{\sqrt{\alpha_t}}(z_t - \epsilon\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}). \tag{11}$$

We remove the variance estimation in Eq. (9):

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}(z_t - \epsilon\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}). \tag{12}$$

Then the noise ϵ is the only uncertain variable. Following previous work [11, 31, 38], we adopt a neural network to predict the noise ϵ on z_t, c and t.

3.4 Training and Inference Strategies

Inspired by [5, 38], we divide the training process into three stages: the first is to train the Latent Encoder with Wavelet-Aware Dynamic Transformer for ground truth feature extraction, the second is to train the Diffusion Model to generate the prior feature, and the third is to jointly train DM and WADT to generate artifact-free and high-realistic videos.

Training Stage One. The first stage aims to train the Latent Encoder (LE) to extract valuable ground truth prior feature z from ground truth (GT). We train LE and WADT as a whole model by inputting z into WADTL instead of z' and ignoring the condition encoder and the diffusion model in Fig. 1. We employ L1 loss and Multi-scale Frequency Reconstruction (MSFR) [10] loss. L1 loss is defined as:

$$L_1 = \sum_{i=1}^{N} \|V_{HQ}^i - V_{GT}^i\|_1, \tag{13}$$

where V_{HQ}^i is the i-th frame of the deblurred HQ video and V_{GT}^i is the i-th frame of the GT video. The MSFR loss measures the L1 distance between multi-scale GT and deblurred frame in the frequency domain as follows:

$$L_{MSFR} = \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{1}{t_k} \|\mathcal{FT}(\hat{S}_k^i) - \mathcal{FT}(S_k^i)\|_1, \tag{14}$$

where $\mathcal{FT}$ denotes the Fast Fourier Transform that transfers image signal to the frequency domain, K denotes the number of different scales and the $\hat{S}_k^i$ indicates the image signal of the i-th frame at the k-th scale. The final optimization function of stage one is determined as follows:

$$L_{deblur} = L_1 + \lambda L_{MSFR}, \tag{15}$$

where we experimentally set $\lambda = 0.1$.

Training Stage Two. In the second stage, we first use a Condition Encoder (CE), whose struct is the same as the Latent Encoder (LE), to generate condition feature c from the blur video. Then, we train a diffusion model (DM) to generate prior feature z' using condition feature c.

We run the complete T iteration reverse processes Eq. (12) to generate prior feature $\hat{z}$. We denoted this loss as:

$$L_{diff} = \|z' - \hat{z}\|_1. \tag{16}$$



Fig. 4: Visual comparison on GoPro [25] dataset. The deblurred results of previous work still contain significant blur effects. Our method generates much clearer frames.

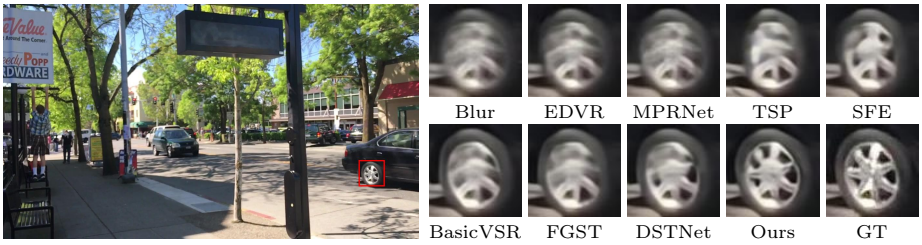


Fig. 5: Visual comparison on DVD [33] dataset. The deblurring effect of our proposed model is significantly better.

Training Stage Three. In Training Stage Two, we train the DM to generate z' to approximate z. However, there inevitably is a slight deviation between the prior features z' and the ground truth prior feature z. To make the DM and WADT work better together to acquire information from different frequency domains for video deblurring and improve the robustness of WADT to the prior feature z', we jointly train DM and WADT in Training Stage Three. The training loss L_{total} is given by the sum of the deblurring loss L_{deblur} and the diffusion loss L_{diff}:

$$L_{total} = L_{deblur} + L_{diff}. \tag{17}$$

Inference. Given a blurry input video $V_{blur} \in R^{T\times 3\times H\times W}$, we first compress V_{blur} into a condition feature $c \in R^{T\times 4C'}$. Then, the diffusion model generates the prior feature z' from pure Gaussian noise ϵ and condition feature c. Finally, VD-diff deblurs the blurry video with z' using WADT.

4 Experiments

4.1 Datasets

GoPro. The GoPro [25] benchmark is composed of over 3,300 blurry-sharp image pairs of dynamic scenes. It is obtained by a high-speed camera. The training and testing subsets are split in proportional to 2:1.

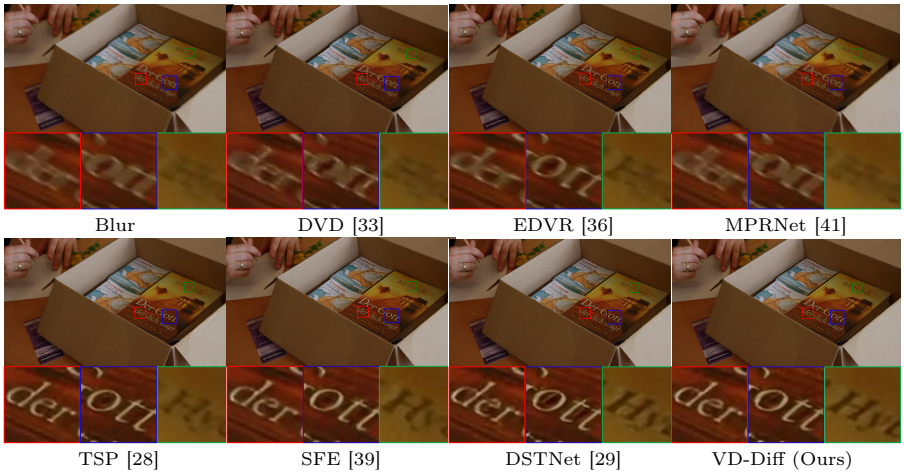


Fig. 6: Visual results of VD-Diff and SOTA methods on the real blurry videos of [33].

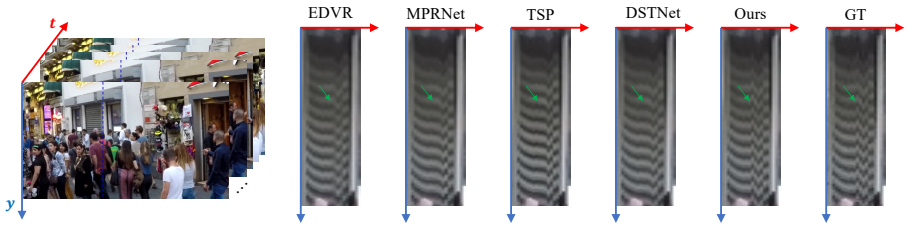


Fig. 7: Comparison of various methods in terms of time dimension.

DVD. The DVD [33] dataset consists of 71 videos with 6,708 blurry-sharp image pairs. It is divided into train/test subsets with 61 videos (5,708 image pairs) and 10 videos (1,000 image pairs). DVD is captured with mobile phones and DSLR at a frame rate of 240 FPS.
BSD. The Beam-Splitter Deblurring Dataset (BSD) [46] consists of 240 videos with 24k blurry-sharp image pairs. Cross-validation experiments between real-world and synthetic datasets demonstrate the high generality of the BSD dataset.
Real-World Videos. We evaluate the generality of our proposed approach on the real blurry datasets [6]. Due to the inaccessibility of the ground truth, we only compare VD-Diff with other models on visual results.

4.2 Implementation Details

We implement VD-Diff in PyTorch [30] using a single NVIDIA RTX A6000 GPU. The training process takes 7 days and 10 hours. We set the batch size as 2, the learning rate as 8×10^{-5} and weight decay as 4×10^{-5}. Patches at the size of 256×256 cropped from training frames are fed into the models. The sequence



Fig. 8: Qualitative comparison with pre-trained large diffusion models [31].

length is set to 15 in training and 16 in testing. Since we use diffusion within a highly compressed latent space, the denoising iteration steps T only needs 4. We set C to 64 and C' to 256, respectively. In the WADT module, we set N_1 and N_2 to 7 and 2, respectively, to balance the calculation amount and the deblurring effect. The horizontal and vertical flips are performed for data augmentation.

4.3 Quantitative Results

We compare our proposed VD-Diff with other SOTA methods, as shown in Tables 1, 2, and 3. As can be seen, our method outperforms SOTA methods on GoPro [25] and DVD [33] benchmark datasets. Specifically, as shown in Table 1, Our method outperforms DSTNet-L, RVRT, and FGST by 0.50dB, 0.63dB, and 2.65dB in PSNR on the GoPro [25] dataset, respectively. Our model achieves SOTA performance on the GoPro [25] dataset. Table 2 shows that our proposed method generates the deblurred videos with higher PSNR and SSIM values on the DVD [33] dataset. As shown in Table 3, we also achieve SOTA on the BSD [46] dataset. As seen from Table 4, our model has a relatively low parameter size and less inference time than other models. Although our model's parameter size and inference time are slightly higher than DSTNet and BasicVSR++, the PSNR metric on the GoPro [25] dataset is at least 1.3dB higher than the two models. As shown in Fig. 9(a), our model achieves favorable results in terms of accuracy and FLOPs. The above results show the effectiveness of our model.

4.4 Qualitative Results

We provide visual comparisons on GoPro [25], DVD [33], and Real-World Video datasets [33] as shown in Fig. 4, 5, and 6, respectively. Some models sacrifice

Method	DVD [33]	EDVR [36]	STFAN [47]	ESTRNN [46]	MPRNet [41]	TSP [28]
PSNR ↑	27.31	26.83	28.59	31.07	32.73	31.67
SSIM ↑	0.8255	0.8426	0.8608	0.9023	0.9366	0.9279
BasicVSR [3]	NAFNet [4]	FGST [22]	DSTNet [29]	DSTNet-L [29]	RVRT [21]	**Ours**
34.01	33.71	32.90	34.16	35.05	34.92	**35.55**
0.9520	0.9668	0.9610	0.9679	0.9733	0.9738	**0.9752**

Table 1: VD-Diff achieves SOTA results on the GoPro dataset [25].

Method	DVD [33]	EDVR [36]	STFAN [47]	ESTRNN [46]	TSP [28]	SEF [39]
PSNR ↑	30.01	28.51	31.15	32.01	32.13	31.71
SSIM ↑	0.8877	0.8637	0.9049	0.9162	0.9268	0.9160
GSTA [34]	MPRNet [41]	ARVo [16]	FGST [22]	DSTNet [29]	RVRT [21]	**Ours**
32.53	32.73	32.80	33.36	33.79	34.30	**34.56**
0.9468	0.9366	0.9352	0.9500	0.9615	0.9655	**0.9684**

Table 2: VD-Diff achieves SOTA results on the DVD dataset [33].

texture and details to maintain overall structure, resulting in over-smooth videos that lose high-frequency detail. Other models ignore global information to preserve high-frequency detail, introducing redundant speckle textures and artifacts that do not match the real video in fast-motion video. In contrast, our VD-Diff has the best visual effect as it can leverage WADT to model long-range dependencies to capture low-frequency information and employ DM to supplement valuable high-frequency information. Therefore, as seen from the visual comparison, our model can restore texture and details while maintaining the overall framework of the video.

Besides, we conduct a qualitative comparison in terms of the time dimension, as shown in Fig. 7. As can be seen, previous models such as EDVR [36], MPRNet [41], TSP [28], and DSTNet [29] all exhibit unclear outlines, indicating that they cannot guarantee temporal consistency. In contrast, our method effectively maintains temporal consistency.

Furthermore, we conduct a comparison with the direct application of a pre-trained large diffusion model (*i.e.*, LDM [31]) to video deblurring, as shown in Fig. 8. We notice that the videos deblurred by LDM contain a significant number of artifacts and inconsistencies with the content of the original videos, which is attributed to motion blur present in the blurred videos. In contrast, our method leverages the diffusion model and WADT to maximally restore the details in the video while preserving the video content.

4.5 Ablation Study

This section explores the significance of each key component of our proposed VD-Diff. All variants are tested on the GoPro [25] dataset.

Effect of Diffusion Model. To validate the effectiveness of the diffusion model, we design a variant that does not utilize DM to generate prior features

Method	DVD [33]	DTBN [14]	Nah et al. [26]	ESTRNN [46]	TSP [28]	DSTNet [29]	**Ours**
PSNR ↑	29.95	31.84	33.00	33.36	32.84	34.45	**35.61**
SSIM ↑	0.8692	0.9170	0.9330	0.9370	0.9398	0.9548	**0.9758**

Table 3: VD-Diff achieves SOTA results on the BSD dataset [46].

Method	DVD	EDVR	TSP	FGST	DSTNet	BasicVSR++	RVRT	DSTNet-L	**Ours**
Parameters (M)	15.30	23.60	16.19	9.70	7.45	9.46	13.60	16.96	11.93
Running time (s)	0.88	2.70	1.90	0.72	0.08	0.10	0.41	0.28	0.31

Table 4: Quantitative evaluations of the video deblurring methods regarding network parameters and running time. The running time is obtained on a machine with a single NVIDIA RTX A6000 GPU. The size of the test images is 1280 × 720 pixels.

Variants	Components				Metrics	
	Diffusion Model	Wavelet	Joint Training	WBPF	PSNR ↑	SSIM ↑
w/o DM	✗	✓	✓	✓	34.38	0.9684
w/o Wavelet	✓	✗	✓	✓	34.75	0.9709
w/o Joint Training	✓	✓	✗	✓	34.92	0.9717
w/o WBPF	✓	✓	✓	✗	33.58	0.9344
Full model	✓	✓	✓	✓	**35.55**	**0.9752**

Table 5: Ablation experiments on various variants of VD-Diif on the GoPro [25] dataset, and the best quantitative results are marked in **bold**.

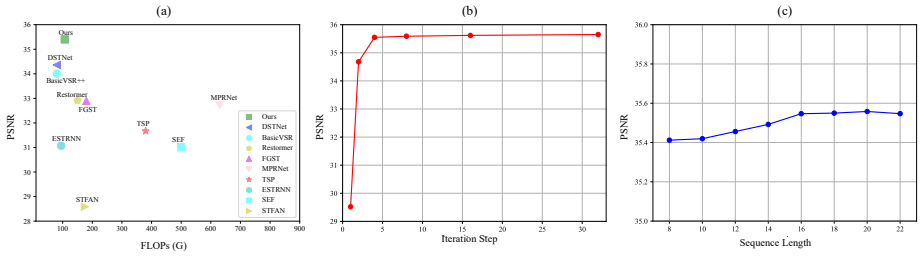


Fig. 9: (a) Floating point operations (FLOPs) *vs* video deblurring performance on the GoPro [25] dataset. Our model achieves favorable results in terms of accuracy and FLOPs. (b) Ablation study of the number of iteration steps. (c) Ablation study of the sequence length during inference.

but solely relies on WADT for video deblurring, named as *w/o* DM. As seen in Table 5, without employing DM, the deblurring performance of the network significantly declines, with a decrease of 1.17 in PSNR, which indicates that DM can provide valuable prior features to enhance the deblurring performance.

Effect of Wavelet Transform. We conduct ablation experiments to study the impact of wavelet transform on VD-Diff. We replace wavelet transform and inverse wavelet transform with bilinear downsample and upsample to evaluate

the effect of wavelet transform, named *w/o* Wavelet. We adopt the same training strategy as VD-Diff. Table 5 shows *w/o* Wavelet decreases by 0.8dB regarding the PSNR metrics. This experiment shows the wavelet transform method for feature extraction is better than the general upsampling and downsampling methods.

Effect of Joint Training Strategy. In Sec. 3.4, we analyze the importance of joint training. To verify our analysis, we remove joint training stage *i.e.*, Training Stage Three, as shown in Table 5. We notice that VD-Diff is significantly better than *w/o* Joint Training with an improvement of 0.63dB in terms of PSNR metrics, which demonstrates that joint training can improve the collaborative working ability of DM and WADT, thereby improving the deblurring effect.

Effect of Wavelet-based Bidirectional Propagation Fuse. To investigate the effect of Wavelet-based Bidirectional Propagation Fuse (WBPF), we design a variant model by removing the WBPF in WADT, termed as *w/o* WBPF. Table 5 shows that WBPF can effectively transmit useful information between frames, improving the model's ability to capture spatio-temporal information.

Effect of Iteration Steps. To explore the impact of the number of iterations on the diffusion model, we set different iteration steps to train the model and then compare the PSNR metrics on the GoPro dataset. The results are shown in Fig. 9(b). The value of PSNR increases as the number of iteration steps increases, but when the number of iterations is greater than 4, the value of PSNR almost remains stable. It shows that in a highly compressed latent space, just setting the number of iteration steps to 4 can generate suitable prior features.

Effect of Sequence Length. Intuitively, the sequence length will affect the deblurring effect of our proposed model. If the sequence length is too short, the model fails to capture essential information from distant frames. Conversely, if the sequence length is too long, the computational load becomes excessive, potentially causing error accumulation. To investigate this, we experimented with various sequence lengths. As shown in Fig. 9(c), the PSNR value increases steadily with the number of frames. However, after 16 frames, the calculation becomes very large as the sequence length increases, and the PSNR does not change significantly. Therefore, we set the sequence length to 16 to reach a near-optimal point between the deblurring effect and computational reduction.

5 Conclusion

In this paper, we propose a brand-new method that innovatively integrates the diffusion model with the Transformer for video deblurring. Specifically, the diffusion model generates prior features containing high-frequency information that conforms to the ground truth (GT) distribution. Meanwhile, the Wavelet-Aware Dynamic Transformer captures low-frequency information and utilizes the prior features generated by the diffusion model to supplement high-frequency information, thereby enhancing the quality of the deblurred video. Comprehensive experimental results show that our proposed method outperforms the state-of-the-art methods.

Acknowledgements

This work was supported in part by National Social Science Foundation Major Project "Research on Virtual Restoration of Tang and Song Painting Colors and Construction of Traditional Color Resource Library from a Digital Perspective" (19ZDA046), Zhejiang Province Program (2022C01222, 2023C03199, 2023C03201), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), Ningbo Science and Technology Plan Project (2022Z167, 2023Z137), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
3. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: On the generalization of basicvsr++ to video deblurring and denoising. arXiv preprint arXiv:2204.05308 (2022)
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European Conference on Computer Vision. pp. 17–33. Springer (2022)
5. Chen, Z., Zhang, Y., Liu, D., Xia, B., Gu, J., Kong, L., Yuan, X.: Hierarchical integration diffusion model for realistic image deblurring. arXiv preprint arXiv:2305.12966 (2023)
6. Cho, S., Wang, J., Lee, S.: Video deblurring for hand-held cameras using patch-based synthesis. ACM Transactions on Graphics (TOG) **31**(4), 1–9 (2012)
7. Chu, T., Chen, J., Sun, J., Lian, S., Wang, Z., Zuo, Z., Zhao, L., Xing, W., Lu, D.: Rethinking fast fourier convolution in image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23195–23205 (2023)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
10. Fan, Y., Hong, C., Wang, X., Zeng, Z., Guo, Z.: Multi-input-output fusion attention module for deblurring networks. In: 2021 IEEE international conference on big data (Big Data). pp. 3176–3182. IEEE (2021)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: Proceedings of the IEEE international conference on computer vision. pp. 1689–1697 (2017)
13. Hyun Kim, T., Mu Lee, K.: Generalized video deblurring for dynamic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5426–5434 (2015)

14. Hyun Kim, T., Mu Lee, K., Scholkopf, B., Hirsch, M.: Online video deblurring via dynamic temporal blending network. In: Proceedings of the IEEE international conference on computer vision. pp. 4038–4047 (2017)
15. Li, A., Zhao, L., Zuo, Z., Wang, Z., Xing, W., Lu, D.: Migt: Multi-modal image inpainting guided with text. Neurocomputing **520**, 376–385 (2023)
16. Li, D., Xu, C., Zhang, K., Yu, X., Zhong, Y., Ren, W., Suominen, H., Li, H.: Arvo: Learning all-range volumetric correspondence for video deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7721–7731 (2021)
17. Li, G., Lyu, J., Wang, C., Dou, Q., Qin, J.: Wavtrans: Synergizing wavelet and cross-attention transformer for multi-contrast mri super-resolution. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 463–473. Springer (2022)
18. Li, G., Rao, C., Mo, J., Zhang, Z., Xing, W., Zhao, L.: Rethinking diffusion model for multi-contrast mri super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11365–11374 (2024)
19. Li, W., Zhao, L., Lin, Z., Xu, D., Lu, D.: Non-local image inpainting using low-rank matrix completion. In: Computer Graphics Forum. vol. 34, pp. 111–122. Wiley Online Library (2015)
20. Li, W., Zhao, L., Xu, D., Lu, D.: Efficient image completion method based on alternating direction theory. In: 2013 IEEE International Conference on Image Processing. pp. 700–703. IEEE (2013)
21. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems **35**, 378–393 (2022)
22. Lin, J., Cai, Y., Hu, X., Wang, H., Yan, Y., Zou, X., Ding, H., Zhang, Y., Timofte, R., Van Gool, L.: Flow-guided sparse transformer for video deblurring. arXiv preprint arXiv:2201.01893 (2022)
23. Liu, X., Park, D.H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., Darrell, T.: More control for free! image synthesis with semantic diffusion guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 289–299 (January 2023)
24. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023)
25. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017)
26. Nah, S., Son, S., Lee, K.M.: Recurrent neural networks with intra-frame iterations for video deblurring. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8102–8111 (2019)
27. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
28. Pan, J., Bai, H., Tang, J.: Cascaded deep video deblurring using temporal sharpness prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3043–3051 (2020)

29. Pan, J., Xu, B., Dong, J., Ge, J., Tang, J.: Deep discriminative spatial and temporal network for efficient video deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22191–22200 (2023)
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
32. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10219–10228 (2023)
33. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1279–1288 (2017)
34. Suin, M., Rajagopalan, A.: Gated spatio-temporal attention-guided video deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7802–7811 (2021)
35. Tashiro, Y., Song, J., Song, Y., Ermon, S.: Csdi: Conditional score-based diffusion models for probabilistic time series imputation. Advances in Neural Information Processing Systems **34**, 24804–24816 (2021)
36. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
37. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17683–17693 (2022)
38. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. arXiv preprint arXiv:2303.09472 (2023)
39. Xiang, X., Wei, H., Pan, J.: Deep video deblurring using sharpness features from exemplars. IEEE Transactions on Image Processing **29**, 8976–8987 (2020)
40. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18456–18466 (June 2023)
41. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021)
42. Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W., Li, H.: Adversarial spatio-temporal learning for video deblurring. IEEE Transactions on Image Processing **28**(1), 291–301 (2018)
43. Zhang, Z., Zhang, Q., Xing, W., Li, G., Zhao, L., Sun, J., Lan, Z., Luan, J., Huang, Y., Lin, H.: Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7396–7404 (2024)
44. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D.: Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: Pro-

ceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5741–5750 (2020)
45. Zheng, B., Yuan, S., Slabaugh, G., Leonardis, A.: Image demoireing with learnable bandpass filters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3636–3645 (2020)
46. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B.: Efficient spatio-temporal recurrent neural network for video deblurring. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 191–207. Springer (2020)
47. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2482–2491 (2019)
48. Zuo, Z., Zhao, L., Li, A., Wang, Z., Zhang, Z., Chen, J., Xing, W., Lu, D.: Generative image inpainting with segmentation confusion adversarial training and contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3888–3896 (2023)