Learning Efficient Deep Discriminative Spatial and Temporal Networks for Video Deblurring

Jinshan Pan, Long Sun, Boming Xu, Jiangxin Dong, and Jinhui Tang

Abstract—How to effectively explore spatial and temporal information is important for video deblurring. In contrast to existing methods that directly align adjacent frames without discrimination, we develop a deep discriminative spatial and temporal network to facilitate the spatial and temporal feature exploration for better video deblurring. We first develop a channel-wise gated dynamic network to adaptively explore the spatial information. As adjacent frames usually contain different contents, directly stacking features of adjacent frames without discrimination may affect the latent clear frame restoration. Therefore, we develop a simple yet effective discriminative temporal feature fusion module to obtain useful temporal features for latent frame restoration. Moreover, to utilize the information from long-range frames, we develop a wavelet-based feature propagation method that takes the discriminative temporal feature fusion module as the basic unit to effectively propagate main structures from long-range frames for better video deblurring. Experimental results show that the proposed method performs favorably against state-of-the-art ones on benchmark datasets in terms of accuracy and model complexity.

Index Terms—Video deblurring, video restoration, deep discriminative learning, spatial and temporal feature fusion.

1 Introduction

With the rapid development of hand-held video capturing devices in our daily life, capturing high-quality clear videos becomes more and more important. However, due to the moving objects, camera shake, and depth variation during the exposure time, the captured videos usually contain significant blur effects. Thus, there is a great need to restore clear videos from blurred ones so that they can be pleasantly viewed on display devices and facilitate the following video understanding problems.

Different from single image deblurring that explores spatial information for blur removal, video deblurring is more challenging as it needs to model both spatial and temporal information. Conventional methods usually use optical flow [1], [2], [3], [4] to model the blur in videos and then jointly estimate optical flow and latent frames under the constraints by some assumed priors. As pointed out by [5], these methods usually lead to complex optimization problems that are difficult to solve. In addition, improper priors will significantly affect the quality of restored videos.

Instead of using assumed priors, lots of methods develop kinds of deep convolutional neural networks (CNNs) to explore spatial and temporal information for video deblurring. Several approaches stack adjacent frames as the input of CNN models [8] or employ spatial and temporal 3D convolution [9] for latent frame restoration. Gast et al. [10] show that using proper alignment strategies in deep CNNs would improve deblurring performance. To this end, several methods introduce alignment modules in deep neural networks. The commonly used alignment modules for video deblurring mainly include optical flow [5], deformable convolution [11], and so on. However, estimating alignment information from blurred adjacent frames is not a trivial task due to the influence of motion blur. In addition, using alignment modules usually leads to large deep CNN models that are difficult to train and computationally expensive. For example, the CDVDTSP method [5] with optical flow as the alignment module has 16.2 million parameters with FLOPs of 357.79G while the EDVR method [11]

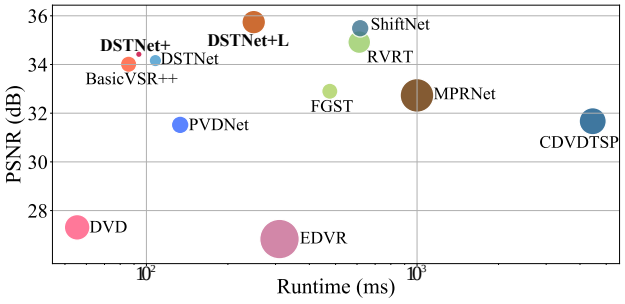


Fig. 1. Comparisons of the network parameters, running time and video deblurring performance on the GoPro dataset [6]. The spatial resolution of the videos is 720×1280 pixels. The area of the circle represents the amount of the network parameters. "DSTNet" denotes the method presented in our conference version [7]. "DSTNet+" denotes the improved proposed method while "DSTNet+L" denotes the large version of the "DSTNet+". Our model has the fewest network parameters and achieves better trade-off between accuracy and model complexity.

J. Pan, L. Sun, B. Xu, J. Dong, and J. Tang are with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China. E-mail: {jspan, cs.longsun, bomingxu1996, jxdong, jinhuitang}@njust.edu.cn.

using the deformable convolution as the alignment module has 23.6 million parameters with FLOPs of 2298.97G. Therefore, it is of great interest to develop a lightweight deep CNN model with lower computational costs to overcome the limitations of existing alignment methods in video deblurring while achieving better performance.

Note that most existing methods restore each clear frame based on limited local frames, where the temporal information from non-local frames is not fully explored. To overcome this problem, several methods employ recurrent neural networks to better model temporal information for video deblurring [12]. However, these methods have limited capacity to transfer the useful information temporally for latent frame restoration as demonstrated in [13]. To remedy this limitation, several methods recurrently propagate information of non-local frames with some proper attention mechanisms [13]. However, if the features of non-local frames are not estimated correctly, the errors will accumulate in the recurrent propagation process, which thus affects video deblurring. Moreover, exploring temporal information from non-local frames usually requires high computational costs, which cannot be applied to some resource-constrained devices. As the temporal information exploration is critical for video deblurring, it is of great need to develop an effective and efficient propagation method that can discriminatively propagate useful information from non-local frames for better video restoration.

In this paper, we develop an effective deep discriminative spatial and temporal network (DSTNet) to distinctively explore useful spatial and temporal information from videos for video deblurring. Motivated by the success of multi-layer perceptron (MLP) models that are able to model global contexts, we first develop a channel-wise gated dynamic network to effectively explore spatial information. In addition, to exploit the temporal information, instead of directly stacking estimated features from adjacent frames without discrimination, we develop a simple yet effective discriminative temporal feature fusion module to fuse the features generated by the channel-wise gated dynamic network so that more useful temporal features can be adaptively explored for video deblurring.

However, the proposed discriminative temporal feature fusion module does not utilize the information from long-range frames. Directly repeating this strategy in a recurrent manner is computationally expensive and may propagate and accumulate the estimation errors of features from long-range frames, leading to adverse effects on the final video deblurring. To solve this problem, we develop a wavelet-based feature propagation method that effectively propagates main structures from long-range frames for better video deblurring. Furthermore, the deep discriminative spatial and temporal network does not require additional alignment modules (e.g., optical flow used in [5], deformable convolution used in [11]) and is thus efficient yet effective for video deblurring as shown in Figure 1 (see results of DSTNet).

The main contributions are summarized as follows:

- We propose a channel-wise gated dynamic network (CWGDN) based on multi-layer perceptron (MLP) models to explore the spatial information. A detailed analysis demonstrates that the proposed CWGDN is more effective for video deblurring.
- We develop a simple yet effective discriminative temporal feature fusion (DTFF) module to explore useful temporal features for clear frame reconstruction.
- We develop a wavelet-based feature propagation (WaveletFP) method to efficiently propagate useful structures from long-range frames and avoid error accumulation for better video deblurring.
- We formulate the proposed network in an end-to-end trainable framework and show that it performs favorably against state-of-the-art methods in terms of accuracy and model complexity.

This manuscript extends its conference version [7] with the following major differences:

- We further analyze the property of CWGDN for spatial feature extraction and develop a lightweight and efficient deep discriminative spatial temporal feature fusion model that explores the properties of both CWGDN and DTFF to reduce the computational cost while achieving better deblurring performance.
- To reduce the computation cost when fusing temporal features, we develop an effective temporal feature fusion method to fuse the features of adjacent frames efficiently.
- We develop a simple yet effective progressive downsamping strategy to estimate dynamic filers effectively.
- We both qualitatively and quantitatively evaluate the proposed method against the method presented in our conference version [7] and state-of-the-art ones and show that it performs better in terms of accuracy and model complexity.

Figure 1 shows that our improved approach, named as DSTNet+, performs better than the DSTNet in terms of accuracy, efficiency and model complexity.

2 Related Work

Hand-crafted prior-based methods. Since the video deblurring problem is ill-posed, conventional methods usually make some assumptions [14], [15], [3], [16], [1], [17], [4] on the motion blur and latent clear frame to make this problem well-posed. However, these methods do not fully exploit the characteristics of clean image data and motion blur, and usually need to solve complicated inference problems.

Deep learning-based methods. Instead of using hand-craft priors, deep learning has been explored to solve video deblurring. In [8], Su et al. take the concatenation of the adjacent frames as the input of a CNN model based on an encoder and decoder architecture to solve video deblurring. Aittala and Durand [18] develop a permutation invariant CNN to solve multi-frame deblurring. However, simply stacking the adjacent frames does not explore the temporal information for video deblurring. In [9], Zhang et al. employ a spatial-temporal 3D convolution to utilize the spatial and temporal features for latent frame restoration. To better model the spatial and temporal information, several methods introduce alignment modules in the CNNs. In [11], Wang et al. use the deformable convolution to achieve the alignment. The optical flow estimation method is widely adopted [5], [19] to align the adjacent frames. Zhou et al. [20] develop an implicit alignment method based on the kernel prediction network [21]. Although using alignment methods improves the deblurring performance, it is not a trivial task to estimate alignment information from blurred adjacent frames due to the influence of motion blur. Moreover, the alignment module usually leads to large models, which are difficult to train.

Several methods aim to explore the properties of video frames for video deblurring. For example, Pan et al. [5] develop a temporal sharpness prior to better guide the network for video deblurring. Son et al. [22] aggregate information from multiple video frames by a blur-invariant motion estimation and pixel volumes. Suin and Rajagopalan [23] select the key frames to facilitate the blur removal. Wang et al. [24] detect the pixel-wise blur level of each frame for video deblurring.

To better utilize temporal information, several methods use recurrent neural networks (RNNs) to solve video deblurring. Wieschollek et al. [25] develop an effective RNN to recurrently use the features from the previous frame in multiple scales. In [12], Kim et al. develop a dynamic temporal blending network based on RNN to solve video deblurring. To better transfer the useful temporal information for latent frames restoration, Zhang et al. [13] develop an efficient spatio-temporal RNN with an attention mechanism for video deblurring. Recurrently propagating information from long-range frames improves the deblurring performance. However, if there exist inaccurately estimated features of long-range frames, the errors will be accumulated, which thus affects video deblurring.

Transformer-based methods. Recently, the Transformer and its variants have been applied to video deblurring. Lin et al. [26] develop an effective flow-guided sparse Transformer for video deblurring. Liang et al. [27] develop a recurrent video restoration transformer. In [28], Pan et al. employ Transformer to extract spatial and temporal features for better video restoration. Cao et al. [29] first use an encoder-decoder Transformer to estimate spatial features and then utilize the Transformer to fuse the spatial features. To reduce the computational cost, several approaches explore the Swin Transformer [30] to solve image restoration [31], [32] In [33], Zamir et al. develop an effective transposed self-attention to solve image restoration efficiently.

Although those above Transformer-based methods achieve decent performance, solving the Transformers needs huge computational costs.

3 Proposed Method

We aim to develop an effective and lightweight deep CNN model to discriminatively explore spatial and temporal information for video deblurring. To this end, we first develop a channel-wise gated dynamic network (CWGDN) to adaptively aggregate the spatial information and then propose a new discriminative temporal feature fusion (DTFF) module to fuse the features generated by the CWGDN so that we can distinctively select the most useful spatial and temporal features from adjacent frames for video deblurring. We further develop an effective wavelet-based feature propagation (WaveletFP) method that takes the DTFF module as the basic unit to better explore long-range information from video frames and avoids the error accumulation during the temporal feature propagation process. Figure 2 summarizes the key components of the proposed method. In the following, we explain the main ideas for each component in detail.

3.1 Channel-wise Gated Dynamic Network

Exploring spatial information is important for video deblurring. Recent methods have shown that using Transformers is able to explore better spatial features for image deblurring [33], [26], [27], [31]. However, they usually compute the self-attention from divided patches of the input features and do not effectively model information within and across patches for video deblurring. Moreover, self-attention estimation usually needs a huge computational cost. In contrast to these methods, we propose a channel-wise gated dynamic network (CWGDN) to discriminatively explore spatial features using a gated dynamic network for video deblurring. The CWGDN is motivated by the gMLP [34]. However, we estimate channel-wise dynamic filters from input features instead of using a static weight that is independent of the input features to generate a spatial gating unit.

Specifically, given N features $\{\mathbf{Y}_i\}_{i=1}^N$ extracted from the consecutive blurred frames $\{\mathrm{B}_i\}_{i=1}^N$, where $\mathbf{Y}_i \in \mathbb{R}^{H\times W\times C}$ has spatial dimension of $H \times W$ and

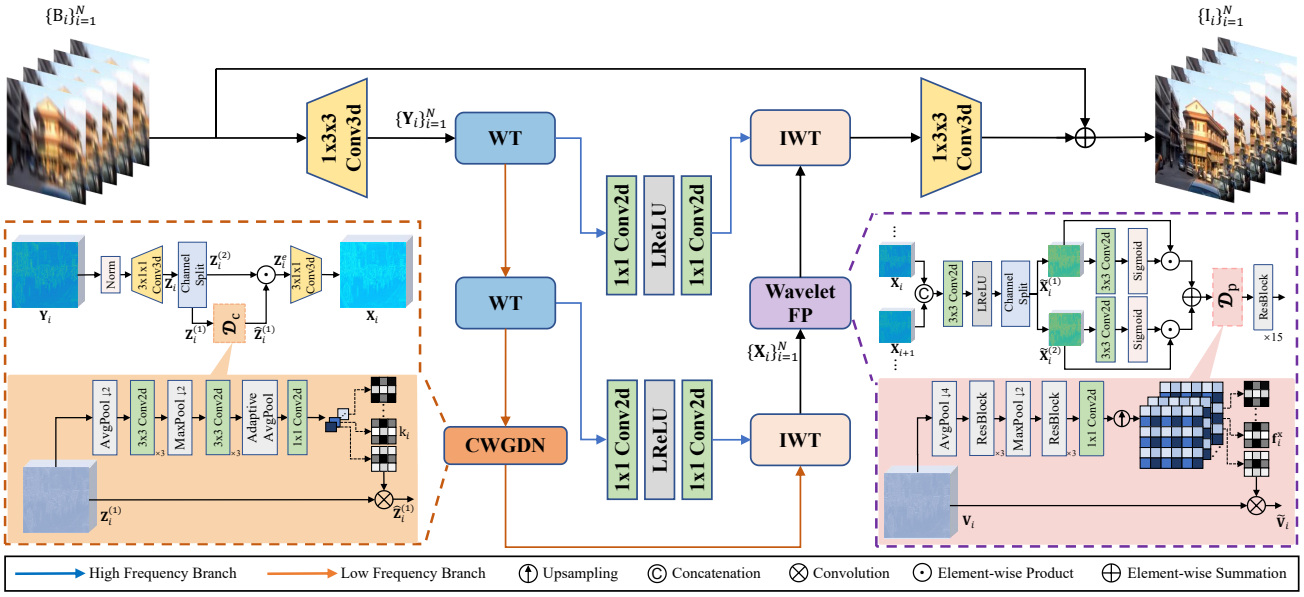


Fig. 2. An overview of the proposed network (DSTNet) for video deblurring. Given a blurred video $\mathbf{B} = \{\mathrm{B}_i\}_{i=1}^N$ with N frames, we first use a feature extraction module to extract features $\{\mathbf{Y}_i\}_{i=1}^N$ from $\mathbf{B}$, where $\mathbf{Y}_i \in \mathbb{R}^{H\times W\times C}$, $H \times W$ denotes the spatial dimension, and C is the number of channels. Then, we apply the channel-wise gated dynamic network (CWGDN) to the low-frequency part of $\{\mathbf{Y}_i\}_{i=1}^N$ by wavelet transformer and use the inverse wavelet transformer to obtain the reconstructed feature $\{\mathbf{X}_i\}_{i=1}^N$ as the input of the wavelet-based feature propagation (WaveletFP) that takes the discriminative temporal feature fusion (DTFF) module as the basic component. With the generated features. Finally, the latent frames are reconstructed based on the features by the wavelet-based feature propagation. "WT" and "IWT" denote the wavelet transform and inverse wavelet transform.

channel dimension of C, we first apply a 3D convolution with filter size of $3\times1\times1$ pixels to $\{\mathbf{Y}_i\}_{i=1}^N$ and obtain the features $\{\mathbf{Z}_i\}_{i=1}^N$, where $\mathbf{Z}_i \in \mathbb{R}^{H\times W\times 8C}$. Then, we split each feature $\mathbf{Z}_i$ into two independent parts $(\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)})$ along the channel dimension. For one of the splitted features, e.g., $\mathbf{Z}_i^{(1)} \in \mathbb{R}^{H\times W\times 4C}$, we develop a simple yet effective channel-wise dynamic network $\mathcal{D}_c$ (see Figure 2 for the detailed network architectures) to generate filters $\{\mathbf{k}_i\}_{i=1}^{4C}$ with the spatial size of $s_c \times s_c$ pixels. We apply the generated filter $\mathbf{k}_i$ to $\mathbf{Z}_i^{(1)}$ by:

$$\hat{\mathbf{Z}}_i^{(1)} = \mathbf{k}_i \otimes \mathbf{Z}_i^{(1)}, \tag{1}$$

where $\otimes$ denotes a convolution operation. Using $\hat{\mathbf{Z}}_i^{(1)}$ as the gate of $\mathbf{Z}_i^{(2)}$, we generate the enhanced features by:

$$\mathbf{Z}_i^{e} = \hat{\mathbf{Z}}_i^{(1)} \odot \mathbf{Z}_i^{(2)}, \tag{2}$$

where $\odot$ denotes the element-wise product operation.

Finally, we apply a 3D convolution operation with filter size of $3\times1\times1$ pixels and filter number of C to $\{\mathbf{Z}_i^e\}$ so that the output $\{\mathbf{X}_i\}_{i=1}^N$ of the CWGDN has the same channel dimension as the input $\{\mathbf{Y}_i\}_{i=1}^N$.

Different from [34], the generated filters $\mathbf{k}_i$ by the proposed method contain global information of the feature in each channel, which can discriminatively explore useful features to facilitate video deblurring. We provide detailed analysis in Section 6.

3.2 Discriminative temporal feature fusion

Given the generated features $\{\mathbf{X}_i\}_{i=1}^N$ by the CWGDN, exiting methods usually simply stack $\{\mathbf{X}_i\}_{i=1}^N$ or the alignment results of $\{\mathbf{X}_i\}_{i=1}^N$ according to some alignment methods for video deblurring. However, if the features $\{\mathbf{X}_i\}_{i=1}^N$ or the alignment results of $\{\mathbf{X}_i\}_{i=1}^N$ are not accurately estimated, directly stacking them would affect the latent frame restoration. Moreover, the contents of various frames are usually different, which may not facilitate the video deblurring. To this end, we propose a discriminative temporal feature fusion (DTFF) module to better explore mutually useful contents from the features of adjacent frames and reduce the influence of inaccurately estimated features. In the following, we first present the method of the fusion of $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ and then apply it to the fusion of $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$.

Specifically, we first apply a convolutional layer with LeakyReLU to the concatenation of $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ and obtain the feature $\tilde{\mathbf{X}}_i$ with the spatial dimension of $H\times W$ and channel dimension of $2C$. Then, we split $\tilde{\mathbf{X}}_i$ into two independent parts $(\tilde{\mathbf{X}}_i^{(1)}, \tilde{\mathbf{X}}_i^{(2)})$ along the channel dimension and obtain the fused feature by:

$$\mathbf{V}_i = \tilde{\mathbf{X}}_i^{(1)} \odot \mathcal{S}(\mathbf{W}_{2d}(\tilde{\mathbf{X}}_i^{(1)})) + \tilde{\mathbf{X}}_i^{(2)} \odot \mathcal{S}(\mathbf{W}_{2d}(\tilde{\mathbf{X}}_i^{(2)})), \tag{3}$$

where $\mathcal{S}$ denotes the Sigmoid function, and $\mathbf{W}_{2d}(\cdot)$ denotes a 2D convolution operation with filter size 3×3 pixels.

To better explore the spatial information of $\mathbf{V}_i$, we further develop a simple yet effective dynamic filter estimation network $\mathcal{D}_p$ (see Figure 2 for the detailed network architecture) to estimate pixel-wise filters from $\mathbf{V}_i$ and apply the estimated pixel-wise filters to $\mathbf{V}_i$:

$$\widetilde{\mathbf{V}}_i(\mathrm{x}) = \mathbf{f}_i^{\mathrm{x}} \otimes P(\mathbf{V}_i(\mathrm{x})), \tag{4}$$

where $\mathbf{f}_i^{\mathrm{x}}$ denotes the estimated filter at the pixel x with the size of $s_p \times s_p$ by $\mathcal{D}_p$, and $P(\mathbf{V}_i(\mathrm{x}))$ denotes a $s_p \times s_p$ patch centered at the pixel x.

Finally, we further employ a network with 15 ResBlocks to refine $\widetilde{\mathbf{V}}_i$ for latent frame restoration.

For simplicity, we use $\mathcal{F}$ to denote the above-mentioned discriminative temporal feature fusion module and refer to the fusion of $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ as the backward temporal feature fusion, which is denoted by:

$$\mathbf{F}_i^b = \mathcal{F}(\mathbf{X}_i, \mathbf{X}_{i+1}), \tag{5}$$

Similarly, the fusion of $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$ is referred to as the forward temporal feature fusion, which is denoted by:

$$\mathbf{F}_i^f = \mathcal{F}(\mathbf{X}_i, \mathbf{X}_{i-1}). \tag{6}$$

3.3 Wavelet-based feature propagation method

Note that the proposed DTFF module only considers two adjacent frames (i.e., $i-1$-th and $i+1$-th frames) when restoring the i-th latent frame, which does not fully explore the information from non-local frames. One straightforward solution is to use (5) and (6) recurrently, which has been also adopted in video deblurring [13] and video super-resolution [35]. However, exploring information from non-local frames requires the DTFF multiple times. If the features, especially the structural details, of non-local frames are not estimated accurately, the errors will be accumulated, which thus affects the video deblurring. Moreover, directly repeating the DTFF using original resolution features needs high computational cost. Thus, to avoid the influence of the inaccurate structural details and reduce computational cost, we develop a wavelet-based feature propagation (WaveletFP) method, which first propagates low-frequency parts of non-local frames and then applies the inverse wavelet transform to the propagated features and high-frequency parts to reconstruct better features for video deblurring.

Specifically, we first apply the Haar transform to the features $\{\mathbf{X}_i\}_{i=1}^N$ and obtain the low-frequency part ($\{\mathbf{X}_i^{LL}\}_{i=1}^N$) and the high-frequency part ($\{\mathbf{X}_i^{LH}\}_{i=1}^N$, $\{\mathbf{X}_i^{HL}\}_{i=1}^N$, and $\{\mathbf{X}_i^{HH}\}_{i=1}^N$).

For the low-frequency part ($\{\mathbf{X}_i^{LL}\}_{i=1}^N$), we adopt the bidirectional approach to propagate the main structures of both local and non-local frames. The backward and forward propagations are achieved by:

$$\begin{aligned} &\mathbf{F}_N^b = \mathbf{X}_N^{LL} \\ &\mathbf{F}_{i-1}^b = \mathcal{F}\left(\mathbf{X}_{i-1}^{LL}, \mathbf{F}_i^b\right), i = N, N-1, ..., 2, \end{aligned} \tag{7}$$

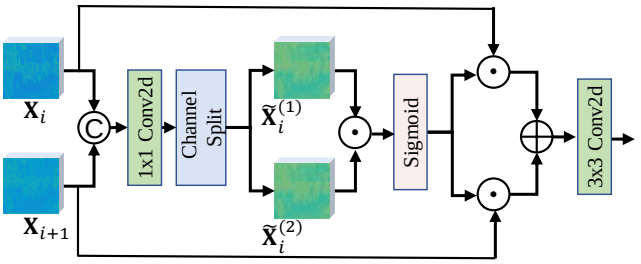


Fig. 3. Improved network architecture for the feature fusion of $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$.

and

$$\begin{aligned} &\mathbf{F}_1^f = \mathbf{F}_1^b \\ &\mathbf{F}_{i+1}^f = \mathcal{F}\left(\mathbf{F}_{i+1}^b, \mathbf{F}_i^f\right), i = 1, 2, ..., N-1 \end{aligned} \tag{8}$$

We then reconstruct the feature by an inverse Haar transform:

$$\widetilde{\mathbf{F}}_i^f = \mathcal{H}^{-1}\left(\mathbf{F}_i^f; \mathcal{N}_h(\mathbf{X}_i^{LH}), \mathcal{N}_h(\mathbf{X}_i^{HL}), \mathcal{N}_h(\mathbf{X}_i^{HH})\right), \tag{9}$$

where $\mathcal{H}^{-1}(\cdot)$ denotes the inverse Haar transform; $\mathcal{N}_h(\cdot)$ is a network containing two convolutional layers with the LeakyReLU in between.

Finally, we restore the latent frame by:

$$\mathrm{I}_i = \mathcal{N}_f\left(\widetilde{\mathbf{F}}_i^f\right) + \mathrm{B}_i, i = 1, 2, ..., N. \tag{10}$$

where $\mathcal{N}_f$ denotes a network with one convolutional layer.

In addition to generating better features for latent frame restoration, the WaveletFP method further reduces the computational cost as the feature propagation is mainly applied to the low-frequency part. We will show its effectiveness and efficiency in Section 6 and supplemental material.

4 Towards Fast and Lightweight Network Designs

Although CWGDN is able to extract features for better video deblurring, it still requires significant computational costs. We note that both CWGDN and DTFF involves a dynamic filter estimation network to adaptively explore spatial information. To simplify the proposed network, we further propose a unified deep discriminative spatial temporal feature fusion model that explores the properties of both CWGDN and DTFF for video deblurring.

We note that the feature fusion by (3) requires significant computational costs as concatenating $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ doubles the channel dimension. To reduce the computational cost and network parameters, we first use one 1×1 convolution without non-linear activation functions instead of using the original 3×3 convolution with LeakyReLU to obtain $\tilde{\mathbf{X}}_i$. We then split $\tilde{\mathbf{X}}_i$ into two independent parts ($\tilde{\mathbf{X}}_i^{(1)}$, $\tilde{\mathbf{X}}_i^{(2)}$) along

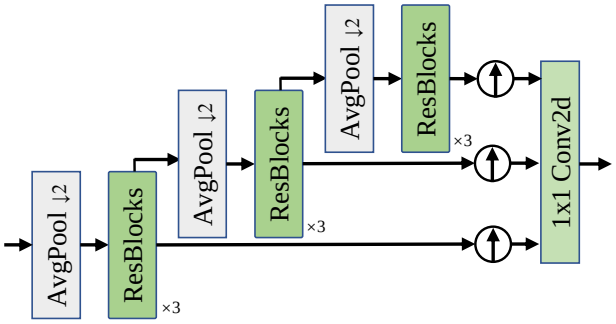


Fig. 4. Improved network architecture for the dynamic filter estimation network $\mathcal{D}_p$.

TABLE 1
Quantitative evaluations on the GoPro dataset [6]. The FLOPs is tested on images with size of 256×256 pixels.

Methods	PSNRs	SSIMs	Parameters (M)	FLOPs (G)
SRN [36]	30.29	0.9014	6.80	1435.0
DVD [8]	27.31	0.8255	15.31	39.69
Wieschollek et al. [25]	25.19	0.7794	-	-
DTBN [12]	26.82	0.8245	-	-
Nah et al. [37]	29.97	0.8497	-	-
EDVR [11]	26.83	0.8426	23.60	194.53
STFAN [20]	28.59	0.8608	5.37	35.40
DVDSEF [38]	31.01	0.9130	75.05	493.97
DUF [39]	28.01	0.8768	5.82	270.46
ESTRNN [13]	31.07	0.9023	2.22	10.46
MPRNet [40]	32.73	0.9366	20.13	760.43
CDVDTSP [5]	31.67	0.9279	16.19	357.46
NAFNet [41]	33.71	0.9668	67.89	63.05
PVDNet [42]	31.52	0.9210	10.51	43.04
BasicVSR++ [43]	34.01	0.9520	9.76	37.69
FGST [26]	32.90	0.9610	9.68	161.94
RVRT [27]	34.92	0.9738	13.57	559.45
CDVDTSPNL [28]	33.00	0.9424	5.49	285.02
ShiftNet [44]	35.49	0.9760	10.51	146.50
DSTNet	34.16	0.9679	7.45	47.87
DSTNet+	34.42	0.9694	3.90	44.61
DSTNet+L	**35.71**	**0.9766**	14.09	158.14

the channel dimension and use one of the splitted features as the gate to keep the most useful information:

$$\hat{\mathbf{X}}_i^e = \hat{\mathbf{X}}_i^{(1)} \odot \mathbf{X}_i^{(2)}, \tag{11}$$

We further apply the Sigmoid function to $\hat{\mathbf{X}}_i^e$ and obtain the improved fused feature by:

$$\mathbf{V}_i = \mathbf{W}_{2d}\left(\mathbf{X}_i^{(1)} \odot \mathcal{S}(\hat{\mathbf{X}}_i^e) + \mathbf{X}_{i+1}^{(1)} \odot (1 - \mathcal{S}(\hat{\mathbf{X}}_i^e))\right), \tag{12}$$

Compared with the feature fusion in (3), the proposed method (12) only requires one 3×3 convolution operation, which simplify the network significantly.

For the dynamic filter estimation network, we note that directly applying downsampling operation with a scale factor of 4 to the input features will lead to significant information loss. To overcome this problem, we adopt a progressive downsamping strategy as shown in Figure 4. Finally, we use the upsampling operation followed by a concatenation with a 1×1 convolution to generate filters. The network details are shown in Figure 4.

TABLE 2
Quantitative evaluations on the DVD dataset [8] in terms of PSNR and SSIM.

Methods	PSNRs	SSIMs
Kim and Lee [1]	26.94	0.8158
Gong et al. [45]	28.27	0.8463
SRN [36]	29.98	0.8842
DVD [8]	30.01	0.8877
DTBN [12]	29.95	0.8692
EDVR [11]	28.51	0.8637
STFAN [20]	31.15	0.9049
DVDSEF [38]	31.71	0.9159
ESTRNN [13]	32.01	0.9162
MPRNet [40]	32.24	0.9253
CDVDTSP [5]	32.13	0.9268
GSTA [23]	32.53	0.9468
PVDNet [42]	32.31	0.9260
FGST [26]	33.36	0.9500
RVRT [27]	34.30	0.9655
CDVDTSPNL [5]	33.71	0.9668
DSTNet	33.79	0.9615
DSTNet+	33.93	0.9618
DSTNet+L	**34.63**	**0.9669**

5 Experimental Results

In this section, we evaluate the effectiveness of the proposed approach and compare it with state-of-the-art methods using public benchmark datasets.

5.1 Datasets and parameter settings

Video deblurring datasets. We evaluate our method on the commonly used video deblurring datasets including the GoPro dataset [6], the DVD dataset [8], and the real-world dataset (BSD) [13], and follow the protocols of these benchmarks for training and test. To evaluate the effect of the proposed network, we adopt the commonly used PSNR and SSIM as the evaluation metrics.

Parameter settings. We implement our method based on the PyTorch and train it from scratch using a machine with 8 NVIDIA GeForce RTX 3090 GPUs. We crop image patches with the spatial size of 256×256 pixels for training. The batch size is set to be 16. We use the AdamW optimizer [46] with default parameter settings as the optimizer. The number of iterations is set to be 600,000. The feature number C is set to be 64. The filter sizes s_c and s_p are set to be 3 empirically. The learning rate is initialized to be 2×10^{-4} and is updated by the Cosine Annealing scheme. We use the same loss function as [47] to constrain the network training. The detailed network architectures of the proposed method and more experimental results are included in the supplemental material. The training code and models are available on our project website: https://github.com/sunny2109/DSTNet-plus/.

In the following, we denote the method presented in the conference version [7] as DSTNet and the improved one as DSTNet+. Similar to [7], we also train a larger model of the DSTNet+ by increasing feature channel numbers and residual blocks of DSTNet+, where we denote this large model as DSTNet+L.

(a) Blurred frame (b) GT (c) DVD [8] (d) EDVR [11]

(e) STFAN [20] (f) ESTRNN [13] (g) CDVDTSP [5] (h) FGST [26]

(i) RVRT [27] (j) DSTNet (k) DSTNet+ (l) DSTNet+L

Fig. 5. Deblurred results on the GoPro dataset [6]. The deblurred results in (c)-(i) still contain blur effects, e.g., the characters and the cars. The proposed method generates much clearer frames as shown in (j)-(l).

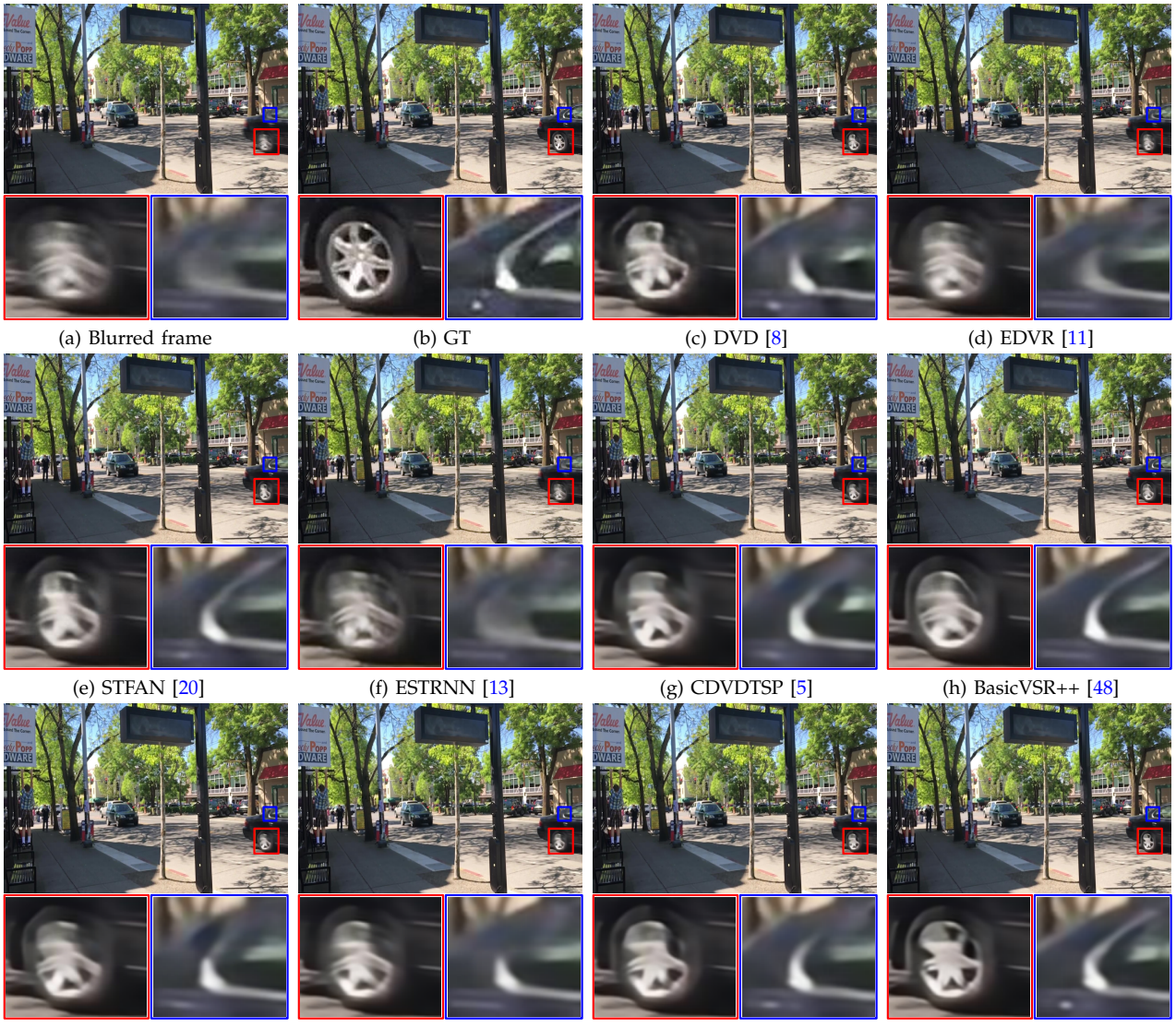
5.2 Comparisons with the state of the art

Evaluations on the GoPro dataset. The GoPro dataset contains 11 videos for test. For fair comparisons, we retrain the deep learning-based methods that are not trained on this dataset using the same protocols. We also compare our method with the video super-resolution method [39] (DUF for short) as it uses the dynamic filters that are similar to $\mathcal{D}_p$ in the DTFF module.

Table 1 summarizes the quantitative evaluation results, where the proposed approach generates high-quality videos with higher PSNR and SSIM values. Although recent proposed method, e.g., ShiftNet [44] performs better than the DSTNet and DSTNet+, our network has fewer parameters and FLOPs values. Also note that our method performs better than ShiftNet [44] when using the similar amounts of parameters (i.e., DSTNet+L in Table 1).

Figure 5 shows some visual comparisons of the evaluated methods on the GoPro dataset. The method [8] concatenates the consecutive frames as the input and does not effectively remove blur (Figure 5(c)), indicating that directly stacking the consecutive frames does not effectively explore the spatial-temporal information for video deblurring. In addition, exploring adjacent frames using alignment methods without discrimination, e.g., [5], [11], does not restore clear frames as the inaccurate alignment will interfere with the deblurring process (Figure 5(d) and (g)). The STFAN method [20] develops an end-to-end-trainable spatial-temporal filter adaptive network to deblur videos. However, this method does not effectively explore temporal information from non-local frames, and the deblurred frame still contains significant blur residual (Figure 5(e)). The method [13] develops an effective spatial and temporal RNN to explore spatial and temporal information from both local and non-local frames. Yet, the temporal information aggregation process does not effectively reduce the influences of inaccurately estimated features, which thus interferes with the final latent frame restoration (see Figure 5(f)). The methods [26], [27] employ Transformers with optical flow as guidance to model spatial and temporal information for video deblurring. However, the deblurred images in Figure 5(h) and (i) still contains blur effects. For example, the characters and boundaries of the cars are not recovered well.

In contrast, our deep discriminative feature propagation network is able to explore spatial and temporal information to reduce the influence of the inaccurately estimated features from non-local frames, and gener-



(a) Blurred frame (b) GT (c) DVD [8] (d) EDVR [11]

(e) STFAN [20] (f) ESTRNN [13] (g) CDVDTSP [5] (h) BasicVSR++ [48]

(i) FGST [26] (j) CDVDTSPNL [28] (k) DSTNet (l) DSTNet+

Fig. 6. Deblurred results on the DVD dataset [8]. The proposed method generates much clearer frames with better wheels and windows of the car as shown in (l).

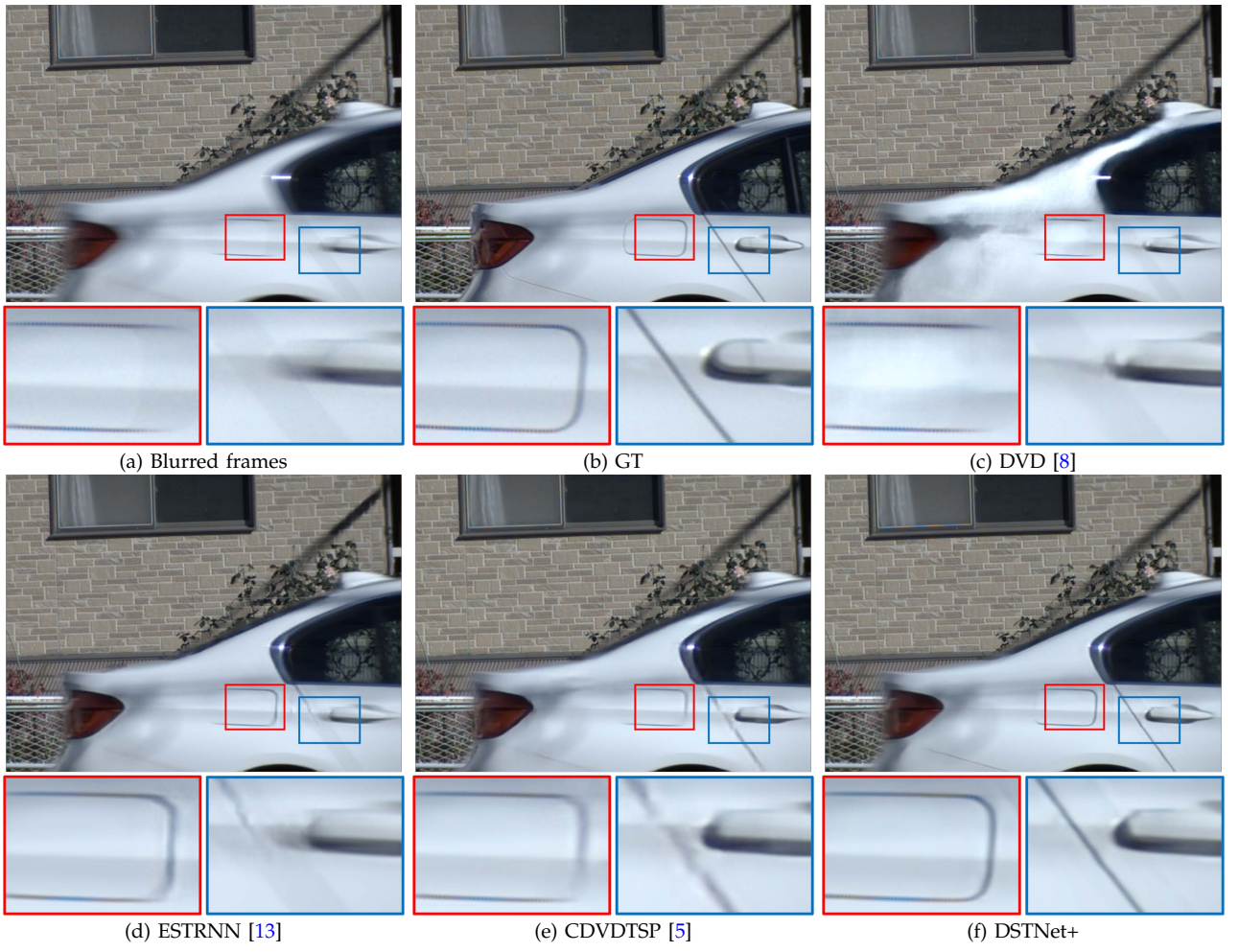
TABLE 3
Quantitative evaluations on the BSD deblurring dataset in terms of PSNR and SSIM.

	DVD [8]	STFAN [20]	ESTRNN [13]	CDVDTSP [5]	BasicVSR++ [43]	ShiftNet [44]	DSTNet [7]	DSTNet+	DSTNet+L
1ms-8ms	33.22 0.9350	32.78 0.9220	33.36 0.9370	33.54 0.9420	34.26 0.9568	34.43 0.9586	34.45 0.9548	34.37 0.9549	**35.18** **0.9618**
2ms-16ms	31.75 0.9220	32.19 0.9190	31.95 0.9250	32.16 0.9260	32.99 0.9526	**33.76** **0.9582**	32.86 0.9481	32.92 0.9502	33.75 0.9581
3ms-24ms	31.21 0.9220	29.47 0.8720	31.39 0.9260	31.58 0.9260	32.70 0.9453	33.41 **0.9529**	32.56 0.9419	32.76 0.9446	**33.48** 0.9503
Average	32.06 0.9263	31.48 0.9043	32.24 0.9293	32.43 0.9313	33.32 0.9516	33.87 0.9566	33.29 0.9483	33.35 0.9499	**34.14** **0.9567**

ates a clearer frame with better details and structures than the state-of-the-art methods. For example, the numbers and boundaries of the cars are close to the ground truth ones (Figure 5(k) and (l)). We note the restored frame by the DSTNet+ is better than the one by the DSTNet. However, the model size of the DSTNet+ is 50% smaller than that of DSTNet, suggesting the efficiency and effectiveness of the proposed DSTNet+.

Evaluations on the DVD dataset. We further evaluate our method on the DVD dataset by Su et al. [8]. Table 2 shows that the proposed method generates the deblurred videos with higher PSNR and SSIM values. Figure 6 shows some visual comparisons of the evaluated methods, where our method generates clearer frames.

Evaluations on the BSD deblurring dataset. As the BSD dataset [13] is a commonly used benchmark for



(a) Blurred frames (b) GT (c) DVD [8]

(d) ESTRNN [13] (e) CDVDTSP [5] (f) DSTNet+

Fig. 7. Qualitative comparisons on the BSD dataset [13]. The structures by the ESTRNN [13] and CDVDTSP [5] are not restored well. In contrast, the proposed method generates much clearer frames, which are visually close to the ground truth ones.

TABLE 4
Quantitative evaluations of the state-of-the-art video denoising methods on the Set8 benchmark dataset [49].

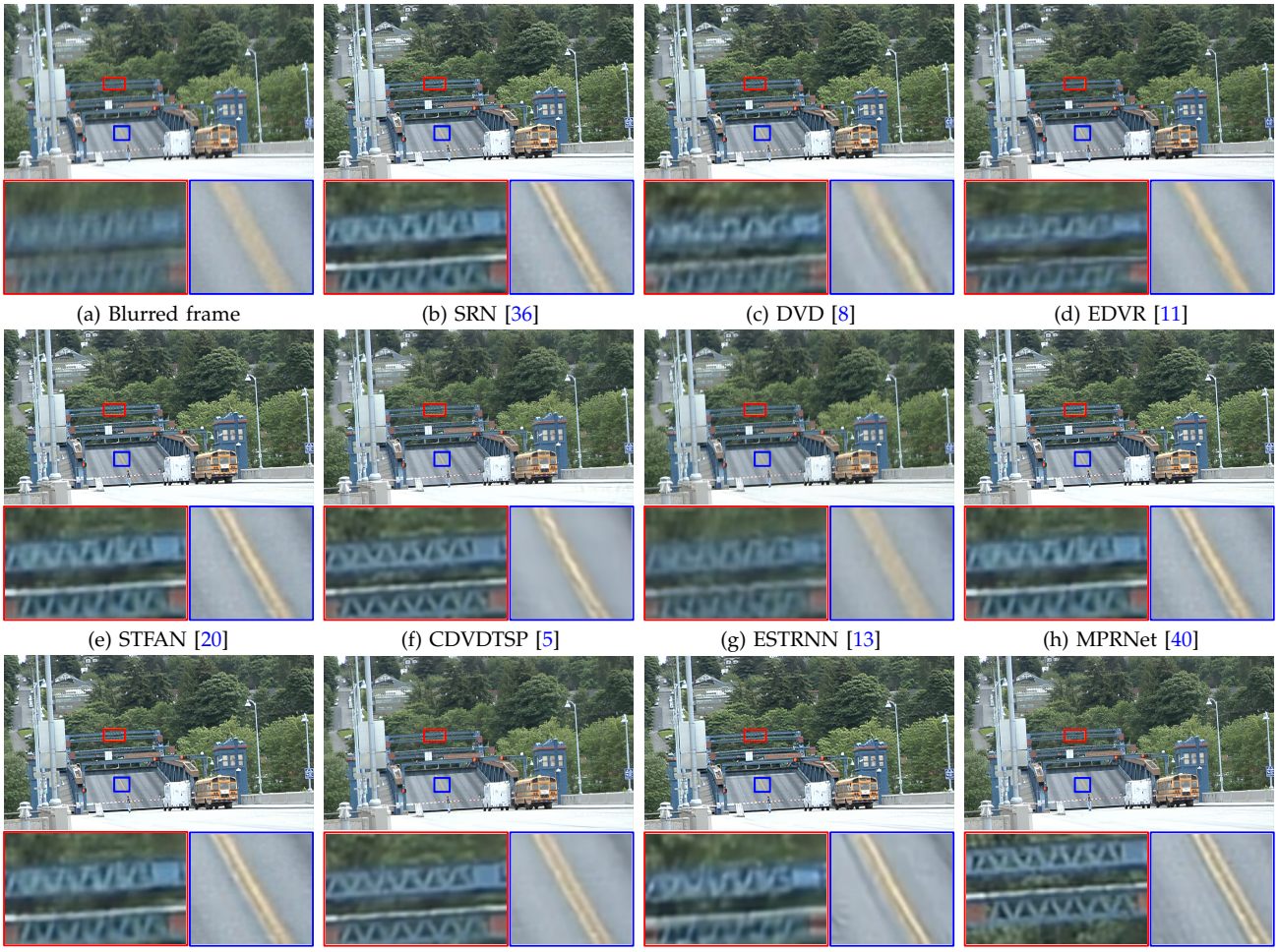
Noise level	DVDNet [49]	FastDVDNet [50]	PaCNet [51]	BasicVSR++ [43]	RVRT [27]	ShiftNet+ [44]	DSTNet+L
10%	36.08/0.9510	36.44/0.9540	37.06/0.9590	36.83/0.9574	37.52/**0.9619**	**37.54**/0.9588	37.28/0.9584
20%	33.49/0.9182	33.43/0.9196	33.94/0.9247	34.15/0.9319	34.79/0.9372	34.91/**0.9379**	**35.03**/0.9374
30%	31.68/0.8862	31.68/0.8889	32.05/0.8921	32.57/0.9095	33.23/0.9155	33.35/0.9166	**33.64**/**0.9186**
40%	30.46/0.8564	30.46/0.8608	30.70/0.8623	31.42/0.8889	32.12/0.8956	32.24/0.8971	**32.62**/**0.9012**
50%	29.53/0.8289	29.53/0.8351	29.66/0.8349	30.49/0.8692	31.22/0.8770	31.42/0.8712	**31.80**/**0.8850**
Average	32.29/0.8881	32.31/0.8917	32.68/0.8946	33.09/0.9114	33.78/0.9174	33.89/0.9163	**34.07**/**0.9201**

video deblurring, we evaluate our method against state-of-the-art ones based on the protocols of [13][1]. Our method performs well on the BSD deblurring dataset as shown in Table 3. Figure 7 shows that the proposed method generates high-quality deblurred frames.

Evaluations on real captured videos. Similar to the existing method [5], we further evaluate our method using real captured videos by Cho et al. [14] and compare the proposed method with state-of-the-art video deblurring methods. Figure 8 shows that the competed methods do not restore the sharp frames well. In contrast, our method generates much clearer frames, where the bridge structures and lines of road are recognizable.

More applications. To examine whether the proposed method can handle other related video restoration problems, we evaluate it on video denoising following the protocols of [27], [44] and compare it with state-of-the-art video denoising methods in Table 4. The proposed method generates favorable results on the video denoising task.

1. We use the batch size as 8 on the BSDS dataset and enlarge the patch size from 256×256 pixels to 400×400 pixels after 600,000 iterations during the training.



(a) Blurred frame (b) SRN [36] (c) DVD [8] (d) EDVR [11]

(e) STFAN [20] (f) CDVDTSP [5] (g) ESTRNN [13] (h) MPRNet [40]

(e) BasicVSR++ [48] (f) CDVDTSPNL [28] (g) DSTNet (h) DSTNet+

Fig. 8. Deblurred results on a real video from [14]. Our method (i.e., DSTNet+) recovers a clearer image with better detailed structures (e.g., bridge structures and lines of road).

6 Analysis and Discussions

To better understand how our method solves video deblurring and demonstrate the effect of its main components, we provide deeper analysis on the proposed method. For the ablation studies in this section, we train the proposed method and all the baselines on the GoPro dataset with 300,000 iterations for fair comparisons.

Effectiveness of CWGDN. The proposed CWGDN is used to explore spatial information for clear frame restoration. To examine whether it facilitates video deblurring, we remove CWGDN from the proposed method and train this baseline using the same settings as ours on the GoPro dataset for fair comparisons.

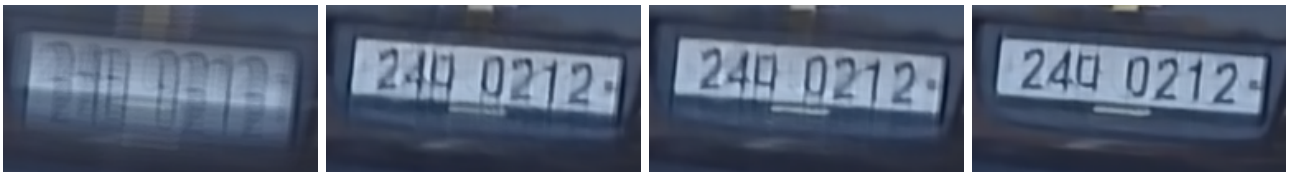
Table 5 shows that using the CWGDN generates better-deblurred frames with higher PSNR and SSIM values, where the average PSNR value of the proposed method is at least 0.15dB higher than the baseline method without using the CWGDN.

As the proposed CWGDN is motivated by gMLP [34], where we estimate channel-wise filters to generate spatial gating units, one may wonder whether deirectly using gMLP generates better results or not. To answer this question, we replace the CWGDN with gMLP in the proposed network and train this baseline method using the same settings as ours for fairness. Table 5 shows that using the original gMLP does not generate favorable results, indicating that using the filters generated by our method can obtain better features for video deblurring.

In addition, Figure 9(b) and (c) shows that the baseline using gMLP and the one without using the CWGDN do not restore clear frames, while the proposed approach using the CWGDN generates much clearer frames (Figure 9(d)).

Effectiveness of DTFF. The proposed DTFF module is used to better explore mutually useful contents and reduce the influences of inaccurately estimated features from adjacent frames. As the proposed DTFF module mainly contains a gated feature fusion (2) and a local spatial information exploration by pixel-wise filters (4), we conduct ablation studies w.r.t. these components to demonstrate their effectiveness on video deblurring.

The gated feature fusion (2) is used to keep the most



(a) Blurred patch (b) DSTNet$_{\text{w/ gMLP [34]}}$ (c) DSTNet$_{\text{w/o CWGDN}}$ (d) DSTNet

Fig. 9. Effectiveness of the CWGDN for video deblurring. Using the gMLP or without using the CWGDN does not restore clear images well.

(a) Blurred patch (b) DSTNet$_{\text{w/ the feature fusion [52]}}$ (c) DSTNet$_{\text{w/o (3)}}$

(d) DSTNet$_{\text{w/o (4)}}$ (e) DSTNet (f) GT

Fig. 10. Effectiveness of the DTFF module for video deblurring. Using the proposed DTFF module is able to remove blur well.

TABLE 5
Effect of the proposed CWGDN on the GoPro dataset. DSTNet$_{\text{w/o CWGDN}}$ denotes the baseline that removes the CWGDN module from DSTNet; DSTNet$_{\text{w/ gMLP [34]}}$ indicates the baseline that replaces CWGDN with gMLP [34].

Methods	DSTNet$_{\text{w/o CWGDN}}$	DSTNet$_{\text{w/ gMLP [34]}}$	DSTNet
PSNRs	33.18	33.16	**33.33**
SSIMs	0.9602	0.9602	**0.9611**
Parameters (M)	3.67	7.73	7.45

useful features from adjacent frames while reducing the influence of inaccurate features from adjacent frames. The results in Table 6 show that our method using (2) obtains better-deblurred results. Note that Park et al. [52] develop an adaptive blending module to fuse the temporal features for video deblurring. However, this method applies the learned weights to the input features $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$, i.e.,

$$\mathbf{V}_i = \mathbf{X}_i \odot \mathcal{S}(\mathbf{W}_{2d}(\tilde{\mathbf{X}}_i^{(1)})) + \mathbf{X}_{i+1} \odot \mathcal{S}(\mathbf{W}_{2d}(\tilde{\mathbf{X}}_i^{(2)})). \quad (13)$$

This fusion method does not generate better results as shown in Table 6.

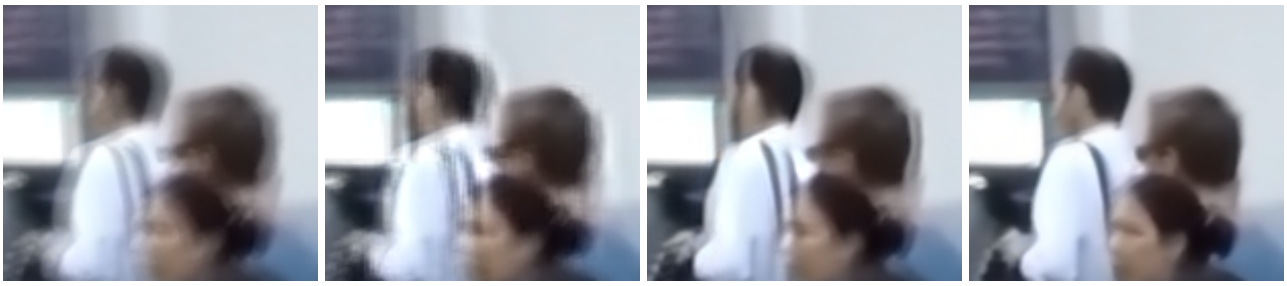
In addition, as shown in Table 6, the PSNR value of our method using the local spatial information exploration by pixel-wise filters (4) is at least 0.78dB higher than that of the baseline without using (4), suggesting that estimating pixel-wise filters further facilitates better estimations of the features for latent clear frame restoration. Figure 10 further demonstrates that using the discriminative temporal feature fusion module is able to facilitate blur removal.

One may wonder whether the performance gains of DTFF are due to the use of a large capacity model. To answer this question, we replace DTFF with a commonly used concatenation followed by 25 ResBlocks so that this baseline model has the similar amount of network parameters to the proposed DSTNet. Table 6 shows that the proposed DSTNet method using the DTFF module still performs better.

Effectiveness of WaveletFP. The proposed WaveletFP is mainly used to avoid the influence of the inaccurate structural details from non-local frames during the feature propagation process. To demonstrate the effectiveness of this module, we further compare with the method without using WaveletFP and train this baseline using the same settings as the proposed

TABLE 6
Effect of the proposed DTFF on the GoPro dataset. $\text{DSTNet}_{\text{w/ the feature fusion [52]}}$ and $\text{DSTNet}_{\text{w/ concatenation\&25 ResBlocks}}$ denote the baseline methods that respectively replace DTFF with the feature fusion [52] and a commonly used concatenation followed by 25 ResBlocks; $\text{DSTNet}_{\text{w/o (3)}}$ and $\text{DSTNet}_{\text{w/o (4)}}$ are the baseline methods that repsectively remove (3) and (4).

Methods	$\text{DSTNet}_{\text{w/ the feature fusion [52]}}$	$\text{DSTNet}_{\text{w/o (3)}}$	$\text{DSTNet}_{\text{w/o (4)}}$	$\text{DSTNet}_{\text{w/ concatenation\&25 ResBlocks}}$	DSTNet
PSNRs	33.00	33.22	32.55	33.08	**33.33**
SSIMs	0.9597	0.9608	0.9504	0.9593	**0.9611**
Parameters (M)	6.63	7.00	6.49	7.54	7.45



(a) Blurred patch (b) $\text{DSTNet}_{\text{w/o WaveletFP}}$ (c) $\text{DSTNet}_{\text{w/ Bilinear in FP}}$ (d) DSTNet

Fig. 11. Effectiveness of the WaveletFP for video deblurring. The method without using WaveletFP or using the Bilinear downsampling and upsampling operations instead of Wavelet transforms does not remove blur well.

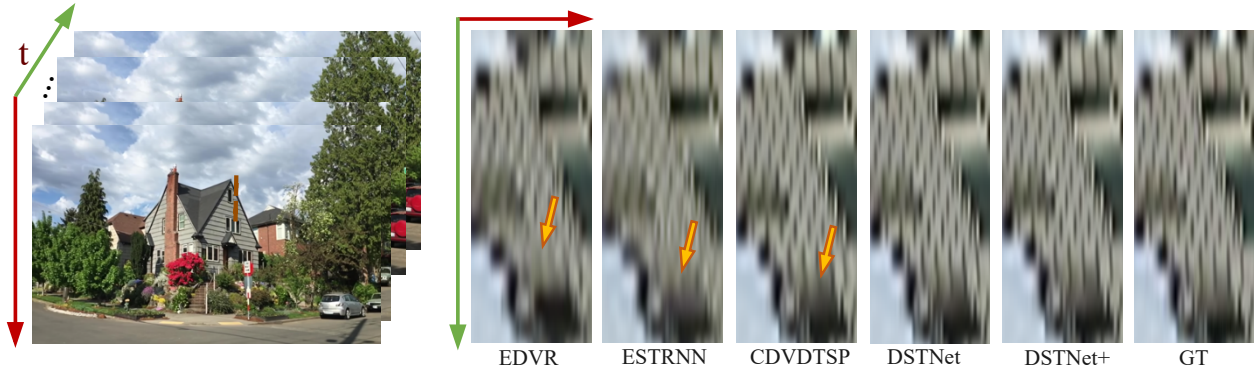


Fig. 12. Visual comparisons of the temporal consistency for restored videos. We visualize the pixels of the selected columns (the dotted line) according to [48].

TABLE 7
Quantitative evaluations of the WaveletFP on the GoPro dataset. $\text{DSTNet}_{\text{w/o WaveletFP}}$ denotes that the DSTNet does not use the WaveletFP; $\text{DSTNet}_{\text{w/ Bilinear in FP}}$ denotes that we use the Bilinear downsampling and upsampling operations instead of the wavelet transform.

Methods	$\text{DSTNet}_{\text{w/o WaveletFP}}$	$\text{DSTNet}_{\text{w/ Bilinear in FP}}$	DSTNet
PSNRs	28.57	32.87	**33.33**
SSIMs	0.9057	0.9587	**0.9611**

method for fair comparisons. Table 7 shows that using the WaveletFP generates better results with higher PSNR and SSIM values, where the PSNR value is 4.76dB higher than the baseline.

In addition, we evaluate the effect of the wavelet transform in the feature propagation. As the low-frequency part can be easily obtained by some sampling methods, e.g., Bilinear interpolation, one may wonder whether using a simple downsampling method generates better results or not. We answer this question by comparing with the method that replaces the wavelet transform and the inverse wavelet transform with the Bilinear downsampling and upsampling operations. Table 7 shows that our method using the wavelet transform in the feature propagation generates better results than the baseline with the Bilinear interpolation, where the PSNR value is 0.46dB higher. Figure 11 also demonstrates that using the WaveletFP generates much clearer frames.

Temporal consistency property. We further evaluate the temporal consistency property of the restored videos. Similar to [48], we show the temporal information of each restored video in Figure 12, where our method generates the videos with a better temporal consistency property. In addition, using the improved lightweight and efficient deep discriminative spatial temporal feature fusion model generates better results than the DSTNet by the conference version [7].

Effectiveness of the improved feature fusion. Compared with the feature fusion (3) that is proposed by our conference version [7], we develop a much more

TABLE 8
Efficiency and effective analysis of the proposed method and the one presented in our conference version [7]. The results are obtained from the GoPro dataset. "Every 5RBs in Figure 2" denotes that we respectively use 5 ResBlocks before and after the "MaxPool$\downarrow_2$" operation in $\mathcal{D}_p$ of Figure 2. The FLOPs is tested on images with size of 256×256 pixels. The running time is tested on the images with size of 1280×720 pixels using RTX 3090GPU.

	CWGDN	Feature fusion (FF)		Dynamic network			PSNRs	Parameters (M)	FLOPs (G)	Runtime (ms)
		(3)	(12)	Figure 2	Every 5RBs in Figure 2	Figure 4				
DSTNet	✔	✔	✗	✔	✗	✗	33.33	7.45	47.87	107.92
DSTNet$_{\text{w/o CWGDN}}$	✗	✔	✗	✔	✗	✗	33.18	3.67	44.80	87.17
DSTNet$_{\text{w/o CWGDN\&w/10RBs in } \mathcal{D}_p}$	✗	✔	✗	✗	✔	✗	33.41	4.24	45.08	88.07
DSTNet+$_{\text{w/10RBs in } \mathcal{D}_p \text{\&w/ FF (12)}}$	✗	✗	✔	✗	✔	✗	33.44	3.90	39.58	82.79
DSTNet+	✗	✗	✔	✗	✗	✔	33.55	3.90	44.61	93.75

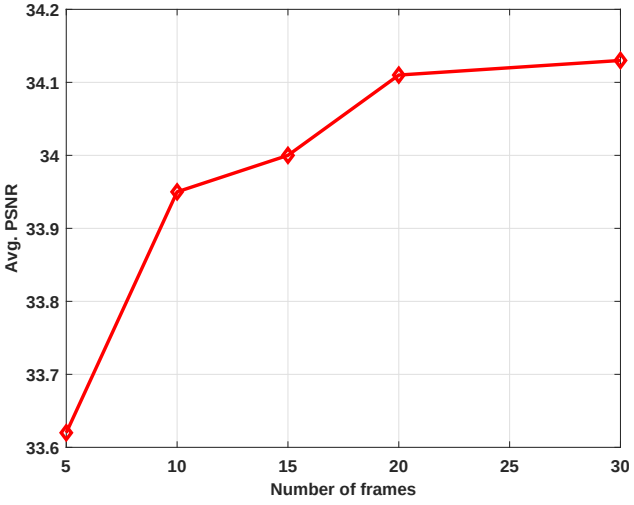


Fig. 13. Effect of the number of frames on video deblurring. The results are obtained from the GoPro dataset [6] using the proposed DSTNet.

efficient feature fusion method (i.e., (12)) for video deblurring. Table 8 shows that using the improved feature fusion method is able to reduce the model complexity while achieving favorable performance compared with [7].

Effectiveness of the improved dynamic filter estimation network. Instead of directly applying downsampling operation with larger scale factors to the input features in the dynamic filter estimation network, we develop a progressive strategy in the dynamic filter estimation network. Table 8 shows that using the progressive strategy generate better deblurred results as more useful information can be better preserved in the downsampling process.

Effect of the number of frames on video deblurring. We quantitatively evaluate the effect of the number of frames on video deblurring using the GoPro dataset. Figure 13 shows that using more frames is able to improve performance. Given the computational cost and accuracy, we use 30 frames in our method.

Further analysis on the large versions of the proposed method and [44]. The large version of [44], i.e., ShiftNet+, is achieved by stacking sophisticated U-Nets in feature extraction and frame reconstruction for better performance improvement. We retrain the ShiftNet+ using the same training datasets as our method for fair comparisons. Table 9 shows that ShiftNet+ generates better results on the GoPro, DVD, and some parts of BSD datasets. However, our method still generates comparable results compared with the ShiftNet+. For example, DSTNet+L generates better results on the BSD (1ms-8ms) dataset (avg. PSNR 35.18dB by DSTNet+L vs avg. 34.73dB by ShiftNet+) and achieves competitive performance on the DVD dataset (avg. PSNR 34.63dB by DSTNet+L vs avg. 34.66dB by ShiftNet+).

In addition, instead of stacking complicated U-Nets, we only use one simple Conv3D operation for feature extraction and frame restoration. Therefore, the proposed DSTNet+L is at least $2.4\times$ and $3.4\times$ faster than ShiftNet and ShiftNet+ as shown in Table 9.

Limitations. Although the proposed method achieves favorable performance on several video deblurring datasets, it cannot effectively handle the scenes with abrupt changes as it is difficult to find useful temporal information from both adjacent and long-range frames. Figure 14 shows an example, where the object and cameras have abrupt motions. Our method does not remove the blur effect well. Future work will consider joint video deblurring and object detection to solve this problem.

7 Conclusion

We have presented an effective lightweight deep discriminative feature propagation network for video deblurring. We develop a channel-wise gated dynamic network to better explore spatial information and propose a discriminative temporal feature fusion module to explore mutually useful contents from frames while reducing the influences of inaccurately estimated features from adjacent frames. To avoid the influence of the inaccurate structural details from non-local frames, we develop a wavelet-based feature propagation method. We formulate each component into an end-to-end trainable deep CNN model and show that our model does not require additional alignment methods and is more compact and efficient for video deblurring. We have analyzed the effect of our method. Both quantitative and qualitative experimental results show that the proposed method

TABLE 9
Quantitative evaluations of DSTNet+L and ShiftNet+ on the datasets of GoPro, DVD, and BSD in terms of PSNR, SSIM, amount of network parameters, FLOPs, and runnning time. The FLOPs is tested on images with the size of 256×256 pixels. The running time is tested on the images with the size of 1280×720 pixels using RTX 3090GPU.

	GoPro		DVD		BSD (1ms-8ms)		BSD (2ms-16ms)		BSD (3ms-24ms)		Model complexity		
	PSNRs	SSIMs	PSNRs	SSIMs	PSNRs	SSIMs	PSNRs	SSIMs	PSNRs	SSIMs	Parameters (M)	FLOPs (G)	Runtime (ms)
ShiftNet [44]	35.49	0.9760	34.58	0.9680	34.43	0.9586	33.76	0.9582	33.41	0.9529	10.51	146.50	616.45
ShiftNet+ [44]	35.88	0.9790	34.69	0.9690	34.73	0.9609	33.90	0.9591	33.48	0.9532	12.30	151.30	853.89
DSTNet+L	35.71	0.9766	34.63	0.9669	35.18	0.9618	33.75	0.9581	33.48	0.9503	14.09	158.14	249.16

(a) Blurred frame (b) Deblurred frame (c) GT

Fig. 14. Limitations of the proposed method. Due to the fast motion of the car and the camera during the exposure time, the captured frame contains significant blur effects, e.g., the wheels of the car. The proposed method does not effectively remove the blur caused by abrupt motions. For example, the wheels are not recovered well as shown in (b).

performs favorably against state-of-the-art methods in terms of accuracy and model complexity.

REFERENCES

[1] T. H. Kim and K. M. Lee, "Generalized video deblurring for dynamic scenes," in *CVPR*, 2015, pp. 5426–5434. 1, 2, 6

[2] L. Bar, B. Berkels, M. Rumpf, and G. Sapiro, "A variational framework for simultaneous motion estimation and restoration of motion-blurred video," in *ICCV*, 2007, pp. 1–8. 1

[3] S. Dai and Y. Wu, "Motion from blur," in *CVPR*, 2008, pp. 1–8. 1, 2

[4] J. Wulff and M. J. Black, "Modeling blurred video with layers," in *ECCV*, 2014, pp. 236–252. 1, 2

[5] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *CVPR*, 2020, pp. 3040–3048. 1, 2, 3, 6, 7, 8, 9, 10

[6] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, 2017, pp. 257–265. 1, 6, 7, 13

[7] J. Pan, B. Xu, J. Dong, J. Ge, and J. Tang, "Deep discriminative spatial and temporal network for efficient video deblurring," in *CVPR*, 2023, pp. 22 191–22 200. 1, 2, 6, 8, 12, 13

[8] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *CVPR*, 2017, pp. 237–246. 1, 3, 6, 7, 8, 9, 10

[9] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE TIP*, vol. 28, no. 1, pp. 291–301, 2019. 1, 3

[10] J. Gast and S. Roth, "Deep video deblurring: The devil is in the details," in *ICCV Workshop*, 2019. 1

[11] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *CVPR Workshops*, 2019, pp. 1954–1963. 1, 2, 3, 6, 7, 8, 10

[12] T. H. Kim, K. M. Lee, B. Schölkopf, and M. Hirsch, "Online video deblurring via dynamic temporal blending network," in *ICCV*, 2017, pp. 4058–4067. 2, 3, 6

[13] Z. Zhong, Y. Gao, Y. Zheng, and B. Zheng, "Efficient spatio-temporal recurrent neural network for video deblurring," in *ECCV*, 2020, pp. 191–207. 2, 3, 5, 6, 7, 8, 9, 10

[14] S. Cho, J. Wang, and S. Lee, "Video deblurring for hand-held cameras using patch-based synthesis," *ACM TOG*, vol. 31, no. 4, pp. 64:1–64:9, 2012. 2, 9, 10

[15] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame video stabilization with motion inpainting," *IEEE TPAMI*, vol. 28, no. 7, pp. 1150–1163, 2006. 2

[16] T. H. Kim and K. M. Lee, "Segmentation-free dynamic scene deblurring," in *CVPR*, 2014, pp. 2766–2773. 2

[17] Y. Li, S. B. Kang, N. Joshi, S. M. Seitz, and D. P. Huttenlocher, "Generating sharp panoramas from motion-blurred videos," in *CVPR*, 2010, pp. 2424–2431. 2

[18] M. Aittala and F. Durand, "Burst image deblurring using permutation invariant convolutional neural networks," in *ECCV*, 2018, pp. 748–764. 3

[19] T. H. Kim, M. S. M. Sajjadi, M. Hirsch, and B. Schölkopf, "Spatio-temporal transformer network for video restoration," in *ECCV*, 2018, pp. 111–127. 3

[20] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter adaptive network for video deblurring," in *ICCV*, 2019, pp. 2482–2491. 3, 6, 7, 8, 10

[21] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *CVPR*, 2018, pp. 2502–2510. 3

[22] H. Son, J. Lee, J. Lee, S. Cho, and S. Lee, "Recurrent video deblurring with blur-invariant motion estimation and pixel volumes," *ACM TOG*, vol. 40, no. 5, pp. 185:1–185:18, 2021. 3

[23] M. Suin and A. N. Rajagopalan, "Gated spatio-temporal attention-guided video deblurring," in *CVPR*, 2021, pp. 7802–7811. 3, 6

[24] Y. Wang, Y. Lu, Y. Gao, L. Wang, Z. Zhong, Y. Zheng, and A. Yamashita, "Efficient video deblurring guided by motion magnitude," in *ECCV*, 2022, pp. 7802–7811. 3

[25] P. Wieschollek, M. Hirsch, B. Schölkopf, and H. P. A. Lensch, "Learning blind motion deblurring," in *ICCV*, 2017, pp. 231–240. 3, 6

[26] J. Lin, Y. Cai, X. Hu, H. Wang, Y. Yan, X. Zou, H. Ding, Y. Zhang, R. Timofte, and L. V. Gool, "Flow-guided sparse transformer for video deblurring," in *ICML*, 2022, pp. 13 334–13 343. 3, 6, 7, 8

[27] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. V. Gool, "Recurrent video restoration transformer with guided deformable attention," in *NeurIPS*, 2022. 3, 6, 7, 9

[28] J. Pan, B. Xu, H. Bai, J. Tang, and M.-H. Yang, "Cascaded deep video deblurring using temporal sharpness prior and non-local spatial-temporal similarity," *IEEE TPAMI*, vol. 45, no. 8, pp. 9411–9425, 2023. 3, 6, 8, 10

[29] M. Cao, Y. Fan, Y. Zhang, J. Wang, and Y. Yang, "VDTR: video deblurring with transformer," *IEEE TCSVT*, vol. 33, no. 1, pp. 160–171, 2023. 3

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 9992–10 002. 3

[31] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *CVPR*, 2022, pp. 17 662–17 672. 3

[32] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCV Workshops*, 2021, pp. 1833–1844. 3

[33] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *CVPR*, 2022, pp. 5718–5729. 3

[34] H. Liu, Z. Dai, D. R. So, and Q. V. Le, "Pay attention to MLPs," in *NeurIPS*, 2021, pp. 9204–9215. 3, 4, 10, 11

[35] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *CVPR*, 2021, pp. 4947–4956. 5

[36] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *CVPR*, 2018, pp. 8174–8182. 6, 10

[37] S. Nah, S. Son, and K. M. Lee, "Recurrent neural networks with intra-frame iterations for video deblurring," in *CVPR*, 2019, pp. 8102–8111. 6

[38] X. Xiang, H. Wei, and J. Pan, "Deep video deblurring using sharpness features from exemplars," *IEEE TIP*, vol. 29, pp. 8976–8987, 2020. 6

[39] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *CVPR*, 2018, pp. 3224–3232. 6, 7

[40] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *CVPR*, 2021, pp. 14 821–14 831. 6, 10

[41] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *ECCV*, 2022. 6

[42] H. Son, J. Lee, J. Lee, S. Cho, and S. Lee, "Recurrent video deblurring with blur-invariant motion estimation and pixel volumes," *ACM TOG*, vol. 40, no. 5, pp. 185:1–185:18, 2021. 6

[43] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "On the generalization of basicvsr++ to video deblurring and denoising," *CoRR*, vol. abs/2204.05308, 2022. 6, 8, 9

[44] D. Li, X. Shi, Y. Zhang, K. C. Cheung, S. See, X. Wang, H. Qin, and H. Li, "A simple baseline for video restoration with grouped spatial-temporal shift," in *CVPR*, 2023, pp. 9822–9832. 6, 7, 8, 9, 13, 14

[45] D. Gong, J. Yang, L. Liu, Y. Zhang, I. D. Reid, C. Shen, A. van den Hengel, and Q. Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *CVPR*, 2017, pp. 3806–3815. 6

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. 6

[47] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *ICCV*, 2021, pp. 4641–4650. 6

[48] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *CVPR*, 2022, pp. 5962–5971. 8, 10, 12

[49] M. Tassano, J. Delon, and T. Veit, "DVDNET: A fast network for deep video denoising," in *ICIP*, 2019, pp. 1805–1809. 9

[50] ——, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *CVPR*, 2020, pp. 1351–1360. 9

[51] G. Vaksman, M. Elad, and P. Milanfar, "Patch craft: Video denoising by deep modeling and patch matching," in *ICCV*, 2021, pp. 2137–2146. 9

[52] J. Park, S. Nah, and K. M. Lee, "Recurrence-in-recurrence networks for video deblurring," in *BMVC*, 2021, p. 20. 11, 12

Jinshan Pan received the Ph.D. degree in computational mathematics from the Dalian University of Technology, China, in 2017. He was a joint-training PhD student in School of Mathematical Sciences, Dalian University of Technology, China, and Electrical Engineering and Computer Science, University of California, Merced, CA. He is a professor of School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interest includes image deblurring, image/video analysis and enhancement, and related vision problems.

Long Sun is currently working toward the PhD degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include image/video super-resolution, deblurring, and other restoration tasks.

Boming Xu is a M.S. student in School of Computer Science and Engineering,Nanjing University of Science and Technology,China. His research interests include video deblurring, image/video enhancement, and related vision problems.

Jiangxin Dong received the Diploma degree in information and computational science from the Dalian University of Technology in 2014 and the PhD degree in computational mathematics from the Dalian University of Technology in 2019. She was a postdoctoral researcher with the Department of Computer Vision and Machine Learning, Max Planck Institute for Informatics, Germany. She is currently a professor with Nanjing University of Science and Technology, China. Her research interests include image restoration and related problems.

Jinhui Tang (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. He has authored more than 200 articles in top-tier journals and conferences. His research interests include multimedia analysis and computer vision. Dr. Tang was a recipient of the Best Paper Awards in ACM MM 2007 and ACM MM Asia 2020, the Best Paper Runner-Up in ACM MM 2015. He has served as an Associate Editor for IEEE TMM, IEEE TKDE, IEEE TNNLS and IEEE TCSVT. He is a Fellow of IAPR.