MULTIPLE VIDEO FRAME INTERPOLATION VIA ENHANCED DEFORMABLE SEPARABLE CONVOLUTION

Xianhang Cheng and Zhenzhong Chen∗
School of Remote Sensing and Information Engineering, Wuhan University

ABSTRACT
Generating non-existing frames from a consecutive video sequence has been an interesting and challenging problem in the video processing field. Recent kernel-based interpolation methods predict pixels with a single convolution process that involves source frames with spatially adaptive local kernels. However, when scene motion is larger than the pre-defined kernel size, these methods are prone to yield less plausible results and they cannot directly generate a frame at an arbitrary temporal position because the learned kernels are tied to the mid-point in time between the input frames. In this paper, we try to solve these problems and propose a novel approach that we refer to as enhanced deformable separable convolution (EDSC) to estimate not only adaptive kernels, but also offsets, masks and biases to make the network obtain information from non-local neighborhood. During the learning process, different intermediate time steps can be involved as a control variable by means of the coord-conv trick, allowing the estimated components to vary with different input temporal information. This makes our method capable to produce multiple in-between frames. Furthermore, we investigate the relationships between our method and other typical kernel- and flow-based methods. Experimental results show that our method performs favorably against the state-of-the-art methods across a broad range of datasets. Code will be publicly available on URL: https://github.com/Xianhang/EDSC-pytorch.

1 INTRODUCTION
Video frame interpolation aims to synthesize middle non-existent frames between the original input video frames, which is a long-studied problem in computer vision. The technology is beneficial to various applications in the field of video processing, ranging from frame rate up-conversion [1, 2], frame recovery and intra prediction in video coding [3, 4], slow motion generation [5, 6, 7] to novel view synthesis [8].

Early proposed methods exploit the motion from time-varying images with 2D flow fields, in which pixel movements are represented by coordinate shifts [9, 10]. Based on the estimated optical flow, frame interpolation algorithms typically warp and blend original frames to produce interpolation results [11, 12]. As the optical flow from the existent frames to target frame can be approximately estimated from the bi-directional flows, intermediate frames with multiple time steps can be generated. However, directly synthesizing the intermediate frames guided by optical flow may produce visual artifacts. In some challenging conditions such as occlusion, large motion, illumination or nonlinear structural changes, the optical flow accuracy decreases, resulting in distortion or artifacts. Recent deep learning approaches towards optical flow estimation have found remarkable success [13, 14, 15, 16, 17, 18, 19]. While the progress has been made to some extent, these methods aim at flow estimation rather than frame interpolation, producing less convincing results [20, 21, 6].

Some recent deep learning methods adopt advanced flow estimation model or its variations as sub-networks to directly synthesizing the interpolation frames in an end-to-end manner [20, 21, 6, 7, 24, 25, 26], where the intermediate frames act as supervision signals for training. Typically, occlusion masks or visibility maps are learned to smoothly transition across images as the synthesis happens in both the “forward” and “backward” direction, simultaneously. However, these approaches heavily depend on the quality of optical flow.

Another major trend which circumvents the flow estimation process in this research is to leverage adaptive convolution for interpolation [27, 28]. For each output pixel, a pair of 2D kernels or four 1D kernels (two for horizontal and the other two for vertical direction) are learned with a neural network. Notably, to handle large motion, large kernel size is required for these kernel-based interpolation methods. Though these methods are able to generate reasonable results, there are some drawbacks: 1) These methods can be problematic since the pre-defined kernel size is certain, which impedes the interpolation results when scene motion is larger than kernel size. 2) It is expensive to consider thousands of pixels to synthesize only one output pixel. 3) These methods cannot produce a frame at an arbitrary time because the kernel parameters are tied to the time step of the intermediate frame.

In this paper, we address these drawbacks mentioned above by presenting a more powerful and effective approach coined Enhanced Deformable Separable Convolution (EDSC). We argue that the limitation of the previous kernel-based interpolation methods [27, 28] is because they process the pixels only in the local neighborhood, which takes no effect on pixels outside the regular grid. Drawing inspiration from the success of deformable convolution networks [29, 30], we propose to learn adaptive kernels, offsets, masks and biases for interpolation, allowing us to use far fewer but more effective pixels to deal with large motion. We further propose to involve different intermediate time steps to make our network capable to estimate a frame at any time instant between two frames. Moreover, we show in detail that conventional flow-based interpolation methods can be regarded as specific instances of our method. Our

This work was supported in part by grants from the National Natural Science Foundation of China under Grant 61771348 and the Fundamental Research Funds for the Central Universities. (Corresponding author: Zhenzhong Chen, E-mail: zzchen@ieee.org)
experiments show that the proposed method achieves the best performance of any existing kernel-based methods and performs favorably against representative state-of-the-art interpolation methods without relying on any other pre-trained components.

Hence, our contributions are:

1. A novel kernel-based method is proposed, which learns not only spatially-adaptive separable convolution kernels, but also deformable offsets, masks, and biases to obtain information in a non-local neighborhood. This model is able to handle different degrees of motion, which is not constrained by the pre-defined kernel size.

2. In our network, different estimators are designed, in which temporal information can be involved as a control variable by means of an extension of coord-conv trick. Such a design enables our network to directly produce a frame at an arbitrary time, without using a recursive manner.

3. From the perspective of convolution, both some flow-based and kernel-based methods are theoretically demonstrated as special cases of our proposed EDSC.

Based on the above contributions, our model performs favorably against the state-of-the-art methods, even though any extra, complex and pre-calculated information (like context, depth, flow and edge information) or post-processing process are not involved in our network.

Please note that, this paper is the extension of our earlier publication [31] in the 34th AAAI Conference on Artificial Intelligence. The changes and improvements are summarized here. First, in the encoder-decoder architecture, heterogeneous convolution (HetConv) [32] is utilized to reduce computation and parameters of the model. In contrast to DSepConv [31], we save about 79.6% FLOPs in computation and 59.6% parameters with no loss in accuracy. Second, an additional bias estimator is introduced to learn residual values to account for pixel synthesis that cannot be well performed by the adaptive convolution. Third, based on the observation that convolutions with extra coordinate channels are particularly beneficial to spatially-conditioned generation tasks [33], we propose to input temporal index as a new control variable. This coord-conv trick enables our model to output different kernels, offsets and masks at different time steps. Alternatively, more comprehensive analysis and evaluations are provided in this paper.

2 RELATED WORK

Various methods that synthesize intermediate video frames have been introduced. In this section, we discuss and provide an overview of recent interpolation methods in the following parts.

2.1 Single Frame Interpolation

Most recently existing interpolation methods are designed specifically for single frame interpolation, which mainly consider the midpoint (in time) between two reference frames. Typically, substantial effort is made to first estimate bi-directional optical flow or its variations and then to synthesize the in-between frame guided by motion. Considering the input frames are not equally informative due to occlusion, mask maps are often estimated together with optical flow for adaptively blending the warped frames. Specifically, Liu et al. [22] proposed a fully-convolutional network DVF to predict 3D flow across space and time. The in-between frame was then generated by trilinear sampling. Liu et al. [23] further improved the performance of DVF by leveraging edge information [34] and a novel cycle consistency loss. Jiang et al. [5] proposed SuperSloMo, which utilized two U-Net architectures to compute bi-directional optical flows and soft visibility maps, respectively. Furthermore, based on SuperSloMo [5], Reda et al. [35] proposed unsupervised techniques to synthesize intermediate frames using cycle consistency. Yuan et al. [36] proposed a model which warped not only input frames, but also their corresponding features extract from ResNet [37].

In order to get more accurate optical flow, some methods utilized off-the-shelf flow estimation architectures with pre-trained parameters as sub-modules in their networks. For instance, Xue et al. [24] proposed ToFlow which utilized SpyNet [15] to estimate optical flow. Niklaus et al. [20, 26] utilized PWC-Net [17] and a modified GridNet [38] to warp and generate interpolated frames. Xu et al. [39] utilized PWC-Net [17] to compute optical flows from four input frames. Haris et al. [40] adopted flow images computed by [41] and refined them for both video frame interpolation and super resolution. Cheng et al. [21] proposed a position feature transform layer, transforming optical flow calculated from PWC-Net [17] into scaling factors to adjust frame interpolation process.

Some methods borrow operations from other image or video processing tasks (e.g., video super resolution) and generate intermediate frames without a component of optical flow computation. For instance, Choi et al. proposed CAIN [42], which employed PixelShuffle [43] and operation with channel attention mechanism [44]. Shen et al. [45] proposed a blurry video frame interpolation (BIN) method for jointly frame interpolation and deblurring. Xiang et al. [46] proposed a one-stage space-time video super-resolution for jointly frame interpolation and super-resolution. Choi et al. [47] proposed to improve the performance of an interpolation algorithm by incorporating meta-learning.

There are some other studies that regard flow estimation as an intermediate step, which can be circumvented with a single convolution process. As a prior of kernel based interpolation methods, AdaConv [27] was proposed to estimate a pair of spatially-adaptive convolution kernels for each output pixel with a neural network. To reduce large memory demand, Niklaus et al. [28] proposed SepConv that separated each 2D convolution kernel into two 1D kernels. Choi et al. [4] further improved the structure of SepConv [28] that both uni-directional and bi-directional prediction were available in video coding. Peleg et al. [48] modified SepConv [28] into a multi-scale architecture and formulated interpolated motion estimation as classification by calculating the center-of-mass of the convolution kernels. Concurrently to our work, Lee et al. [49] proposed a new warping module AdaCoF with a similar motivation to ours. They further introduce a dual-frame adversarial loss to improve their performance. Moreover, Bao et al. [6, 7] combined the advantages of flow based and kernel based methods, proposed an adaptive warping layer that warps images or features based on the given optical flow and learned local convolution kernels.
2.2 Multiple Frame Interpolation

A straightforward way to generate multiple intermediate frames is to recursively apply a single frame video interpolation method. However, this manner is not flexible enough and error would accumulate during the recursive process. Some flow based interpolation methods [20, 7, 35, 26, 25] are also well-suited for multi-frame interpolation while the other are not. The difference among these methods is whether the occlusion reasoning is tied to an arbitrary time step [5, 35] and whether motion compensation is performed before synthesizing the output frame [20, 7, 26, 25, 39].

Several methods utilize phase information to learn the motion relationship for multiple video frame interpolation. Meyer et al. [50] proposed the phase-based method which utilized phase information across the levels of a multi-scale pyramid. Furthermore, combined with CNNs, PhaseNet [51] was proposed with a better performance. Another related problem is video frame inpainting, which focus on the intersection of general video inpainting, frame interpolation and video prediction. Szeto et al. [52] devised a method bi-TAI that was composed of a bidirectional video prediction module and a temporally-aware frame interpolation module, achieving impressive inpainting results.

In relation to kernel based interpolation methods [27, 28, 4, 31, 49], to our best knowledge, none of these methods can directly generate frames at an arbitrary temporal position. In light of this limitation, we suggest to solve the under-explored problem.

3 Proposed Method

In this section, we introduce our proposed algorithm for video frame interpolation, including the details of our network architecture and our training details. The notations are provided in Table 1 for clarity.

3.1 Problem Statement

To explore the relationships among kernel and flow based methods, we introduce our frame interpolation algorithm for single or multiple time steps individually.

3.1.1 Single Frame Interpolation

Assume that there are two temporally neighboring frames \( I_1 \) and \( I_2 \), our purpose is to interpolate frame \( \hat{I} \) that is located in the midpoint of the them. For each pixel \( \hat{I}(x, y) \) to be synthesized, the widely used kernel based interpolation model \([4, 28, 27]\) learns a pair of convolution kernels and uses them to convolve the local patches \( P_1(x, y) \) and \( P_2(x, y) \). This process can be formulated as

\[
\hat{I}(x, y) = \mathbf{K}_1(x, y) * P_1(x, y) + \mathbf{K}_2(x, y) * P_2(x, y),
\]

where * means convolution operation and \( \mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n \times n} \) represent \( n \times n \) 2D convolution kernels. Figure 1(a) illustrates this kind of method. For standard local convolution, \( n \) has to be big enough to capture large motion. For instance, in AdaConv [27], the kernel size \( n \) equals to 41. However, estimating such an amazing number of kernels (41 \( \times \) 41) simultaneously entails heavy computational load. In [28], each 2D kernel is approximated with two 1D kernels \((\mathbf{k}_{1,i}, \mathbf{k}_{1,h})\) or \((\mathbf{k}_{2,i}, \mathbf{k}_{2,h})\) with formulation:

\[
\mathbf{K}_1(x, y) = \mathbf{k}_{1,v}(x, y) \cdot \mathbf{k}^T_{1,h}(y, x), \\
\mathbf{K}_2(x, y) = \mathbf{k}_{2,v}(x, y) \cdot \mathbf{k}^T_{2,h}(y, x),
\]

which helps to reduce the memory consumption from \( O(n^2) \) to \( O(2n) \). Nonetheless, despite thousands of pixels have been considered, these methods are limited to motions up to \( n \) pixels between two input frames.

To solve this problem, we propose to use much smaller convolution kernel size and learn additional offsets and masks, allowing us to focus on fewer but more relevant pixels rather than all the pixels in a large neighborhood. Towards this end, the patches that filled with local pixels should be resampled by those pixels which mostly contribute to the final value.

Let \( p_{j,i} \) denote the pre-specified offset for the \( j \)-th (\( j \in [1, n^2] \)) location in a specific patch and \( i \) represents either of the two input frames. With learned offset \( \Delta p_{j,i} \) and modulation scalar \( \Delta m_{j,i} \), the pixels in a resampled patch \( P \) can be expressed as

\[
\begin{align*}
P'_1(x, y; p_{1,j}) &= P_1(x, y; p_{1,j} + \Delta p_{1,j}) \cdot \Delta m_{1,j}, \\
P'_2(x, y; p_{2,j}) &= P_2(x, y; p_{2,j} + \Delta p_{2,j}) \cdot \Delta m_{2,j}.
\end{align*}
\]
As the learned offsets are typically fractional, pixels located at non-integral coordinates are bilinearly sampled. Moreover, 1D separable kernels are used to calculate 2D convolution kernels in Eq. (2) and we further introduce to learn pixel-wise residual values \( \Delta b(x, y) \) in case that the convolution kernels are less accurate. Therefore, our final interpolation process is expressed as

\[
\hat{I}(x, y) = k_1 \cdot B_1(x, y) P_1'(x, y) + k_2 \cdot B_2(x, y) P_2'(x, y) + \Delta b(x, y) + K_2(x, y) P_2(x, y) + K_2(x, y) P_2(x, y) + \Delta b(x, y).
\] (4)

### 3.1.2 Relationships with Kernel and Flow Based Methods

In our method, both previous kernel-based methods [28, 4] and conventional flow-based methods can be seen as specific instances of our approach. In Eqs. (3) and (4), it is easy to make out that when \( \Delta p = 0, \Delta m = 1, \Delta b = 0 \), the interpolation process is the same as those proposed in [4, 28].

As for flow-based method, typically, the warping function can be formulated as

\[
\hat{I}(x, y) = k_1 \cdot I_1(x + u_1, y + v_1) + k_2 \cdot I_2(x + u_2, y + v_2),
\] (5)

where \( (u_1, v_1) \) and \( (u_2, v_2) \) denote the bidirectional optical flow values and \( k_1, k_2 \) represent occlusion masks.

In relation to the bilinear interpolation process in those flow based methods, as shown in Figure 1(c), we redefine the warping operation in Eq.(5) as a \( 2 \times 2 \) pixel-wise convolution process with the formulation:

\[
\hat{I}(x, y) = k_1 \cdot B_1(x, y) P_1'(x, y) + k_2 \cdot B_2(x, y) P_2'(x, y),
\] (6)

where \( B \) denote fixed bilinear interpolation coefficients and \( P_i'(x, y) \) is calculated by:

Table 2: A list of conditions in which our method can be equivalent or similar to the other kinds of algorithms.

| Type | Condition | Relationship |
|------|-----------|--------------|
| Kernel based [4, 28] | \( n = 51, \Delta p = 0, \Delta m = 1, \Delta b = 0 \) | Equivalence |
| Flow based [24, 22] | \( n = 1, \Delta m = 1, \Delta b = 0 \) | Equivalence |
| Our previous work [31] | \( \Delta b = 0 \) | Equivalence |
| Adaptive warping [6, 7] | \( n = 4, \Delta m = B \) | Resemblance |
variable in pixel synthesis. Thus, the formulation with respect to arbitrary time frame interpolation is:

\[
P_i^\prime(x, y) = P_i(x, y; p_{ij} + \Delta p_{ij}), i \in [1, 2], j \in [1, 4],
\]

where \([\cdot]\) represents floor operation.

In Eqs. (3) and (4), if we set \(n = 1, \Delta m = 1\) and \(\Delta b = 0\), our interpolation process is the same as the one in Eq. (5), indicating that the flow based method can be a specific case of our method.

We further show the adaptive warping method proposed in [6, 7] in Figure 1(d). When \(\Delta m\) in Eq. (3) equals to the bilinear interpolation coefficients \(B\), our method bears some resemblance to the operation of adaptive warping. The only difference is that the locations of pixels used to resample the convolutional patches \(P\) can be dispersed, while those used in \(P'\) are not.

In Table 2, we summarize the main relationships and the conditions between our method and previous kernel based [4, 28], flow based [24, 22], adaptive warping based [6, 7] algorithms as well as our prior work [31].

### 3.1.3 Multiple Frame Interpolation

So far, none of the kernel-based interpolation methods can directly generate in-between frames at an arbitrary temporal position. This is because the pixels chosen for the final adaptive convolution are tied to a specific time step \(t = 0.5\). A possible solution is to resample the pixels based on \(t\), which is easy for methods with learned offsets. For instance, given the model trained for \(t = 0.5\), we can respectively multiply the learned offsets \(\Delta p_{1, j}\) and \(\Delta p_{2, j}\) by \(t/0.5\) and \((1 - t)/(1 - 0.5)\) in Eq. (3) to shift the locations of the reference pixels, producing an intermediate frame at arbitrary time \(t\). However, as shown in Figure 2, this solution is problematic since the occlusion is still handled for \(t = 0.5\), indicating that the learned masks and kernels should also be controlled by \(t\) for multiple video frame interpolation.

Followed a similar route in Eq. (4), kernels, masks, offsets and biases are needed for multiple frame interpolation. The only difference is that intermediate time step \(t\) is a crucial control variable in pixel synthesis. Thus, the formulation with respect to arbitrary time frame interpolation is:

\[
\hat{I}(x, y, t) = k_{1, f}(x, y, t) \cdot k_{1, b}(x, y, t) \ast P_1(x, y, t) + k_{2, f}(x, y, 1 - t) \cdot k_{2, b}(x, y, 1 - t) \ast P_2'(x, y, 1 - t) + \Delta b(x, y).
\]
Figure 3: Illustration of the architecture of our proposed EDSC network, which includes five major components: an encoder-decoder architecture and a set of kernel, mask, offset and bias estimators.

3.2.3 Deformable Convolution

The deformable convolution utilizes the estimated kernels, offsets and masks to adaptively convolve input frames, yielding an intermediate interpolation result. This operation is adaptive and does not totally resemble the process which shares the same kernel weights described in [30]. In the right part of Figure 3, the frames generated from deformable convolution look dimmer than the final interpolation result in brightness except area with occlusion (e.g., area around the red ball), suggesting the effectiveness of our method to handle motion and occlusion.

3.3 Training

3.3.1 Loss functions

We consider two kinds of loss functions to penalize the interpolated frame \( \hat{I} \) that is not similar to the ground-truth \( I_{GT} \).

The first loss measures the difference between the interpolated pixel color and the ground-truth color with the function:

\[
L_C = \rho(I - I_{GT}),
\]

where \( \rho(x) = \sqrt{x^2 + \epsilon^2} \).

The second type of loss functions aims to penalize results that are not perceptually similar to ground truth by additionally defining a distance measure between high-level features extracted from a pre-trained network. The definition is as follows.

\[
L_F = ||\phi(\hat{I}) - \phi(I_{GT})||_2^2,
\]

where \( \phi \) represents the feature extractor based on the relu4_4 layer of the pre-trained VGG-19 network [55]. When training with this loss function, we use the model pre-trained on \( L_C \) loss function and then fine tune it with both \( L_C \) loss and \( L_F \) loss for 2 more epochs.

3.3.2 Training Details.

We trained two versions of our model: one produces only single midpoint in time of the frames (EDSC_s) and another generates
A wide variety of datasets are involved to evaluate our method.

**4.1 Experimental Setup**

**4.1.1 Datasets**

A wide variety of datasets are involved to evaluate our method. We use Vimeo90K-Septuplet dataset, which consists of 91,701 sequences with 51,312 triplets. The Vimeo90K dataset consists of 51,312 triplets with a resolution of 256 × 448 pixels. The Vimeo90K-Interp dataset is used instead because it contains 3,782 triplets with a resolution of 256 × 448 pixels for video frame interpolation. The Middlebury dataset consists of 1,478 pairs of images with high frame rate videos including videos from GOPRO test set and YouTube. The evaluation set contains four subsets: Easy, Medium, Hard and Extreme with different degrees of motion, each of which consists 310 triplets. The maximum resolution of this dataset is 1280 × 720 pixels.

**4.1.2 Metrics**

For quantitative evaluation, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [59] and Learned Perceptual Image Patch Similarity (LPIPS) [60] metrics. In addition, we report the average Interpolation Error (IE) on the Middlebury dataset. Bigger PSNR and SSIM indicates better performance, while for LPIPS and IE, the smaller, the better.

**4.1.3 Baselines**

We compare and analyze our method with most of the recent state-of-the-art interpolation methods since 2017. We divide these methods into four categories: 1) those with a component of learned convolutional kernel estimation (kernel based); 2) those using the entire frames with learning rates of 1.25; 3) those with multiple intermediate frames (EDSC_m) at arbitrary in-between time. The only difference between them is whether time information is involved in the estimators as detailed in section 3.2.2. In addition, two kinds of loss functions were used for both models.

For EDSC_s, we use Vimeo90K-Interp dataset, which contains 51,312 triplets with a resolution of 256 × 448 pixels. The triplets were randomly flipped horizontally or vertically for data augmentation. In the context of EDSC_m, Vimeo90K-Septuplet dataset is used instead because more consecutive frames are desired for multiple time step frame generation. The Vimeo90K-Septuplet dataset consists of 91,701 sequences with a resolution of 256 × 448 pixels, each of which contains 7 consecutive frames. When training EDSC_m, five target frames \( \hat{I}_t, t \in \{1, 0.667, 0.333, 0.5, 0.167, 0.833\} \) are randomly generated.

The models were trained using Adam optimizer [56]. We first trained our network for 120 epochs using a learning rate schedule of 1e-4, dropping by half every 40 epochs. The training patch size was randomly cropped into 256 × 256 pixels and the batch size was 4. Notice that some previous works trained their networks with large patches [20, 6, 7], we fine-tuned our network using the entire frames with learning rates of 1.25e-5 for another 10 epochs. This makes us use smaller batch size (which equals to 2) to deal with the increasing memory footprint.

**4 Experiments**

In this section, we first introduce the evaluation datasets and metrics. We then compare the proposed method with state-of-the-art algorithms. Finally, we perform comprehensive ablation studies to analyze the contribution of some important components.
with components of both optical flow and learned convolutional kernel estimation (adaptive warping based); 3) those with a component of optical flow or its variations' estimation (flow based); 4) those without any components of optical flow or adaptive convolutional kernel estimation.

For the first category, we typically choose kernel-based interpolation methods, including AdaConv [27], SepConv [28], IM-Net [48], DSepConv [31] and AdaCoF [49]. Our method belongs to this category as well. The second category contains MEMC-Net* [6] and DAIN [7]. The third category includes DVF [22], SuperSlomo [5], CtxSyn [20], ToFlow [24], CyclicGen [23], MS-PFT [21], STAR-TIR [40] as well as SoftSplat [26]. Additionally, we include CAIN [42] which makes use of PixelShuffle and attention mechanism in the fourth category.

Notably, considering that some methods provide more than one version of the same model, we report all their performances and treat them differently (e.g., CtxSyn, SepConv and SoftSplat are trained with two kinds of loss functions. CyclicGen and AdaCoF provides two models).

### 4.2 Comparisons with state-of-the-arts

Since most of the baselines focus on single-frame interpolation, we here first discuss our EDSC\_s model in section 4.2.2 and the EDSC\_m model will be discussed in section 4.2.3.

#### 4.2.1 Network setting comparisons

We analyse and report different network settings contributed to interpolation algorithms with following components: training dataset, sub-networks and model parameters shown in Table 3. The Middlebury dataset [12] is abbreviated by M.B. for the sake of simplicity. We further divide the sub-networks into different parts: flow, kernel, mask, context estimation networks as well as post-processing networks (abbreviated by post-proc.). Specially, networks for learning information that falls outside the mentioned five modules will be categorized as “Other” class.

In Table 3, the column “Flow” specifies which methods are based on a pre-inferred displacement fields such as SpyNet [15], PWC-Net [17], FlowNetS [13] and Liu’s method [41]. A self-defined encoder-decoder structure is abbreviated by Enc-Dec.

### Figure 5: Interpolation error comparisons among kernel based methods on the Middlebury Evaluation set [12]. Lower bars represent better performances.

**Table 4: Quantitative comparisons against methods using adaptive convolutional kernels.** The numbers in **bold** and with an underline represent the first and the second best performance, respectively.

| Methods          | Training dataset | UCF101 [57, 22] | Vimeo90K [24] | M.B.-Other [12] | Parameters (million) |
|------------------|------------------|----------------|--------------|----------------|---------------------|
|                  |                  | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | IE ↓ | LPIPS ↓ |
| AdaConv [27]     | proprietary      | —     | —     | —     | 33.80 | 0.970 | 0.027 | 2.27 | 0.017 | 21.6 |
| SepConv-\_L\_1 [28] | proprietary     | 34.79 | 0.967 | 0.029 | 33.45 | 0.967 | 0.019 | 2.44 | 0.029 | 21.8 |
| SepConv-\_L\_F [28] | proprietary     | 34.69 | 0.966 | 0.024 | 34.73 | 0.974 | 0.028 | 2.06 | 0.022 | 21.8 |
| IM-Net [48]      | proprietary      | —     | —     | —     | 33.50 | —     | —     | —    | —     | —    |
| DSepConv [31]    | Vimeo90K         | 35.08 | **0.969** | 0.030 | 34.73 | 0.974 | 0.028 | 2.06 | 0.022 | 21.8 |
| AdaCoF [49]      | Vimeo90K         | 34.91 | 0.968 | 0.029 | 34.27 | 0.971 | 0.031 | 2.31 | 0.029 | 21.8 |
| AdaCoF+ [49]     | Vimeo90K         | 34.90 | 0.968 | 0.030 | 34.47 | 0.973 | 0.029 | 2.23 | 0.026 | 22.9 |
| MEMC-Net\* [6]   | Vimeo90K         | 35.01 | 0.968 | 0.030 | 34.40 | 0.974 | 0.027 | 2.10 | 0.020 | 70.3 |
| DAIN [7]         | Vimeo90K         | 35.00 | 0.968 | 0.028 | 34.72 | **0.976** | 0.022 | 2.04 | 0.017 | 24.0 |
| EDSC\_s-\_L\_C  | Vimeo90K         | **35.13** | 0.968 | 0.029 | **34.84** | 0.975 | 0.026 | **2.02** | 0.020 | 8.9  |
| EDSC\_s-\_L\_F  | Vimeo90K         | 34.78 | **0.967** | **0.023** | 34.49 | 0.972 | **0.016** | 2.15 | **0.010** | 8.9  |

†: Results copied from [24]. ‡: Results copied from [48].
We also submit the interpolation results of our approaches which incorporate adaptive convolutional kernels. As shown in Table 5, we can learn the attention equipped method CAIN cannot blend the content of the attention map with the context, flow or edge information. In addition to STAR-T\textsubscript{HR} [40] and SoftSplat [26], our $L_C$-trained model convincingly outperforms the other methods in terms of most of the PSNR, SSIM and LPIPS, whilst our $L_F$-trained model performs the best in terms of most of the NIE among all published algorithms at the time of submission. We specifically compare approaches which incorporate adaptive kernel estimations and the results are shown in Table 4. Among all the methods, our $L_C$-trained model achieves the best performance in terms of PSNR and IE and our $L_F$-trained model performs the best in terms of LPIPS. In particular, we achieve 0.12 dB and 0.13 dB gain in terms of PSNR on the UCF101 and Vimeo90K datasets compared to DAIN [7], without relying on pre-trained sub-models like PWC-Net [17] and MegaDepth [63]. Additionally, we can see that our $L_C$-trained model outperforms AdaCoF$^*$ [49] by 0.23 dB on UCF101 and 0.37 dB on Vimeo90K in terms of PSNR, whilst requiring 61% fewer parameters.

We also submit the interpolation results of our $L_C$-trained model on Evaluation set to the Middlebury benchmark\footnote{http://vision.middlebury.edu/flow/eval/results/results-i1.php}. According to the feedback from the benchmark organizer, our approach ranks 3\textsuperscript{rd} in terms of IE and 2\textsuperscript{nd} in terms of NIE among all published algorithms at the time of submission. We specifically show the comparisons among kernel based methods which do not rely on any other information in Figure 5. Among these methods, our model performs the best on 5 out of 8 sequences and achieves the best performance on average, which demonstrates the good generalization ability of our method.

In what follows we compare methods that make no use of adaptive convolutional kernels. As shown in Table 5, we can learn the fact that the usage of off-the-shelf and pre-trained model provides good performance. It is true that our method is inferior to STAR-T\textsubscript{HR} [40] and SoftSplat [26]. Combined the summary reported in Table 3, STAR-T\textsubscript{HR} [40] additionally utilizes pre-inferred displacement fields [41] and pre-trained RBPN [62], leading to a reasonable performance with enormous parameters (which is more than 11X bigger than our method). For SoftSplat [26], they have reached the best performance so far due to their effectively handling cases where multiple source pixels map to the same target location, conditioned on pre-calculated optical flow [17]. Nonetheless, we can see that a good kernel learner competitive without relying on any other extra information (like context, flow or edge information). In addition to STAR-T\textsubscript{HR} [40] and SoftSplat [26], our $L_C$-trained model convincingly outperforms the other methods in terms of most of the PSNR, SSIM and IE whereas our $L_F$-trained model performs the best in terms of LPIPS. For qualitative comparisons, we compare our method against interpolation methods published since the year of 2019, including MEMC-Net$^*$ [6], CyclicGen [23], ToFlow [24], DAIN [7], STAR-T\textsubscript{HR} [40], CAIN [42], DSepConv [31] as well as AdaCoF [49].

In Figure 6, we show an example of a skateboarder playing in front of a building. From the overlaid frame in Figure 6(a), we can see that only one leg is shown in one frame while both legs can be seen in the other, making it difficult to estimate optical flow accurately. Therefore, methods that make use of optical flow (like MEMC-Net$^*$, ToFlow, DAIN, STAR-T\textsubscript{HR}) generate visible blur or artifacts. Since the scene is also complex, the attention equipped model CAIN cannot blend the content of source images well and loses some information around the left shin. The results from DSepConv and AdaCoF contain exhibit bluriness as a result of inaccurate kernel learning. Our $L_C$-trained model suffers from some information lost, whereas our result from $L_F$-trained model appears clear with fewer visual distortions.

Figure 7 shows an example of rotation motion around the knee joint (the blue rectangle) and occlusion (the yellow rectangle).

Table 5: Quantitative comparisons against methods without using adaptive convolutional kernels. The numbers in \textbf{bold} and with an underline represent the first and the second best performance, respectively.

| Methods          | Training dataset | UCF101 [57, 22] | Vimeo90K [24] | M.B.-Other [12] | Parameters (million) |
|------------------|------------------|----------------|---------------|-----------------|----------------------|
|                  |                  | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | IE ↓ | LPIPS ↓ |
| DVF [22]         | UCF101           | 34.12  | 0.963 | —      | 31.54  | 0.946 | —      | 1.04  | —      |
|                  | Adobes240        | 34.75  | 0.968 | —      | 33.15  | 0.966 | —      | 1.28  | —      |
| CtxSyx-L\textsubscript{C} [20] | proprietary      | 34.62  | 0.931 | —      | 34.39  | 0.942 | 0.024  | —      | 0.016  |
| CtxSyx-L\textsubscript{F} [20] | proprietary      | 34.01  | 0.924 | —      | 33.76  | 0.960 | 0.017  | —      | 0.013  |
| ToFlow [24]      | Vimeo90K         | 34.58  | 0.967 | 0.027  | 33.73  | 0.968 | 0.027  | 2.51  | 0.024  |
| CyclicGen [23]   | UCF101           | 35.11  | 0.968 | 0.030  | 32.10  | 0.949 | 0.058  | 2.86  | 0.046  |
| CyclicGen+ [23]  | UCF101, M.B.     | 34.69  | 0.968 | 0.034  | 31.46  | 0.940 | 0.060  | 3.04  | 0.053  |
| MS-PFT [21]      | Vimeo90K         | 34.70  | 0.967 | 0.023  | 34.26  | 0.971 | 0.020  | 2.28  | 0.014  |
| STAR-T\textsubscript{HR} [40] | Vimeo90K       | 35.17  | \textbf{0.969} | 0.030  | 35.14  | \textbf{0.976} | 0.026  | \textbf{1.95} | —      | 111.6  |
| SoftSplat-L\textsubscript{Lap} [26] | Vimeo90K       | \textbf{35.39} | —      | 0.033  | \textbf{36.10} | —      | 0.021  | —      | 0.016  |
| SoftSplat-L\textsubscript{F} [26] | Vimeo90K       | \textbf{35.10} | \textbf{0.022} | —      | \textbf{35.58} | —      | \textbf{0.013} | —      | \textbf{0.008} |
| CAIN [42]        | Vimeo90K         | 34.91  | \textbf{0.969} | 0.032  | 34.65  | 0.973 | 0.031  | 2.28  | 0.025  |
| EDSC\_s-L\textsubscript{C} | Vimeo90K       | 35.13  | 0.968 | 0.029  | 34.84  | 0.975 | 0.026  | 2.02  | 0.020  |
| EDSC\_s-L\textsubscript{F} | Vimeo90K       | 34.78  | 0.967 | 0.023  | 34.49  | 0.972 | 0.016  | 2.75  | 0.010  |

\(*$: Results copied from [42], \(\dagger\$: Results copied from [26].)
The CyclicGen, ToFlow and DAIN produce broken results on the man’s skin due to the usage of inaccurate optical flow and most of the methods lose information of the head area and appear blurry. In contrast, both our methods handle these situations better than the others.

The example in Figure 8 is subject to explicit camera motion.
We observe that the interpolation results from MEMC-Net∗, CyclicGen, ToFlow, DAIN and STAR-T_{HR} fail to reconstruct the bottle clearly because both the bottle areas from the two source frames are wrongly estimated as occlusion. On the contrary, our two results are sharp and free from blurriness, with the \( \mathcal{L}_F \)-trained model retaining more high-frequency details. Additionally, compared to the other kernel-based methods DSepConv and AdaCoF, the proposed method produces more complete result. We attribute this to the use of bias estimator, which learns residual information for better pixel reconstruction.

We further show an example where the motion is discontinuous in Figure 9. From the overlayed frame in 9(a) we can observe that the motion is continuous except the sign lighted with yellow rectangle. This discontinuity makes it hard to estimate optical flow accurately, causing ghosting artifacts for those methods strictly relying on optical flow (MEMC-Net∗, ToFlow, DAIN, STAR-T_{HR}). In this example, the other methods, including ours, perform well.

4.2.3 Arbitrary-position frame interpolation

We perform a quantitative evaluation on the Vimeo90K-Septuplet test set [24]. Specifically, we interpolate frame 2 through 6 from frame 1 and frame 7 on all its 7,824 sequences to generate \( \times 6 \) slow motion frames. We also compare our method to DAIN [7], which can interpolate arbitrary in-between frames. The PSNR scores at each frame index are shown in Figure 10, it can be clearly seen that our method outperforms DAIN for each individual in-between time step. In spite of the usage of adaptive convolutional kernels, the sampling locations for each synthesized pixel of DAIN heavily depend on optical flow, thus little inaccuracy may result in less plausible results. On the contrary, we learn which pixels to reference without a strict guidance (optical flow).

In Figure 11, we show a set of interpolation results at \( t = 0.1, 0.3, 0.5, 0.7 \) and 0.9. We also visualize the effective sampling locations of a pixel (indicated by the red +), which locates at the same position of the synthesized frames. First, despite some time steps (e.g., \( t = 0.1, 0.3, 0.7, 0.9 \)) are not involved during the training process, our method can generate plausible results. Second, our method is aware of the inequality of information between the two input frames when producing intermediate frames with different temporal positions. To be detailed, when \( t < 0.5 \), our model mainly takes information from the first frame, whereas for \( t > 0.5 \), the non-zero elements are mainly in the second frame. This is in line with the assumption that the former frame is more reliable in synthesis for earlier time steps and so is the latter for later time steps. Third, our method is aware of the motion between the two input frames. For instance, the non-zero elements are spatially farther away from the center in the first kernels when \( t \) getting bigger, while those move in opposite direction in the second kernels. This phenomenon shows that the learned offsets vary from different \( t \), indicating the effectiveness of the usage of analogous coord-conv trick to deal with temporal consistency.

4.3 Model Analysis

4.3.1 Effect of dealing with different motion degrees

We investigate the ability of different algorithms to handle different motion degrees. Typically, methods with a component of optical flow estimation can capture large motion as long as it is accurately computed. That is why it is so popular to make use of the off-the-shelf optical flow estimators and further perform fine-tuning. However, there is no optical flow utilized in kernel based methods. The capacity to deal with large motion hinges on kernel estimates. Therefore, for a fair comparison we evaluate the performance with respect to the amount of motion among kernel based methods on a more comprehensive dataset SNU-FILM [42]. As shown in Table 6, our \( \mathcal{L}_C \)-trained model achieves the best performance on the Easy, Medium and Hard sets in terms of PSNR and SSIM, while is marginally worse than AdaCoF+ which learns kernels with a larger size 11 × 11.
Table 6: Quantitative comparisons against kernel-based methods on SNU-FILM [42] dataset (abbreviated by S.F.). The numbers in bold represent the best performance.

| Methods      | Kernel (size) | S.F.-Easy | S.F.-Medium | S.F.-Hard | S.F.-Extreme |
|--------------|---------------|-----------|-------------|-----------|--------------|
|              |               | PSNR↑/SSIM↑/LPIPS↓ | PSNR↑/SSIM↑/LPIPS↓ | PSNR↑/SSIM↑/LPIPS↓ | PSNR↑/SSIM↑/LPIPS↓ |
| SepConv-\(L_{1}\) [28] | learned(51)   | 39.47 / 0.999 / 0.017 | 34.98 / 0.976 / 0.032 | 29.35 / 0.925 / 0.075 | 24.31 / 0.845 / 0.154 |
| SepConv-\(L_{F}\) [28] | learned(51)   | 39.33 / 0.989 / 0.012 | 34.79 / 0.979 / 0.024 | 29.10 / 0.921 / 0.057 | 24.10 / 0.837 / 0.124 |
| DSepConv [31] | learned(5)    | 39.94 / 0.990 / 0.019 | 35.30 / 0.977 / 0.035 | 29.56 / 0.925 / 0.074 | 24.34 / 0.840 / 0.149 |
| AdaCoF [49]   | learned(5)    | 39.43 / 0.990 / 0.020 | 34.90 / 0.975 / 0.037 | 29.41 / 0.924 / 0.076 | 24.29 / 0.844 / 0.149 |
| AdaCoF+ [49]  | learned(11)   | 39.53 / 0.990 / 0.020 | 34.99 / 0.976 / 0.036 | 29.50 / 0.925 / 0.074 | 24.45 / 0.845 / 0.146 |
| Ours-\(L_{C}\) | learned(5)    | 40.01 / 0.990 / 0.019 | 35.37 / 0.978 / 0.034 | 29.59 / 0.926 / 0.074 | 24.39 / 0.843 / 0.145 |
| Ours-\(L_{F}\) | learned(5)    | 39.50 / 0.990 / 0.013 | 35.02 / 0.976 / 0.024 | 29.33 / 0.921 / 0.055 | 24.12 / 0.834 / 0.121 |

Figure 12: By convolving the patches with corresponding kernels we can get the final synthesized frame. We show the effective sampling locations of occlusion area centered at yellow + in the synthesized frame. The second row provides the magnified views of non-zero kernel values, in which the black rectangle represents the local kernel and greener regions indicate higher absolute values. By learning offsets, our method can obtain information outside the regular local kernel.

Figure 13: Effective sampling locations of occlusion area centered at yellow + in the synthesized frame. The reference patches are padded to align the kernels. Caused by large motion, the occlusion is handled by taking pixels mainly from one of the reference patches.

Besides, our \(L_{F}\)-trained model performs favorably against the others on most of the subsets in terms of LPIPS. The main reason of these improvements is that our model better captures the content of source images by joint learning kernels, offsets, masks and biases. Notice that despite 51 \times 51 pixels are involved in SepConv for each pixel’s synthesis, the available information is constrained in a local neighborhood and thousands of unrelated pixels make it prone to inaccuracies.

4.3.2 Effect of occlusion handling

Here, we use our \(L_{F}\)-trained model to explain how our method handles occlusion and show two kinds of representative examples in Figures 12 and 13, respectively.

It is noteworthy that each pair of 1D kernels is convolved to produce its equivalent 2D kernel for a better understanding. We also multiply the mask values by the kernel weights to empha-
For simplicity. Bias values are omitted because they are not fit for pixel-level visualization.

In Figure 12, we show a pixel from the background which is occluded by the elbow moving right. Despite that this pixel can be only seen in Patch 2, our method produces kernels that choose pixels with similar appearance from both the patches.

Figure 13 shows a pixel that moves outside the second frame, which always locates at the boundary areas. In this case, the pixel is only visible in Patch 1, and the generated kernels choose to sample corresponding pixels mainly from one of the patches (Patch 1).

We further compare our approach with methods which utilize warping or adaptive warping operations [6, 24, 7] based on off-the-shelf optical flow estimators [15, 17] to see their abilities to handle occlusion. Since there is no labeled occlusion regions from the test datasets, we use brightness constancy as a measure of occlusion [12, 26]. Additionally, we report the average IE from boundary 10 pixels wide of the synthesized frames, a special region where obvious occlusion often occurs due to camera motion. As shown in Table 7, all the three methods perform worse than ours in terms of both occlusion IE and boundary IE. In particular, we achieve considerable gains in boundary IE. This is because the warped frames guided by optical flow are prone to be inaccurate especially in occluded regions, making it more difficult for later post-processing to improve the quality. In Figure 14, when significant occlusion occurs in the boundary, our approach is able to produce better results with less blur.

4.3.3 Effect of loss functions

We use two versions of loss functions to train our model by minimizing color and perceptual difference, respectively. Moreover, we can achieve a continuous transition between the effects of two loss functions by using Deep Network Interpolation (DNI) methodology [64]. To be more detailed, the model parameters of a new interpolated model can be derived by:

$$\theta_{interp} = (1 - \alpha)\theta_{L_C} + \alpha \theta_{L_F},$$  (12)

where $\theta$ represents network parameters and $\alpha \in [0, 1]$ denotes the interpolation coefficient. As shown in Figure 15, $L_F$-trained model recovers the details well whist model trained with $L_C$ does not. By adjusting $\alpha$, the imagery effects change smoothly.

We further perform quantitative comparisons with different $\alpha$ values on the three datasets shown in Table 8. Bigger $\alpha$ leads to better performance in terms of LPIPS whereas performs worse in terms of PSNR, which indicates that we can balance distortion and perceptual quality by simply changing $\alpha$ to meet different requirements of users.

4.3.4 Execution speed

Table 9 shows the runtime of each component of our method on a single NVIDIA Titan X GPU using sequences with different resolutions. Enc-Dec. is short for the encoder-decoder architecture and D.C. is short for the deformable convolution process which utilizes the learned components. We notice that the deformable convolution process occupies most of the runtime. This is because we need to first sample and expand the source frames into $n$ (kernel size) times their size and then perform convolution with a stride of $n$. Other
Table 11: Ablation experiments to quantitatively analyze the effect of different kernel size.

| Size | UCF101 PSNR↑/SSIM↑ | Vimeo90K PSNR↑/SSIM↑ | M.B. IE | FLOPs (G) | Runtime (s) |
|------|-------------------|----------------------|-------|----------|-------------|
| 1x1  | 34.83 / 0.967     | 33.47 / 0.965        | 2.65  | 11.4     | 0.068       |
| 3x3  | 34.99 / 0.968     | 34.59 / 0.973        | 2.16  | 11.8     | 0.182       |
| 5x5  | 35.13 / 0.968     | 34.84 / 0.975        | 2.02  | 13.8     | 0.557       |

Figure 16: Synthesizing frames with different kernel size n. We show the overlayed frame and input patches centered at the yellow + in (a). Synthesized frames and effective sampling locations are shown in (b), (c) and (d). Please zoom in the figures for a better view.

4.4 Ablation study

In this section, we perform comprehensive ablations to analyse the major components of our method, including the settings of the encoder-decoder architecture, different sizes of the estimated kernels and the usage of mask and bias estimators.

4.4.1 Encoder-decoder architecture

In typical kernel based interpolation methods, the encoder-decoder architecture occupies most of the network parameters (e.g., 97.7% in SepConv [28] and 97.1% in DSepConv [31]). To reduce model parameters, we replace some of the 3×3 filters into 1×1 in each convolution layer by using HetConv [32], leaving only a specific rate (1/P in [32]) of 3×3 kernels out of total kernels. As shown in Table 10, the FLOPs and the number of network parameters decrease when the rate getting smaller. The model performs the best when rate equals to 1/4, indicating that we can find a good balance between accuracy and computation, which is in line with the findings in [32] for the task of classification.

4.4.2 Generated kernel size

For each pixel to be synthesized, the generated kernel size n indicates how many pixels in the non-regular grid augmented with offsets could be used. Larger n enables the network to reference more pixels but it inevitably has more FLOPs and runtime. As shown in Table 11, the performance improves but the computation (FLOPs) and runtime increase when using larger kernel sizes. Please note that we do not recommend to use network with kernel size larger than 5 (e.g., n = 7, 9, 11) because they increase by 39.8%, 128.1% and 294.8% in terms of FLOPs and runtime increase by 92.0%, 210.6% and 357.8% in terms of runtime, compared to n = 5. Figure 16 shows an example of the effect of different kernel sizes. We also choose pixels from the synthesized frames indicated by the yellow + and visualize the effective sampling locations. We see that the proposed model with n = 5 can correctly use pixels from lower right of the first patch and upper left of the second, producing the best result.

Since the flow based methods can be seen as a special case of our method when n = 1 in section 3.1.2, we further visualize the offsets which are equivalent to optical flow in Figure 17. Without an explicit training phase for optical flow, our method learns meaningful information about motion between the frames for the task of frame interpolation.

4.4.3 Mask estimator

To examine the effectiveness of the mask estimator in our network, we trained a network without estimating masks. As shown in Table 12, the mask estimator significantly improves the performances on different datasets, especially in terms of PSNR and SSIM. On one hand, the learned masks help modulate the sample pixels guided by offsets, which allows the network to vary the spatial distribution and change the relative influence of the reference pixels [30]. On the other hand, masks reduce the burden of estimating separable convolution kernels, making the network better handle challenging cases such as occlusion.

4.4.4 Bias estimator

We compare the performance between model without bias estimator and the full model in Table 12. By introducing new bias values for each pixel’s synthesis, the performance saturates in terms of SSIM while improves by 0.1 dB and 0.04 dB in terms of PSNR.
of PSNR on UCF101 and Vimeo90K datasets, respectively. The learned bias values help to better model the linear relationship between the sampled pixels and corresponding kernels, which is in line with the flexible usage in common convolutional layers.

5 Discussions and limitations

By extending the approaches from [28, 31], our proposed EDSC_s achieves the best performance and the EDSC_m model is the first to able to produce an in-between frame at arbitrary time steps among all the kernel-based interpolation methods. However, our method has some limitations. First, despite we prove theoretically the optical flow based interpolation methods to be specific instances of our method when \( n = 1 \), the estimated bi-directional optical flows and generated frames fail to reach the same level of them. This is because the network is really simple and the offsets (which is equivalent to optical flow) are learned in an unsupervised manner, unlike those which utilize off-the-shelf flow estimation networks with a good initialization. Second, for multiple frame interpolation, our EDSC_m model is not so flexible as the methods in [20, 7, 26] that explicitly warp pixels and features before generating the output frame. We need to train from scratch and supervise the model at different time steps \( t \) while they do not.

Some recent researches enhance the performance of interpolation by making use of auxiliary information (e.g., more reference frames [39, 47] and high frame rate video with low spatial resolution [65]). Besides, a good initialization and fine-tuning of pre-trained sub-networks (such as PWC-Net [17], RBPN [62], Megadepth [63]) can greatly help to produce high quality interpolation results. Although the well-known kernel based methods, including ours, do not utilize any of them, it would be interesting to explore its use and extend our method to generate frames with higher quality. Another direction in recent research is joint video enhancement problem [45, 40, 46]. In the future, we plan to extend our approach to fix more tasks in the area of video processing.

6 Conclusion

In this paper, we have presented an enhanced deformable separable network for video frame interpolation. Our method improves the performance of kernel-based methods with fewer parameters by processing the information in a non-local neighborhood with learned adaptive offsets, kernels, masks and biases. And we provide the first kernel-based method that can generate as many intermediate frames as needed between two consecutive frames. Further, as demonstrated theoretically, both kernel- and flow-based methods can be regarded as special cases of our method. Comprehensive experiments show that our method performs favorably against state-of-the-art methods.

References

[1] R. Castagno, P. Haavisto, and G. Ramponi, “A method for motion adaptive frame rate up-conversion,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 5, pp. 436–446, 1996.

[2] W. Bao, X. Zhang, L. Chen, L. Ding, and Z. Gao, “High-order model and dynamic filtering for frame rate up-conversion,” IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3813–3826, 2018.

[3] J. Wu, C. Yuen, N. Cheung, J. Chen, and C. W. Chen, “Modeling and optimization of high frame rate video transmission over wireless networks,” IEEE Transactions on Wireless Communications, vol. 15, no. 4, pp. 2713–2726, 2016.

[4] H. Choi and I. V. Bajić, “Deep frame prediction for video coding,” IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[5] H. Jiang, D. Sun, V. Jampani, M. Yang, E. Learned-Miller, and J. Kautz, “Super slomo: High quality estimation of multiple intermediate frames for video interpolation,” in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9000–9008.

[6] W. Bao, W. Lai, X. Zhang, Z. Gao, and M. Yang, “MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[7] W. Bao, W. Lai, C. Ma, X. Zhang, Z. Gao, and M. Yang, “Depth-aware video frame interpolation,” in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3703–3712.

[8] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, “Deep stereo: Learning to predict new views from the world’s imagery,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5515–5524.

[9] J. Konrad and E. Dubois, “Bayesian estimation of motion vector fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 9, pp. 910–927, 1992.

[10] L. Xu, J. Jia, and Y. Matsushita, “Motion detail preserving optical flow estimation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1744–1757, 2012.

[11] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in European Conference on Computer Vision, 2004, pp. 25–36.

[12] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, “A database and evaluation methodology for optical flow,” International Journal of Computer Vision, vol. 92, no. 1, pp. 1–31, 2011.

[13] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, “FlowNet: Learning optical flow with convolutional networks,” in IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.

[14] 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 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647–1655.

[15] A. Ranjan and M. J. Black, “Optical flow estimation using a spatial pyramid network,” in IEEE Conference on
16. T. Hui, X. Tang, and C. C. Loy, “LiteFlowNet: A lightweight convolutional neural network for optical flow estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8981–8989.

17. D. Sun, X. Yang, M. Liu, and J. Kautz, “PWC-Net: Cnns for optical flow using pyramid, warping, and cost volume,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

18. D. Sun, X. Yang, M. Liu, and J. Kautz, “Models matter, so does training: An empirical study of cnns for optical flow estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

19. C. Bailer, B. Taetz, and D. Stricker, “Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1879–1892, 2019.

20. S. Niklaus and F. Liu, “Context-aware synthesis for video frame interpolation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1701–1710.

21. X. Cheng and Z. Chen, “A multi-scale position feature transform network for video frame interpolation,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

22. Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, “Video frame synthesis using deep voxel flow,” in *IEEE International Conference on Computer Vision*, 2017, pp. 4473–4481.

23. Y. Liu, Y. Liao, Y. Lin, and Y. Chuang, “Deep video frame interpolation using cyclic frame generation,” in *AAAI Conference on Artificial Intelligence*, 2019, pp. 8794–8802.

24. T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

25. S. Yu, B. Park, and J. Jeong, “Posnet: 4x video frame interpolation using position-specific flow,” in *IEEE International Conference on Computer Vision Workshops*. IEEE, 2019, pp. 3503–3511.

26. S. Niklaus and F. Liu, “Softmax splatting for video frame interpolation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

27. S. Niklaus, L. Mai, and F. Liu, “Video frame interpolation via adaptive convolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2270–2279.

28. S. Niklaus, L. Mai, and F. Liu, “Video frame interpolation via adaptive separable convolution,” in *IEEE Conference on International Conference on Computer Vision*, 2017, pp. 261–270.

29. J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

30. X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable convnets V2: more deformable, better results,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.

31. X. Cheng and Z. Chen, “Video frame interpolation via deformable separable convolution,” in *AAAI Conference on Artificial Intelligence*, 2020, pp. 10 607–10 614.

32. P. Singh, V. K. Verma, P. Rai, and V. P. Namboodiri, “Het-conv: Heterogeneous kernel-based convolutions for deep cnns,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4835–4844.

33. R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, “An intriguing failing of convolutional neural networks and the coordconv solution,” in *Advances in Neural Information Processing Systems*, 2018, pp. 9628–9639.

34. S. Xie and Z. Tu, “Holistically-nested edge detection,” *International Journal of Computer Vision*, vol. 125, no. 1-3, pp. 3–18, 2017.

35. F. A. Reda, D. Sun, A. Dundar, M. Shoeybi, G. Liu, K. J. Shih, A. Tao, J. Kautz, and B. Catanzaro, “Unsupervised video interpolation using cycle consistency,” *CoRR*, vol. abs/1906.05928, 2019.

36. L. Yuan, Y. Chen, H. Liu, T. Kong, and J. Shi, “Zoom-in-to-check: Boosting video interpolation via instance-level discrimination,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 183–12 191.

37. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

38. D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Trémeau, and C. Wolf, “Residual conv-deconv grid network for semantic segmentation,” in *British Machine Vision Conference*, 2017.

39. X. Xu, L. Si-Yao, W. Sun, Q. Yin, and M. Yang, “Quadratic video interpolation,” in *Advances in Neural Information Processing Systems*, 2019, pp. 1645–1654.

40. M. Haris, G. Shakhnarovich, and N. Ukita, “Space-time-aware multi-resolution video enhancement,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

41. C. Liu et al., “Beyond pixels: exploring new representations and applications for motion analysis,” Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

42. M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, “Channel attention is all you need for video frame interpolation,” in *AAAI Conference on Artificial Intelligence*, 2020, pp. 10 663–10 671.

43. W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

44. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel
attention networks,” in *European Conference on Computer Vision*, vol. 11211, 2018, pp. 294–310.

[45] W. Shen, W. Bao, G. Zhai, L. Chen, X. Min, and Z. Gao, “Blurry video frame interpolation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[46] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, “Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[47] M. Choi, J. Choi, S. Baik, T. H. Kim, and K. M. Lee, “Scene-adaptive video frame interpolation via meta-learning,” 2020.

[48] T. Peleg, P. Szekely, D. Sabo, and O. Sendik, “Im-net for high resolution video frame interpolation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2398–2407.

[49] H. Lee, T. Kim, T. Chung, D. Pak, Y. Ban, and S. Lee, “Adacof: Adaptive collaboration of flows for video frame interpolation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1410–1418.

[50] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung, “Phase-based frame interpolation for video,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1410–1418.

[51] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. H. Gross, and C. Schroers, “Phasenet for video frame interpolation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 498–507.

[52] R. Szeto, X. Sun, K. Lu, and J. J. Corso, “A temporally-aware interpolation network for video frame inpainting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[53] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of International Conference on Machine Learning*, 2010, pp. 807–814.

[54] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, “Two deterministic half-quadratic regularization algorithms for computed imaging,” in *International Conference on Image Processing*, 1994, pp. 168–172.

[55] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.

[56] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, 2014.

[57] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” *CoRR*, 2012.

[58] S. Nah, T. H. Kim, and K. M. Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 257–265.

[59] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.