Deep Discriminative Spatial and Temporal Network for Efficient Video Deblurring

Jinshan Pan[1*], Boming Xu[1*], Jiangxin Dong[1†], Jianjun Ge[2], and Jinhui Tang[1†]
[1]Nanjing University of Science and Technology [2]China Electronics Technology Group Corporation

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. We show that the proposed method does not require additional alignment methods and performs favorably against state-of-the-art ones on benchmark datasets in terms of accuracy and model complexity.

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 [2, 9, 14, 36] 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 [26], 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.

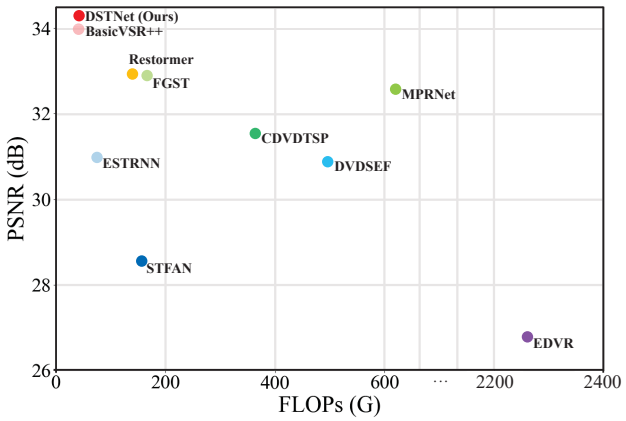


Figure 1. Floating point operations (FLOPs) vs. video deblurring performance on the GoPro dataset [24]. Our model achieves favorable results in terms of accuracy and FLOPs.

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 [29] or employ spatial and temporal 3D convolution [40] 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 [26], deformable convolution [32], 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

*Co-first authorship
†Corresponding author

are difficult to train and computationally expensive. For example, the CDVDTSP method [26] with optical flow as the alignment module has 16.2 million parameters with FLOPs of 357.79G while the EDVR method [32] 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 [15]. However, these methods have limited capacity to transfer the useful information temporally for latent frame restoration as demonstrated in [41]. To remedy this limitation, several methods recurrently propagate information of non-local frames with some proper attention mechanisms [41]. 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. As the temporal information exploration is critical for video deblurring, it is a great need to develop an effective 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 [26], deformable convolution used in [32]) and is thus efficient yet effective for video deblurring as shown in Figure 1.

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.

2. Related Work

Hand-crafted prior-based methods. Since the video deblurring problem is ill-posed, conventional methods usually make some assumptions [7, 9, 13, 14, 18, 22, 36] 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 [29], 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 [1] 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 [40], 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 [32], Wang et al. use the deformable convolution to achieve the alignment. The optical flow estimation method is widely adopted [16, 26] to align the adjacent frames. Zhou et al. [42] develop an implicit alignment method based on the kernel prediction network [23]. 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. [26] develop a temporal sharpness prior to better guide the network for video deblurring. Son et al. [28] aggregate information from multiple video frames by a blur-invariant motion estimation and pixel volumes. Suin and Rajagopalan [30] select the key frames to facilitate the blur removal. Wang et al. [33] 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. [35] develop an effective RNN to recurrently use the features from the previous frame in multiple scales. In [15], 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. [41] 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. [20] develop an effective flow-guided sparse Transformer for video deblurring. Liang et al. [19] develop a recurrent video restoration transformer. Although decent performance has been archived, 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. 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 [19, 20, 34, 38]. 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 [21]. 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.

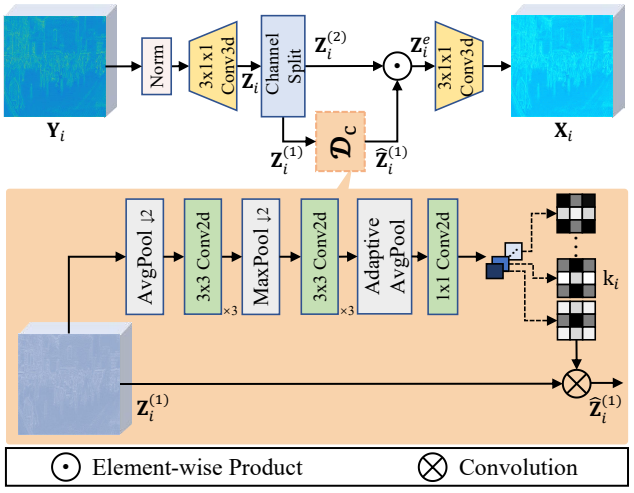


Figure 2. Network details of the proposed CWGDN.

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 channel dimension of C, we first apply a 3D convolution with filter size of $3 \times 1 \times 1$ 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 [21], the generated filters $\mathbf{k}_i$ by the proposed method contain global information of the feature in

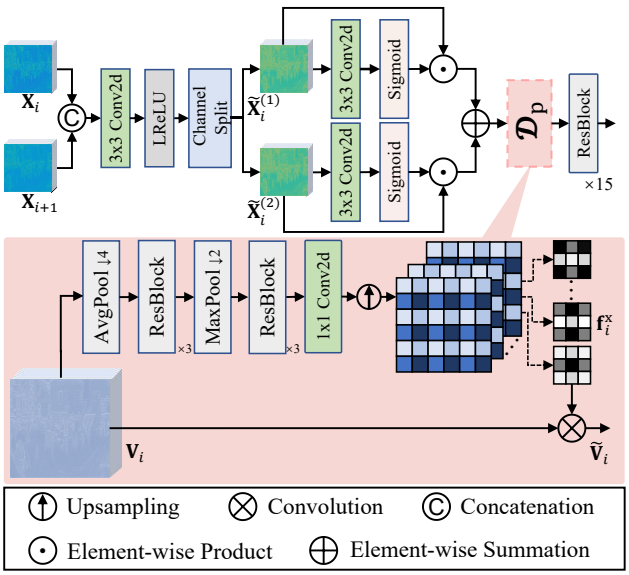


Figure 3. Network details of the proposed DTFF module.

each channel, which can discriminatively explore useful features to facilitate video deblurring. We provide detailed analysis in Section 5.

3.2. Discriminative temporal feature fusion module

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:

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

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 3 for the detailed network architecture) to estimate pixel-wise filters from $\mathbf{V}_i$ and apply the estimated pixel-wise filters to $\mathbf{V}_i$:

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

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:

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

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:

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

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 [41] and video super-resolution [3]. 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:

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

and

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

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 5 and supplemental material.

4. 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.

4.1. Datasets and parameter settings

Video deblurring datasets. We evaluate our method on the commonly used video deblurring datasets including the GoPro dataset [24], the DVD dataset [29], and the real-world dataset (BSD) [41], 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 [17] 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 [8] to constrain the network training. The detailed network architectures of the proposed method and more experimental results are included in the supplemental material due to the page limit. The training code and models are available on our project website: https://github.com/xuboming8/DSTNet.

4.2. 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 [12] (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 RVRT [19] performs better than our method, our network has fewer parameters and is at least $5\times$ faster than RVRT [19] (Table 7). Also note that our method performs better when using the similar amounts of parameters (i.e., Ours-L in Table 1).

Figure 4 shows some visual comparisons of the evaluated methods on the GoPro dataset. The method [29] concatenates the consecutive frames as the input and does not effectively remove blur (Figure 4(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., [26, 32], does not restore clear frames as the inaccurate alignment will interfere with the deblurring process (Figure 4(d) and (g)). The STFAN method [42] 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 4(e)). The method [41] 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 4(f)). The recent method [20] employs Transformers with optical flow as guidance to model spatial and temporal information for video deblurring. However, the deblurred image in Figure 4(i) still contains blur effects.

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 generates 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 4(j)).

Evaluations on the DVD dataset. We further evaluate our method on the DVD dataset by Su et al. [29]. Table 2 shows that the proposed method generates the deblurred videos with higher PSNR and SSIM values.

Evaluations on the BSD deblurring dataset. As the BSD dataset [41] is a commonly used benchmark for video deblurring, we evaluate our method against state-of-the-art ones based on the protocols of [41]. Our method generates high-quality deblurred results as shown in Table 3.

Evaluations on real captured videos. Similar to the existing method [26], we further evaluate our method using

Table 1. Quantitative evaluations on the GoPro dataset [24]. "Ours-L" denotes a large model, where we use 96 features and 30 ResBlocks in the DTFF module.

Methods	SRN [31]	DVD [29]	Wieschollek et al. [35]	DTBN [15]	Nah et al. [25]	EDVR [32]	STFAN [42]	DVDSEF [37]	DUF [12]
PSNRs	30.29	27.31	25.19	26.82	29.97	26.83	28.59	31.01	28.01
SSIMs	0.9014	0.8255	0.7794	0.8245	0.8947	0.8426	0.8608	0.9130	0.8768
Methods	ESTRNN [41]	MPRNet [39]	CDVDTSP [26]	NAFNet [6]	BasicVSR++ [5]	FGST [20]	RVRT [19]	Ours	Ours-L
PSNRs	31.07	32.73	31.67	33.71	34.01	32.90	34.92	34.16	**35.05**
SSIMs	0.9023	0.9366	0.9279	0.9668	0.9520	0.9610	**0.9738**	0.9679	0.9733

Table 2. Quantitative evaluations on the DVD dataset [29] in terms of PSNR and SSIM.

Methods	Kim and Lee [14]	Gong et al. [11]	SRN [31]	DVD [29]	DTBN [15]	EDVR [32]	STFAN [42]
PSNRs	26.94	28.27	29.98	30.01	29.95	28.51	31.15
SSIMs	0.8158	0.8463	0.8842	0.8877	0.8692	0.8637	0.9049
Methods	DVDSEF [37]	ESTRNN [41]	MPRNet [39]	CDVDTSP [26]	GSTA [30]	FGST [20]	Ours
PSNRs	31.71	32.01	32.24	32.13	32.53	33.36	**33.79**
SSIMs	0.9159	0.9162	0.9253	0.9268	0.9468	0.9500	**0.9615**

(a) Blurred frame (b) GT (c) DVD [29] (d) EDVR [32] (e) STFAN [42]

(f) ESTRNN [41] (g) CDVDTSP [26] (h) DVDSEF [37] (i) FGST [20] (j) Ours

Figure 4. Deblurred results on the GoPro dataset [24]. The deblurred results in (c)-(i) still contain significant blur effects. The proposed method generates much clearer frames.

real captured videos by Cho et al. [7] and compare the proposed method with state-of-the-art video deblurring methods [5, 26, 29, 32, 41, 42]. Figure 5 shows that the competed methods do not restore the sharp frames well. In contrast, our method generates much clearer frames, where the characters are recognizable.

5. 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 4 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 [21], 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 4 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.

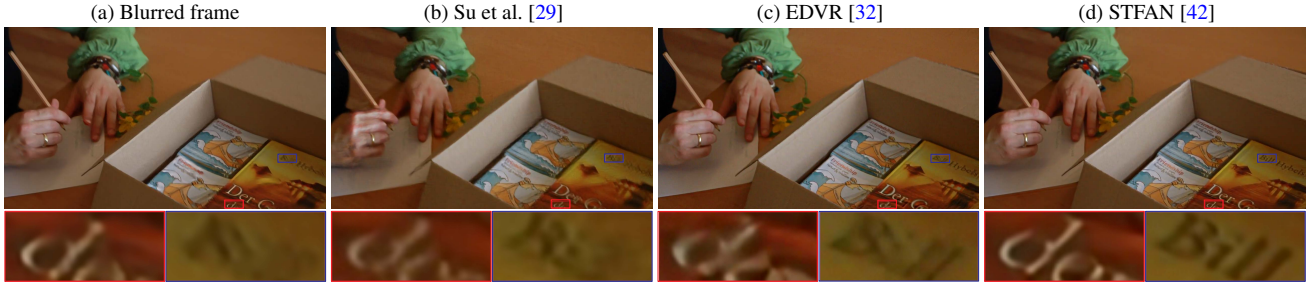
In addition, Figure 6(b) and (c) shows that the baseline using gMLP and the one without using the CWGDN do not restore clear frames, while the proposed approach using the CWGDN generates much clearer frames (Figure 6(d)).

Effectiveness of DTFF. The proposed DTFF module is

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

Methods	DVD [29]	DTBN [15]	Nah et al. [25]	ESTRNN [41]	CDVDTSP [26]	Ours
PSNRs	29.95	31.84	33.00	33.36	32.84	**34.45**
SSIMs	0.8692	0.9170	0.9330	0.9370	0.9398	**0.9548**

(a) Blurred frame (b) Su et al. [29] (c) EDVR [32] (d) STFAN [42]



(e) CDVDTSP [26] (f) ESTRNN [41] (g) BasicVSR++ [5] (h) Ours

Figure 5. Deblurred results on a real video from [7]. The evaluated video deblurring methods do not recover the clear frames. In contrast, our method recovers a clearer image with recognizable characters (e.g., "Bill").

Table 4. Effect of the proposed CWGDN on the GoPro dataset.

Methods	w/ gMLP [21]	w/o CWGDN	w/ CWGDN (Ours)
PSNRs	33.16	33.18	**33.33**
SSIMs	0.9602	0.9602	**0.9611**

(a) Blurred patch (b) w/ gMLP (c) w/o CWGDN (d) w/ CWGDN

Figure 6. Effectiveness of the CWGDN for video deblurring.

Table 5. Effect of the proposed DTFF on the GoPro dataset.

Methods	w/ the feature fusion [27]	w/o (3)	w/o (4)	Ours
PSNRs	33.00	33.22	32.55	**33.33**
SSIMs	0.9597	0.9608	0.9504	**0.9611**

Table 6. Quantitative evaluations of the WaveletFP on the GoPro dataset. "FP" denotes the abbreviation of feature propagation.

Methods	w/o WaveletFP	w/ Bilinear in FP	Ours
PSNRs	28.57	32.87	**33.33**
SSIMs	0.9057	0.9587	**0.9611**

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 (3) 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 (3) is used to keep the most useful features from adjacent frames while reducing the influence of inaccurate features from adjacent frames. The results in Table 5 show that our method using (3) obtains better-deblurred results. Note that Park et al. [27] 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.,

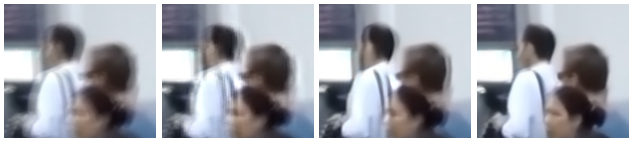
$$\mathbf{V}_i = \mathbf{X}_i \odot \mathcal{S}(\mathbf{W}_{2d}(\tilde{\mathbf{X}}_i^{(1)})) + \mathbf{X}_{i+1} \odot \mathcal{S}(\mathbf{W}_{2d}(\tilde{\mathbf{X}}_i^{(2)})). \quad (11)$$

This fusion method does not generate better results as

(a) Blurred patch (b) w/ [27] (c) w/o (3)

(d) w/o (4) (e) w/ (3) and (4) (f) GT

Figure 7. Effectiveness of the DTFF module for video deblurring.



(a) Blurred patch (b) w/o WaveletFP (c) w/ Bilinear (d) Ours

Figure 8. Effectiveness of the WaveletFP for video deblurring.

Table 7. Quantitative evaluations of the video deblurring methods with better accurate performance on the GoPro dataset in terms of network parameters and running time. The running time is obtained on a machine with an NVIDIA RTX A6000 GPU. The size of the test images is 1280×720 pixels.

Methods	DVD [29]	EDVR [32]	DVDSEF [37]	CDVDTSP [26]	FGST [20]	BasicVSR++ [5]	RVRT [19]	Ours	Ours-L
Network parameters (M)	15.30	23.60	75.05	16.19	9.70	9.46	13.60	**7.45**	16.96
Running time (/s)	0.88	2.70	0.52	1.90	0.72	0.10	0.41	**0.08**	0.28

shown in Table 5.

In addition, as shown in Table 5, 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 7 further demonstrates that using the discriminative temporal feature fusion module is able to facilitate blur removal.

As the contents of various frames are different, we split the features and estimate the individual weight to respectively utilize useful information from each feature in the DTFF module. We demonstrate the effectiveness of the splitting operation and provide more visual comparisons in the supplemental material due to the page limit.

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 method for fair comparisons. Table 6 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 6 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 8 also demonstrates that using the WaveletFP generates much clearer frames.

We include further analysis on the WaveletFP in the supplemental material due to the page limit.

Temporal consistency property. We further evaluate the temporal consistency property of the restored videos. Similar to [4], we show the temporal information of each restored video in Figure 9, where our method generates the videos with a better temporal consistency property.

Model complexity. In addition to the FLOPs in Figure 1, we further examine the model complexity of the proposed method and state-of-the-art ones in terms of the model parameters and running time. Table 7 shows that the proposed method has fewer model parameters and is relatively faster than the evaluated methods.

Figure 9. Visual comparisons of the temporal consistency for restored videos. We visualize the pixels of the selected columns (the dotted line) according to [4].

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 (see the supplemental material for examples). Future work will consider joint video deblurring and object detection to solve this problem.

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

Acknowledgements. This work has been partly supported by the National Key R&D Program of China (No. 2018AAA0102001), the National Natural Science Foundation of China (Nos. U22B2049, 62272233, 61922043, 61925204, 61872421), and the Fundamental Research Funds for the Central Universities (No. 30920041109).

References

[1] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *ECCV*, pages 748–764, 2018. 2

[2] Leah Bar, Benjamin Berkels, Martin Rumpf, and Guillermo Sapiro. A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In *ICCV*, pages 1–8, 2007. 1

[3] Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*, pages 4947–4956, 2021. 4

[4] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, pages 5962–5971, 2022. 8

[5] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. On the generalization of basicvsr++ to video deblurring and denoising. *CoRR*, abs/2204.05308, 2022. 6, 7, 8

[6] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 6

[7] Sunghyun Cho, Jue Wang, and Seungyong Lee. Video deblurring for hand-held cameras using patch-based synthesis. *ACM TOG*, 31(4):64:1–64:9, 2012. 2, 6, 7

[8] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, pages 4641–4650, 2021. 5

[9] Shengyang Dai and Ying Wu. Motion from blur. In *CVPR*, pages 1–8, 2008. 1, 2

[10] Jochen Gast and Stefan Roth. Deep video deblurring: The devil is in the details. In *ICCV Workshop*, 2019. 1

[11] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian D. Reid, Chunhua Shen, Anton van den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *CVPR*, pages 3806–3815, 2017. 6

[12] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, pages 3224–3232, 2018. 5, 6

[13] Tae Hyun Kim and Kyoung Mu Lee. Segmentation-free dynamic scene deblurring. In *CVPR*, pages 2766–2773, 2014. 2

[14] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *CVPR*, pages 5426–5434, 2015. 1, 2, 6

[15] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Schölkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *ICCV*, pages 4058–4067, 2017. 2, 3, 6, 7

[16] Tae Hyun Kim, Mehdi S. M. Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, pages 111–127, 2018. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[18] Yunpeng Li, Sing Bing Kang, Neel Joshi, Steven M. Seitz, and Daniel P. Huttenlocher. Generating sharp panoramas from motion-blurred videos. In *CVPR*, pages 2424–2431, 2010. 2

[19] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *NeurIPS*, 2022. 3, 5, 6, 8

[20] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. In *ICML*, pages 13334–13343, 2022. 3, 5, 6, 8

[21] Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay attention to MLPs. In *NeurIPS*, pages 9204–9215, 2021. 3, 6, 7

[22] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *IEEE TPAMI*, 28(7):1150–1163, 2006. 2

[23] Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, pages 2502–2510, 2018. 2

[24] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 257–265, 2017. 1, 5, 6

[25] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*, pages 8102–8111, 2019. 6, 7

[26] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *CVPR*, pages 3040–3048, 2020. 1, 2, 3, 5, 6, 7, 8

[27] JoonKyu Park, Seungjun Nah, and Kyoung Mu Lee. Recurrence-in-recurrence networks for video deblurring. In *BMVC*, page 20, 2021. 7

[28] Hyeongseok Son, Junyong Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *ACM TOG*, 40(5):185:1–185:18, 2021. 3

[29] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, pages 237–246, 2017. 1, 2, 5, 6, 7, 8

[30] Maitreya Suin and A. N. Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *CVPR*, pages 7802–7811, 2021. 3, 6

[31] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 6

[32] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, pages 1954–1963, 2019. 1, 2, 5, 6, 7, 8

[33] Yusheng Wang, Yunfan Lu, Ye Gao, Lin Wang, Zhihang Zhong, Yinqiang Zheng, and Atsushi Yamashita. Efficient video deblurring guided by motion magnitude. In *ECCV*, pages 7802–7811, 2022. 3

[34] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17662–17672, 2022. 3

[35] Patrick Wieschollek, Michael Hirsch, Bernhard Schölkopf, and Hendrik P. A. Lensch. Learning blind motion deblurring. In *ICCV*, pages 231–240, 2017. 3, 6

[36] Jonas Wulff and Michael Julian Black. Modeling blurred video with layers. In *ECCV*, pages 236–252, 2014. 1, 2

[37] Xinguang Xiang, Hao Wei, and Jinshan Pan. Deep video deblurring using sharpness features from exemplars. *IEEE TIP*, 29:8976–8987, 2020. 6, 8

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729, 2022. 3

[39] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 6

[40] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Wei Liu, and Hongdong Li. Adversarial spatio-temporal learning for video deblurring. *IEEE TIP*, 28(1):291–301, 2019. 1, 2

[41] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*, pages 191–207, 2020. 2, 3, 4, 5, 6, 7

[42] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, pages 2482–2491, 2019. 2, 5, 6, 7