ERDN: Equivalent Receptive Field Deformable Network for Video Deblurring

Bangrui Jiang[1,2], Zhihuai Xie[2*], Zhen Xia[2], Songnan Li[2], and Shan Liu[2]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
jbangrui@gmail.com
[2] Tencent Media Lab, Shenzhen, China
{zhihuaixie,zhenxia,sunnysnli,shanl}@tencent.com

Abstract. Video deblurring aims to restore sharp frames from blurry video sequences. Existing methods usually adopt optical flow to compensate misalignment between reference frame and each neighboring frame. However, inaccurate flow estimation caused by large displacements will lead to artifacts in the warped frames. In this work, we propose an equivalent receptive field deformable network (ERDN) to perform alignment at the feature level without estimating optical flow. The ERDN introduces a dual pyramid alignment module, in which a feature pyramid is constructed to align frames using deformable convolution in a cascaded manner. Specifically, we adopt dilated spatial pyramid blocks to predict offsets for deformable convolutions, so that the theoretical receptive field is equivalent for each feature pyramid layer. To restore the sharp frame, we propose a gradient guided fusion module, which incorporates structure priors into the restoration process. Experimental results demonstrate that the proposed method outperforms previous state-of-the-art methods on multiple benchmark datasets. The code is made available at: https://github.com/TencentCloud/ERDN.

Keywords: Video deblurring, deformable convolution, receptive field

1 Introduction

The goal of video deblurring is to recover high quality consecutive frames from their blurry counterparts. Camera shake and object motion, which are common when capturing videos with hand-held devices, serve as the main causes of video blur. Video deblurring is a fundamental problem in the community of computer vision and can be beneficial to many high-level tasks including video segmentation and video understanding. However, it is an ill-posed problem since each blurry frame may have multiple sharp solutions.

Compared to image deblurring, video deblurring can use neighboring frames for restoration. A vital issue in video deblurring is how to align the neighboring frames with the reference frame, since the neighboring frames may not be naturally aligned due to motions of objects. Previous methods [21, 9, 29, 18, 12]

* Corresponding Author

usually exploit optical flow to predict motion fields, which can be used to warp neighboring frames to the reference frame. However, these methods perform flow estimation explicitly and wrap frames at the image level, which may introduce artifacts in regions with large displacements. ARVo [12] further estimates pairwise image correspondence to compensate misalignment caused by inaccurate flow estimation, while extra computational cost and storage space are required.

Besides flow-based alignment, deformable convolution [7] is widely adopted as another alignment method and achieve remarkable performance in video restoration [24, 13, 26]. When deformable convolution is applied in temporal alignment, the displaced kernels on neighboring frames will be used to align intermediate features from several locations, while optical flow only samples from one location. Because of the diverse sampling, deformable convolution tends to perform better than flow-based alignment [3]. However, existing works [24, 26] usually predict offsets in a limited receptive field, which leads to relatively local offset predictions and fail to achieve comparable performance.

To alleviate the problem, we propose an equivalent receptive field deformable network (ERDN) that performs temporal alignment using deformable convolution. A dual pyramid alignment module is developed to construct a feature pyramid for reference frame and each neighboring frame. In particular, the module first aligns high level features with coarse offset estimations and then propagates the offsets and the aligned features to lower levels, in which way the offsets can be refined in a coarse-to-fine manner. However, different from [26], we propose a novel dilated spatial pyramid block to predict offsets for each feature pyramid layer. The block uses dilated convolution to guarantee each layer having equivalent receptive field, with the principle that the offset refinement is more effective to be performed in an equivalent region. To compensate information loss caused by dilated convolution, the block further concatenates several convolution layers with different dilation rates, which constitute a spatial receptive field pyramid.

In order to preserve structures for further enhancing the deblurring performance, we introduce a gradient guided fusion module, in which a gradient branch converts the gradient maps of blurry frames to the sharp ones. The recovered gradients can be integrated into the deblurring branch with a series of Spatial Feature Transform (SFT) layers [25], allowing our method to incorporate structure priors. To the best of our knowledge, we are the first to introduce gradient branch in video deblurring and achieve success.

The main contributions can be summarized as follows:

1. We propose a novel equivalent receptive field deformable network (ERDN) for video deblurring, which performs alignment using deformable convolution without optical flow estimation.
2. To handle large displacement, we propose a dual pyramid alignment module, which predicts offsets in a cascaded manner within equivalent receptive field.
3. We introduce a gradient branch to incorporate structure priors as a guidance for the frame restoration.
4. Extensive experiments show that our method achieves superior performance over previous state-of-the-art methods quantitatively and qualitatively.

2 Related Works

2.1 Video Deblurring

Early video deblurring approaches [1, 28, 20] mostly estimate blur kernels and restore sharp frames by applying deconvolution with the estimated kernels. Some works [1, 28] utilize segmentation map for blur kernel estimation, and process blur caused by moving objects. More recently, Ren et al. [20] propose a pixel-wise non-linear kernel model, where the blur kernel is estimated from optical flow. However, estimating spatially-varying blur kernels is a severely ill-posed problem. These approaches may fail when applied to videos that include complex blur.

To overcome the above problems, several methods explicitly align frames using optical flow. Su et al. [21] restore the deblurred frame using channel-concatenated neighboring frames, which are aligned by an external optical flow module. Pan et al. [18] plug an optical flow estimation module into an end-to-end model while using temporal sharpness prior. Based on [18], Li et al. [12] further incorporate a correlation volume pyramid to learn spatial correspondence between pixel pairs in the feature space. The spatial correspondence can serve as a complement to the optical flow alignment. Most of the optical flow-based methods align frames at the image level, while incorrect flow estimation may introduce image artifacts.

By contrast, a few methods adopt implicitly alignment using Recurrent Neural Network (RNN) for its excellent performance on time-series signal [5, 23]. Wieschollek et al. [27] introduce a novel recurrent encoder-decoder network and transfer temporal feature between subsequent iterations. Hyun Kim et al. [10] incorporate a dynamic temporal blending mechanism that enable adaptive information propagation. The recent method [31] uses a recurrent cell based on residual dense blocks and a global spatio-temporal attention module. The above works have achieved considerable success, but are still insufficient to deal with fast motion and large displacement. In this work, we alternatively use deformable convolution for implicitly temporal alignment.

2.2 Deformable Convolution

Dai et al. [7] first propose deformable convolution, which is originally applied to high-level vision tasks, such as object detection [2] and crowd understanding [14]. By learning an additional offset, deformable convolution has spatial flexibility to obtain information from several locations. Due to its sampling diversity, deformable convolution has been creatively utilized in low-level video restoration. In video super-resolution, Tian et al. [24] first adopt deformable convolution to align frames at the feature level. Yue et al. [30] successfully apply deformable alignment to video denoising. Lin et al. [13] further integrate optical flow into deformable convolution in order to add explicit motion constraints.

However, deformable convolution is rarely explored in video deblurring. Wang et al. [26] propose a pyramid deformable architecture for video super-resolution, and attempt to transfer it into video deblurring without specific designs. Since

the offsets are predicted by typical convolution layers in [26], the theoretical receptive field is limited, which may lead to performance drop in occurrence of fast motion and large displacement. In contrast, our work develops a well-designed deblurring framework based on deformable convolution, and proposes a novel dual pyramid alignment module that introduces a larger receptive field.

3 Method

3.1 Overview

Let $I_t \in \mathbb{R}^{H\times W\times C}$ be the t-th blurry video frame and $I_t^s \in \mathbb{R}^{H\times W\times C}$ be the ground-truth sharp frame, where $H \times W$ denotes the frame size, and C refers to channel number. Our goal is to restore the sharp video frame I_t^r from the consecutive $2N+1$ frames $\{I_i\}_{i=t-N}^{t+N}$. I_t^r should be as similar to I_t^s as possible.

The overall framework is depicted as Fig. 1. It consists of two modules: 1) Dual pyramid alignment module that aligns neighboring frames with reference frame. 2) Gradient guided fusion module that reconstructs the sharp frame.

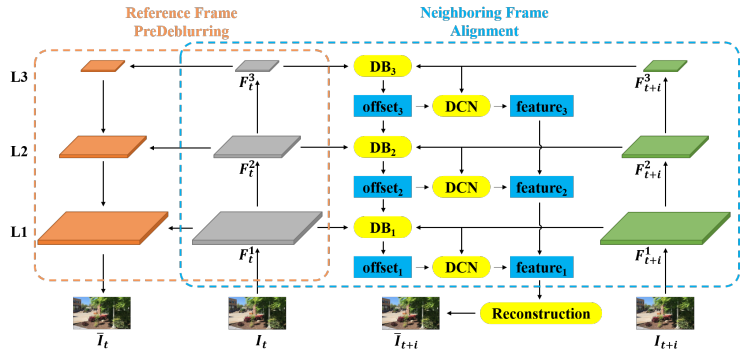
3.2 Dual Pyramid Alignment Module

Motivation: Cascading refinement has been well-established in optical flow estimation and achieves remarkable performance [11, 8, 22]. In PWC-Net [22], the main principle for optical flow refinement can be summarized as three steps: warping, cost volume computation and optical flow estimation. The PWC-Net first warps features of the target image toward the reference image using the coarse optical flow. Next, a cost volume, which stores the matching costs for associating a pixel with its corresponding pixels, is constructed between features of the reference image and warped features of the target image. Finally, a refined optical flow is estimated based on the cost volume and the coarse optical flow. The cost volume is explicitly computed to indicate errors of coarse estimation.

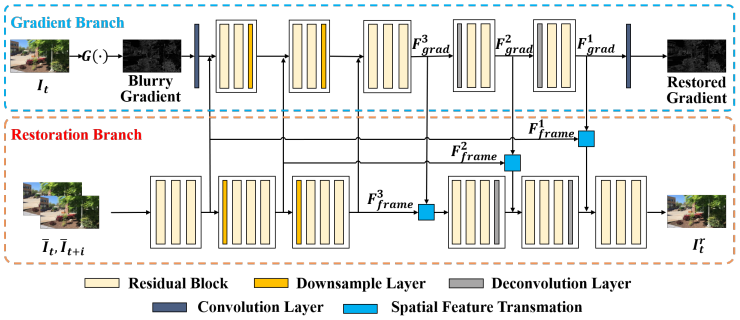
However, the principle of cascading refinement has not been thoroughly studied in the deformable alignment, which serves as another important alignment method in video restoration [26, 30]. The EDVR [26] predicts offsets at each pyramid level using several convolution layers. Then, the offsets predicted at higher level will be propagated to lower level, and fused with offsets from lower level. During the whole procedure, refinements are conducted via the fusion of offsets from different levels. However, the offsets of higher level are predicted in relatively larger receptive fields compare to those of lower level, which may lead to inconsistency of offset scales. Therefore, the refinement may fail since offsets in larger scale can not facilitate offset estimation in smaller scale.

In this work, we guarantee equivalent receptive field in different levels so that offset in deeper level can guide offset prediction in lower level. In other words, our framework first detect region with similar structure, then detect area with similar details among that region. More details are described in the following.

Alignment in Feature Pyramid: The alignment module aims to align neighboring frames to reference frame, while pre-deblurring the reference frame.



(a) Dual Pyramid Alignment Module



(b) Gradient Guided Fusion Module

Fig. 1: Overall framework of our ERDN method. The dual pyramid alignment module aims to align neighboring frames using deformable convolution in a cascaded manner. The module adopts dilated spatial pyramid blocks (DB) to predict offsets within equivalent receptive fields. A gradient guided fusion module is utilized to restore frame, taking reference frame and aligned neighboring frames as input. For simplicity, we only present the feature pyramid with three layers.

The proposed module constructs a feature pyramid and aligns features in each layer using deformable convolution. As for a deformable convolution kernel with K sample positions, we note the learned offsets for location p as $\Delta P_{t+i}(p)_k$. Then aligned feature F^a_{t+i} can be obtained using deformable convolution:

$$F^a_{t+i}(p) = \sum_{k=1}^{K} \omega_k \cdot F_{t+i}(p + p_k + \Delta P_{t+i}(p)_k) \tag{1}$$

where F_{t+i} represents features extracted from neighboring frame I_{t+i}, while ω_k and p_k represent weight and pre-specified offset respectively. The learned offsets ΔP_{t+i} can be predicted from the features of reference frame F_t and the features of neighboring frame F_{t+i}.

Inspired by [26, 30], we align frames in a cascaded manner. As shown in Fig. 1, we use strided convolution with factor 2 to generate F^l_{t+i} at the l-th level, obtaining a four-level pyramid of feature representation. At the l-th level, offsets and aligned features are predicted using the l-th features together with the upsampled offsets and the aligned features from the $(l+1)$-th level:

$$\Delta P^l_{t+i} = DB([F^l_{t+i}, F^l_t],\ (\Delta P^{l+1}_{t+i})^{\uparrow 2}) \tag{2}$$

$$(F^a_{t+i})^l = f(DCN(F^l_{t+i}, \Delta P^l_{t+i}), ((F^a_{t+i})^{l+1})^{\uparrow 2}) \tag{3}$$

where DCN is deformable convolution described in Eqn. 1 and f represent several convolution layers, while $(\cdot)^{\uparrow 2}$ refers to $\times 2$ upsampling. Specially, we utilize a dilated spatial pyramid block (DB) to predict offsets for each layer. After cascaded alignment, $(F^a_{t+i})^1$ is further applied into several residual blocks to reconstruct aligned frame $\bar{I}_{t+i}$, while a decoder is used to restore reference frame to obtain pre-deblurring frame $\bar{I}_t$.

Dilated Spatial Pyramid Block: A vital step in the cascaded alignment is offset refinement. However, previous methods [26, 30] predict offsets using several convolution layers at different levels, which leads to inconsistency of receptive fields, as shown in Fig. 2 (a). Due to the inconsistency, offsets predicted at the higher level may exceed receptive field at the lower level, so that the offsets will not be refined at lower levels.

A simple solution is to replace typical convolution with dilation convolution for offset prediction, making each level to have equivalent receptive field. However, for a pixel p in a dilated convolution layer, the information that contributes to pixel p comes from a nearby $k_d \times k_d$ region centered at p. Since dilated convolution introduces zeros in the convolution kernel, the actual pixels that participate in the computation from the $k_d \times k_d$ region are just $k \times k$. As a result, pixel p can only view information from limited location which can be irrelevant across large distances, and lose a large portion of information.

To overcome the previous issue, we propose a dilated spatial pyramid block, which concatenates several convolution layers with different dilation rates.

A dilated layer is defined as follows:

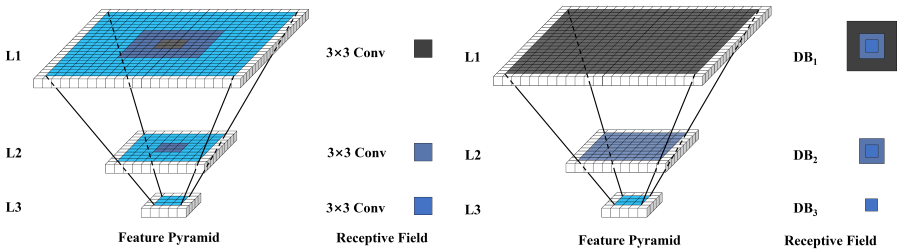
$$C_i(F) = DConv_d(Conv_k(F)) \tag{4}$$

where $DConv_d$ represents dilated convolution with dilation rate d and kernel size 3×3, and $Conv_k$ represents typical convolution with kernel size $k \times k$.

In this work, we adopt 3×3 convolution layers with stride 2 for downsampling. Therefore, a $k \times k$ theoretical receptive field at the l-th level correspond to $(2k+2) \times (2k+2)$ theoretical receptive field at the $(l-1)$-th level. In practice, we use receptive field with size $(2k+3) \times (2k+3)$. Therefore, we set d and k in Eqn. 4 as $d = k = 2^i - 1$, where the theoretical receptive field is $(2^{i+1} + 2^i - 3) \times (2^{i+1} + 2^i - 3)$.

At l-th level of a N-layer feature pyramid, the dilated spatial pyramid block is built as:

$$DB_l(F, \Delta P) = f([Conv_1([C_1(F), \cdots, C_{N-l+1}(F)]), \Delta P]) \tag{5}$$



(a) EDVR (b) Ours

Fig. 2: EDVR predicts offsets using 3×3 convolution at every scales, which will lead to inconsistency of receptive fields due to downsample operation as shown in (a). The proposed method introduces dilated spatial pyramid blocks, which guarantees predicting offsets within equivalent receptive field. The dilated spatial pyramid block constructs a receptive field spatial pyramid as shown in (b). The receptive fields of DB_3, DB_2 and DB_1 are 3×3, 9×9 and 21×21 respectively.

where f represents several convolution layers and $[\cdot]$ represents concatenation.

Similar to ASPP [4] and RFB [15], the dilated spatial pyramid block makes use of multi-branch pooling with varying kernels corresponding to receptive fields of different sizes. Specially, the proposed dilated spatial pyramid block (DB) compensates inconsistency of receptive field, as shown in Fig. 2 (b).

3.3 Gradient Guided Fusion Module

The target of the fusion network is to restore sharp frame from the aligned neighboring frames and the pre-deblurred reference frame. Inspired by [16], we introduce a simple but effective gradient branch, which is to translate gradient maps from the blurry modality to the sharp one. The gradient map for an image I is obtained by computing the difference between adjacent pixels:

$$\begin{aligned} G_x(\mathbf{x}) &= I(x+1, y) - I(x-1, y) \\ G_y(\mathbf{x}) &= I(x, y+1) - I(x, y-1) \\ G(I) &= \sqrt{G_x^2 + G_y^2} \end{aligned} \tag{6}$$

Where $G(\cdot)$ stands for the operation to extract gradient map for pixels with coordinates $\mathbf{x} = (x, y)$. As shown in Fig. 1, the proposed module first extracts gradient map from reference frame, which is applied to gradient branch. The gradient branch incorporates several intermediate features from the restoration branch, because the restoration branch carries structure information which is beneficial to the restoration of gradient maps. Then, several spatial feature transform layers are utilized to effectively incorporate structure priors, which can implicitly reflect whether a region should be sharp or smooth. Specially, the

gradient information is integrated in different scales, taking into account both structure information at high level and detailed texture information at low level.

In our task, both gradient branch and restoration branch adopt encoder-decoder structures. In each scale, we integrate gradient features $\boldsymbol{F}_{grad}$ using a pair of affine transformation parameters (α, β), which is learned using several convolutional layers. After that, the modulation is carried out by scaling and shifting the frame feature $\boldsymbol{F}_{frame}$:

$$\begin{aligned} \alpha,\ \beta = f_1(\boldsymbol{F}_{grad}),\ f_2(\boldsymbol{F}_{grad}) \\ \boldsymbol{F}_{output} = SFT(\boldsymbol{F}_{frame}|\alpha, \beta) = \alpha \odot \boldsymbol{F}_{frame} + \beta \end{aligned} \tag{7}$$

Although gradient branch has been used in [16], the original gradient map in video deblurring is more difficult to be reconstructed due to the blur structure. The proposed gradient branch has a more effective structure, which performs feature fusion at three scales, taking into account local texture information and global structural information, and adopts SFT for feature fusion.

3.4 Training Strategy

Cascaded Training: Following [18] and [12], we train the proposed method using a cascaded strategy. At stage n, we restore the frame $I^r_{t,n}$ from three consecutive frames $\{I^r_{i,n-1}\}^{t+1}_{i=t-1}$, which are the outputs of stage $n-1$. In particular, the proposed method takes $\{I_i\}^{t+1}_{i=t-1}$ and outputs $I^r_{t,1}$ for the first stage. Therefore, a N stage training strategy will take $2N+1$ blurry frames as input.

Then, the overall loss function $L_{overall}$ for N stage strategy can be formulated as follows:

$$L_{overall} = \sum_{n=1}^{N} \sum_{i=t-N+n}^{t+N-n} L_{stage}(\{I^r_{j,n-1}\}^{i+1}_{j=i-1}, I^s_i) \tag{8}$$

where L_{stage} represents loss function for recovering center frames from three consecutive frames and the final restoration result is $I^r_{t,N}$.

In this work, we use five consecutive frames for restoration with two stages as the trade-off between efficiency and performance. Then the overall loss function can be specified as:

$$L_{overall} = \sum_{i=t-1}^{t+1} L_{stage}(\{I_j\}^{i+1}_{j=i-1}, I^s_i) + L_{stage}(\{I^r_{j,1}\}^{t+1}_{j=t-1}, I^s_t) \tag{9}$$

Loss Functions: The training objective L_{stage} consists of reconstruction loss L_{rec}, alignment loss L_{align} and gradient loss L_{grad}.

We adopt the widely-used L1 loss as our reconstruction loss, defined as follows:

$$L_{rec} = \|I^r_t - I^s_t\|_1 \tag{10}$$

To stabilize the training of dual pyramid alignment module, we add an additional alignment loss:

$$L_{align} = \frac{1}{2N+1} \sum_{i=t-1}^{t+1} \|\bar{I}_i - I_i^s\|_1 \tag{11}$$

Specially, we argue that the alignment module is used to align neighboring frames with sharp target frame rather than blurry reference counterpart.

We further design a gradient loss to penalize the difference between reconstructed gradient map and gradient map of sharp frame:

$$L_{grad} = \|G_t - G(I_t^s)\|_1 \tag{12}$$

where G_t represents recovered gradient maps by gradient branch. The overall objective is defined as follows:

$$L_{stage} = L_{rec} + L_{grad} + \lambda L_{align} \tag{13}$$

where λ represents weight of alignment loss.

4 Experiments

4.1 Implementation Details

Datasets and Evaluation Metrics: Following previous deblurring method [32, 18, 12], we adopt Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as the evaluation metrics. We evaluate the proposed method DVD [21], GOPRO [17] and HFR-DVD [12]. We provide a brief introduction here.

• *DVD* contains 71 videos (6,708 pairs) captured at 240 fps, splitting into 61 training videos (5708 pairs) and 10 testing videos (1000 pairs).

• *GOPRO* is composed of 33 videos (3214 frame pairs) captured at 240 fps, of which 22 videos (2103 pairs) are used for training and 11 videos (1111 pairs) are used for testing.

• *HFR* contains 120 videos for training and 30 videos for testing, each with 90 frames. It is a newly released dataset that captures videos at 1000 fps.

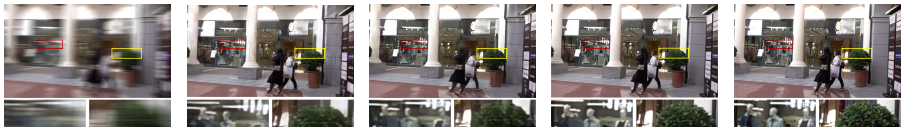
Training Details: As for the trade-off between performance and efficiency, we use a four-layer feature pyramid in the dual pyramid alignment module. We apply patches of size 256×256 for training, while adopting flip and rotation as data augmentation. Our method is implemented on Pytorch [19]. We set the training rate as 1×10^{-4} and reduce it to half every 200 epochs.

4.2 Comparisons with the State-of-the-art

Quantitative Comparison: We compare our method quantitatively with previous video deblurring methods including EDVR [26], STFAN [32], CDVD-TSP [18] and ARVo [12]. Results of PSNR and SSIM values are presented in Table 1.

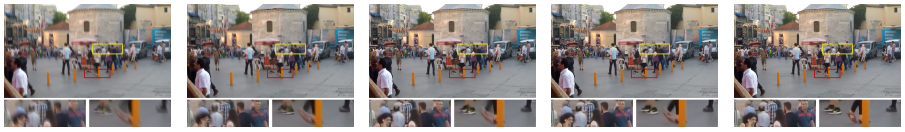
Table 1: Quantitative comparison with state-of-the-art methods. The best performance is denoted in **red**, while second best performance in blue.

Dataset	Metric	EDVR [26] (7 Frames)	STFAN [32] (2 Frames)	TSP [18] (5 Frames)	ARVo [12] (5 Frames)	Ours (3 Frames)	**Ours (5 Frames)**
DVD	PSNR	28.51	31.15	32.13	32.80	32.91	**33.31**
	SSIM	0.8637	0.9049	0.9268	0.9352	0.9334	**0.9395**
GOPRO	PSNR	26.83	28.59	31.67	-	32.20	**32.48**
	SSIM	0.8426	0.8608	0.9279	-	0.9288	**0.9329**
HFR	PSNR	29.15	28.48	29.71	31.15	32.01	**32.14**
	SSIM	0.8733	0.8560	0.8822	0.9063	0.9156	**0.9173**



(a) Blurry Input (b) EDVR [26] (c) TSP [18] (d) Ours (e) GT

Fig. 3: Qualitative comparisons on the DVD-HFR dataset [12]



(a) Blurry Input (b) EDVR [26] (c) TSP [18] (d) Ours (e) GT

Fig. 4: Qualitative comparisons on the GOPRO dataset [17].

In each row, the best result is highlighted in red while the second best is in blue. In all datasets, our proposed method achieves the best performance using five frames, while achieving the second best performance in most metrics using three frames. In comparision with the CDVD-TSP and ARVo on the DVD dataset, the ERDN yields significant improvements with 1.18 dB and 0.51 dB increases in the PSNR, respectively. On the GOPRO dataset, our ERDN achieves a 1.17 dB improvement in the PSNR and a tiny margin of improvement in the SSIM. Notably, compared with the ARVo on both the PSNR and SSIM metrics, the ERDN achieves considerable improvements of 0.99 dB and 0.011 on the HFR dataset, respectively. The results demonstrate that the ERDN has superior robustness.

Specifically, the EDVR is originally designed for video super-resolution and transferred to video deblurring without specific design. The method first applies downsampling to the input frames, which will lead to significant information loss. The TSP aligns frames using optical flow at the image level, which is less effective in occurrence of fast object motions. The ARVo learns spatial correspondence between pixel pairs as a complement to optical flow-based alignment.

(a) Blurry Input (b) TSP [18] (c) ARVo [12] (d) Ours (e) GT

Fig. 5: Qualitative comparisons on the DVD dataset [21].

(a) Blurry Input (b) Cho et al. [6] (c) EDVR [26] (d) TSP [18] (e) Ours

Fig. 6: Qualitative comparisons on the real blurry frames from [6].

(a) Frame 1 (b) Frame 2 (c) Frame 3 (d) Frame 3 (e) Frame 4

Fig. 7: Qualitative comparisons on consecutive frames from DVD dataset[21]

However, artifacts caused by inaccurate flow will affect the performance of the whole model. In contrast, our model adopts deformable convolution for better alignment at the feature level without estimating optical flow. With the well-designed dual pyramid alignment module, the proposed model can predict offsets in a larger receptive field, which is effective to handle large displacement.

Qualitative Comparison: We also conduct visual comparison on the three datasets, as shown in Fig. 3, 4 and 5. We observe that previous methods generate obvious artifacts and suffer from incomplete deblurring. In contrast, the proposed method is capable to restore clearer frames. Specially, since the HFR dataset [12] exhibits more fast motions, our method outperforms previous methods more significantly, which exhibits the strength of handling fast motions.

To further evaluate the effectiveness of our proposed method, we conduct experiment on the real video deblurring dataset released by [6]. As shown in

Table 2: Network ablation analysis on DVD dataset [21]. DCN, ERF, DB and GB represent deformable convolution, equivalent receptive field, dilated spatial pyramid block and gradient branch respectively.

	DCN	ERF	DB	GB	PSNR	SSIM	Param.
Baseline					32.13	0.9268	16.19M
Net-1	✓				32.82	0.9328	16.33M
Net-2	✓	✓			33.02	0.9359	16.33M
Net-3	✓	✓	✓		33.17	0.9378	17.54M
Net-4	✓	✓	✓	✓	**33.31**	**0.9395**	**22.84M**

Table 3: Effects of using different layers of feature pyramids on the video deblurring result. The final setting is highlighted.

Layers	1	2	3	**4**	5
PSNR	31.89	32.41	33.05	**33.31**	33.20
SSIM	0.9168	0.9256	0.9359	**0.9395**	0.9382

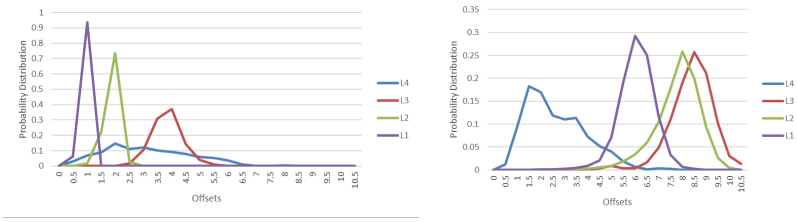
Fig. 6, our method restores better detailed structure with sharper outlines, which demonstrates promising generalization ability to handle real-world blurry videos.

As the temporal consistency property is very important for video deblurring, we also present deblurring results on a series of consecutive frames as shown in Fig. 7. It is obvious that inconsistent bands and artifacts appear near edges in TSP [18], while our proposed method demonstrates better consistency.

4.3 Ablation Studies

We conduct more experiments on different models to validate the necessity of each part in our proposed framework. Since we adopt the architecture of CDVD-TSP [18], we use it as the baseline model. The Net-1 replaces the optical flow estimation module with deformable convolution architecture similar to EDVR [26]. Compared to the Net-1, the Net-2 replaces typical convolution with dilated convolution without increasing parameters. The Net-3 adopts dilated spatial pyramid block instead of a single dilation convolution layer. The Net-4 further incorporates gradient branch. Quantitative comparison is presented in Table 2.

Effects of the Dual Pyramid Alignment Module: As shown in Table 2, the replacement of flow-based alignment with deformable alignment significantly improves deblurring performance, which reveals the effectiveness of deformable convolution for alignment in video deblurring. Furthermore, we enlarge receptive field with dilated convolution at low level, achieving an improvement of 0.2 dB without any extra parameters. The improvement of performance straightly demonstrates that refinement of offsets can be conducted more effectively within equivalent receptive field. With our proposed dilated spatial pyramid block, the Net-3 is nearly 0.2 dB better than the Net-2, since features in the receptive field is better explored compared to a single dilated convolution layer.



(a) EDVR (b) Ours

Fig. 8: Quantitative comparison of the offset distribution on video with fast motion. L1, L2, L3 and L4 represent offsets predicted at different feature pyramid levels, denoted the same as Fig. 1. Both the proposed method and EDVR predict offsets in a cascaded manner. Due to the inconsistency of receptive field, EDVR fails to refine offsets as shown in (a), while offsets in lower level are significantly smaller than those at higher level. Our method otherwise holds comparable offsets for each level.

(a) Blurry Gradient (b) Sharp Gradient (c) Restored Gradient

Fig. 9: Visualization of gradient maps. The sharp gradient map has larger intensity than the blurry counterparts. Our gradient branch is capable to recover gradient map with pleasant structures.

We further analyse the offsets using different methods quantitatively in Fig. 8. In EDVR, offsets predicted at different layers have different scales obviously, which reveals the failure of offset refinement when processing video with fast motion. In contrast, our method holds comparable offset scale and samples in a significant larger region at $L1$ layer. The major difference is that previous method predicts offsets with several convolution layers, while our method adopts dilated spatial pyramid block with larger receptive field at lower level.

We propose to construct a feature pyramid for alignment at different scales. To analyze the effect, we conduct experiments with different numbers of pyramid layers. As shown in Table 3, we observe that the restoration quality achieves improvement by adding pyramid layers. This is because enlarging receptive field helps to capture large displacement. However, we notice a little performance drop when using pyramid with five layers. This probably implies that the receptive field is too large to capture useful information.

(a) w/o GB (b) GB

Fig. 10: Qualitative comparison of the models with and without the gradient branch. Frame restored by the model with gradient branch is clearer with detailed structures.

Effects of the gradient branch: As shown in Table 2, the model with gradient branch achieves better results. This is because the gradient branch incorporates structure priors into restoration process as an implicitly guidance.

In order to further reveal the effectiveness of the gradient branch, we visualize the gradient maps in Fig. 9. The gradient map extracted from the blurry frame commonly have thick outlines, while the gradient map from the sharp counterpart have clear outlines and larger intensity. From the output gradient map in Fig. 9, we can see that the proposed gradient branch successfully recover gradient map similar to the sharp gradient map.

Moreover, the restoration results are shown in Fig. 10. The boundaries restored by the complete model are more sharper than those recovered by the model without gradient branch. The change of detailed textures reveals that the gradient branch can help preserve structure.

5 Conclusion

In this paper, we propose an effective model for video deblurring using deformable alignment. The model develops a novel dual pyramid alignment module, which constructs a feature pyramid to align frames using deformable convolution in a coarse-to-fine manner. Based on the feature pyramid, we further incorporate dilated spatial pyramid blocks to predict offsets within equivalent receptive fields for every feature pyramid layers, which guarantee temporal compensatory information can be sampled in a large region from the neighboring frames. To restore sharp frame, we introduce a gradient guided fusion module providing implicit structure guidance to alleviate geometric distortion. The proposed model has shown its effectiveness and outperforms previous state-of-the-art methods on several benchmarks.

Bibliography

[1] L. Bar, B. Berkels, M. Rumpf, and G. Sapiro. A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 3

[2] G. Bertasius, L. Torresani, and J. Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018. 3

[3] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy. Understanding deformable alignment in video super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 973–981, 2021. 2

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3

[6] S. Cho, J. Wang, and S. Lee. Video deblurring for hand-held cameras using patch-based synthesis. *ACM Transactions on Graphics (TOG)*, 31(4):1–9, 2012. 11

[7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2, 3

[8] T.-W. Hui, X. Tang, and C. C. Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018. 4

[9] T. Hyun Kim and K. Mu Lee. Generalized video deblurring for dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5426–5434, 2015. 1

[10] T. Hyun Kim, K. Mu Lee, B. Scholkopf, and M. Hirsch. Online video deblurring via dynamic temporal blending network. In *Proceedings of the IEEE international conference on computer vision*, pages 4038–4047, 2017. 3

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 4

[12] D. Li, C. Xu, K. Zhang, X. Yu, Y. Zhong, W. Ren, H. Suominen, and H. Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7721–7731, 2021. 1, 2, 3, 8, 9, 10, 11

[13] J. Lin, Y. Huang, and L. Wang. Fdan: Flow-guided deformable alignment network for video super-resolution. *arXiv preprint arXiv:2105.05640*, 2021. 2, 3

[14] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3225–3234, 2019. 3

[15] S. Liu, D. Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 385–400, 2018. 7

[16] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7769–7778, 2020. 7, 8

[17] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 9, 10

[18] J. Pan, H. Bai, and J. Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3043–3051, 2020. 1, 3, 8, 9, 10, 11, 12

[19] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 9

[20] W. Ren, J. Pan, X. Cao, and M.-H. Yang. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1077–1085, 2017. 3

[21] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017. 1, 3, 9, 11, 12

[22] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 4

[23] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 3

[24] Y. Tian, Y. Zhang, Y. Fu, and C. Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3360–3369, 2020. 2, 3

[25] X. Wang, K. Yu, C. Dong, and C. C. Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 2

[26] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3, 4, 6, 9, 10, 11, 12

[27] P. Wieschollek, M. Hirsch, B. Scholkopf, and H. Lensch. Learning blind motion deblurring. In *Proceedings of the IEEE international conference on computer vision*, pages 231–240, 2017. 3

[28] J. Wulff and M. J. Black. Modeling blurred video with layers. In *European Conference on Computer Vision*, pages 236–252. Springer, 2014. 3

[29] X. Xiang, H. Wei, and J. Pan. Deep video deblurring using sharpness features from exemplars. *IEEE Transactions on Image Processing*, 29:8976–8987, 2020. 1

[30] H. Yue, C. Cao, L. Liao, R. Chu, and J. Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2301–2310, 2020. 3, 4, 6

[31] Z. Zhong, Y. Gao, Y. Zheng, and B. Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 3

[32] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE international conference on computer vision*, pages 2482–2491, 2019. 9, 10