# VDFlow: Joint Learning for Optical Flow and Video Deblurring

Yanyang Yan[1,2], Qingbo Wu[1,2], Bo Xu[3], Jingang Zhang[2]*, and Wenqi Ren[1]
[1]SKLOIS, IIE, CAS [2]University of Chinese Academy of Sciences [3]Yunnan Normal University

## Abstract

*Video deblurring is a challenging task as the blur in videos is caused by the combination of camera motion, object moving and depth variation. Recent deep neural networks improve the performance of video deblurring by using the concatenated neighboring frames to estimate the latent images directly. In this paper, we propose a united end-to-end network, called VDFlow, for both optical flow estimation and video deblurring simultaneously. The VDFlow contains two branches where feature representations are bi-directional propagated. The deblurring branch employs an encoder-decoder style network while the optical flow branch is based on the FlowNet network. The optical flow is no longer a tool for alignment but serves as an information carrier of motion trajectories, which helps to restore the latent sharp frames. Extensive experiments demonstrate that the proposed method performs favorably against the state-of-the-art video deblurring approaches on challenging blurry videos and improves the performance of optical flow estimation as well.*

## 1. Introduction

Camera shakes, object motions, and depth variations may introduce blur in videos, which affects many high-level applications. Therefore, deblurring attracts considerable research attention in computer vision community. However, most extensive methods are designed for single image deblurring [22, 3], but instead pay less attention to videos [34, 12, 24, 27], where blur are more easily to be caused.

The reconstruction of a sharp frame from the corresponding blurry observation is a highly ill-posed problem due to the fact that both the image and the blur kernel are unkonwn. The most common deblurring approaches are based on deconvolution algorithms, which initially determine the blur kernel and then solve the sharp image. However, the space-variant blur kernel is directly connected with many unknown factors such as camera shakes, scene depth and segmentation boundaries of objects, making it a great challenge to recover the sharp image. In addition, given the blury frame and kernel information, non-blind methods may also introduce undesired artifacts. Therefore, in this paper, we directly estimate the latent sharp frame by using a deep network formulation.

Unlike single image deblurring, the abundant information from neighboring frames can be leveraged to sharpen blurry ones in video deblurring. However, aggregating the information from neighboring frames remains a challenging problem. In the previous work, Su et al. [29] introduce an encoder-decoder network to learn how to accumulate information and deblur videos. They stack the neighboring frames together with various alignment types: no-alignment, frame-wise homography alignment, and optical flow alignment. Although optical flow is typically useful to improve the deblur performance, flow estimation is a challenging problem and thus the motion information does not always helpful. In addition, performing frame-to-frame alignment with optical flow directly often introducing additional warping artifacts.

In this paper, the optical flow is regarded as an information carrier about motion trajectories but not directly used for image alignment. We present an unified end-to-end network, named VDFlow, which consist of two branches. One branch is designed to deblur and the other one aims to estimate the forward and backward optical flows for the central frame. We simultaneously learn the feature representations for each task and and bi-directionally propagate the learned features to help each other. We compare quantitatively with the ground truth sharp frames in data set and qualitatively to videos previously used for video deblurring. To evaluate the optical flow performance of our proposed method, we compare results on the Sintel [1], KITTI [8] and Flying Chairs [7] datasets. Extensive experimental results demonstrate the proposed algorithm performs favorably against the state-of-the-art methods on challenging blurry videos.

The main contributions are summarized as follows.

1. We utilize the bi-directional optical flow as an information carrier about motion trajectory of blurry frames in video deblurring.

2. We propose an end-to-end trainable framework for simultaneously estimating the sharp frames and optical

*corresponding author

flow in blurry videos, in which the feature representations are bi-directional propagated to help each other task.

3. We employ an iterative training strategy a method for our proposed model to relax the constrain that both the sharp image and optical flow ground truths should exist.

The rest of the paper is organized as follows. In Section 2, we briefly review the related work in several relevant respects. An overview of our model is given in Section 3, including the details about deblurring and optical flow branches and our analysis about employed the bi-directional propagation, followed by the implementation and training strategy in Section 4. Extensive experimental results are presented in Section 5 and Section 6 draws conclusions.

## 2. Related Work

**Deblurring based on multi-image aggregation** Deblurring methods based on multi-image aggregation rely on the observation that not all frames are equally blurrred. Multiple images are combined directly in either spatial or frequency domain. For example, sharper pixels of neighboring frames are transferred and then interpolated to deblur the target frame in [19]. Cho et al. [5] further extend this approach by detecting clear regions to restore blurry regions with the same content in neighboring frames, which improves the robustness against moving objects. However, the patch matching process is computationally expensive and large depth variations can not be handled. In [30], Sunkavalli et al. propose a multi-image enhancement method based on a unified Bayesian framework to establish correspondence among nearby frames. Recently, Delbracio and Sapiro [6] aggregate multiple frames, which are aligned using optical flow, in the Fourier domain and show effective and computationally efficient deblurring. All above approaches do not solve any inverse problem and rely on sharp pixels or patches from neighboring frames which may not exist.

**Deblurring based on deconvolution** Recent years has witnessed many successful single image deblurring methods jointly estimating the blurring kernel and the sharp image based on deconvolution [20, 36]. For video, additional information are utilized to alleviate the ill-posedness of single image deblurring. For instance, Wulff and Black [34] segment images into foreground as well as background regions and model the global motion blur kernels using affine motion. However, deblurring methods based on segmentation depend strongly on the accurate segments and not perform well when segments are poor. To solve this problem, Zhang and Yang [38] use a projection path model to estimate blur kernels. However, the projective motion path model is designed for global camera motions, which is insufficient to complex object motion as well as depth variations. In [12], a segmentation-free approach is proposed by Kim and Lee using optical flow to improve estimating motion blurs and latent frames for dynamic scene deblurring. However, all the above approaches are based on the assumed image degradation model and thus may be inefficient when deal with real-world blurry data. In addition, suboptimal estimations of the models will also influence the performance.

**Deblurring based on deep learning.** Deep learning has shown its remarkable performance in various computer vision tasks given enough labeled data [15], such as object detection [18, 35, 16], image denoising [32, 33], dehazing [25, 39, 40], deraining [17], face restoration [26, 37], and deblurring [28, 4]. Schuler et al. [28] introduce an end-to-end deep neural network using synthetic training data to estimate the blur kenel for blind deconvolution. Their performance degrades if the blur kernel have large size. In [4], Chakrabarti et al. propose to learn the Fourier coefficients of a deconvolution filter and then apply to estimating the sharp image patches. For video deblurring, an encoder-decoder style network is developed by Su et al. to address real-world video deblurring. They stack the neighboring frames together to aggregate the information and achieves state-of-the-art performance. Three versions of dataset are created to train their work with varying degrees of alignment, including with no alignment, using optical flow to align, and using hamography. The network whose input is aligned with optical flow has higher PSNR than the others but fails when inputs frame are much blurrier, which may due to optical flow errors when the input auality is good. In this paper, we use an encoder-decoder style network to jointly estimate the clear frames and optical flow.

**Optical Flow.** Varieties of approaches have been came up with to estimate the optical flow since the concept is proposed by Horn and Schunck [10]. In [31], sparse convolutions and max-pooling are used to aggregate the feature information from fine to coarse. However their handcrafted parameters are fixed, which leads to lack of universality. Dosovitskiy et al. [7] proposed an end-to-end optical flow estimation method with convolutional networks. Two network architectures: FlowNetSimple and FlowNetCorr are produced based on encoder-decoder network architecture. Ilg et al. [11] advance the architectures by employing a warping operation during and stacking multiple diverse networks for flow refinement. Their FlowNet 2.0 achieves retrieval of fine structures, highlight motion boundaries, and robustness to compression artifacts. In this paper, for convenience, we employ the FlowNetS architecture, which is referred as FlowNetSimple in [7], to calculate the corresponding feature representations of optical flows.

## 3. Proposed Method

The fully convolutional neural networks have shown good ability at learning input/output relations when large training datasets exist. In this paper, an end-to-end learning approach is proposed for video deblurring and optical flow estimation: given datasets consisting of blurry video frames and ground truth ones, we train a nerual network to predict the sharp frames directly from the blurry ones. The input is a stack of neighboring frames while the output is the deblurred image and the bi-directional optical flow corresponding to the central frame. A unified model named VDFlow is constructed which has two branches, a video deblurring branch based on the fully convolutional network, and an optical flow estimation branch based on the FlowNetS. In the following, we first describe our neural network architecture, then perform a number of experiments to anlyze its efficiency and comparing with existing methods. The architecture of our proposed VDFlow is illustrated in Figure 1. The key advantage of our approach is the feature representations for video deblurring and optical flow is propagated bidirectionally to help each other.

### 3.1. Deblurring Branch

To make full use of both low-level and high-level features, we choose an encoder-decoder style network, which has shown good ability for generative tasks [21, 23]. The encoder is used to capture the context of the fused neighboring frames and produce a latent feature representation while the decoder takes this feature representation and produces the sharp image content.

Similar to [29], symmetric skip connections are added every a few layers between corresponding layers in encoder and decoder halves of the network, which are illustrated as black dotted lines in Figure 1. The response from a convolutional layer is directly propagated to the corresponding mirrored deconvolutional layer, both forwardly and backwardly. In this paper, the features are added elementwise. These skip connections not only significantly help to pass the information to the top layers and back-propagate the gradients to bottom layers, which help generate much sharper video frames, but also greatly accelerates the convergence. To optimize the network, the deblurring branch uses a $L_2$ distance to the ground truth sharp frame as the reconstruction loss $\mathbf{L}_d$:

$$\mathbf{L}_d = \sum_{i,j}(I(i,j) - G(i,j))^2, \tag{1}$$

where $I(i,j)$ and $G(i,j)$ denote the pixel value at position $(i,j)$ of the estimated image and ground truth, respectively.

### 3.2. Optical Flow Branch

Our major objective is to produce the sharp image of the central frame. Therefore, we try to estimate the corresponding forward and backward optical flows to model the motion trajectory, which can approximate the blur kernel information of the central frame. Although the FlowNet 2 [11] achieves several improvements to Flownet [7], considering the convenience and efficiency, we employ the FlowNetS architecture to calculate the corresponding feature representations of optical flows. In addition, our proposed deblurring branch also helps to improve the accuracy of optical flow in the optical flow branch. Similar to deblurring branch, the FlowNetS architecture also employs an encoder-decoder style architecture. Additional skip connections also exist for feature concatenations between the encoder and decoder, which are illustrated as black dotted lines in top half in Figure 1. On the other side, the decoder part includes the upsampled coarser flow prediction with sizes of $1/16$, $1/8$, $1/4$ of the input image size, which do not exist in the deblurring branch. As shown in Figure 1, both the optical flow branch and deblurring branch have feature representations with sizes of $1/8$ and $1/4$ of the input image size. Therefore, the bi-directional propagation of features are operated in these two scales, which will be introduced in the next subsection.

For labeled data, to optimize our network, the optical flow branch uses the endpoint error (EPE) loss:

$$\mathbf{L}_{f\_EPE}(f, \hat{f}) = \sqrt{(u-\hat{u})^2 + (v-\hat{v})^2}, \tag{2}$$

where $f = [u, v]$ represents the predicted optical flow field and the ground truth optical flow is $\hat{f} = [\hat{u}, \hat{v}]$. However, when training our proposed VDFlow using the blurry videos dataset, we have no access to the ground truth optical flow. To solve this problem, the image warping loss is employed for unlabeled data. Given the input image pair $I_1, I_2$ the image warping loss can be defined as:

$$\mathbf{L}_{f\_warp}(I_1, I_2, f) = \rho(I_1 - W(I_2, f)) \tag{3}$$

where $\rho(\cdot)$ is the robust penalty function, $W(I_2, f)$ denotes the warping operation which warps $I_2$ accoding to the optical flow $f$. One intance of the flow warp error is shown in Figure 2. Note that minimizing $\mathbf{L}_{f\_warp}(I_1, I_2, f)$ can make the flow warp error close to zero at each pixel, which is consistent with the one of ground truth.

For unified expression, the loss function of optical flow becomes:

$$\mathbf{L}_f = \sum_{I_1, I_2 \in D_l} \sqrt{(u-\hat{u})^2 + (v-\hat{v})^2} + \sum_{I_1, I_2 \in D_u} \rho(I_1 - W(I_2, f)) \tag{4}$$

where $D_l$ and $D_u$ denote the labeled and unlabeled datasets, respectively.

### 3.3. Bi-directional Propagation

In [29], the deep deblurring network with aligned inputs by optical flow achieves the highest PSNR than the other

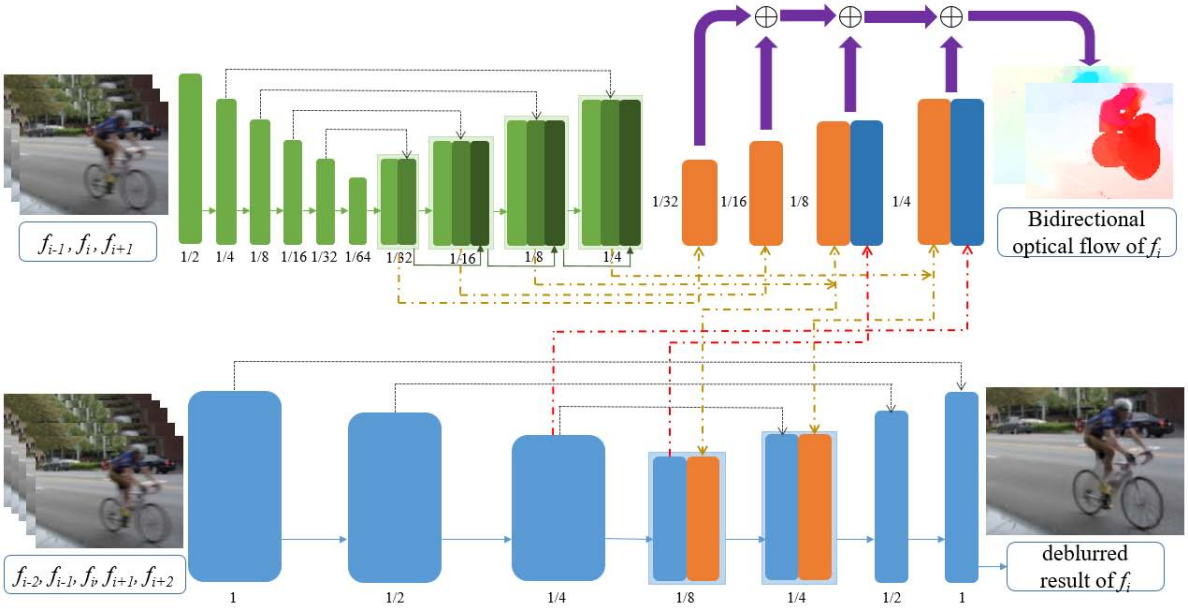



Figure 1. Network architecture of the proposed VDFlow model. Our VDFlow contains two branches: the encoder-decoder style deblurring branch and the optical flow branch based on the FlowNetS architecture. Our network combine feature representations between two branches for both tasks naturally using the bi-directionally propagations at different scales, the size of $1/8$ and $1/4$ of the input image size.

types of deep deblurring network. However, directly aligning the input with optical flow cannot make full use of the information optical flow. Instead, we try to use optical flow information in the feature space. In this paper, we aim to simultaneously learn the helpful motion representations to improve the deblurring performance though bi-directional propagation. Therefore, the input of our network is only the stack of blurry frames. We first analysis the practicability of bi-directional propagation. According to the definition, the feature representation of optical flow represent the motion information, which is closely related to blur kernel for each frame. In addition, both the deblurring and optical flow branches employ similar encoder-decoder style architecture and have feature representations in the same scales: the size of $1/8$ and $1/4$ of the input image size, which enables the bi-directional propagation.

Since the output of deblurring branch is the latent image of the central frame $f_t$, we use the frames at time $t-1$ and $t$ to estimate the backward optical flow of frame $f_t$, and use frames at time $t$ and $t+1$ to compute the forward one. For convenience, only one network architecture is shown in Figure 1. Our architecture combines the feature representations between two branches naturally using the bi-directionally propagations at different scales, the size of $1/8$ and $1/4$ of the input image size, which are illustrated as yellow and red dotted lines in Figure 1. For example, the concatenated feature maps with the size of $1/8$ of the input image size in optical branch are concatenated to the feature maps in deblurring branch with the same scale. Then, the fused feature maps are utilized for further predictions in deblurring branch. Comparing to [29], our method add optical flow branch for estimation, which does not increase too much model capacity. With bi-directional propagation, two branches can communicate with each other. Their shared feature representations make it possible to improve the performance of both tasks. To optimize the network, our final loss function is defined as a weighted sum of the deblurred loss $\mathbf{L}_d$ loss and the optical flow loss $\mathbf{L}_f$,

$$\begin{aligned}\mathbf{L}_{total} &= \mathbf{L}_d + \lambda_f \mathbf{L}_f \\ &= \sum_{i,j}(I(i,j) - G(i,j))^2 + \sum_{I_1, I_2 \in D_l} \sqrt{(u-\hat{u})^2 + (v-\hat{v})^2} \\ &\quad + \sum_{I_1, I_2 \in D_u} \rho(I_1 - W(I_2, f)),\end{aligned} \tag{5}$$

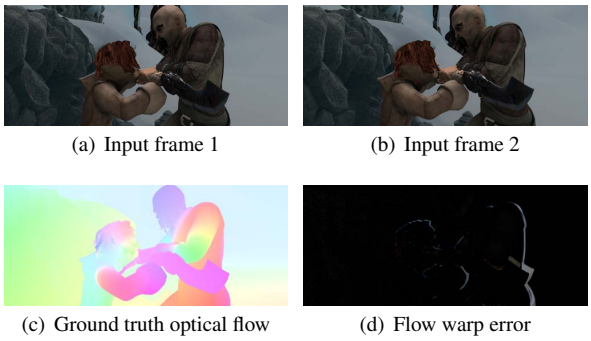
where $\lambda_f$ is the balanced weight.

## 4. Implementation and Training

In this section, we present more details about how to train our proposed VDFlow including the training scheme, datasets and implementation details.

### 4.1. Training Scheme

In the training stage, the united model requires ground truths of both the deblurring and optical flow. However, it is difficult to construct a large-scale dataset with such ground truths. Therefore, we employ a iterative training scheme



(a) Input frame 1 (b) Input frame 2

(c) Ground truth optical flow (d) Flow warp error

Figure 2. Flow warp error.

instead. The optical flow branch is initialized using the weights from FlowNetS in [7] while no existing models is used to initialize the deblurring branch. We first learn the parameters of the deblurring branch with the optical flow branch fixed. After training 12000 iterations, we switch to train the optical flow branch with with the deblurring branch fixed. The process repeats until the proposed loss function reach a convergence. In this way, only one of the ground truths of the deblurring and optical flow is needed to optimize the network gradually. During the training process, the network focuses on one branch to useful feature representations for current task. The learned feature representations are also utilized for another task through bi-directional propagation. Therefore, our iterative training scheme not only solves the demand of ground truths, but also can learn useful representations foir both tasks.

### 4.2. Training Datasets

Generating realistic training data is a great challenge for video deblurring since the ground truth sharp frames cannot be obtained easily and the blur is complex. Su et al. construct a benchmark which contains videos captured at 240fps with various devices like iPhone 6s, GoPro Hero 4 Black, and Canon 7D. The ground truth sharp videos are generated by subsampling every eighth frame. This benchmark dataset includes two sub-datasets: quantitative and qualitative ones. The quantitative subset consists of 6708 blurry frames and their corresponding ground-truth sharp frames from 71 videos. The qualitative one contains 2714 blurry frames of 22 scenes without ground truth. In this paper, the quantitative subset is spilt in to 61 videos for training and 10 videos for testing, which is the same to [29].

To train the optical flow branch, we first use the same traing dataset with the deblurring branch. Since no optical flow ground turth exist, the image warping loss plays the role in the optimization. In addition, we also train our model on the benchmark dataset: MPI-Sintel [2], which is a synthetic dataset with dense ground truth flow. The MPI-Sintel dataset provides two versions, Clean and Final, that both contain 1041 images. On the one hand, both versions contain small displacements and large motion. On the other hand, the Final version of images contain complicated environment variables like motion blur and atmospheric effects while the Clean version of images not. Since the Final version of MPI-Sintel dataset is closely related to the motion blur. The feature representations learnt from the optical flow branch can be helpful for the deblurring branch through bi-directional propagation. Therefore, the Final version of MPI-Sintel dataset is also employed as our training dataset.

For data augmentation, we adopt affine transformations (i.e., scaling, rotating, flipping) and then crop from images to generate various patch samples. Note that the affine transformations and crop coordinates keep the same among the frames of each input sample.

### 4.3. Implementation details

When traing VDFlow, we use a batch size of 10, and patches of $256 \times 256$. We use the msra distribution [9] to initialize the deblurring weights and ADAM [13] for minimzing the loss. The learning rate starts from $0.0001$ and is divided by 2 when the error plateaus. We employ a weight decay of $0.0004$ and momentums of $\beta 1 = 0.9, \beta 2 = 0.999$. For all the results reported in the paper we train the network for 36000 iterations. It takes about 12 hours on an NVidia $K80$ GPU, which indicates that our model has the ability to deal with large data. At the test time, our method deblurs images in a single forward pass, which is computationally very efficient. We can process a $720p$ frame within 20s on an NVidia $K80$ GPU. While the method in [29] is also computationally efficient,other previous approaches took more than 1 hour [4] and several minutes [12] per frame GPUs.

## 5. Experimental Results

In this section, we conduct experiments to evaluate the effectiveness of the VDFlow model. For delurring task, we first show the improvement by using the optical flow branch, and then shown comparisons to existing methods quantitatively and qualitatively. For the optical flow branch, we compare the proposed algorithm with the baseline FlownetS. As for the evaluated datasets, we first compare our method with existing approaches on the 10 test quantitative videos and other qualitative videos in [29] to show the deblurring performance of the proposed algorithm. We then use the Clean version of MPI-Sintel dataset [2] to demonstrate the effectiveness of optical flow estimation. The Peak-Signal-to-Noise-Ratios (PSNR) and the Mean Structural SIMilarity (MSSIM) are used as the performance evaluation standards in deblurring comparisons while the average endpoint errors (EPE) are employed for optical flow comparisons since the ground truth flow available.

(a) Ground truth

(b) Blurry (28.03 dB, 27.43dB, 26.05dB, 24.12dB)

(c) DBN+NOALIGN (28.39dB, 27.94dB, 26.32dB, 24.61dB)

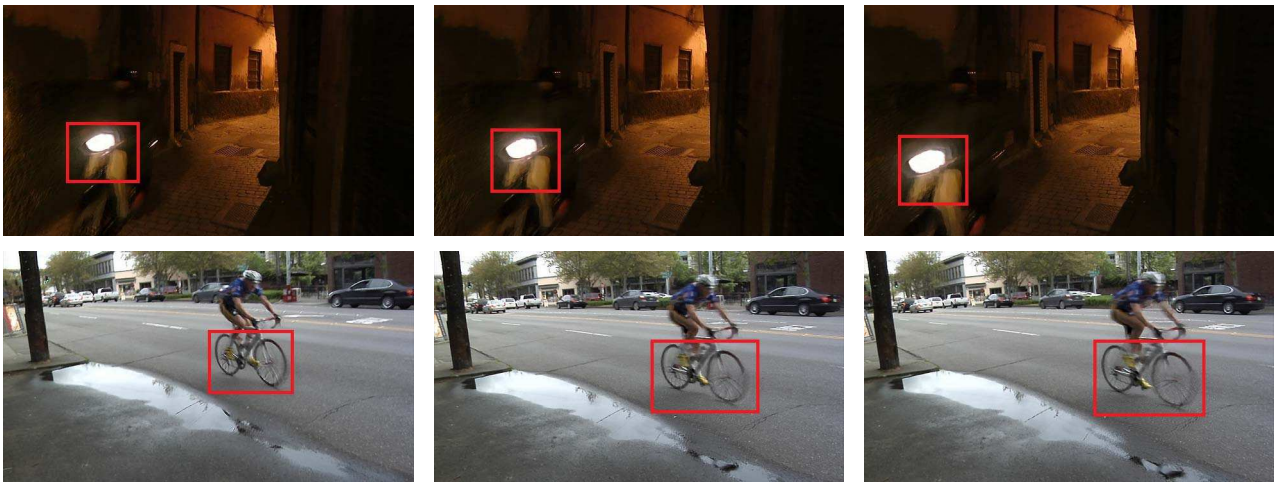
(d) VDFlow (29.30dB, 28.94dB, 26.98dB, 25.59dB)

Figure 3. Visual and quantitative comparison of VDFlow and DBN+NOALIGN. The PSNRs of the whole images and the selected patchs are labeled.

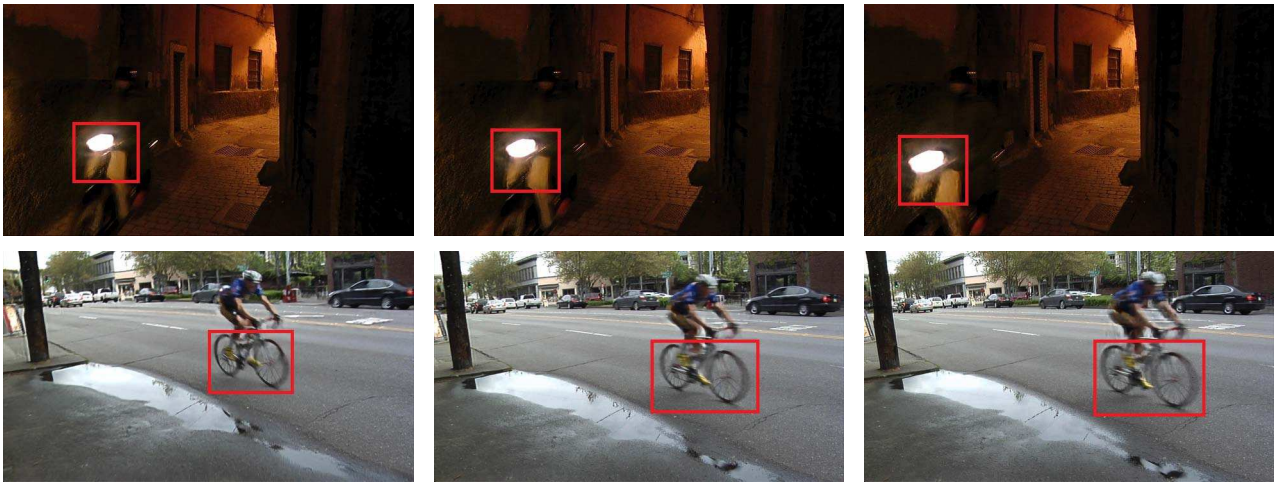
### 5.1. Deblurring Results

**Effects of optical flow branch.** We analyze the contribution of using the optical flow branch as the information carrier about motion trajectory. Note that we do not use the frame alignment, which is essential to most pervious methods for video deblurring. In [29], DBN+NOALIGN also achieves comparable results to previous work without alignment. Therefore, we compare the proposed algorithm with DBN+NOALIGN. On the one hand, both our approach and DBN+NOALIGN are able to handle the input frames. On the other hand, our proposed VDFlow outperforms DBN+NOALIGN, such as the instance in the examples in Figure 3. with more clearer details. In addition, we label the PSNRs of the whole images and the selected patchs as well. This phenomenon is mainly due to the differenece between various motion degrees. The relative motions are small in some cases, so the displacement is also small between neighboring frames which are captured. Therefore, there is little need to use the alignment between neighboring frames. While in other cases the displacements are hardly to be ignored, which makes deblurring methods without alignment, like DBN+NOALIGN, failed. Although our proposed VDFlow does not perform alignment as well, the feature information of optical flow is utilized through the bi-directional propagation and influences the following convolutional operations in the deblurring branch. Therefore, the optical flow branch can be regarded as a substitution of the alignment in the feature level. From this point of view, using the nerual network to learn the relationship between the optical flows and blurry frames is more reasonable than directly aligning.

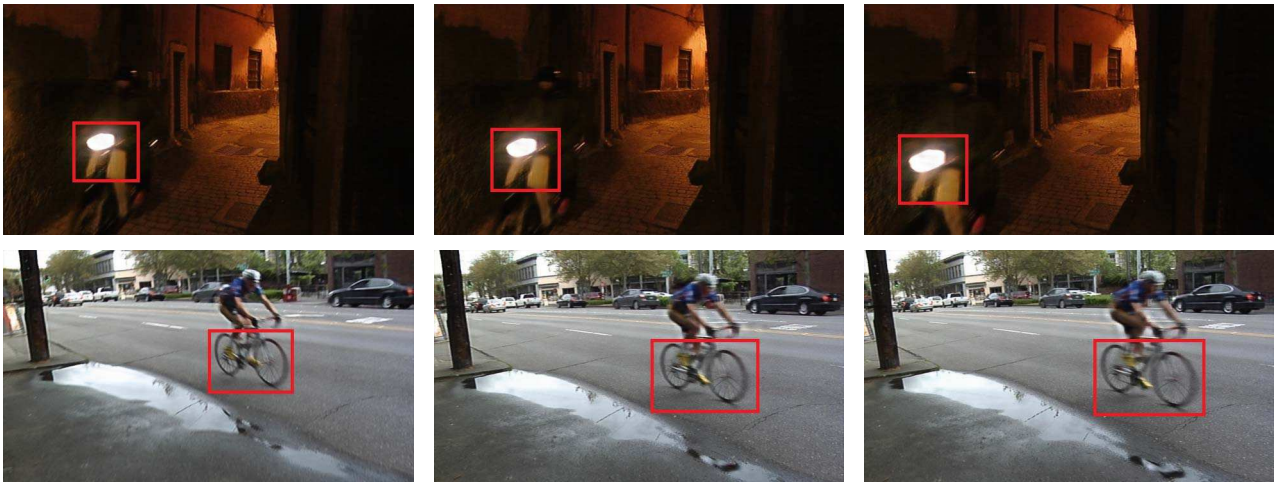
**Comparisons to existing methods.** We compare the proposed VDFlow with the state-of-the-art methods including DBN+NOALIGN [29] and DBN+SINGLE, on the quantitative dataset in [29]. DBN+NOALIGN shows the state-of-the-art performance on the video deblurring dataset and



(a) PWNLK [24]



(b) DBN+NOALIGN [29]



(c) VDFlow

Figure 4. Visual comparisons of PWNLK [24], DBN+NOALIGN [29], and the proposed VDFlow.

Table 1. Quantitative results on the dataset [29] with comparisons to DBN+SINGLE and DBN+NOALIGN. Average PSNR/MSSIM [14] measurements are calculated for 10 test datasets ($\#1 \rightarrow \#10$).

| Methods | DBN+SINGLE | | DBN+NOALIGN | | VDFlow | |
|---|---|---|---|---|---|---|
| Evaluation | PSNR | MSSIM | PSNR | MSSIM | PSNR | MSSIM |
| #1 | 25.33 | 0.884 | 25.45 | 0.886 | 25.98 | 0.897 |
| #2 | 30.26 | 0.958 | 30.53 | 0.961 | 31.03 | 0.963 |
| #3 | 28.56 | 0.925 | 28.84 | 0.927 | 29.84 | 0.941 |
| #4 | 28.14 | 0.907 | 28.23 | 0.908 | 28.69 | 0.915 |
| #5 | 22.71 | 0.864 | 22.82 | 0.866 | 23.12 | 0.874 |
| #6 | 29.22 | 0.953 | 29.36 | 0.954 | 29.90 | 0.958 |
| #7 | 27.83 | 0.947 | 27.97 | 0.948 | 28.42 | 0.952 |
| #8 | 24.02 | 0.911 | 24.13 | 0.913 | 24.71 | 0.923 |
| #9 | 30.65 | 0.977 | 30.91 | 0.978 | 31.82 | 0.981 |
| #10 | 26.26 | 0.928 | 26.44 | 0.930 | 27.04 | 0.937 |
| Average | 27.30 | 0.925 | 27.47 | 0.927 | 28.17 | 0.934 |

Table 2. Average endpoint errors for optical flow.

| Methods | Sintel $Clean$ train | Sintel $Clean$ test |
|---|---|---|
| FlowNetS | 4.50 | 7.42 |
| VDFlow | 4.38 | 7.26 |

DBN+SINGLE is a variant of DBN which replicate the central reference frame 5 times instead of a stack of neighboring frames. Both DBN+SINGLE and DBN+NOALIGN are trained using the same data and training iteration with our VDFlow. Results for quantitative comparisons are shown in Table 1. Our VDFlow performs better DBN+SINGLE by up to 0.87dB in terms of PSNR, and by up to 0.70dB than DBN+NOALIGN on average. In terms of MSSIM, these three methods are qualitatively equivalent. Therefore, We also evaluate the proposed algorithm against the state-of-the-art video deblurring approaches on real challenging sequences. Figure 4 shows results of two examples which contain high-speed moving objects. Since large the motion blur exist in these sequences, it is difficult to aggregate the information between frames. As shown in Figure 4, our deblurred frames contain fewer artifacts than other methods, which demonstrate the effectiveness of the proposed VDFlow model. Therefore, our proposed approach generalizes well in real scenes.

### 5.2. Optical Flow Results

Since our model employs five neighbor frames as the input, some traditional optical flow datasets such as KITTI 2012 dataset [8] are not meet the condition. We compare our method against our baseline FlowNetS, which is employed in the flow branch, on the $Clean$ version of MPI-Sintel dataset. Table 2 shows the average endpoint error of our VDFlow model and our baseline FlowNetS. Note that our VDFlow model achieves lower endpoint errors against FlowNetS, which validates the benefit of incorporating the information from the deblurring branch.

## 6. Conclution

In this paper, we establish a united encoder-decoder style network to estimate the sharp frames and optical flow in videos simultaneously. The feature representations of the deblurring branch and the optical flow branch are bi-directional propagated to help each other task. In this way, the proposed model aggregates the information from neighboring frames and optical flows together in feature level, which is more reasonable than directly using the optical flows to align frames. Extensive experimental results on challenging videos show that the proposed algorithm performs favorably existing against the state-of-the-art methods.

**Acknowledgments.** This work was supported by the Joint Foundation Program of the Chinese Academy of Sciences for equipment prefeasibility study (141A01011601), the National Natural Science Foundation of China (Nos. 61775219, 61771369, 61640422, and 61802403), the Equipment Research Program of the Chinese Academy of Sciences (Nos.YJKYYQ20180039 and Y70X25A1HY), the CCF-Tencent Open Fund, and Zhejiang Lab's International Talent Fund for Young Professionals.

## References

[1] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 1

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 5

[3] Xiaochun Cao, Wenqi Ren, Wangmeng Zuo, Xiaojie Guo, and Hassan Foroosh. Scene text deblurring using text-specific multiscale dictionaries. *TIP*, 24(4):1302–1314, 2015. 1

[4] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *ECCV*, pages 221–235, 2016. 2, 5

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

[6] Mauricio Delbracio and Guillermo Sapiro. Hand-held video deblurring via efficient fourier aggregation. *TCI*, 1(4), 2015. 2

[7] Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 1, 2, 3, 5

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 8

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 5

[10] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3), 1981. 2

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *CoRR*, abs/1612.01925, 2016. 2, 3

[12] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *CVPR*, 2015. 1, 2, 5

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014. 5

[14] Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *ECCV*, 2012. 8

[15] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single image dehazing and beyond. *TIP*, 2018. 2

[16] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *TIP*, 2020. 2

[17] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *CVPR*, 2019. 2

[18] Yu Liu, Guanlong Zhao, Boyuan Gong, Yang Li, Ritu Raj, Niraj Goel, Satya Kesav, Sandeep Gottimukkala, Zhangyang Wang, Wenqi Ren, et al. Improved techniques for learning to dehaze and beyond: A collective study. *arXiv preprint arXiv:1807.00202*, 2018. 2

[19] Yasuyuki Matsushita, Eyal Ofek, Xiaoou Tang, and Heung Yeung Shum. Full-frame video stabilization. In *CVPR*, 2005. 2

[20] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *ECCV*, 2014. 2

[21] Wenqi Ren and Xiaochun Cao. Deep video dehazing. In *Pacific rim conference on multimedia*, 2017. 3

[22] Wenqi Ren, Xiaochun Cao, Jinshan Pan, Xiaojie Guo, Wangmeng Zuo, and Ming-Hsuan Yang. Image deblurring via enhanced low rank prior. *TIP*, 25(7):3426–3437, 2016. 1

[23] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, 2018. 3

[24] Wenqi Ren, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In *ICCV*, 2017. 1, 7

[25] Wenqi Ren, Jinshan Pan, Hua Zhang, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks with holistic edges. *IJCV*, pages 1–20, 2019. 2

[26] Wenqi Ren, Jiaolong Yang, Senyou Deng, David Wipf, Xiaochun Cao, and Xin Tong. Face video deblurring using 3d facial priors. In *ICCV*, 2019. 2

[27] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *TIP*, 28(4):1895–1908, 2018. 1

[28] Christian J. Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. Learning to deblur. *TPAMI*, 38(7):1439–1451, 2016. 2

[29] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring. *CoRR*, abs/1611.08387, 2016. 1, 3, 4, 5, 6, 7, 8

[30] Kalyan Sunkavalli, Neel Joshi, Sing Bing Kang, Michael F Cohen, and Hanspeter Pfister. Video snapshots: Creating high-quality images from video clips. *TVCG*, 18(11):1868–79, 2012. 2

[31] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, 2013. 2

[32] Qingbo Wu, Wenqi Ren, and Xiaochun Cao. Learning interleaved cascade of shrinkage fields for joint image dehazing and denoising. *TIP*, 29:1788–1801, 2019. 2

[33] Qingbo Wu, Jingang Zhang, Wenqi Ren, Wangmeng Zuo, and Xiaochun Cao. Accurate transmission estimation for removing haze and noise from a single image. *TIP*, 2019. 2

[34] Jonas Wulff and Michael Julian Black. Modeling blurred video with layers. In *ECCV*, 2014. 1, 2

[35] Yanyang Yan, Wenqi Ren, and Xiaochun Cao. Recolored image detection via a deep discriminative model. *TIFS*, 14(1):5–17, 2018. 2

[36] Yanyang Yan, Wenqi Ren, Yuanfang Guo, Rui Wang, and Xiaochun Cao. Image deblurring via extreme channels prior. In *CVPR*, 2017. 2

[37] Ye Yuan, Wenhan Yang, Wenqi Ren, Jiaying Liu, Walter J Scheirer, and Zhangyang Wang. Ug2+ track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments. *arXiv preprint arXiv:1904.04474*, 2019. 2

[38] Haichao Zhang and Jianchao Yang. Intra-frame deblurring by leveraging inter-frame camera motion. In *CVPR*, 2015. 2

[39] Shengdong Zhang, Fazhi He, Wenqi Ren, and Jian Yao. Joint learning of image detail and transmission map for single image dehazing. *The Visual Computer*, pages 1–12, 2018. 2

[40] Shengdong Zhang, Wenqi Ren, and Jian Yao. Feed-net: Fully end-to-end dehazing. In *ICME*, 2018. 2