Understanding Deformable Alignment in Video Super-Resolution

Kelvin C.K. Chan,[1*] **Xintao Wang,**[2*†] **Ke Yu,**[3] **Chao Dong,**[4,5] **Chen Change Loy**[1‡]

[1]S-Lab, Nanyang Technological University [2]Applied Research Center, Tencent PCG
[3]CUHK – SenseTime Joint Lab, The Chinese University of Hong Kong
[4]Shenzhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[5]SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
{chan0899, ccloy}@ntu.edu.sg, xintao.wang@outlook.com, yk017@ie.cuhk.edu.hk, chao.dong@siat.ac.cn

Abstract

Deformable convolution, originally proposed for the adaptation to geometric variations of objects, has recently shown compelling performance in aligning multiple frames and is increasingly adopted for video super-resolution. Despite its remarkable performance, its underlying mechanism for alignment remains unclear. In this study, we carefully investigate the relation between deformable alignment and the classic flow-based alignment. We show that deformable convolution can be decomposed into a combination of spatial warping and convolution. This decomposition reveals the commonality of deformable alignment and flow-based alignment in formulation, but with a key difference in their offset diversity. We further demonstrate through experiments that the increased diversity in deformable alignment yields better-aligned features, and hence significantly improves the quality of video super-resolution output. Based on our observations, we propose an offset-fidelity loss that guides the offset learning with optical flow. Experiments show that our loss successfully avoids the overflow of offsets and alleviates the instability problem of deformable alignment. Aside from the contributions to deformable alignment, our formulation inspires a more flexible approach to introduce offset diversity to flow-based alignment, improving its performance.

Introduction

Video super-resolution (SR) aims at recovering high-resolution consecutive frames from their low-resolution counterparts. The key challenge of video SR lies in the effective use of complementary details from adjacent frames, which can be misaligned due to camera and object motions. To establish inter-frame correspondence, early methods (Caballero et al. 2017; Liu et al. 2017; Sajjadi, Vemulapalli, and Brown 2018; Tao et al. 2017; Xue et al. 2019) employ optical flow for explicit frame alignment. They warp neighboring frames to the reference one and pass these images to Convolutional Neural Networks (CNNs) for super-resolution. Recent studies (Tian et al. 2020; Wang et al. 2019a,b) perform alignment implicitly via deformable convolution and show superior performance. For instance, the winner of NTIRE 2019 video restoration challenges (Nah et al. 2019a,b), EDVR (Wang et al. 2019b), significantly outperforms previous methods with coarse-to-fine deformable convolutions.

These two kinds of methods are generally regarded as orthogonal approaches and are developed independently. It is of great interest to know (1) *the relationship between explicit and implicit alignments*, and (2) *the source of improvement brought by implicit modeling*. As there are no related works, we bridge the gap by exploring the intrinsic connections of two representative methods – flow-based alignment (explicit alignment with optical flow) and deformable alignment (implicit alignment with deformable convolution). Studying their relation not only helps us understand the working mechanism of deformable alignment, but also inspires a more general design of video SR approaches.

Deformable convolution (Dai et al. 2017; Zhu et al. 2019) (DCN) is originally designed for spatial adaption in object detection. The key idea is to displace the sampling locations of standard convolution by some learned offsets. When DCN is applied in temporal alignment, the displaced kernels on neighboring frames will be used to align intermediate features. On the face of it, this procedure is different from flow-based methods, which align adjacent frames by flow-warping. To reveal their relationship, we show that deformable alignment can be formulated as a combination of feature-level flow-warping and convolution. This intuitive decomposition indicates that these two kinds of alignment intrinsically share the same formulation but differ in their *offset diversity*. Specifically, flow-based alignment only learns one offset at each feature location, while deformable alignment introduces multiple offsets, the number of which is in proportion to the kernel size of DCN.

Under this relation, we systematically investigate the effects of offset diversity and gain two interesting insights. First, the learned offsets in deformable alignment have similar patterns as optical flow, suggesting that deformable and flow-based alignments are strongly correlated in both concepts and behaviors. Second, diverse offsets achieve better restoration quality than a single offset. As different offsets are complementary to each other, they can effectively allevi-

*Both authors contributed equally to this work.
†This work was done during his study in CUHK.
‡Corresponding author.

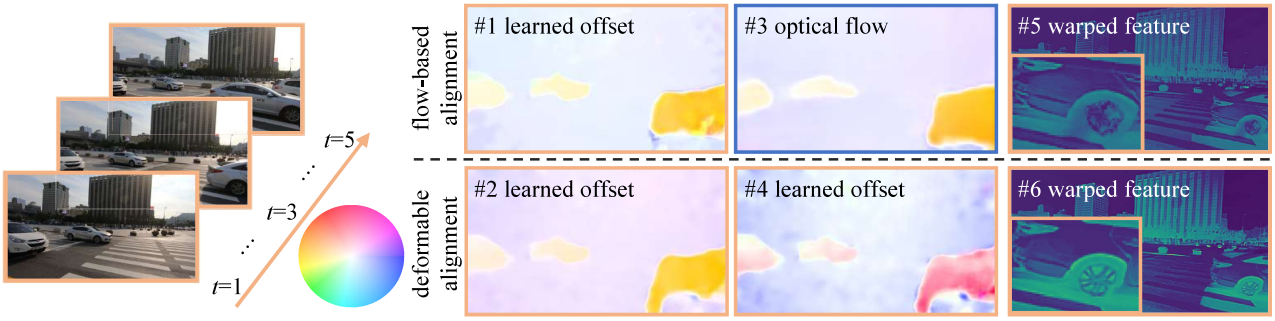


Figure 1: The learned offsets in both flow-based alignment (#1) and deformable alignment (#2, #4) have similar patterns as optical flow obtained using a deep learning-based optical flow estimator (Sun et al. 2018) (#3). The offset diversity allows deformable alignment to learn complementary offsets (#4), which effectively alleviate the occlusion problem and reduce warping errors. As a result, the warped feature after deformable alignment (#6) contains more details than that with flow-based alignment (#5) (see details of the car wheel).

ate the occlusion problem and reduce warping errors caused by large motions. Figure 1 depicts the comparisons of these two methods in their learned offsets and feature patterns.

With a more profound understanding of their relationship, we decided to use the widely-adopted optical flow technique to benefit the training of deformable convolution. It is known that the training of deformable alignment is unstable and the overflow of offsets could severely degrade the performance (Wang et al. 2019b). We propose an offset-fidelity loss that adopts optical flows to guide the offset learning of DCN while preserving offset diversity. Our experiments show that the proposed strategy successfully stabilizes the training process of deformable alignment.

Apart from the contributions to deformable alignment, our decomposition of DCN is also beneficial to flow-based alignment approaches. Specifically, in our formulation, the number of offsets is not necessarily equal to the square of kernel size. Compared to deformable convolution, our formulation provides a more flexible means for increasing offset diversity in flow-based alignment approaches.

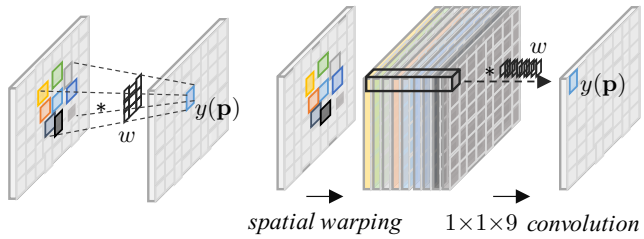
Our **contributions** are summarized as follows: (1) While deformable alignment has been shown a compelling alternative to the conventional flow-based alignment for motion compensation, its link with flow-based alignment is only superficially discussed in the literature. This paper is the first study that establishes the relationship between the two important concepts formally. (2) We systematically investigate the benefits of offset diversity. We show that offset diversity is the key factor for improving both the alignment accuracy and SR performance. (3) Based on our studies, we propose an offset-fidelity loss in deformable alignment to stabilize training while preserving offset diversity. An improvement of up to 1.7 dB is observed with our loss. (4) Our formulation inspires a more flexible approach to increase offset diversity in flow-based alignment methods.

Related Work

Different from single image SR (Chan et al. 2020a; Dai et al. 2019; Dong et al. 2014; Haris, Shakhnarovich, and Ukita 2018; He et al. 2019; Ledig et al. 2017; Lim et al. 2017; Liu et al. 2020; Mei et al. 2020; Wang et al. 2018b,a; Zhang et al. 2018; Zhang, Gool, and Timofte 2020), an additional challenge of video SR (Chan et al. 2020b; Dai et al. 2015; Huang, Wang, and Wang 2015; Liu and Sun 2014; Yi et al. 2019; Li et al. 2020; Isobe et al. 2020a,b) is to align multiple frames for the construction of accurate correspondences. Based on whether optical flow is explicitly estimated, existing motion compensation approaches in video SR can be mainly divided into two branches – *explicit* methods and *implicit* methods.

Most existing methods adopt an explicit motion compensation approach. Earlier works of this approach (Kappeler et al. 2016; Liao et al. 2015) first use a fixed and external optical flow estimator to estimate the flow fields between the reference and its neighboring frames, and then learn a mapping from the flow-warped inputs to the high-resolution output. Such two-stage methods are time-consuming and tend to fail when the flow estimation is not accurate. Several follow-up studies (Caballero et al. 2017; Liu et al. 2017; Sajjadi, Vemulapalli, and Brown 2018; Tao et al. 2017; Xue et al. 2019) incorporate the flow-estimation component into the SR pipeline. For instance, TOFlow (Xue et al. 2019) points out that the optimal flow is task-specific in video enhancement including video SR, and thus a trainable motion estimation component is more effective than a fixed one. Nevertheless, all these methods explicitly perform flow estimation and warping in the *image domain*, which may introduce artifacts around image structures (Tian et al. 2020).

Several recent methods perform motion compensation implicitly and show superior performance. For instance, DUF (Jo et al. 2018) learns an upsampling filter for each pixel location, and a few other methods (Tian et al. 2020; Wang et al. 2019a,b) incorporate deformable convolution into motion compensation. Deformable convolution (Dai et al. 2017) is capable of predicting additional offsets that offer spatial flexibilities to a convolution kernel. This differs from a standard convolution, which is restricted to a regular neighborhood. TDAN (Tian et al. 2020) applies deformable



(a) DCN (b) DCN decomposition

Figure 2: Deformable convolution with a 3×3 kernel can be decomposed into nine spatial warpings and one 3D convolution. Kernel weights are represented as w.

convolutions for temporal alignment in video SR. Following the structure design in flow-estimation methods (Dosovitskiy et al. 2015; Ranjan and Black 2017; Sun et al. 2018), EDVR (Wang et al. 2019b) adopts deformable alignment in a pyramid and cascading architecture, achieving state-of-the-art performance in video SR.

Although deformable alignment and the classic flow-based alignment look unconnected at first glance, they are indeed highly related. In this study, we delve deep into the connections between them. Based on our analyses, we propose an offset-fidelity loss to stabilize the training and improve the performance of deformable alignment.

Unifying Deformable and Flow-Based Alignments

Deformable Convolution Revisited

We start with a brief review of deformable convolution (DCN) (Dai et al. 2017), which was originally proposed to accommodate geometric variations of objects in the tasks of object detection (Bertasius, Torresani, and Shi 2018) and image segmentation (Dai et al. 2017). Let $\boldsymbol{p}_k$ be the k-th sampling offset in a standard convolution with kernel size $n \times n$. For example, when $n=3$, we have $\boldsymbol{p}_k \in \{(-1,-1),(-1,0),\cdots,(1,1)\}$. We denote the k-th additional learned offset at location $\boldsymbol{p}+\boldsymbol{p}_k$ by $\Delta\boldsymbol{p}_k$. A deformable convolution can be formulated as:

$$y(\boldsymbol{p}) = \sum_{k=1}^{n^2} w(\boldsymbol{p}_k) \cdot x(\boldsymbol{p}+\boldsymbol{p}_k+\Delta\boldsymbol{p}_k), \tag{1}$$

where x and y represent the input and output features, respectively. The kernel weights are denoted by w. As illustrated in Fig. 2(a), unlike standard convolution, a deformable convolution has more flexible sampling locations. In practice, one can divide the C channel features into G groups of features with C/G channels, and $n^2 \times G$ offsets are learned for each spatial location. In DCNv2 (Zhu et al. 2019) a modulation mask is introduced to further strengthen the capability in manipulating spatial support regions. A detailed analysis of the mask is included in the supplementary material[1].

[1]Please refer to https://arxiv.org/abs/2009.07265

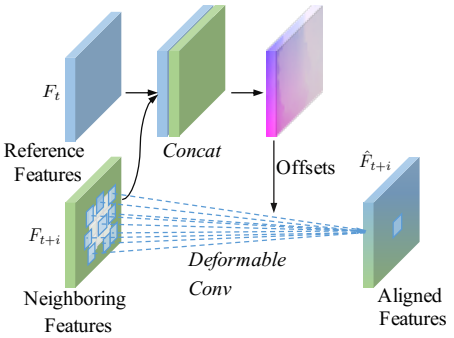


Figure 3: Deformable alignment applies deformable convolution to align neighboring features to the reference feature. The offsets are predicted by a few convolutions with both the reference and neighboring features as the inputs. The reference feature is used only to predict the offsets, and is not directly involved in the convolution.

Deformable Alignment

In video SR, it is crucial to establish correspondences between consecutive frames for detail extraction and fusion. Recent studies (Tian et al. 2020; Wang et al. 2019a,b) go beyond the traditional way of flow-warping and apply deformable convolution for feature alignment, as shown in Fig. 3.

Let F_t and F_{t+i} be the intermediate features of the reference and neighboring frames, respectively. In deformable alignment, a deformable convolution is used to align F_{t+i} to F_t. Mathematically, we have:

$$\hat{F}_{t+i}(\boldsymbol{p}) = \sum_{k=1}^{n^2} w(\boldsymbol{p}_k) \cdot F_{t+i}(\boldsymbol{p}+\boldsymbol{p}_k+\Delta\boldsymbol{p}_k), \tag{2}$$

where $\hat{F}_{t+i}$ represents the aligned feature. The offsets $\Delta\boldsymbol{p}_k$ are predicted by a few convolutions with both F_t and F_{t+i} as the inputs. The reference feature F_i is used only to predict the offsets and is not directly involved in the convolution.

Relation between Deformable Alignment and Flow-Based Alignment

There is an intuitive yet less obvious connection between deformable and flow-based alignments. The connection is rarely discussed in previous works. Instead of treating them as orthogonal approaches, we unify these two important concepts in this paper. Now, we discuss the connection between deformable alignment and flow-based alignment by showing that DCN can be decomposed into spatial warping and standard convolution.

Let x be the input feature, and $\boldsymbol{p}_k+\Delta\boldsymbol{p}_k$ $(k=1,\cdots,n^2)$ be the k-th offset for location $\boldsymbol{p}$. Next, denote the feature warped by the k-th offset by $x_k(\boldsymbol{p}) = x(\boldsymbol{p}+\boldsymbol{p}_k+\Delta\boldsymbol{p}_k)$. From Eqn. (1), we have:

$$y(\boldsymbol{p}) = \sum_{k=1}^{n^2} w(\boldsymbol{p}_k) \cdot x_k(\boldsymbol{p}), \tag{3}$$

which is equivalent to a $1\times1\times n^2$ standard 3D convolution. Hence, we see that a deformable convolution with kernel size $n\times n$ is equivalent to n^2 individual spatial warpings followed by a standard 3D convolution with kernel size $1\times1\times n^2$. The illustration is shown in Fig. 2(b).

Remarks:

1. By replacing n^2 with $N\in\mathbb{N}$ in Eqn. (3) , this decomposition *generalizes* DCN by removing the constraint that the number of offsets within each group must be equal to n^2. Therefore, in the remaining sections, we denote the number of offsets per group by N.
2. By stacking the N warped features in the channel dimension, the $1\times1\times N$ 3D convolution can be implemented as a 1×1 2D convolution. In other words, DCN is equivalent to N separate spatial warpings followed by a 1×1 2D convolution.

From Eqn. (3), we see that the special case of $n{=}1$ is equivalent to a spatial warping followed by a 1×1 convolution. In the context of motion compensation, this special case corresponds to a flow-based alignment. In other words, deformable and flow-based alignments share the same formulation but with a difference in the *offset diversity*.

Discussion. The aforementioned analysis leads to a few interesting explorations:

1. *Where does deformable alignment gain the extra performance in comparison to flow-based alignment?* The analysis points to offset diversity, and we verify this hypothesis in our experiments.
2. *Is higher offset diversity always better?* We demonstrate that although the output quality increases with offset diversity in general, a performance plateau is observed when the number of offsets gets larger. Hence, indefinitely increasing the number of offsets could lower the efficiency of the model without significant performance gain. In practice, one should balance the performance and computational efficiency by choosing a suitable number of offsets.
3. *Can we increase the offset diversity of flow-based alignment?* Unlike deformable alignment, where the number of offsets must be equal to the square of kernel size, our formulation generalizes deformable alignment with an arbitrary number of offsets. As a result, it provides a more flexible approach to introduce offset diversity to flow-based alignment. We show in the experiments that increasing offset diversity helps a flow-based network to achieve better SR performance.

Offset-fidelity Loss

In this section, motivated by our decomposition, we demonstrate how optical flow can benefit deformable alignment through the our proposed offset fidelity loss.

Due to its unclear offset interpretability, deformable alignment is usually trained from scratch with random initializations. With increased network capacities, the training of deformable alignment becomes unstable, and the overflow of offsets severely degrades the model performance[2]. In contrast, in flow-based alignment, various training strategies are developed to improve alignment accuracy and speed of convergence, such as the adoption of flow network structure (Haris, Shakhnarovich, and Ukita 2019; Xue et al. 2019), flow guidance loss (Liu et al. 2017), and flow pretraining (Caballero et al. 2017; Tao et al. 2017; Xue et al. 2019).

Given the relation between spatial warping and deformable convolution, we propose to use optical flow to guide the training of offsets. Specifically, we propose an offset-fidelity loss to constrain the offsets so that they do not deviate much from the optical flow. Furthermore, to facilitate the learning of optimal and diverse offsets for video SR, Heaviside step function is incorporated. More specifically, we augment the data-fitting loss as follows:

$$\hat{L} = L + \lambda \sum_{n=1}^{N} L_n, \tag{4}$$

where L is the data-fitting loss (*e.g.*, Charbonnier loss in (Wang et al. 2019b)) and

$$L_n = \sum_i \sum_j H\left(|x_{n,ij} - y_{ij}| - t\right) \cdot |x_{n,ij} - y_{ij}|, \tag{5}$$

where y and $x_{n,ij}$ denote the optical flow (computed by an off-the-shelf optical flow method) and the n-th learned offsets at location (i, j), respectively. $H(\cdot)$ represents the Heaviside step function. Here λ and t are hyper-parameters controlling the diversity of the offsets. The quantities are robust to changes, and $\lambda{=}1$, $t{=}10$ is a reasonable setting. As shown in our experiments, our loss is able to stabilize the training and avoid the offset overflow in large models.

Analysis

We conduct experiments to reveal the connections and differences between deformable and flow-based alignments in video SR. Unless specified, EDVR-M[3] is adopted for analyses as it maintains a good balance between training efficiency and performance. Moreover, to decouple the complex relation among different components in deformable alignment, we use the non-modulated DCN. The experimental details are provided in the supplementary material.

Deformable Alignment vs. Optical Flow

By setting $G{=}N{=}1$ (*i.e.*, group=1 and the number of offsets per group=1), the offset learned by deformable alignment resembles the optical flow as in a flow-based alignment approach. Specifically, when there is only one offset to learn, the model automatically learns to align features based on the motions between frames. As shown in Fig. 4, the offsets are highly similar to the optical flows estimated by PWC-Net (Sun et al. 2018).

[2]The instability of EDVR is observed in (Wang et al. 2019b) and also in our experiments.

[3]EDVR-M is a moderate version of EDVR provided by the official implementation (Wang et al. 2019b).

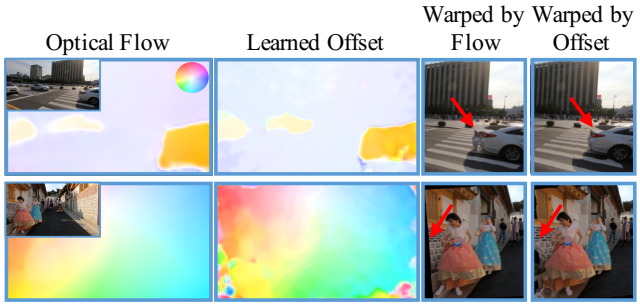


Figure 4: While the learned offsets are highly similar to optical flows, their disparity is non-negligible because optical flow may not be optimal for video SR. Dark borders and ghosting regions appear in the images warped by optical flow. In contrast, those warped by learned offsets are clearer.

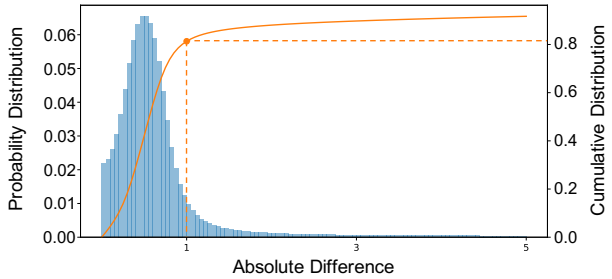


Figure 5: Over 80% of the estimations have a difference to optical flow smaller than one pixel (orange dot). This demonstrates that in the case of $G{=}N{=}1$, deformable alignment is indeed equivalent to flow-based alignment.

Despite their high similarity, the disparity between learned offsets and optical flows is non-negligible due to the fundamental difference of the task nature (Xue et al. 2019). Specifically, while PWC-Net is trained to describe the motions between frames, our baseline is trained for video SR, in which optical flow may not be the optimal representation of frame correspondences. From Fig. 4, we see that the image warped by the learned offsets clearly preserves more scene contents. In contrast, a dark region and a ghosting region are seen in the images warped by optical flow. Note that the offsets are learned for warping the features, and the warped images in Fig. 4 are solely for illustrative purposes.

We quantitatively study the correlation between the offsets and optical flows, by computing their pixelwise difference. As shown in Fig. 5, over 80% of the estimations have a difference smaller than one pixel from the optical flow. This demonstrates that in the case of $G{=}N{=}1$, deformable alignment is indeed highly similar to flow-based alignment. In the following analyses, we will adopt this model as our approximation to the flow-based alignment baseline.

Feature Warping. The aforementioned flow-based alignment baseline performs feature warping. This differs from a majority of flow-based methods that learn flows for image warping (Liu et al. 2017; Xue et al. 2019). In those methods, the flows contain fractional values and hence interpolation is required during warping. This inevitably introduces information loss, particularly high-frequency details. Consequently, the blurry aligned images yield suboptimal SR results. Recent deformable alignment methods (Tian et al. 2020; Wang et al. 2019a,b) attempt to perform alignment at the feature level and achieve remarkable results. We inspect the contribution of feature-level warping by replacing the feature alignment module in our flow-based baseline with an image alignment module. Surprisingly, despite the closeness of the architecture, image alignment leads to a drop of 0.84 dB. This indicates that feature-level warping is beneficial to flow-based alignment. More comparisons are provided in the supplementary material.

	$G{=}1, k{=}1$	$G{=}1, k{=}3$	$G{=}8, k{=}1$	$G{=}8, k{=}3$
DCN	29.979	30.199	30.183	30.264
Our Decomp.	29.992	30.240	30.179	30.231
Difference	+0.013	+0.041	-0.004	-0.033

Table 1: PSNR of two instantiations of EDVR-M on REDS4. The similar performance verifies our claim that DCN can be decomposed into spatial warping and convolution. G and k represent the number of groups and the kernel size used in DCN, respectively.

Offset Diversity

Decomposition Equivalence. In this section, we use our decomposition in this paper in place of DCN since it provides a more flexible choice of the number of offsets. To verify their equivalence, we train two instantiations – original DCN and our decomposition. As shown in Table 1, our experiments show that the two instantiations achieve similar performance, corroborating our hypothesis.

Learned Offsets. Given that the primary difference between flow-based alignment and deformable alignment is the number of offsets N, it is natural to question *the roles and characteristics of the additional offsets in deformable alignment (i.e. $N{>}1$).* To answer this, we fix $G{=}1$ and compare the performance in the cases of $N{=}1$ and $N{=}15$.

We sort the 15 offsets according to their l_1-distance to the optical flow, and an example is shown in Fig. 6. On the one hand, there exists offsets that closely resemble the optical flow. On the other hand, some offsets have different estimated directions compared to the optical flows; although these offsets are also able to separate the motions of different objects, as the optical flow does, their directions do not correspond to the actual camera and object motions.

We further visualize the diversity of the offsets, which is measured by the pixelwise standard deviation of the offsets. We observe that the offsets tend to have a larger diversity in regions that optical flows do not work well for alignment. For instance, as shown in the heatmap of Fig. 6, the standard deviation tends to be larger in image boundaries, where unseen regions are common. Although diverse offsets with different estimated directions are obtained, they are analogous

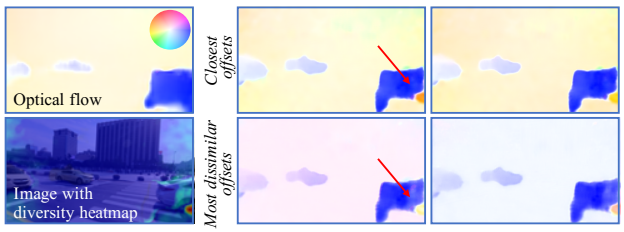


Figure 6: While the closest offsets are highly similar to optical flow with slight differences (top row), the most dissimilar offsets have different estimated directions (bottom row). Moreover, the offsets tend to have a larger diversity in regions that optical flows do not work well for alignment (see the diversity heatmap and regions marked by red arrows). (Zoom-in for best view)

TDAN			Flow-based		
N=1	N=9	N=25	N=1	N=2	N=3
33.313	33.483	**33.540**	32.835	32.973	**33.017**

Table 2: PSNR on Vimeo-90K-T (Xue et al. 2019) using two additional models. For both models, the performance increases with increased number of offsets.

to optical flow in terms of their overall shape. This suggests that the motion between frames is still an important clue in deformable alignment, as in flow-based alignment. More qualitative results are shown in the supplementary material.

Contributions of Diversity. We are also interested in *whether the diverse flow-like offsets are beneficial to video SR*. This motivates us to inspect the aligned features and the corresponding performance. With a single offset, the aligned features suffer from the warping error induced by unseen regions and inaccurate motion estimation. The inaccurately aligned features inevitably hinder the aggregation of information and therefore harm the subsequent restoration. In contrast, with multiple offsets, the independently warped features are reciprocal and provide better-aligned features during fusion, hence alleviating the inaccurate alignment by a single offset. An example of the aligned features are visualized in Fig. 7. It is observed that with a single offset, the aligned features are less coherent. For instance, in the image boundaries, which correspond to the regions that do not exist in the neighboring frame, the feature warped by a single offset contains a large area of dark regions. Contrarily, with 15 offsets, the complementary warped features provide additional information for fusion, resulting in features that are more coherent and preserve more details.

Increasing Offset Diversity. We then examine the performance gain by gradually increasing the number of offsets and attempt to examine *if more offsets will always lead to a better performance*.

The qualitative and quantitative comparison with different N are shown in Fig. 8 and Fig. 9, respectively. In particular, as the number of offsets increases from 1 to 5, the PSNR increases rapidly. When N further increases, the PSNR saturates at about 30.23 dB. This result indicates that the performance reaches a plateau when the number of offsets gets larger. As a result, simply increasing the number of offsets could lower the computational efficiency without significant performance gain. It is noteworthy that it is infeasible to balance the performance and computational efficiency in deformable alignment since the number of offsets must be equal to the square of kernel size. Our formulation, contrarily, generalizes deformable alignment with an arbitrary number of offsets, thus providing a more flexible approach to introducing offset diversity.

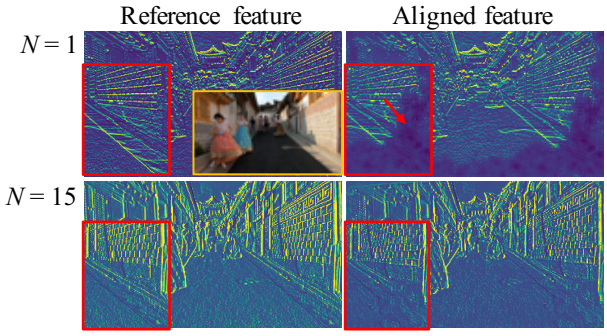


Figure 7: Aligned features with N=1 and N=15. With a single offset, the model lacks the ability to handle occlusion and inaccurate motion estimation (red arrows). With increased number of offsets, the features are better aligned, and more details are better preserved. (Zoom-in for best view)

We also inspect the correlation between offset diversity and PSNR performance. We measure the offset diversity by the pixelwise standard deviation of all offsets. As shown in Fig. 9, the performance of the model positively correlates with the offset diversity (Pearson Correlation Coefficient=0.9418 based on these six data points). This result implies that offset diversity indeed contributes to the performance gain.

To further support our conclusion, we additionally test the improvement brought by offset diversity using TDAN (Tian et al. 2020) and a flow-based network[4]. As shown in Table 2, the PSNR of the two models improves by up to 0.23 dB. Besides, an improvement of 0.18 dB is observed in the flow-based network, suggesting that the offset diversity not only improves feature alignment but is also constructive in image alignment.

Besides increasing the number of offsets, diversity can also be achieved by increasing the number of deformable groups G. Interestingly, the above conclusions are also applicable to G. A more detailed analysis is included in the supplementary material.

[4]We use an architecture similar to TOFlow (Xue et al. 2019) except that spatial warping is done on LR space instead of HR space for computational efficiency. We increase the number of offsets by using multiple SPyNet (Ranjan and Black 2017).

Figure 8: While the quality improves markedly when N increases from 1 to 3, further improvement at N=25 is relatively small.

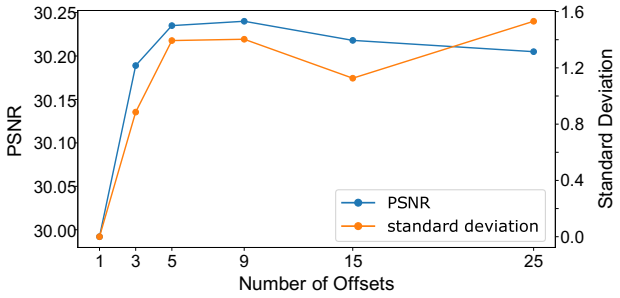
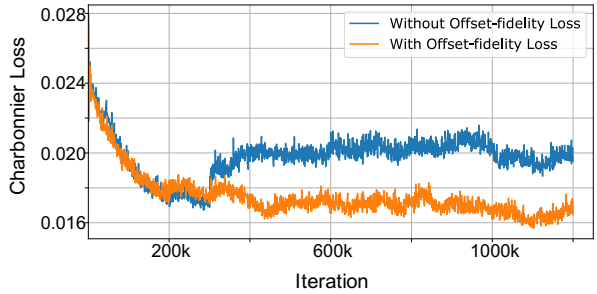


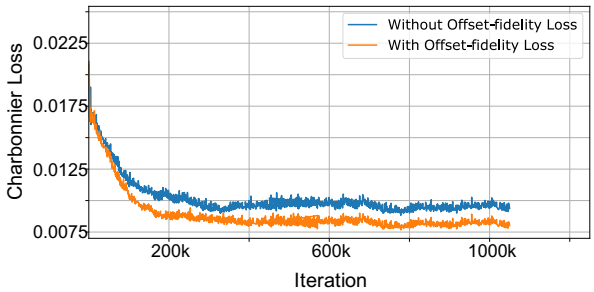
Figure 9: The performance of the models positively correlates with the offset diversity. When N increases from 1 to 5, both the PSNR and standard deviation of the offsets increase rapidly. When N further increases, both the PSNR and the standard deviation of the offsets saturate.

	REDS4	Vimeo-90K-T
without offset-fidelity loss	28.753	33.632
with offset-fidelity loss	30.480	35.223
Difference	+1.727	+1.591

Table 3: Quantitative comparison (PSNR) on REDS4 and Vimeo-90K-T for 4× video super-resolution. Results are evaluated on RGB channels.



(a) REDS



(b) Vimeo90K

Figure 10: (a) When training on REDS, the model trained without the offset-fidelity loss becomes unstable at about 300K, where the loss increases and is consistently greater than that trained with our loss thereafter. (b) When training on Vimeo-90K, the model trained with offset-fidelity loss is able to reach a lower training loss.

Offset-fidelity Loss

We train EDVR-L with the official training scheme. As the network capacity increases, the training of deformable alignment becomes unstable. Without the offset-fidelity loss, the overflow of offsets produces a zero feature map after deformable alignment. As a result, EDVR essentially becomes a single image SR model. On the contrary, our loss penalizes the offsets when they deviate from the optical flow, resulting in much more interpretable offsets and better performance. As shown in Fig. 10, EDVR converges with a lower training loss with our offset-fidelity loss. Note that in Fig. 10(a), the training loss increases at about 300K, which is the time when offsets overflow. In Table 3 we see that our loss introduces an additional improvement of up to 1.73 dB. The qualitative results are provided in the supplementary material.

Conclusion

The success of deformable alignment in video super-resolution has aroused great attention. In this study, we uncover the intrinsic connection in both concepts and behaviors between deformable alignment and flow-based alignment. For flow-based alignment, our work relaxes the constraint of deformable convolution on the number of offsets. It allows a more flexible way to increase the offset diversity in flow-based alignment approaches, improving the output quality. As for deformable alignment, our investigation empowers us to understand its underlying mechanism, potentially inspiring new alignment approaches. Motivated by our analysis, we propose an offset-fidelity loss to mitigate the stability problem during training.

Acknowledgments

This research was conducted in collaboration with SenseTime. This work is supported by A*STAR through the Industry Alignment Fund - Industry Collaboration Projects Grant. It is also partially supported by Singapore MOE AcRF Tier 1 (2018-T1-002-056), NTU SUG, and the National Natural Science Foundation of China (61906184).

References

Bertasius, G.; Torresani, L.; and Shi, J. 2018. Object detection in video with spatiotemporal sampling networks. In *ECCV*.

Caballero, J.; Ledig, C.; Andrew, A.; Alejandro, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation. In *CVPR*.

Chan, K. C.; Wang, X.; Xu, X.; Gu, J.; and Loy, C. C. 2020a. GLEAN: Generative Latent Bank for Large-Factor Image Super-Resolution. *arXiv preprint arXiv:2012.00739* .

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2020b. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. *arXiv preprint arXiv:2012.02181* .

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *ICCV*.

Dai, Q.; Yoo, S.; Kappeler, A.; and Katsaggelos, A. K. 2015. Dictionary-based multiple frame video super-resolution. In *ICIP*.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order Attention Network for Single Image Super-Resolution. In *CVPR*.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a Deep Convolutional Network for Image Super-resolution. In *ECCV*.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning Optical Flow With Convolutional Networks. In *ICCV*.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep Back-Projection Networks For Super-Resolution. In *CVPR*.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2019. Recurrent Back-Projection Network for Video Super-Resolution. In *CVPR*.

He, X.; Mo, Z.; Wang, P.; Liu, Y.; Yang, M.; and Cheng, J. 2019. ODE-Inspired Network Design for Single Image Super-Resolution. In *CVPR*.

Huang, Y.; Wang, W.; and Wang, L. 2015. Bidirectional Recurrent Convolutional Networks for Multi-Frame Super-Resolution. In *NeurIPS*.

Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; and Tian, Q. 2020a. Video Super-Resolution with Recurrent Structure-Detail Network. In *ECCV*.

Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.; Xu, C.; Li, Y.-L.; Wang, S.; and Tian, Q. 2020b. Video Super-Resolution with Temporal Group Attention. In *CVPR*.

Jo, Y.; Wug Oh, S.; Kang, J.; and Joo Kim, S. 2018. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. In *CVPR*.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. K. 2016. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging* 2(2): 109–122.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A. P.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *CVPR*.

Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. MuCAN: Multi-Correspondence Aggregation Network for Video Super-Resolution. In *ECCV*.

Liao, R.; Tao, X.; Li, R.; Ma, Z.; and Jia, J. 2015. Video super-resolution via deep draft-ensemble learning. In *CVPR*.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Lee, K. M. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *CVPRW*.

Liu, C.; and Sun, D. 2014. On Bayesian Adaptive Video Super Resolution. *TPAMI* .

Liu, D.; Wang, Z.; Fan, Y.; Liu, X.; Wang, Z.; Chang, S.; and Huang, T. 2017. Robust Video Super-Resolution with Learned Temporal Dynamics. In *ICCV*.

Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; and Wu, G. 2020. Residual Feature Aggregation Network for Image Super-Resolution. In *CVPR*.

Mei, Y.; Fan, Y.; Zhou, Y.; Huang, L.; Huang, T. S.; and Shi, H. 2020. Image Super-Resolution with Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining. In *CVPR*.

Nah, S.; Timofte, R.; Baik, S.; Hong, S.; Moon, G.; Son, S.; and Mu Lee, K. 2019a. NTIRE 2019 Challenge on Video Deblurring: Methods and Results. In *CVPRW*.

Nah, S.; Timofte, R.; Baik, S.; Hong, S.; Moon, G.; Son, S.; and Mu Lee, K. 2019b. NTIRE 2019 Challenge on Video Super-Resolution: Methods and Results. In *CVPRW*.

Ranjan, A.; and Black, M. J. 2017. Optical Flow Estimation using a Spatial Pyramid Network. In *CVPR*.

Sajjadi, M. S. M.; Vemulapalli, R.; and Brown, M. 2018. Frame-Recurrent Video Super-Resolution. In *CVPR*.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *CVPR*.

Tao, X.; Gao, H.; Liao, R.; Wang, J.; and Jia, J. 2017. Detail-revealing Deep Video Super-resolution. In *CVPR*.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. TDAN: Temporally Deformable Alignment Network for Video Super-Resolution. In *CVPR*.

Wang, H.; Su, D.; Liu, C.; Jin, L.; Sun, X.; and Peng, X. 2019a. Deformable Non-Local Network for Video Super-Resolution. *IEEE Access* .

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Loy, C. C. 2019b. EDVR: Video Restoration With Enhanced Deformable Convolutional Networks. In *CVPRW*.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018a. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform. In *CVPR*.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Loy, C. C.; Qiao, Y.; and Tang, X. 2018b. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *ECCVW*.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video Enhancement with Task-Oriented Flow. *IJCV* .

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive Fusion Video Super-Resolution Network via Exploiting Non-Local Spatio-Temporal Correlations. In *ICCV*.

Zhang, K.; Gool, L. V.; and Timofte, R. 2020. Deep Unfolding Network for Image Super-Resolution. In *CVPR*.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *ECCV*.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable ConvNets v2: More Deformable, Better Results. In *CVPR*.